Eastern Economy Edition

FOURTH EDITION

Introductory Methods of Numerical Analysis

S.S. Sastry

Presented By: http://www.ebooksuit.com

Copyrighted material

Introductory Methods of Numerical Analysis

Fourth Edition

S.S. SASTRY
Formerly, Scientist/Engineer SF
Vikram Sarabhai Space Centre
Trivandrum

Prentice-Hall of India Private Limited
New Delhi-110001
2006

This One
W2H6-ABZ-G08F
Presented By: http://www.ebooksuit.com

Rs. 195.00

INTRODUCTORY METHODS OF NUMERICAL ANALYSIS, 4th ed.
S.S. Sastry

© 2005 by Prentice-Hall of India Private Limited, New Delhi. All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publisher.

ISBN-81-203-2761-6

The export rights of this book are vested solely with the publisher.

Thirty-seventh Printing (Fourth Edition) ... ... August, 2006

Published by Asoke K. Ghosh, Prentice-Hall of India Private Limited, M-97, Connaught Circus, New Delhi-110001 and Printed by Rajkamal Electric Press, B-35/9, G.T. Kamal Road Industrial Area, Delhi-110033.

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Preface

This volume contains ten chapters on numerical methods which could be gainfully employed by scientists and engineers to solve the problems arising in research and industry. It also covers the syllabus prescribed for engineering studies at the undergraduate level. An essential feature of the present edition is that it provides information about the readily available computer program packages for implementing the numerical methods described in the book. These include references to MATLAB, IMSL and *Numerical Recipes* program libraries. Several problems have been set as exercises to illustrate the use of these libraries. Nevertheless, for a better understanding of these methods, the readers are advised to develop their own programs in any computer language of their choice.

More than two decades have elapsed since the first appearance of this book and, quite naturally, there have been many changes in the presentation of the material as also new additions of topics to meet the changing requirements of students in various universities. Thus, topics like curve fitting procedures, cubic spline methods, approximation of functions, numerical solution of integral equations, Graeffe's root-squaring method, weighted least-squares approximations, B-splines, Householder and QR methods, singular value decomposition, shooting method, the ADI method and the finite element method were gradually added to enhance the utility of the book. In the present edition, most sections have been rewritten to provide a better understanding of the topics. Thus the section on cubic splines has been rewritten with the inclusion of linear and quadratic splines, and a new section on surface fitting by cubic splines has been added. Similarly, a new section on Fourier transforms has also been included. The worked examples have been modified, new problems have been introduced and the number of

Presented By: http://www.ebooksuit.com

worked examples and homework problems has been significantly increased. This edition therefore contains about 500 problems including the illustrative examples and exercises for homework. Answers have been provided to some selected end-of-chapter exercises.

The author is very much obliged to the students and teachers of various universities who have been using this book during the last several years. Grateful thanks are due to Prof. I. Chandra Mohan, Sri Venkateswara University, Tirupati, for his suggestion to derive the general formulae in predictor-corrector methods in Chapter 7. Any suggestions towards the improvement of the book will be gratefully accepted. Special thanks are due to Sri Asoke K. Ghosh, Chairman and Managing Director, Prentice-Hall of India, New Delhi, for his courteous cooperation in bringing out this edition.

Chennai
February, 2005

S.S. SASTRY

Presented By: http://www.ebooksuit.com

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

Presented By: http://www.ebooksuit.com

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) $\frac{dy}{dx}, \frac{d^2y}{dx^2}, \ldots,$ 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 finite element method for solving 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

Presented By: http://www.ebooksuit.com

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

Presented By: http://www.ebooksuit.com

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

Presented By: http://www.ebooksuit.com

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

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).

Presented By: http://www.ebooksuit.com

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)*

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

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

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

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

Presented By: http://www.ebooksuit.com

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, ..., $\sqrt{2}$, π, 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

Presented By: http://www.ebooksuit.com

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

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

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

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

Presented By: http://www.ebooksuit.com

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,

$$E_A = aE_A^2 + bE_A^1 + E_A^1 E_A^2$$

$$= bE_A^1 + aE_A^2, \text{ approximately} \tag{1.6}$$

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

$$\frac{a + E_A^1}{b + E_A^2} - \frac{a}{b} = \frac{bE_A^1 - aE_A^2}{b(b + E_A^2)}$$

$$= \frac{bE_A^1 - aE_A^2}{b^2(1 + E_A^2/b)}$$

$$= \frac{bE_A^1 - aE_A^2}{b^2}, \text{ assuming that } E_A^2/b \text{ is small in comparison with 1}$$

$$= \frac{a}{b}\left(\frac{E_A^1}{a} - \frac{E_A^2}{b}\right). \tag{1.7}$$

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 $\Delta X = 0.005$, and the relative accuracy is given by $0.005/0.51 = 0.98\%$.

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

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

Presented By: http://www.ebooksuit.com

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 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

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

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 is 0.05. Hence, we round-off all the other number to two decimal digits. These are:

Presented By: http://www.ebooksuit.com

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

The Sum S is given by

$$S = 305.1 + 143.3 + 0.15 + 15.45 + 0.00 + 8.12 + 0.02 + 0.64 + 0.17$$

$$= 472.59$$

$$= 472.6$$

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

$$E_A = 2(0.05) + 7(0.005)$$

$$= 0.1 + 0.035$$

$$= 0.135$$

$$= 0.14$$

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,

$$S = 472.6 \pm 0.15.$$

Example 1.8 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 with the greatest absolute error is 2.5. Hence we round-off the second number to three significant digits, i.e. 48.3. Their product is given by

$$P = 48.3 \times 2.5 = 120.75 = 1.2 \times 10^2,$$

where we have 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

In this section, we derive a general formula for the error committed in using a certain formula or a functional relation. Let

$$u = f(x_1, x_2, \ldots, x_n) \tag{1.8}$$

be a function of several variables x_i $(i = 1, 2, \ldots, n)$, and let the error in each x_i be Δx_i. Then the error Δu in u is given by

$$u + \Delta u = f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n). \tag{1.9}$$

Expanding the right-hand-side by Taylor's series (see Theorem 1.9), we obtain

Presented By: http://www.ebooksuit.com

$$u + \Delta u = f(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i + \text{terms involving } (\Delta x_i)^2. \quad (1.10)$$

Assuming that the errors in x_i are small and that $(\Delta x_i)/x_i << 1$, so that the squares and higher powers of Δx_i can be neglected, the above relation yields

$$\Delta u \approx \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i = \frac{\partial f}{\partial x_1} \Delta x_1 + \frac{\partial f}{\partial x_2} \Delta x_2 + \cdots + \frac{\partial f}{\partial x_n} \Delta x_n \quad (1.11)$$

We observe that this formula has the same form as that for the total differential of u. The formula for the relative error follows immediately:

$$E_R = \frac{\Delta u}{u} = \frac{\partial u}{\partial x_1} \frac{\Delta x_1}{u} + \frac{\partial u}{\partial x_2} \frac{\Delta x_2}{u} + \cdots + \frac{\partial u}{\partial x_n} \frac{\Delta x_n}{u} \quad (1.12)$$

The following example illustrates the use of this formula.

Example 1.9 Let $u = 5xy^2/z^3$.

Then

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

and

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

In general, the errors $\Delta x, \Delta y$ and Δz may be positive or negative, and hence we take the absolute values of the terms on the right side. This gives

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

Now, let $\Delta x = \Delta y = \Delta z = 0.001$ 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 truncation 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

Presented By: http://www.ebooksuit.com

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

where

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

In (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 (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

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

Eq. (1.13) may be written as

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

where $O(h^{n+1})$ 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:

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

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

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

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 1.10 Evalute $f(1)$ 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.

Presented By: http://www.ebooksuit.com

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

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

$f^{iv}(x)$ 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 (i), the zero-order approximation is given by

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

and therefore

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

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

For the first approximation, we have

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

and therefore

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

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

Again, the second-order Taylor approximation is given by

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

and therefore

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

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

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

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

Presented By: http://www.ebooksuit.com

and therefore

$$f(1) = f(0) + hf'(x_0) + \frac{h^2}{2} f''(x_0) + \frac{h^3}{6} f'''(x_0)$$

$$\approx -10 + 5 + \frac{1}{2}(-6) + \frac{1}{6}(6)$$

$$= -7,$$

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 1.11 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

$$f\left(\frac{\pi}{3}\right) = f\left(\frac{\pi}{6}\right) + hf'\left(\frac{\pi}{6}\right) + \frac{h^2}{2} f''\left(\frac{\pi}{6}\right) + \frac{h^3}{6} f'''\left(\frac{\pi}{6}\right) + \frac{h^4}{24} f^{iv}\left(\frac{\pi}{6}\right)$$

$$+ \frac{h^5}{120} f^{v}\left(\frac{\pi}{6}\right) + \frac{h^6}{720} f^{vi}\left(\frac{\pi}{6}\right) + \frac{h^7}{5040} f^{vii}\left(\frac{\pi}{6}\right) + O(h^8) \qquad \text{(i)}$$

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

$$\sin\left(\frac{\pi}{3}\right) \approx \sin\left(\frac{\pi}{6}\right) + \frac{\pi}{6}\cos\left(\frac{\pi}{6}\right) + \frac{1}{2}\left(\frac{\pi}{6}\right)^2\left(-\sin\frac{\pi}{6}\right) + \frac{1}{6}\left(\frac{\pi}{6}\right)^3\left(-\cos\frac{\pi}{6}\right)$$

$$+ \frac{1}{24}\left(\frac{\pi}{6}\right)^4\left(\sin\frac{\pi}{6}\right) + \frac{1}{120}\left(\frac{\pi}{6}\right)^5\left(\cos\frac{\pi}{6}\right) + \frac{1}{720}\left(\frac{\pi}{6}\right)^6\left(-\sin\frac{\pi}{6}\right)$$

$$+ \frac{1}{5040}\left(\frac{\pi}{6}\right)^7\left(-\cos\frac{\pi}{6}\right).$$

$$= 0.5 + \frac{\pi}{12}\sqrt{3} - \frac{1}{4}\frac{\pi^2}{36} - \frac{\sqrt{3}}{12}\left(\frac{\pi}{6}\right)^3 + \frac{1}{48}\left(\frac{\pi}{6}\right)^4 + \frac{\sqrt{3}}{240}\left(\frac{\pi}{6}\right)^5 - \frac{1}{1440}\left(\frac{\pi}{6}\right)^6$$

$$- \frac{\sqrt{3}}{10080}\left(\frac{\pi}{6}\right)^7.$$

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

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

Presented By: http://www.ebooksuit.com

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 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

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

Presented By: http://www.ebooksuit.com

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 1.12 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.

EXERCISES

1.1. 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.

Presented By: http://www.ebooksuit.com

1.4. If $u = 3v^7 - 6v$, find the percentage error in u at $v = 1$, if the error in v is 0.05.

1.5. Define the term *absolute error*. Given that

$$a = 10.00 \pm 0.05$$

$$b = 0.0356 \pm 0.0002$$

$$c = 15300 \pm 100$$

$$d = 62000 \pm 500,$$

find the maximum value of the absolute error in

(a) $a + b + c + d$ (b) $a + 5c - d$ and (c) c^3.

1.6. Obtain the range of values within which the exact value of

$$\frac{1.265(10.21 - 7.54)}{47}$$

lies, if all the numerical quantities are rounded-off.

1.7. What is meant by *absolute* and *relative errors*? If

$$y = \frac{0.31x + 2.73}{x + 0.35},$$

where the coefficients are rounded-off, find the absolute and relative errors in y when $x = 0.5 \pm 0.1$.

1.8. Find the sum of the numbers 105.5, 27.25, 6.56, 0.1568, 0.000256, 208.6, 0.0235, 0.538 and 0.0571, where each number is correct to the digits given. Estimate the absolute error in the sum.

1.9. Find the product of the numbers 56.54 and 12.4 which are both correct to the significant digits given.

1.10. Find the quotient $q = x/y$, where $x = 4.536$ and $y = 1.32$, both x and y being correct to the digits given. Find also the relative error in the result.

1.11. Prove that the relative error of a product of three non-zero numbers does not exceed the sum of the relative errors of the given numbers.

1.12. Find the number of terms of the exponential series such that their sum gives the value of e^x correct to five decimal places for all values of x in the range $0 \le x \le 1$.

1.13. The function $f(x) = \tan^{-1}x$ can be expanded as

$$\tan^{-1}x = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^{n-1}\frac{x^{2n-1}}{2n-1} + \cdots$$

Find n such that the series determines $\tan^{-1}1$ correct to eight significant digits.

Presented By: http://www.ebooksuit.com

1.14. Derive the series

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

and use it to compute the value of $\log_e(1.2)$ correct to seven decimal places. Determine the number of terms required if the series for $\log_e(1+x)$ were used instead.

1.15. Write down the Taylor series expansion of $f(x)=\cos x$ at $x=\pi/3$ in terms of $f(x)$ and its derivatives at $x=\pi/4$. Compute the approximations from the zero-order to the fifth order and also state the absolute error in each case.

1.16. 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 $\sin 25°$ with an accuracy of 0.001.

Presented By: http://www.ebooksuit.com

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 otherhand, 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 (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 (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.

Presented By: http://www.ebooksuit.com

2.2 THE 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

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

Inequality (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:

Presented By: http://www.ebooksuit.com

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

Presented By: http://www.ebooksuit.com

$$x_0 = \frac{2+3}{2} = 2.5$$

Since $f(x_0) = 5.6250$, we choose [2, 2.5] as the new interval. Then

$$x_1 = \frac{2+2.5}{2} = 2.25 \quad \text{and} \quad f(x_1) = 1.890625$$

Proceeding in this way, the following table is obtained.

n	a	b	x	$f(x)$
1	2	3	2.5	5.6250
2	2	2.5	2.25	1.8906
3	2	2.25	2.125	0.3457
4	2	2.125	2.0625	–0.3513
5	2.0625	2.125	2.09375	–0.0089
6	2.09375	2.125	2.10938	0.1668
7	2.09375	2.10938	2.10156	0.07856
8	2.09375	2.10156	2.09766	0.03471
9	2.09375	2.09766	2.09570	0.01286
10	2.09375	2.09570	2.09473	0.00195
11	2.09375	2.09473	2.09424	–0.0035
12	2.09424	2.09473		

At $n = 12$, it is seen that the difference between two successive iterates is 0.0005, which is less than 0.001. Thus this result agrees with condition given in (2.4).

Example 2.3 Find a positive root of the equation $xe^x = 1$, which lies between 0 and 1.

Let $f(x) = xe^x - 1$. Since $f(0) = -1$ and $f(1) = 1.718$, it follows that a root lies between 0 and 1. Thus, $x_0 = 0.5$. Since $f(0.5)$ is negative, it follows that the root lies between 0.5 and 1. Hence the new root is 0.75, i.e., $x_1 = 0.75$. Using the values of x_0 and x_1, we calculate ε_1:

$$\varepsilon_1 = \left| \frac{x_1 - x_0}{x_1} \right| \times 100 = 33.33\%.$$

Again, we find that $f(0.75)$ is positive and hence the root lies between 0.5 and 0.75, i.e. $x_2 = 0.625$. Now, the error is

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

Proceeding in this way, the following table is constructed where only the sign of the function value is indicated. The prescribed tolerance is 0.05%.

Presented By: http://www.ebooksuit.com

Iteration	a	b	x_r	Sign of $f(x_r)$	E_r (%)
1	0	1	0.5	negative	—
2	0.5	1	0.75	positive	33.33
3	0.5	0.75	0.625	positive	20.00
4	0.5	0.625	0.5625	negative	11.11
5	0.5625	0.625	0.59375	positive	5.263
6	0.5625	0.59375	0.5781	positive	2.707
7	0.5625	0.5781	0.5703	positive	1.368
8	0.5625	0.5703	0.5664	negative	0.688
9	0.5664	0.5703	0.5684	positive	0.352
10	0.5664	0.5684	0.5674	positive	0.176
11	0.5664	0.5674	0.5669	negative	0.088
12	0.5669	0.5674	0.5671	negative	0.035

Thus, after 12 iterations, the error, ε_r, finally satisfies the prescribed tolerance, viz., 0.05%. Hence the required root is 0.567 and it is easily seen that this value is correct to three decimal places.

2.3 THE 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 (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 (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

Presented By: http://www.ebooksuit.com

a graphical representation of the method. The error criterion (2.5) can be used in this case also.

Figure 2.2 Method of false position.

Example 2.4 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 2.5 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

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

Presented By: http://www.ebooksuit.com

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.

2.4 THE ITERATION METHOD

We have so far discussed root-finding methods, which require the interval in which the root lies. We now describe methods which require one or more starting values of x. 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 the roots 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

$$x^3 + x^2 - 1 = 0$$

can be expressed as either of the forms:

$$x = (1+x)^{-1/2}, \qquad x = (1-x^3)^{1/2}, \qquad x = (1-x^2)^{1/3}, \ldots.$$

Let x_0 be an approximate value of the desired root ξ. Substituting it for x on the right side of (2.8), we obtain the first approximation

$$x_1 = \phi(x_0)$$

The successive approximations are then given by

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

A number of questions now arise:

(i) Does the sequence of approximations $x_0, x_1, \ldots, x_n$, always converge to some number ξ?

(ii) If it does, will ξ be a root of the equation $x = \phi(x)$?

(iii) How should we choose ϕ in order that the sequence $x_0, x_1, \ldots, x_n$ converges to the root?

The answer to the first question is negative. As an example, we consider the equation

$$x = 10^x + 1.$$

Presented By: http://www.ebooksuit.com

If we take $x_0 = 0$, $x_1 = 2$, $x_2 = 101$, $x_3 = 10^{101} + 1$, etc., and as n increases, x_n increases without limit. Hence, the sequence $x_0, x_1, x_2, \ldots, x_n$ does not always converge and, in Theorem 2.1 below, we state the conditions which are sufficient for the convergence of the sequence.

The second question is easy to answer, for consider the equation

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

which gives the relation between the approximations at the nth and $(n+1)$th stages. As n increases, the left side tends to the root ξ, and if ϕ is continuous the right side tends to $\phi(\xi)$. Hence, in the limit, we have $\xi = \phi(\xi)$ which shows that ξ is a root of the equation $x = \phi(x)$.

The answer to the third question is contained in the following theorem:

Theorem 2.1 Let $x = \xi$ be a root of $f(x) = 0$ and let I be an interval containing the point $x = \xi$. Let $\phi(x)$ and $\phi'(x)$ be continuous in I, where $\phi(x)$ is defined by the equation $x = \phi(x)$ which is equivalent to $f(x) = 0$. Then if $|\phi'(x)| < 1$ for all x in I, the sequence of approximations $x_0, x_1, x_2, \ldots, x_n$ defined by (2.9) converges to the root ξ, provided that the initial approximation x_0 is chosen in I.

Proof Since ξ is a root of the equation $x = \phi(x)$, we have

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

From (2.9),

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

Subtraction gives

$$\xi - x_1 = \phi(\xi) - \phi(x_0)$$

By using the mean value theorem (see Theorem 1.3), the right-hand side can be written as $(\xi - x_0)\phi'(\xi_0)$, $x_0 < \xi_0 < \xi$. Hence we obtain

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

Similarly we obtain

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

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

$$\vdots$$

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

If we let

$$|\phi'(\xi_i)| \le k < 1, \quad \text{for all } i \tag{2.16}$$

Presented By: http://www.ebooksuit.com

then Eqs. (2.12)–(2.15) give

$$|\xi - x_1| \le |\xi - x_0|, \qquad |\xi - x_2| \le |\xi - x_1|, \ldots,$$

which show that each successive approximation remains in I provided that the initial approximation is chosen in I. Now, multiplying Eqs. (2.12) to (2.15) and simplifying, we obtain

$$\xi - x_{n+1} = (\xi - x_0)\,\phi'(\xi_0)\,\phi'(\xi_1)\ldots\phi'(\xi_n). \tag{2.17}$$

Since $|\phi'(\xi_i)| < k$, the above equation becomes

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

As $n \to \infty$, the right-hand side of (2.18) tends to zero, and it follows that the sequence of approximations $x_0, x_1, \ldots,$ converges to the root ξ if $k < 1$. The method can be represented graphically as follows. By sketching the line $y = x$ and the curve $y = \phi(x)$ and considering the way in which the approximations x_i are obtained, a geometrical significance of the method is obtained and this is shown in Figs. 2.3–2.6.

Figure 2.3 Convergence of $x_{n+1} = \phi(x_n)$, when $|\phi'(x)| < 1$.

Figure 2.4 $|\phi'(x)| < 1$ but $\phi'(x) < 0$. The process is convergent but the approximations oscillate about the exact value.

Presented By: http://www.ebooksuit.com

Figure 2.5 $\phi'(x) > 1$; the process is divergent.

Figure 2.6 $|\phi'(x)| > 1$; the process is divergent.

The root so obtained is *unique*. To prove this, let ξ_1 and ξ_2 be two roots of Eq. (2.8), i.e. let $\xi_1 = \phi(\xi_1)$ and $\xi_2 = \phi(\xi_2)$. Then we obtain

$$|\xi_1 - \xi_2| = |\phi(\xi_1) - \phi(\xi_2)| = |\phi'(\eta)||\xi_1 - \xi_2|, \qquad \eta \in (\xi_1, \xi_2)$$

which further simplifies to

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

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

Again,

$$\frac{d}{dx}[x - \phi(x)] = 1 - \phi'(x)$$

which is positive, since $\phi'(x) < 1$ in the interval I. This shows that the root obtained by this method is a *simple* root.

To estimate the error of the approximate root obtained, we have

$$\begin{aligned} |\xi - x_n| &= |\phi(\xi) - \phi(x_{n-1})| \le k|\xi - x_{n-1}| \\ &= k|\xi - x_n + x_n - x_{n-1}| \\ &\le k|\xi - x_n| + k|x_n - x_{n-1}|, \end{aligned}$$

Presented By: http://www.ebooksuit.com

which gives

$$|\xi - x_n| \le \frac{k}{1-k}|x_n - x_{n-1}| \le \frac{k^n}{1-k}|x_1 - x_0|. \tag{2.20}$$

In general, the speed of the iteration depends on the value of k; the smaller the value of k, the faster would be the convergence. If ε is the specified accuracy, i.e. if

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

then formula (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 iterates necessary to achieve a specified accuracy. The following examples illustrate the application of this method.

Example 2.6 Find a real root of the equation $x^3 + x^2 - 1 = 0$ on the interval $[0, 1]$ with an accuracy of 10^{-4}.

To find this root, we rewrite the given equation in the form

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

Thus

$$\phi(x) = \frac{1}{\sqrt{x+1}}, \qquad \phi'(x) = -\frac{1}{2}\frac{1}{(x+1)^{3/2}}$$

and

$$\max_{[0,1]} |\phi'(x)| = \frac{1}{2\sqrt{8}} = k = 0.17678 < 0.2.$$

Using (2.21) we then obtain

$$|x_n - x_{n-1}| < \frac{0.0001 \times 0.8}{0.2} = 0.0004.$$

Hence when the absolute value of the difference does not exceed 0.0004, the required accuracy will be achieved and then the iteration can be terminated.

Starting with $x_0 = 0.75$, we obtain the following table:

n	x_n	$\sqrt{x_n+1}$	$x_{n+1} = 1/\sqrt{x_n+1}$
0	0.75	1.3228756	0.7559289
1	0.7559289	1.3251146	0.7546517
2	0.7546517	1.3246326	0.7549263

Presented By: http://www.ebooksuit.com

At this stage, we find that

$$|x_{n+1} - x_n| = 0.7549263 - 0.7546517 = 0.0002746,$$

which is less than 0.0004. The iteration is therefore terminated and the root to the required accuracy is 0.7549.

Example 2.7 Find the root of the equation $2x = \cos x + 3$ correct to three decimal places.

We rewrite the equation in the form

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

so that

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

and

$$|\phi'(x)| = \left|\frac{\sin x}{2}\right| < 1.$$

Hence the iteration method can be applied to the eq. (i) and we start with $x_0 = \pi/2$. The successive iterates are

$$x_1 = 1.5, \qquad x_2 = 1.535, \qquad x_3 = 1.518,$$
$$x_4 = 1.526, \qquad x_5 = 1.522, \qquad x_6 = 1.524,$$
$$x_7 = 1.523, \qquad x_8 = 1.524.$$

Hence we take the solution as 1.524 correct to three decimal places.

Example 2.8 Use the method of iteration to find a positive root, between 0 and 1, of the equation $xe^x = 1$.

Writing the equation in the form

$$x = e^{-x} \tag{i}$$

We find that $\phi(x) = e^{-x}$ and so $\phi'(x) = -e^{-x}$.

Hence $|\phi'(x)| < 1$ for $x < 1$, which assures that the iterative process defined by the equation $x_{n+1} = \phi(x_n)$ will be convergent.

Starting with $x_0 = 1$, we find that the successive iterates are given by

$$x_1 = 1/e = 0.3678794, \qquad x_2 = 0.6922006,$$
$$x_3 = 0.5004735, \qquad x_4 = 0.6062435,$$
$$x_5 = 0.5453957, \qquad x_6 = 0.5796123,$$
$$x_7 = 0.5601154, \qquad x_8 = 0.5711431,$$
$$x_9 = 0.5648793, \qquad x_{10} = 0.5684287,$$

Presented By: http://www.ebooksuit.com

$x_{11} = 0.5664147,$ $x_{12} = 0.5675566,$

$x_{13} = 0.5669089,$ $x_{14} = 0.5672762,$

$x_{15} = 0.5670679,$ $x_{16} = 0.567186,$

$x_{17} = 0.567119,$ $x_{18} = 0.567157,$

$x_{19} = 0.5671354,$ $x_{20} = 0.5671477.$

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 Section 2.4, 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 (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.

Presented By: http://www.ebooksuit.com

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

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

Example 2.9 We consider again Example 2.7, 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

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

Presented By: http://www.ebooksuit.com

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

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

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

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

which is the *Newton–Raphson formula*.

If we compare Eq. (2.24) 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 by Theorem 2.1, the Newton–Raphson formula (2.24) 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 formula (2.24). 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

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

Presented By: http://www.ebooksuit.com

From (2.24) 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. 2.7. It can be used for solving both algebraic and transcendental equations and it can also be used when the roots are complex.

Figure 2.7 Newton–Raphson method.

Example 2.10 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. (2.24) 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 (i), we obtain

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

Presented By: http://www.ebooksuit.com

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 2.11 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 2.12 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

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

Presented By: http://www.ebooksuit.com

so that

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

Proceeding in this way, we obtain

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

Generalized Newton's method

If ξ is a root of $f(x) = 0$ with multiplicity p, then the iteration formula corresponding to (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 (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 2.13 Find a double root of the equation $f(x) = x^3 - x^2 - x + 1 = 0$.

Choosing $x_0 = 0.8$, we have

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

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

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

Presented By: http://www.ebooksuit.com

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 method* which can be used to determine the smallest root of the equation

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

where $f(x)$ is of the form

$$f(x) = 1 - (a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \cdots). \tag{2.32}$$

For smaller values of x, we can write

$$[1 - (a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \cdots)]^{-1} = b_1 + b_2x + b_3x^2 + \cdots \tag{2.33}$$

Expanding the left-hand side by binomial theorem, we obtain

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

Comparing the coefficients of like powers of x on both sides of (2.34), we get

$$\left.\begin{aligned} b_1 &= 1, \\ b_2 &= a_1 = a_1b_1, \\ b_3 &= a_1^2 + a_2 = a_1b_2 + a_2b_1, \\ &\vdots \\ b_n &= a_1b_{n-1} + a_2b_{n-2} + \cdots + a_{n-1}b_1 \qquad n = 2, 3, \ldots \end{aligned}\right\} \tag{2.35}$$

Without any proof, Ramanujan states that the successive convergents, viz., b_n/b_{n+1}, approach a root of the equation $f(x) = 0$, where $f(x)$ is given by (2.32). The following examples illustrate the application of this method.

Example 2.14 Find the smallest root of the equation

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

*See Berndt [1985], p.41.

Presented By: http://www.ebooksuit.com

We have

$$\left(1-\frac{11x-6x^2+x^3}{6}\right)^{-1}=b_1+b_2x+b_3x^2+\cdots \qquad \text{(ii)}$$

Here

$$a_1=\frac{11}{6}, \qquad a_2=-1, \qquad a_3=\frac{1}{6}, \qquad a_4=a_5=\cdots=0$$

Hence,

$$b_1=1;$$

$$b_2=a_1=\frac{11}{6};$$

$$b_3=a_1b_2+a_2b_1=\frac{121}{36}-1=\frac{85}{36};$$

$$b_4=a_1b_3+a_2b_2+a_3b_1=\frac{575}{216};$$

$$b_5=a_1b_4+a_2b_3+a_3b_2+a_4b_1=\frac{3661}{1296};$$

$$b_6=a_1b_5+a_2b_4+a_3b_3+a_4b_2+a_5b_1=\frac{22631}{7776};$$

Therefore,

$$\frac{b_1}{b_2}=\frac{6}{11}=0.54545$$

$$\frac{b_2}{b_3}=\frac{66}{85}=0.7764705$$

$$\frac{b_3}{b_4}=\frac{102}{115}=0.8869565$$

$$\frac{b_4}{b_5}=\frac{3450}{3661}=0.9423654$$

$$\frac{b_5}{b_6}=\frac{3138}{3233}=0.9706155$$

The smallest root of the given equation is unity and it can be seen that the successive convergents approach this root.

Presented By: http://www.ebooksuit.com

Example 2.15 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

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

We then have

$$b_1 = 1;$$

$$b_2 = a_2 = 1;$$

$$b_3 = a_1 b_2 + a_2 b_1 = 1 + 1 = 2;$$

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

$$b_5 = a_1 b_4 + a_2 b_3 + a_3 b_2 + a_4 b_1 = \frac{7}{2} + 2 + \frac{1}{2} + \frac{1}{6} = \frac{37}{6};$$

$$b_6 = a_1 b_5 + a_2 b_4 + a_3 b_3 + a_4 b_2 + a_5 b_1 = \frac{37}{6} + \frac{7}{2} + 1 + \frac{1}{6} + \frac{1}{24} = \frac{261}{24};$$

Therefore,

$$\frac{b_2}{b_3} = \frac{1}{2} = 0.5$$

$$\frac{b_3}{b_4} = \frac{4}{7} = 0.5714$$

$$\frac{b_4}{b_5} = \frac{21}{37} = 0.56756756$$

$$\frac{b_5}{b_6} = \frac{148}{261} = 0.56704980$$

It can be seen that Newton's method (see Example 2.12) gives the value 0.5671433 to this root.

Example 2.16 Find a root of the equation $\sin x = 1 - x$.

Using the expansion of $\sin x$, the given equation may be written as

Presented By: http://www.ebooksuit.com

$$x + x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = 1.$$

Hence we write

$$\left[1 - \left(2x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \cdots\right)\right]^{-1} = b_1 + b_2 x + b_3 x^2 + \cdots$$

Here

$$a_1 = 2, \qquad a_2 = 0, \qquad a_3 = -\frac{1}{6}, \qquad a_4 = 0,$$

$$a_5 = \frac{1}{120}, \qquad a_6 = 0, \qquad a_7 = -\frac{1}{5040}, \cdots$$

We then obtain

$$b_1 = 1;$$

$$b_2 = a_1 = 2;$$

$$b_3 = a_1 b_2 + a_2 b_1 = 4;$$

$$b_4 = a_1 b_3 + a_2 b_2 + a_3 b_1 = 8 - \frac{1}{6} = \frac{47}{6};$$

$$b_5 = a_1 b_4 + a_2 b_3 + a_3 b_2 + a_4 b_1 = \frac{46}{3};$$

$$b_6 = a_1 b_5 + a_2 b_4 + a_3 b_3 + a_4 b_2 + a_5 b_1 = \frac{3601}{120};$$

Therefore,

$$\frac{b_1}{b_2} = \frac{1}{2};$$

$$\frac{b_2}{b_3} = \frac{1}{2};$$

$$\frac{b_3}{b_4} = \frac{24}{27} = 0.5106382$$

$$\frac{b_4}{b_5} = \frac{47}{92} = 0.5108695$$

$$\frac{b_5}{b_6} = \frac{1840}{3601} = 0.5109691$$

The root, correct to four decimal places, is 0.5110.

Presented By: http://www.ebooksuit.com

Example 2.17 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

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

To apply Ramanujan's method, we write

$$\left\{1-\left[x-\frac{x^2}{(2!)}+\frac{x^3}{(3!)^2}-\frac{x^4}{(4!)^2}+\cdots\right]\right\}^{-1}=b_1+b_2x+b_3x^2+\cdots \qquad \text{(ii)}$$

Here

$$a_1=1, \qquad a_2=-\frac{1}{(2!)^2}, \qquad a_3=\frac{1}{(3!)^2},$$

$$a_4=-\frac{1}{(4!)^2}, \qquad a_5=\frac{1}{(5!)^2}, \qquad a_6=-\frac{1}{(6!)^2},\cdots$$

Hence, we obtain

$$b_1=1,$$

$$b_2=a_1=1,$$

$$b_3=a_1b_2+a_2b_1=1-\frac{1}{(2!)^2}=\frac{3}{4};$$

$$b_4=a_1b_3+a_2b_2+a_3b_1$$

$$=\frac{3}{4}-\frac{1}{(2!)^2}+\frac{1}{(3!)^2}=\frac{3}{4}-\frac{1}{4}+\frac{1}{36}$$

$$=\frac{19}{36},$$

$$b_5=a_1b_4+a_2b_3+a_3b_2+a_4b_1$$

$$=\frac{19}{36}-\frac{1}{4}\times\frac{3}{4}+\frac{1}{36}\times 1-\frac{1}{576}$$

$$=\frac{211}{576},$$

$$\vdots$$

Presented By: http://www.ebooksuit.com

It follows

$$\frac{b_1}{b_2} = 1;$$

$$\frac{b_2}{b_3} = \frac{4}{3} = 1.333\ldots;$$

$$\frac{b_3}{b_4} = \frac{3}{4} \times \frac{36}{19} = \frac{27}{19} = 1.4210\ldots,$$

$$\frac{b_4}{b_5} = \frac{19}{36} \times \frac{576}{211} = 1.4408\ldots,$$

where the last result is correct to three significant figures.

This example demonstrates that Ramanujan's method is preferable when the given function consists of an infinite series.

2.7 THE 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 2.18 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

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

Presented By: http://www.ebooksuit.com

Putting $i = 0$ in (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 (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 (2.37), and simplifying, we get $x_3 = 2.094824145$, which is correct to three significant figures.

2.8 MULLER'S METHOD

In this method, $f(x)$ 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, converges almost quadratically, and can be used to obtain complex roots.

Let x_{i-2}, x_{i-1}, x_i be three distinct approximations to a root of $f(x) = 0$. Let y_{i-2}, y_{i-1}, y_i be the corresponding values of $y = f(x)$. Now, any second-degree curve passing through the point (x_i, y_i) may be written as

$$p(x) = A(x - x_i)^2 + B(x - x_i) + y_i. \tag{2.38}$$

Since the curve also passes through (x_{i-2}, y_{i-2}) and (x_{i-1}, y_{i-1}), we obtain

$$y_{i-2} = A(x_{i-2} - x_i)^2 + B(x_{i-2} - x_i) + y_i, \tag{2.39}$$

and

$$y_{i-1} = A(x_{i-1} - x_i)^2 + B(x_{i-1} - x_i) + y_i. \tag{2.40}$$

Solving (2.39) and (2.40), we get

$$A = \frac{y_{i-1} - y_i}{(x_{i-1} - x_{i-2})(x_{i-1} - x_i)} + \frac{y_{i-2} - y_i}{(x_{i-2} - x_i)(x_{i-2} - x_{i-1})} \tag{2.41}$$

and

$$B = \frac{y_{i-1} - y_i}{x_{i-1} - x_i} - A(x_{i-1} - x_i) \tag{2.42}$$

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

Presented By: http://www.ebooksuit.com

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

A direct solution from (2.43) leads to inaccurate results and therefore we take the equivalent form :

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

In (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) then gives the next approximation to the root. The following example illustrates the application of this method.

Example 2.19 Find the root of the equation $y(x) = x^3 - 2x - 5 = 0$, which lies between 2 and 3.

Let $x_{i-2} = 1$, $x_{i-1} = 2$ and $x_i = 3$. Then $y_{i-2} = -6$, $y_{i-1} = -1$ and $y_i = 16$. Hence

$$A = \frac{-17}{1(-1)} + \frac{-22}{(-2)(-1)} = 17 - 11 = 6,$$

and

$$B = \frac{-17}{-1} - 6(-1) = 23.$$

The quadratic equation is given by

$$6(x-3)^2 + 23(x-3) + 16 = 0,$$

which gives the next approximation:

$$x_{i+1} = 3 - \frac{2(16)}{23 + \sqrt{(23)^2 - 4(6)(16)}},$$

where the positive sign is chosen since B is positive.

Hence

$$x_{i+1} = 3 - \frac{32}{23 + \sqrt{145}} = 2.086799548.$$

The error in the above result is given by

$$\left| \frac{2.086799548 - 3}{2.086799548} \right| \times 100\%,$$

which simplifies to 43.76%. Since this is quite large, we proceed to the next iteration with

$$x_{i-2} = 1, \qquad x_{i-1} = 2, \qquad x_i = 2.086799548.$$

Presented By: http://www.ebooksuit.com

The corresponding values of y are

$$y_{i-2} = -6, \qquad y_{i-1} = -1, \qquad y_i = -0.086145588.$$

Using these values in Eqs. (2.41) and (2.42), we obtain

$$A = 5.086799558 \quad \text{and} \quad B = 10.96986336.$$

The next approximation x_{i+1} is obtained as

$$x_{i+1} = 2.086799548 + \frac{2(0.086145588)}{22.01933047} = 2.09462409.$$

The error in this result is 0.3735535 19%.

2.9 GRAEFFE'S ROOT-SQUARING METHOD

This is another method usually recommended for the numerical solution of polynomial equations. It, however, suffers from the disadvantage of having a numerically complicated procedure.

Let $P_n(x)$ be a polynomial of degree n. Graeffe's method consists in transforming $P_n(x)$ into another polynomial, say $Q_n(z)$, of the same degree but whose roots are the squares of the roots of the original polynomial. The process is repeated so that the roots of the new polynomial are distributed more spaciously. This is possible provided that the roots of the original polynomial are all real and distinct. The roots are finally computed directly from the coefficients. We consider, for example,

$$P_3(x) = (x+1)(x-2)(x+3) \tag{2.45}$$

Then

$$P_3(-x) = (-x+1)(-x-2)(-x+3) \tag{2.46}$$

$$= (-1)^3 (x-1)(x+2)(x-3). \tag{2.47}$$

Hence,

$$P_3(x)\, P_3(-x) = (-1)^3 (x^2-1)(x^2-4)(x^2-9). \tag{2.48}$$

Thus, we obtain

$$Q_3(z) = (z-1)(z-4)(z-9), \tag{2.49}$$

where

$$z = x^2.$$

It is seen that the roots of (2.49) are the squares of the roots of the original polynomial $P_3(x)$.

Again,

$$Q_3(-z) = (-z-1)(-z-4)(-z-9) \tag{2.51}$$

$$= (-1)^3 (z+1)(z+4)(z+9). \tag{2.52}$$

Presented By: http://www.ebooksuit.com

Hence

$$Q_3(z)\,Q_3(-z) = (-1)^3\,(z^2-1)\,(z^2-16)\,(z^2-81) \tag{2.53}$$

The next new polynomial is therefore given by

$$S_3(u) = (u-1)\,(u-16)\,(u-81), \tag{2.54a}$$

where

$$u = z^2. \tag{2.54b}$$

Suppose that the above procedure is repeated m times (i.e. the squaring is done m times successively); then the roots are estimated from the formula:

$$\left|\frac{a_i}{a_{i-1}}\right|^{1/m}, \ldots \qquad (i = 1, 2, \ldots, n), \tag{2.55}$$

where the a_i's are the coefficients of the new polynomial (obtained after squaring m times) and n is the degree of the original polynomial.

It is clear from the above discussion that this method gives approximations to the magnitudes of the roots. To determine the signs of the roots, we substitute each 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 method of application is illustrated below.

Example 2.20 Find the real roots of the equation $x^3 - 6x^2 + 11x - 6 = 0$.

Let

$$P_3(x) = x^3 - 6x^2 + 11x - 6. \tag{i}$$

Then

$$\begin{aligned} P_3(-x) &= (-x)^3 - 6(-x)^2 + 11(-x) - 6 \\ &= (-1)^3\,(x^3 + 6x^2 + 11x + 6). \end{aligned} \tag{ii}$$

Hence

$$P_3(x)\,P_3(-x) = (-1)^3\,(x^6 - 14x^4 + 49x^2 - 36). \tag{iii}$$

Let

$$Q_3(z) = z^3 - 14z^2 + 49z - 36, \tag{iv}$$

where $z = x^2$. Therefore, roots of $P_3(x) = 0$ are given by

$$\sqrt{\frac{36}{49}} = 0.857, \qquad \sqrt{\frac{49}{14}} = 1.871, \qquad \sqrt{14} = 3.741. \tag{v}$$

Again

$$Q_3(-z) = (-1)^3\,(z^3 + 14z^2 + 49z + 36). \tag{vi}$$

Presented By: http://www.ebooksuit.com

Hence,

$$Q_3(z)\,Q_3(-z) = (-1)^3\,(z^6 - 98z^4 + 1393z^2 - 1296). \qquad \text{(vii)}$$

It follows that

$$S_3(u) = u^3 - 98u^2 + 1393u - 1296, \qquad \text{(viii)}$$

where $u = z^2 = x^4$. From the roots of the new polynomial $S_3(u) = 0$, we obtain the approximation to the roots of $P_3(x) = 0$ as

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

The convergence to the exact roots, 1, 2 and 3 is quite clear.

2.10 LIN–BAIRSTOW'S METHOD

A method which is often useful in finding quadratic factors of polynomials is *Lin-Bairstow*'s method and this is briefly described below:

Let the polynomial be given by

$$f(x) = A_3x^3 + A_2x^2 + A_1x + A_0 = 0. \qquad (2.56)$$

Let $x^2 + Rx + S$ be a quadratic factor of $f(x)$ and also let an approximate factor be $x^2 + rs + s$. Usually, first approximations of r and s can be obtained from the last-three terms of the given polynomial. Thus,

$$r = \frac{A_1}{A_2} \quad \text{and} \quad s = \frac{A_0}{A_2} \qquad (2.57)$$

Let

$$\begin{aligned} f(x) &= (x^2 + rx + s)(B_2x + B_1) + Cx + D \\ &= B_2x^3 + (B_2r + B_1)x^2 + (C + B_1r + sB_2)x + (B_1s + D), \end{aligned} \qquad (2.58)$$

where the constants B_1, B_2, C and D have to be determined. Equating the coefficients of the like powers of x in Eqs. (2.56) and (2.58), we obtain

$$\left.\begin{aligned} B_2 &= A_3 \\ B_1 &= A_2 - rB_2 \\ C &= A_1 - rB_1 - sB_2 \\ D &= A_0 - sB_1 \end{aligned}\right\} \qquad (2.59)$$

From (2.59), it is clear that the coefficients B of the factored polynomial and also the coefficients C and D are functions of r and s. Since $x^2 + Rx + S$ is a factor of the given polynomial, it follows that

Presented By: http://www.ebooksuit.com

$$C(R, S) = 0 \quad \text{and} \quad D(R, S) = 0 \tag{2.60}$$

Letting

$$R = r + \Delta r \quad \text{and} \quad S = s + \Delta s \tag{2.61}$$

Eqs. (2.60) can be expanded by Taylor's series

$$\left.\begin{aligned} C(R, S) &= C(r, s) + \Delta r \frac{\partial C}{\partial r} + \Delta s \frac{\partial C}{\partial s} = 0 \\ D(R, S) &= D(r, s) + \Delta r \frac{\partial D}{\partial r} + \Delta s \frac{\partial D}{\partial s} = 0, \end{aligned}\right\} \tag{2.62}$$

where the derivatives are to be computed at r and s.

Equations (2.62) can then be solved for Δr and Δs. Use of these values in (2.61) will give the next approximation to R and S, respectively. The process can be repeated until successive values of R and S show no significant change. The following example illustrates the application of this method.

Example 2.21 Find the quadratic factor of the polynomial given by $f(x) = x^3 - 2x^2 + x - 2$.

We have $A_3 = 1$, $A_2 = -2$, $A_1 = 1$ and $A_0 = -2$. It is easily seen that $r = -1/2$ and $s = 1$. Equations (2.59) give

$$B_2 = 1$$

$$B_1 = -2 - r;$$

$$C = 1 - r(-2 - r) - s = 1 + 2r + r^2 - s,$$

$$D = -2 - s(-2 - r) = -2 + 2s + rs.$$

Also,

$$[C(r, s)]_{(-1/2, 1)} = 1 - 1 + \frac{1}{4} - 1 = -\frac{3}{4};$$

$$[D(r, s)]_{(-1/2, 1)} = -2 + 2 - \frac{1}{2} = -\frac{1}{2};$$

$$\left(\frac{\partial C}{\partial r}\right)_{(-1/2, 1)} = 2 + 2r = 1;$$

$$\left(\frac{\partial C}{\partial s}\right)_{(-1/2, 1)} = -1;$$

$$\left(\frac{\partial D}{\partial r}\right)_{(-1/2, 1)} = s = 1;$$

$$\left(\frac{\partial D}{\partial s}\right)_{(-1/2, 1)} = 2 + r = \frac{3}{2}.$$

Presented By: http://www.ebooksuit.com

Equations (2.62) give:

$$\Delta r - \Delta s = \frac{3}{4} \quad \text{and} \quad \Delta r + \frac{3}{2}\Delta s = \frac{1}{2}.$$

Hence

$$\Delta r = \frac{13}{20} \quad \text{and} \quad \Delta s = -\frac{1}{10}.$$

Thus

$$R = -\frac{1}{2} + \frac{13}{20} = \frac{3}{20} = 0.15$$

and

$$S = 1 - \frac{1}{10} = \frac{9}{10} = 0.9.$$

It follows that the quadratic factor is $x^2 + 0.15x + 0.9$. Thus, for the second approximation, the approximate quadratic factor is $x^2 + 0.15x + 0.9$ so that $r = 0.15$ and $s = 0.9$. Now,

$$C = 1 + 2.15(0.15) - 0.9 = 0.4225,$$

$$D = -2 + 2.15(0.9) = 0.065,$$

$$\partial C/\partial r = 2 + 2(0.15) = 2.30,$$

$$\partial C/\partial s = -1,$$

$$\partial D/\partial r = 0.9,$$

$$\partial D/\partial s = 2 + r = 2.15.$$

Hence, Eq. (2.62) give

$$2.3\Delta r - \Delta s = -0.4225 \quad \text{and} \quad 0.9\Delta r + 2.15\Delta s = -0.065$$

Solving the above equations, we obtain

$$\Delta r = -0.1665312, \qquad \Delta s = 0.0394783.$$

The second approximations are therefore given by

$$R = 0.15 - 0.1665312 = -0.0165312$$

$$S = 0.9 + 0.0394783 = 0.9394783$$

Thus the second approximation to the quadratic factor of $f(x)$ is $x^2 - 0.0165312x + 0.9394783$, and it can be seen that R and S are approaching their actual values 0 and 1, respectively.

Presented By: http://www.ebooksuit.com

2.11 THE QUOTIENT-DIFFERENCE METHOD

This is a general method to obtain the approximate roots of polynomial equations and is originally due to Rutishauser [1954]. We adopt here the notation and methodology described in Henrici [1958]. The procedure is quite general and is illustrated here with a cubic polynomial. Let the given cubic equation be

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

and let x_1, x_2 and x_3 be its roots such that $0 < |x_1| < |x_2| < |x_3|$. Then $f(x)$ can be expressed in the following way:

$$\frac{1}{f(x)} = \sum_{r=1}^{3} \frac{p_r}{x - x_r} = \sum_{r=1}^{\infty} -p_r \left(\frac{1}{x_r} + \frac{x}{x_r^2} + \frac{x^2}{x_r^3} + \cdots \right)$$

$$= \sum_{i=0}^{\infty} \alpha_i x_i, \tag{2.64}$$

where

$$\alpha_i = -\sum_{r=1}^{3} \frac{p_r}{x_r^{i+1}}. \tag{2.65}$$

The method derives its name to the *quotients* $q^{(i)}$ and the *differences* $\Delta^{(i)}$ defined by the relations

$$q_1^{(i)} = \frac{\alpha_i}{\alpha_{i-1}} \tag{2.66}$$

and

$$\Delta_1^{(i)} = q_1^{(i+1)} - q_1^{(i)} \tag{2.67}$$

We have already seen that

$$\lim_{i \to \infty} \frac{\alpha_{i-1}}{\alpha_i} \tag{2.68}$$

tends to the smallest root of the equation $f(x) = 0$ (see, Section 2.6). The quotient-difference method is therefore an extension of Ramanujan's method and determines all the real roots of a polynomial equation. Using the definitions (2.66) and (2.67), starting values of $q_1^{(i)}$ and $\Delta_1^{(i)}$ can be found and these are used to generate a table of quotients and differences from the general formulae:

$$\Delta_r^{(i)} q_{r+1}^{(i)} = \Delta_r^{(i+1)} q_r^{(i+1)} \tag{2.69}$$

Presented By: http://www.ebooksuit.com

and

$$\Delta_r^{(i)} + q_r^{(i)} = \Delta_{r-1}^{(i+1)} + q_r^{(i+1)} \tag{2.70}$$

$\Delta_0^{(i)} = \Delta_n^{(i)} = 0$, for all i, n being the degree of the polynomial. These formulae can easily be established using the definitions and their proofs are left as exercises to the reader (*see*, Henrici [1974]). A typical quotient-difference, table is given in Table 2.1.

Table 2.1 A Typical Quotient-difference Table

Δ_0	q_1	Δ_1	q_2	Δ_2	q_3	Δ_3
	$q_1^{(0)}$		$q_2^{(-1)}$		$q_3^{(-2)}$	
$\Delta_0^{(1)}$		$\Delta_1^{(0)}$		$\Delta_2^{(-1)}$		$\Delta_3^{(-2)}$
	$q_1^{(1)}$		$q_2^{(0)}$		$q_3^{(-1)}$	
$\Delta_0^{(2)}$		$\Delta_1^{(1)}$		$\Delta_2^{(0)}$		$\Delta_3^{(-1)}$
	$q_1^{(2)}$		$q_2^{(1)}$		$q_3^{(0)}$	
$\Delta_0^{(3)}$		$\Delta_1^{(2)}$		$\Delta_2^{(1)}$		$\Delta_3^{(0)}$

If the first-two rows are known, the succeeding rows can be generated using the formulae (2.69) and (2.70) alternately. To determine the differences $\Delta_r^{(i+1)}$, formula (2.69) is used whereas formula (2.70) determines the quotients $q_r^{(i+1)}$. As the building-up of Table 2.5 proceeds, the quantities $q_1^{(i)}$, $q_2^{(i)}$ and $q_3^{(i)}$, tend to the reciprocals of the roots of the cubic equation (2.63). However, instead of (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). The following example illustrates the method of procedure :

Example 2.22 Find the real roots of the equation $x^3 - 6x^2 + 11x - 6 = 0$.

Let

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

To obtain the roots directly, we consider the transformed equation

$$-6x^3 + 11x^2 - 6x + 1 = 0 \tag{ii}$$

Presented By: http://www.ebooksuit.com

Let

$$\frac{1}{-6x^3+11x^2-6x+1}=\sum_{i=0}^{\infty}\alpha_i x^i=\alpha_0+\alpha_1 x+\alpha_2 x^2+\alpha_3 x^3+\cdots$$

Hence,

$$(-6x^3+11x^2-6x+1)(\alpha_0+\alpha_1 x+\alpha_2 x^2+\alpha_3 x^3+\cdots) \qquad \text{(iii)}$$

Comparing the coefficients of like powers of x on both sides of (iii), we obtain

$$\alpha_0=1,\ \alpha_1-6\alpha_0=0,\ \alpha_2-6\alpha_1+11\alpha_0=0,\ \alpha_3-6\alpha_2+11\alpha_1-6\alpha_0=0, \text{ etc}$$

The above equations give

$$\alpha_1=6, \qquad \alpha_2=25, \qquad \alpha_3=90,\ldots$$

Hence

$$q_1^{(1)}=\alpha_1/\alpha_0=6.00000,$$

$$q_1^{(2)}=\alpha_2/\alpha_1=25/6=4.16667,$$

$$q_1^{(3)}=\alpha_3/\alpha_2=90/25=3.60000$$

Again,

$$\Delta_1^{(1)}=q_1^{(2)}-q_1^{(1)}=-1.83333,$$

and

$$\Delta_1^{(2)}=q_1^{(3)}-q_1^{(2)}=-0.56667.$$

Now, in the rhombus (see Table 2.1),

$$\begin{array}{ccc} & \Delta_1^{(1)} & \\ q_1^{(2)} & & q_2^{(1)} \\ & \Delta_1^{(2)} & \end{array}$$

all the quantities except $q_2^{(1)}$ are known. Taking $r=1$, and $i=1$, formula (2.69) gives

$$q_2^{(1)}=\frac{\Delta_1^{(2)}q_1^{(2)}}{\Delta_1^{(1)}}=1.28789.$$

Next, we need to determine the value of $\Delta_2^{(0)}$. For this, we consider the rhombus

$$\begin{array}{ccc} & q_2^{(0)} & \\ \Delta_1^{(1)} & & \Delta_2^{(0)} \\ & q_2^{(1)} & \end{array}$$

Presented By: http://www.ebooksuit.com

in which all the quantities, except $\Delta_2^{(0)}$, are known. Using formula (2.70), we obtain

$$\begin{aligned}\Delta_2^{(0)} &= q_2^{(1)} + \Delta_1^{(1)} - q_2^{(0)} \\ &= 1.28789 - 1.83333, \qquad \text{since } q_2^{(0)} = 0 \\ &= -0.54544.\end{aligned}$$

Also,

$$\Delta_3^{(-1)} = 0.$$

Thus, all the quantities in the first-two rows are known and the succeeding rows can be built-up, using the formulae (2.69) and (2.70), alternately. The numerical results are shown in the following table:

Δ_0	q_1	Δ_1	q_2	Δ_2	q_3	Δ_3
	6.00000		0		0	
0		−1.83333		−0.54544		0
	4.16667		1.28789		0.54544	
0		−0.56667		−0.23100		0
	3.60000		1.62356		0.77644	
0		−0.25556		−0.11047		0
	3.34444		1.76865		0.88691	
0		−0.13515		−0.05540		0
	3.20929		1.8484		0.94231	
0		−0.07784		−0.02824		0
	3.13145		1.898		0.9706	

It is evident that q_1, q_2 and q_3 are gradually converging to the roots 3, 2 and 1, respectively.

2.12 SOLUTION TO SYSTEMS OF NONLINEAR EQUATIONS

In this section, we consider the solution to simultaneous nonlinear equations by two methods: (i) the method of iteration and (ii) Newton–Raphson method. In both the cases, the first approximations are usually obtained from a graph. For simplicity, we consider the case of two equations in two unknowns.

2.12.1 The Method of Iteration

Let the equations be given by

$$f(x, y) = 0, \qquad g(x, y) = 0, \tag{2.72}$$

Presented By: http://www.ebooksuit.com

whose real roots are required within a specified accuracy. As in the method of iteration for a single equation (see Section 2.4), we assume that the equations in (2.72) may be written in the form

$$x = F(x, y), \qquad y = G(x, y), \tag{2.73}$$

where the functions F and G satisfy the conditions:

$$\left|\frac{\partial F}{\partial x}\right| + \left|\frac{\partial F}{\partial y}\right| < 1 \quad \text{and} \quad \left|\frac{\partial G}{\partial x}\right| + \left|\frac{\partial G}{\partial y}\right| < 1 \tag{2.74}$$

in the neighbourhood of the root.

Let (x_0, y_0) be the initial approximation to a root (ξ, η) of the system (2.72). We then construct the successive approximations according to the following formulae:

$$\left.\begin{array}{ll} x_1 = F(x_0, y_0), & y_1 = G(x_0, y_0) \\ x_2 = F(x_1, y_1), & y_2 = G(x_1, y_1) \\ x_3 = F(x_2, y_2), & y_3 = G(x_2, y_2) \\ \vdots & \vdots \\ x_{n+1} = F(x_n, y_n), & y_{n+1} = G(x_n, y_n) \end{array}\right\} \tag{2.75}$$

For faster convergence, recently computed values of x_i may be used in the evaluation of y_i in (2.75). If the iteration process (2.75) converges, then we obtain

$$\xi = F(\xi, \eta) \quad \text{and} \quad \eta = G(\xi, \eta) \tag{2.76}$$

in the limit. Thus, ξ and η are the roots of the system (2.73) and hence, also, of the system (2.72). Conditions (2.74) are sufficient for the convergence of the iteration (2.75). We state the following theorem:

Theorem 2.2 Let $x = \xi$ and $y = \eta$ be one pair of roots of the system (2.73) in the closed neighbourhood R. If the functions F and G and their first partial derivatives are continuous in R,

$$\left|\frac{\partial F}{\partial x}\right| + \left|\frac{\partial F}{\partial y}\right| < 1 \quad \text{and} \quad \left|\frac{\partial G}{\partial x}\right| + \left|\frac{\partial G}{\partial y}\right| < 1 \tag{2.77}$$

for all (x, y) in R, and the initial approximation (x_0, y_0) is chosen in R, then the sequence of approximations given by (2.75) converges to the roots $x = \xi$ and $y = \eta$ of the system (2.73).

The proof of this theorem is similar to that of Theorem 2.1 and is omitted here. The method can obviously be generalized to any number of equations.

Presented By: http://www.ebooksuit.com

Example 2.23 Find a real root of the equations:

$$x = 0.2x^2 + 0.8, \qquad y = 0.3xy^2 + 0.7$$

We have

$$F(x, y) = 0.2x^2 + 0.8, \qquad G(x, y) = 0.3xy^2 + 0.7$$

Then

$$\frac{\partial F}{\partial x} = 0.4x, \qquad \frac{\partial F}{\partial y} = 0$$

$$\frac{\partial G}{\partial x} = 0.3y^2, \quad \frac{\partial G}{\partial y} = 0.6xy.$$

It is easy to see that $x = 1$ and $y = 1$ are the roots of the system. Choosing $x_0 = y_0 = 1/2$, we find that

$$\left|\frac{\partial F}{\partial x}\right|_{(x_0, y_0)} + \left|\frac{\partial F}{\partial y}\right|_{(x_0, y_0)} = 0.2 < 1$$

and

$$\left|\frac{\partial G}{\partial x}\right|_{(x_0, y_0)} + \left|\frac{\partial G}{\partial y}\right|_{(x_0, y_0)} = 0.225 < 1$$

Thus, conditions (2.77) are satisfied. Hence,

$$x_1 = F(x_0, y_0) = \frac{0.2}{4} + 0.8 = 0.85$$

and

$$y_1 = G(x_0, y_0) = \frac{0.3}{8} + 0.7 = 0.74$$

For the second approximation, we obtain

$$x_2 = F(x_1, y_1) = 0.2 \times (0.85)^2 + 0.8 = 0.9445$$

$$y_2 = G(x_1, y_1) = 0.3 \times (0.85) \times (0.74)^2 + 0.7 = 0.81.$$

Convergence to the root (1, 1) is obvious. In computing the y values in the above, we could have used the recently computed values of x for a faster convergence. For example,

$$y_1 = G(x_1, y_0) = 0.3 \times \frac{1}{4} \times (0.85) = 0.764,$$

which is a better approximation than the previously computed one.

Presented By: http://www.ebooksuit.com

2.12.2 Newton–Raphson Method

Let (x_0, y_0) be an initial approximation to the root of the system (2.72). If $(x_0 + h, y_0 + k)$ is the root of the system, then we must have

$$f(x_0 + h, y_0 + k) = 0, \qquad g(x_0 + h, y_0 + k) = 0 \tag{2.78}$$

Assuming that f and g are sufficiently differentiable, we expand (2.78) by Taylor's series (see Theorem 1.6, Section 1.2) to obtain

$$\left.\begin{aligned} f_0 + h\frac{\partial f}{\partial x_0} + k\frac{\partial f}{\partial y_0} + \cdots = 0 \\ g_0 + h\frac{\partial g}{\partial x_0} + k\frac{\partial g}{\partial y_0} + \cdots = 0, \end{aligned}\right\} \tag{2.79}$$

where

$$\frac{\partial f}{\partial x_0} = \left[\frac{\partial f}{\partial x}\right]_{x=x_0}, \qquad f_0 = f(x_0, y_0), \text{ etc.}$$

Neglecting the second- and higher-order terms, we obtain the following system of linear equations:

$$\left.\begin{aligned} h\frac{\partial f}{\partial x_0} + k\frac{\partial f}{\partial y_0} = -f_0 \\ h\frac{\partial g}{\partial x_0} + k\frac{\partial g}{\partial y_0} = -g_0 \end{aligned}\right\} \tag{2.80}$$

and

If the Jacobian

$$J(f, g) = \begin{vmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \\ \dfrac{\partial g}{\partial x} & \dfrac{\partial g}{\partial y} \end{vmatrix} \tag{2.81}$$

does not vanish, then the linear Eqs. (2.80) possess a unique solution given by

$$h = \frac{1}{J(f, g)}\begin{vmatrix} -f & \dfrac{\partial f}{\partial y} \\ -g & \dfrac{\partial g}{\partial y} \end{vmatrix} \quad \text{and} \quad k = \frac{1}{J(f, g)}\begin{vmatrix} \dfrac{\partial f}{\partial x} & -f \\ \dfrac{\partial g}{\partial x} & -g \end{vmatrix} \tag{2.82}$$

The new approximations are then given by

$$x_1 = x_0 + h, \qquad y_1 = y_0 + k \tag{2.83}$$

Presented By: http://www.ebooksuit.com

The process is to be repeated till we obtain the roots to the desired accuracy. If the iteration converges, it does so quadratically. Theorem 2.3 below (stated without proof) gives the conditions which are sufficient for convergence:

Theorem 2.3 Let (x_0, y_0) be an approximation to a root (ξ, η) of the system (2.72) in the closed neighbourhood R containing (ξ, η). If (a) f, g and all their first and second derivatives are continuous and bounded in R, and (b) $J(f, g) \neq 0$ in R, then the sequence of approximations given by

$$x_{i+1} = x_i - \frac{1}{J(f,g)} \begin{vmatrix} f & g \\ \dfrac{\partial f}{\partial y} & \dfrac{\partial g}{\partial y} \end{vmatrix} \quad \text{and} \quad y_{i+1} = y_i - \frac{1}{J(f,g)} \begin{vmatrix} g & f \\ \dfrac{\partial g}{\partial x} & \dfrac{\partial f}{\partial x} \end{vmatrix} \tag{2.84}$$

converges to the root (ξ, η) of the system (2.72).

Example 2.24 Find a real root of the equations $x^2 - y^2 = 3$ and $x^2 + y^2 = 13$.

For starting the solution, we take $y = x$ as our first approximation. This gives

$$x_0 = y_0 = \sqrt{6.5} = 2.54951,$$

and therefore

$$f_0 = -3 \quad \text{and} \quad g_0 = 0,$$

where

$$f = x^2 - y^2 - 3 \quad \text{and} \quad g = x^2 + y^2 - 13.$$

Further,

$$\frac{\partial f}{\partial x_0} = 2x_0 = 5.09902, \qquad \frac{\partial g}{\partial x_0} = 2x_0 = 5.09902,$$

$$\frac{\partial f}{\partial y_0} = -2y_0 = -5.09902, \qquad \frac{\partial g}{\partial y_0} = 2y_0 = 5.09902.$$

Hence

$$\begin{vmatrix} \dfrac{\partial f}{\partial x_0} & \dfrac{\partial f}{\partial y_0} \\ \dfrac{\partial g}{\partial x_0} & \dfrac{\partial g}{\partial y_0} \end{vmatrix} = \begin{vmatrix} 5.09902 & -5.09902 \\ 5.09902 & 5.09902 \end{vmatrix} \neq 0;$$

and therefore the convergence criterion is satisfied. We then have

$$h(5.09902) + k(-5.09902) = 3 \quad \text{and} \quad h(5.09902) + k(5.09902) = 0.$$

These equations give

$$h = 0.29417 \quad \text{and} \quad k = -0.29417.$$

Presented By: http://www.ebooksuit.com

Hence the first approximation to the root is given by

$$x_1 = x_0 + h = 2.54951 + 0.29417 = 2.84368$$

$$y_1 = y_0 + k = 2.54951 - 0.29417 = 2.25534.$$

For the second approximation, we have

$$f_1 = f(x_1, y_1) = -0.000042573$$

$$g_1 = g(x_1, y_1) = 0.173074458.$$

Then

$$\frac{\partial f}{\partial x_1} = 2x_1 = 5.68736, \qquad \frac{\partial f}{\partial y_1} = -2y_1 = -4.51068,$$

$$\frac{\partial g}{\partial x_1} = 2x_1 = 5.68736, \qquad \frac{\partial g}{\partial y_1} = 2y_1 = 4.51068.$$

Clearly, the condition of convergence is satisfied and we have the simultaneous equations

$$h(5.68736) + k(-4.51068) = 0.000042573$$

and

$$h(5.68736) + k(4.51068) = -0.17307.$$

Solving the above equations, we obtain

$$h = -0.01521 \quad \text{and} \quad k = -0.01919.$$

the second approximation is therefore given by

$$x_2 = 2.84368 - 0.01521 = 2.82847,$$

and

$$y_2 = 2.25534 - 0.01919 = 2.23615.$$

The above values may be compared with the true values, which are given by

$$x = \sqrt{8} = 2.82843 \quad \text{and} \quad y = \sqrt{5} = 2.23607.$$

EXERCISES

Obtain a root, correct to three decimal places, for each of the following equations using the bisection method (Problems 1–10):

2.1. $x^3 + x^2 + x + 7 = 0$

2.2. $x^3 - 4x - 9 = 0$

2.3. $x^3 - x - 4 = 0$

2.4. $x^3 - 18 = 0$

2.5. $x^3 - x^2 - 1 = 0$

2.6. $x^3 + x^2 - 1 = 0$

2.7. $x^3 - 3x - 5 = 0$

2.8. $x^3 - x - 1 = 0$

2.9. $x^3 - 5x + 3 = 0$

2.10. $x^3 + x - 1 = 0.$

Presented By: http://www.ebooksuit.com

Use the method of false position to obtain a root, correct to three decimal places, of each of the following equations (Problems 11–15):

2.11. $x^3 + x^2 + x + 7 = 0$

2.12. $x^3 - x - 4 = 0$

2.13. $x^3 - x^2 - 1 = 0$

2.14. $x^3 - x - 1 = 0$

2.15. $x^3 + x - 1 = 0.$

Use the iteration method to find, correct to four significant figures, a real root of each of the following equations (Problems 16–25):

2.16. $\cos x = 3x - 1$

2.17. $x = 1/(x+1)^2$

2.18. $x = (5-x)^{1/3}$

2.19. $\sin x = 10(x-1)$

2.20. $e^{-x} = 10x$

2.21. $x \sin x = 1.0$

2.22. $\sin^2 x = x^2 - 1$

2.23. $e^x = \cot x$

2.24. $1 + x^2 = x^3$

2.25. $5x^3 - 20x + 3 = 0.$

2.26. Compute a root of the equation $e^x = x^2$ to an accuracy of 10^{-5}, using the iterative method.

Use Newton–Raphson method to obtain a root, correct to three decimal places, of the following equations (Problems 27–36):

2.27. $x^{\sin 2} - 4 = 0$

2.28. $\sin x = 1 - x$

2.29. $x^3 - 5x + 3 = 0$

2.30. $x^4 + x^2 - 80 = 0$

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

2.32. $4(x - \sin x) = 1$

2.33. $x - \cos x = 0$

2.34. $\sin x = (1/2)x$

2.35. $x + \log x = 2$

2.36. $xe^{-2x} = (1/2)\sin x$

2.37. Establish the formula

$$x_{i+1} = \frac{1}{2}\left(x_i + \frac{N}{x_i}\right)$$

and hence compute the value of $\sqrt{2}$ correct to six decimal places.

2.38. Find the least positive root of the equation $\tan x = x$ to an accuracy of 0.0001 by Newton–Raphson method.

2.39. Obtain, to four decimal places, the root between 1 and 2 of the equation $x^3 - 2x^2 + 3x - 5 = 0$ by (a) Regula–Falsi, (b) Newton–Raphson method.

2.40. Using Ramanujan's method, obtain the first-eight convergents of the equation $x + x^3 = 1$.

Presented By: http://www.ebooksuit.com

2.41. Using iteration method, find the real root of the equation

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

Solve the same equation using Ramanujan's method and compare the results.

2.42. Using iteration method, find the real root of the equation

$$x - \frac{x^3}{3} + \frac{x^5}{10} - \frac{x^7}{42} + \frac{x^9}{216} - \cdots = 0$$

2.43. Use the secant method to determine the root, lying between 5 and 8, of the equation $x^{2.2} = 69$. Compare your result with that obtained in Example 2.5.

2.44. Determine the real root of the equation $xe^x = 1$ using the secant method. Compare your result with the true value of $x = 0.567143\ldots$

2.45. Apply Graeffe's root-squaring method to determine the approximate roots of the equations:

(a) $x^3 - 2x^2 - x + 2 = 0$ (b) $x^3 - 7x^2 + 10x - 2 = 0$.

2.46. Use Muller's method to find a root of the equations:

(a) $x^3 - x - 1 = 0$ (b) $x^3 - x^2 - x - 1 = 0$.

2.47. Using Lin–Bairstow's method, obtain the quadratic factors of the following equations:

(a) $x^3 - 2x^2 + x - 2 = 0$ (b) $x^4 + 5x^3 + 3x^2 - 5x - 9 = 0$.

2.48. Apply the quotient-difference method to obtain the approximate roots of the equations:

(a) $x^3 - x^2 - 2x + 1 = 0$ (b) Problem 45(b).

2.49. Prove that

(a) $\lim\limits_{i\to\infty} \dfrac{\alpha_i}{\alpha_{i-1}} = \dfrac{1}{x_1}$

(b) $\lim\limits_{i\to\infty} \dfrac{1/x_1 - q_1^{(i)}}{(x_1/x_2)^i} = \dfrac{1}{x_1}\dfrac{p_2}{p_1}\left(1 - \dfrac{x_1}{x_2}\right)$

(c) $\lim\limits_{i\to\infty} \dfrac{1/x_1 - q_1^{(i+1)}}{(x_1/x_2)^i} = \dfrac{1}{x_2}\dfrac{p_2}{p_1}\left(1 - \dfrac{x_1}{x_2}\right)$,

in the notation of Section 2.11.

2.50. Develop a subprogram (in any computer language of your choice) for the Newton–Raphson method indicating the maximum number of iterations and also the tolerance for the percentage error in your solution. Test your program on Problem 36.

Presented By: http://www.ebooksuit.com

2.51. Develop a subprogram for computing roots by Muller's method. Use Problem 46(b) to test your program.

2.52. Use a package of your choice to determine all the roots of the polynomial equation

$$f(x) = (x+1)(x-2)(x-4)(x+6) = 0.$$

2.53. Develop a subprogram to implement Bairstow's method and test it on Problem 47(b).

2.54. MATLAB has a subprogram called roots (c) for calculating the roots of a polynomial, whose coefficients are stored in the vector (c) and to return the roots to a vector (r). Use this program to obtain the roots of the equation

$$x^4 - 5x^3 + 9.25x^2 - 7.75x + 2.5 = 0.$$

2.55. The following equation occurs in rocket dynamics:

$$m_0\left[1 - e^{-(v+gt)/v_r}\right] = u_f t,$$

where m_0 is the mass of the rocket at time $t = 0$, v is its upward velocity at time t seconds, v_r is the relative velocity at which the fuel is ejected, u_f is the fuel consumption rate and g is the acceleration due to gravity (= 9.8 m/sec^2). Determine t (to within 1% of the true value) when $v = 1500$ m/sec, $m_0 = 200{,}000$ kg, $v_r = 2500$ m/sec and $u_f = 3000$ kg/sec, using any subprogram of your choice.

Solve the following systems of nonlinear equations by any suitable method (Problems 56–58):

2.56. $x^2 - y^2 = 4,\ x^2 + y^2 = 16.$

2.57. $x^2 + y = 11,\ y^2 + x = 7.$

2.58. $x^2 = 3xy - 7,\ y = 2(x+1).$

Presented By: http://www.ebooksuit.com

CHAPTER

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

$$|f(x) - P(x)| < \varepsilon, \quad \text{for all } x \text{ in } (x_0, x_n).$$

Presented By: http://www.ebooksuit.com

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 (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.

Presented By: http://www.ebooksuit.com

Let this point be given by $x = \xi$, $x_0 < \xi < x_n$. On differentiating (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 (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 $y_1 - y_0, y_2 - y_1, \ldots, y_n - y_{n-1}$ are called the *differences* of y. Denoting these differences by $\Delta y_0, \Delta y_1, \ldots, \Delta y_{n-1}$ 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 $\Delta y_0, \Delta y_1, \ldots,$ are called *first forward differences*. The differences of the first forward differences are called *second forward differences* and are denoted by $\Delta^2 y_0, \Delta^2 y_1, \ldots$. Similarly, one can define *third forward differences*, *fourth forward differences*, etc. Thus,

Presented By: http://www.ebooksuit.com

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = y_2 - y_1 - (y_1 - y_0)$$
$$= y_2 - 2y_1 + y_0,$$
$$\Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0 = y_3 - 2y_2 + y_1 - (y_2 - 2y_1 + y_0)$$
$$= y_3 - 3y_2 + 3y_1 - y_0$$
$$\Delta^4 y_0 = \Delta^3 y_1 - \Delta^3 y_0 = y_4 - 3y_3 + 3y_2 - y_1 - (y_3 - 3y_2 + 3y_1 - y_0)$$
$$= y_4 - 4y_3 + 6y_2 - 4y_1 + y_0.$$

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	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						

3.3.2 Backward Differences

The differences $y_1 - y_0, y_2 - y_1, \ldots, y_n - y_{n-1}$ are called *first backward differences* if they are denoted by $\nabla y_1, \nabla y_2, \ldots, \nabla y_n$ respectively, so that $\nabla y_1 = y_1 - y_0$, $\nabla y_2 = y_2 - y_1, \ldots, \nabla y_n = y_n - y_{n-1}$, where ∇ is called the *backward difference operator*. In a similar way, one can define backward differences of higher orders. Thus we obtain

$$\nabla^2 y_2 = \nabla y_2 - \nabla y_1 = y_2 - y_1 - (y_1 - y_0) = y_2 - 2y_1 + y_0,$$
$$\nabla^3 y_3 = \nabla^2 y_3 - \nabla^2 y_2 = y_3 - 3y_2 + 3y_1 - y_0, \text{ etc.}$$

With the same values of x and y as in Table 3.1, a backward difference table can be formed:

Presented By: http://www.ebooksuit.com

Table 3.2 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 can be formed:

Table 3.3 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 the three 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

$$\Delta y_0 = \nabla y_1 = \delta y_{1/2}, \qquad \Delta^3 y_2 = \nabla^3 y_5 = \delta^3 y_{7/2}, \ldots$$

Presented By: http://www.ebooksuit.com

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:

$$\left.\begin{aligned} &\nabla = 1 - E^{-1}, \\ &\delta = E^{1/2} - E^{-1/2}, \\ &\mu = (1/2)(E^{1/2} + E^{-1/2}),\ \mu^2 = 1 + (1/4)\,\delta^2 \\ &\Delta = \nabla E = \delta E^{1/2}. \end{aligned}\right\} \tag{3.8b}$$

As an example, we prove the relation $\mu^2 \equiv 1 + (1/4)\,\delta^2$. We have, by definition,

$$\begin{aligned} \mu y_r &= \frac{1}{2}(y_{r+1/2} + y_{r-1/2}) \\ &= \frac{1}{2}(E^{1/2} y_r + E^{-1/2} y_r) \end{aligned}$$

*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.

Presented By: http://www.ebooksuit.com

$$= \frac{1}{2}(E^{1/2} + E^{-1/2})y_r.$$

Hence

$$\mu = \frac{1}{2}(E^{1/2} + E^{-1/2})$$

and

$$\mu^2 = \frac{1}{4}(E^{1/2} + E^{-1/2})^2$$

$$= \frac{1}{4}(E + E^{-1} + 2)$$

$$= \frac{1}{4}[(E^{1/2} - E^{-1/2})^2 + 4]$$

$$= \frac{1}{4}(\delta^2 + 4).$$

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

$$Ey(x) = \left(1 + hD + \frac{h^2D^2}{2!} + \frac{h^3D^3}{3!} + \cdots\right)y(x).$$

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

$$\Delta^n u_{x-n} = u_x - nu_{x-1} + \frac{n(n-1)}{2}u_{x-2} + \cdots + (-1)^n u_{x-n}.$$

Presented By: http://www.ebooksuit.com

To prove this result, we start with the right-hand side. Thus,

$$u_x - nu_{x-1} + \frac{n(n-1)}{2}u_{x-2} + \cdots + (-1)^n u_{x-n}$$

$$= u_x - nE^{-1}u_x + \frac{n(n-1)}{2}E^{-2}u_x + \cdots + (-1)^n E^{-n}u_x$$

$$= \left[1 - nE^{-1} + \frac{n(n-1)}{2}E^{-2} + \cdots + (-1)^n E^{-n}\right]u_x$$

$$= (1 - E^{-1})^n u_x$$

$$= \left(1 - \frac{1}{E}\right)^n u_x$$

$$= \left(\frac{E-1}{E}\right)^n u_x$$

$$= \frac{\Delta^n}{E^n}u_x$$

$$= \Delta^n E^{-n} u_x$$

$$= \Delta^n u_{x-n},$$

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,

$$e^x\left(u_0 + x\,\Delta u_0 + \frac{x^2}{2!}\Delta^2 u_0 + \cdots\right) = e^x\left(1 + x\Delta + \frac{x^2\Delta^2}{2!} + \cdots\right)u_0$$

$$= e^x e^{x\Delta}u_0 = e^{x(1+\Delta)}u_0$$

$$= e^{xE}u_0$$

$$= \left(1 + xE + \frac{x^2E^2}{2!} + \cdots\right)u_0$$

$$= u_0 + xu_1 + \frac{x^2}{2!}u_2 + \cdots,$$

which is the required result.

Presented By: http://www.ebooksuit.com

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 3.4.

Table 3.4 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.

Presented By: http://www.ebooksuit.com

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 3.4 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

$$y(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_n.$$

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 $a_0n(n-1)h^2$.

Presented By: http://www.ebooksuit.com

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., $(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, 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 formula (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 (3.11) is useless in practice. Nevertheless, if $y^{(n+1)}(x)$

Presented By: http://www.ebooksuit.com

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 (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 $x_n, x_{n-1}, \ldots, x_2, x_1, x_0$, 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

$$y(x) - y_n(x) = \frac{p(p+1)(p+2)\ldots(p+n)}{(n+1)!}h^{n+1}y^{(n+1)}(\xi), \tag{3.15}$$

where $x_0 < \xi < x_n$ and $x = x_n + ph$.

Presented By: http://www.ebooksuit.com

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, formula (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

$$S_{n+1} - S_n = (n+1)^3,$$

Presented By: http://www.ebooksuit.com

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,

$$S_1 = 1, \qquad \Delta S_1 = 8, \qquad \Delta^2 S_1 = 19, \qquad \Delta^3 S_1 = 18, \qquad \Delta^4 S_1 = 6.$$

Formula (3.10) gives

$$S_n = 1 + (n-1)(8) + \frac{(n-1)(n-2)}{2}(19) + \frac{(n-1)(n-2)(n-3)}{6}(18)$$

$$+ \frac{(n-1)(n-2)(n-3)(n-4)}{24}(6)$$

$$= \frac{1}{4}n^4 + \frac{1}{2}n^3 + \frac{1}{4}n^2$$

$$= \left[\frac{n(n+1)}{2}\right]^2.$$

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 $\sin 38°$.

Presented By: http://www.ebooksuit.com

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 formula (3.14), we obtain

$$\begin{aligned} y(38) &= 0.6427876 - 0.4\,(0.0692112) + \frac{-0.4(-0.4-1)}{2}(-0.0043652) \\ &\quad + \frac{(-0.4)(-0.4+1)(-0.4+2)}{6}(-0.0005599) \\ &\quad + \frac{(-0.4)(-0.4+1)(-0.4+2)(-0.4+3)}{24}(0.0000289) \\ &\quad + \frac{(-0.4)(-0.4+1)(-0.4+2)(-0.4+3)(-0.4+4)}{120}(0.0000041) \\ &= 0.6427876 - 0.02768448 + 0.00052382 + 0.00003583 \\ &\qquad - 0.00000120 \\ &= 0.6156614. \end{aligned}$$

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 $3^3 = 27$?

Presented By: http://www.ebooksuit.com

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 formula (3.10) gives

$$\begin{aligned}\tan(0.12) &= 0.1003 + 0.4(0.0508) + \frac{0.4(0.4-1)}{2}(0.0008) \\ &\quad + \frac{0.4(0.4-1)(0.4-2)}{6}(0.0002) \\ &\quad + \frac{0.4(0.4-1)(0.4-2)(0.4-3)}{24}(0.0002) \\ &= 0.1205.\end{aligned}$$

Presented By: http://www.ebooksuit.com

(b) To find tan (0.26), we have $0.26 = 0.30 + p(0.05)$, which gives $p = -0.8$. Hence formula (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

Gauss' forward formula

We consider the following difference table in which the central ordinate is taken for convenience as y_0 corresponding to $x = x_0$.

Presented By: http://www.ebooksuit.com

The differences used in this formula lie on the line shown in Table 3.5. 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 $G_1, G_2, \ldots$ 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 3.5 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 (3.16) can also be expressed in terms of y_0, Δy_0 and higher-order differences. We have

$$\Delta^2 y_{-1} = \Delta^2 E^{-1} y_0$$

$$= \Delta^2 (1+\Delta)^{-1} y_0$$

$$= \Delta^2 (1 - \Delta + \Delta^2 - \Delta^3 + \cdots) y_0$$

$$= \Delta^2 (y_0 - \Delta y_0 + \Delta^2 y_0 - \Delta^3 y_0 + \cdots)$$

$$= \Delta^2 y_0 - \Delta^3 y_0 + \Delta^4 y_0 - \Delta^5 y_0 + \cdots$$

Presented By: http://www.ebooksuit.com

$$\Delta^3 y_{-1} = \Delta^3 y_0 - \Delta^4 y_0 + \Delta^5 y_0 - \Delta^6 y_0 + \cdots$$

$$\begin{aligned}\Delta^4 y_{-2} &= \Delta^4 E^{-2} y_0 \\ &= \Delta^4 (1+\Delta)^{-2} y_0 \\ &= \Delta^4 (y_0 - 2\Delta y_0 + 3\Delta^2 y_0 - 4\Delta^3 y_0 + \cdots) \\ &= \Delta^4 y_0 - 2\Delta^5 y_0 + 3\Delta^6 y_0 - 4\Delta^7 y_0 + \cdots\end{aligned}$$

Hence (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 $\Delta y_0, \Delta^2 y_0, \Delta^3 y_0$, etc., on both sides of (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 3.6.

Presented By: http://www.ebooksuit.com

Table 3.6 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

$$1.17 = 1.15 + p(0.05),$$

Presented By: http://www.ebooksuit.com

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

$$e^{1.17} = 3.1582 + \frac{2}{5}(0.1619) + \frac{(2/5)(2/5-1)}{2}(0.0079)$$

$$+ \frac{(2/5+1)(2/5)(2/5-1)}{6}(0.0004)$$

$$= 3.1582 + 0.0648 - 0.0009$$

$$= 3.2221.$$

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 (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.

Presented By: http://www.ebooksuit.com

$\vdots$	$\vdots$						
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

$$y_p = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!} \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \frac{p(p-1)(p-1/2)}{3!} \Delta^3 y_{-1}$$

$$+ \frac{(p+1)\, p(p-1)(p-2)}{4!} \frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \cdots \tag{3.24}$$

Presented By: http://www.ebooksuit.com

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

$$y_p = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!}\,\frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \frac{p(p-1)(p-1/2)}{3!}\Delta^3 y_{-1}$$

$$+ \frac{(p+1)\,p(p-1)(p-2)}{4!}\,\frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \cdots$$

Presented By: http://www.ebooksuit.com

$$= y_0 + p(y_1 - y_0) + \frac{p(p-1)}{2!}\frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \frac{p(p-1)(p-1/2)}{3!}(\Delta^2 y_0 - \Delta^2 y_{-1})$$

$$+ \frac{(p+1)\,p(p-1)(p-2)}{4!}\frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \cdots$$

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) If interpolation is desired near the beginning or end of a table, there is no alternative to Newton's forward and backward difference formulae, simply because higher-order central differences do not exist at the beginning or end of a table of values.

(ii) For interpolation near the middle of a table, Stirling's formula gives the most accurate result for $-1/4 \le p \le 1/4$, and Bessel's formula is most efficient near $p = 1/2$, say $1/4 \le p \le 3/4$. But in the case where a series of calculations have to be made, it would be inconvenient to use both these formulae, and a choice must be made between them. The choice depends on the *order* of the highest difference that could be neglected so that contributions from it and further differences would be less than half a unit in the last decimal place. If this highest difference is of odd order. Stirling's formula is recommended; if it is of even order, Bessel's formula might be preferred. Estimation of the maximum value of a difference of any order in an interpolation formula is not difficult. Thus, in Stirling's formula (3.21), the term containing the third differences, viz.,

$$\frac{p(p^2-1)}{6}\frac{\Delta^3 y_{-1} + \Delta^3 y_{-2}}{2}$$

Presented By: http://www.ebooksuit.com

may be neglected if its contribution to the interpolate is less than half a unit in the last place. This means that

$$\left| \frac{p(p^2-1)}{6} \frac{\Delta^3 y_{-1} + \Delta^3 y_{-2}}{2} \right| < \frac{1}{2},$$

for all p in the range $0 \le p \le 1$.

But the maximum value of $p(p^2-1)/6$ is 0.064 and so we have

$$\left| 0.064 \frac{\Delta^3 y_{-1} + \Delta^3 y_{-2}}{2} \right| < \frac{1}{2},$$

which gives

$$\left| \frac{\Delta^3 y_{-1} + \Delta^3 y_{-2}}{2} \right| < 8.$$

If we consider Bessel's formula (3.24), the contribution from the term containing the third difference will be less than half a unit in the last place provided that

$$\left| \frac{p(p-1)(p-1/2)}{6} \Delta^3 y_{-1} \right| < \frac{1}{2}.$$

But the maximum value of

$$\frac{p(p-1)(p-1/2)}{6}$$

is 0.008, and so $|\Delta^3 y_{-1}| < 60$. In other words, if we neglect the third differences, Bessel's formula is about seven times more accurate than Stirling's formula. If the third differences need to be retained (i.e. when they are more than 60 in magnitude), then Everett's formula may be gainfully employed for the aforesaid reason, viz., Everett's formula with second differences is equivalent to Bessel's formula with third differences. The following examples illustrate the use of the central difference formulae.

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

Presented By: http://www.ebooksuit.com

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

$$y(0.644) = 1.896481 + 0.4\frac{0.018871 + 0.019060}{2} + \frac{0.16}{2}(0.000189)$$

$$= 1.896481 + 0.0075862 + 0.00001512$$

$$= 1.904082,$$

whilst 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.

Presented By: http://www.ebooksuit.com

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				

Presented By: http://www.ebooksuit.com

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

Presented By: http://www.ebooksuit.com

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 $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ 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 (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 (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

Presented By: http://www.ebooksuit.com

$$L_2(x) = \sum_{i=0}^{2} l_i(x)\, y_i$$

$$= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} y_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} y_1 + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} y_2, \tag{3.33}$$

where the $l_i(x)$ satisfy the conditions given in (3.31) and (3.32).

To derive the general formula, let

$$L_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \tag{3.34}$$

be the desired polynomial of the nth degree such that conditions (3.28) (called the *interpolatory conditions*) are satisfied. Substituting these conditions in (3.34), we obtain the system of equations

$$\left.\begin{aligned} y_0 &= a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n \\ y_1 &= a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n \\ y_2 &= a_0 + a_1 x_2 + a_2 x_2^2 + \cdots + a_n x_2^n \\ &\vdots \\ y_n &= a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_n^n. \end{aligned}\right\} \tag{3.35}$$

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 $a_0, a_1, \ldots, a_n$ from Eqs. (3.34) and (3.35), we obtain

$$\begin{vmatrix} L_n(x) & 1 & x & x^2 & \cdots & x^n \\ y_0 & 1 & x_0 & x_0^2 & \cdots & x_0^n \\ y_1 & 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ y_n & 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix} = 0, \tag{3.37}$$

Presented By: http://www.ebooksuit.com

which shows that $L_n(x)$ is a linear combination of $y_0, y_1, y_2, \ldots, y_n$. 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

$$\left.\begin{aligned} l_i(x_j) &= 0 \qquad \text{if} \quad i \neq j \\ l_j(x_j) &= 1 \qquad \text{for all } j \end{aligned}\right\},$$

which are the same as (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 (3.39) becomes

$$l_i(x) = \frac{\pi_{n+1}(x)}{(x-x_i)\,\pi'_{n+1}(x_i)}. \tag{3.42}$$

Hence (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 (3.39), are called *Lagrange interpolation coefficients*. Interchanging x and y in (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

$$\bar{L}_n(x_i) = y_i, \qquad i = 0, 1, 2, \ldots, n.$$

Presented By: http://www.ebooksuit.com

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 (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 (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 formula (3.44), we have

$$\begin{aligned}x &= \frac{(-5)(-12)}{(-8)(-15)}(1) + \frac{(3)(-12)}{(8)(-7)}(3) + \frac{(3)(-5)}{(15)(7)}(4) \\ &= \frac{1}{2} + \frac{27}{14} - \frac{4}{7} \\ &= 1.86.\end{aligned}$$

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

Presented By: http://www.ebooksuit.com

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

$$\begin{aligned} L_2(x) &= (2x^2 - 11x + 15)(0.69315) - (4x^2 - 20x + 24)(0.91629) \\ &\quad + (2x^2 - 9x + 10)(1.09861) \\ &= -0.08164x^2 + 0.81366x - 0.60761, \end{aligned}$$

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

$$\begin{aligned} \sin\frac{\pi}{6} &\approx \frac{(\pi/6-0)(\pi/6-\pi/2)}{(\pi/4-0)(\pi/4-\pi/2)}(0.70711) + \frac{(\pi/6-0)(\pi/6-\pi/4)}{(\pi/2-0)(\pi/2-\pi/4)}(1) \\ &= \frac{8}{9}(0.70711) - \frac{1}{9} \\ &= \frac{4.65688}{9} \\ &= 0.51743. \end{aligned}$$

Example 3.17 Using Lagrange's interpolation formula, find the form of the function $y(x)$ from the following table

Presented By: http://www.ebooksuit.com

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.

Presented By: http://www.ebooksuit.com

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

Presented By: http://www.ebooksuit.com

$$H_{2n+1}(x) = \sum_{i=0}^{n} u_i(x) y_i + \sum_{i=0}^{n} v_i(x) y_i', \tag{3.50}$$

where $u_i(x)$ and $v_i(x)$ are polynomials in x of degree $(2n+1)$. Using conditions (3.49), we obtain

$$\left.\begin{aligned} u_i(x_j) &= \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}; \quad v_i(x) = 0, \text{ for all } i \\ u_i'(x) &= 0, \text{ for all } i; \quad v_i'(x_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j. \end{cases} \end{aligned}\right\} \tag{3.51}$$

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 (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 (3.51) in (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 Eqs. (3.54), we deduce

$$\left.\begin{aligned} a_i &= -2l_i'(x_i), \quad b_i = 1 + 2x_i\ l_i'(x_i) \\ c_i &= 1, \qquad\qquad d_i = -x_i. \end{aligned}\right\} \tag{3.55}$$

Hence Eqs. (3.53) become

$$\begin{aligned} u_i(x) &= [-2x\, l_i'(x_i) + 1 + 2x_i l_i'(x_i)]\,[l_i(x)]^2 \\ &= [1 - 2\,(x - x_i)\, l_i'(x_i)]\,[l_i(x)]^2 \end{aligned} \tag{3.56a}$$

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 (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*.

Presented By: http://www.ebooksuit.com

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, m_i), $i = 0, 1$.

Putting $n = 1$ in Hermite's formula (3.57), we obtain

$$H_3(x) = [1 - 2(x - x_0)\, l_0'(x_0)]\,[l_0(x)]^2\, y_0 + [1 - 2\,(x - x_1)\, l_1'(x_1)]\,[l_1(x)]^2\, y_1$$
$$+ (x - x_0)\,[l_0(x)]^2\, y_0' + (x - x_1)\,[l_1(x)]^2\, y_1'. \qquad \text{(i)}$$

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, (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

$$l_0'(x_0) = -3, \qquad l_1'(x_1) = 0, \qquad l_2'(x_2) = 3$$

Presented By: http://www.ebooksuit.com

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:

$$\left.\begin{aligned} [x_0, x_1] &= \frac{y_1 - y_0}{x_1 - x_0}, \\ [x_0, x_1, x_2] &= \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0} \\ &\vdots \\ [x_0, x_1, \ldots, x_n] &= \frac{[x_1, x_2, \ldots, x_n] - [x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}. \end{aligned}\right\} \qquad (3.58)$$

Presented By: http://www.ebooksuit.com

Even if the arguments are equal, the divided differences may still have a meaning. We then set $x_1 = x_0 + \varepsilon$ so that

$$[x_0, x_1] = \lim_{\varepsilon \to 0} [x_0, x_0 + \varepsilon]$$

$$= \lim_{\varepsilon \to 0} \frac{y(x_0 + \varepsilon) - y(x_0)}{\varepsilon}$$

$$= y'(x_0), \quad \text{if } y(x) \text{ is differentiable.}$$

Similarly,

$$\underbrace{[x_0, x_0, \ldots, x_0]}_{(r+1)\text{ arguments}} = \frac{y^r(x_0)}{r!}. \tag{3.59}$$

From (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,

$$[x_0, x_1, x_2] = \frac{1}{x_2 - x_0}\left(\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}\right)$$

$$= \frac{1}{x_2 - x_0}\left[\frac{y_2}{x_2 - x_1} - y_1\left(\frac{1}{x_2 - x_1} + \frac{1}{x_1 - x_0}\right) + \frac{y_0}{x_1 - x_0}\right]$$

$$= \frac{y_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{y_1}{(x_1 - x_0)(x_1 - x_2)}$$

$$+ \frac{y_2}{(x_2 - x_0)(x_2 - x_1)}. \tag{3.60}$$

Similarly it can be shown that

$$[x_0, x_1, \ldots, x_n] = \frac{y_0}{(x_0 - x_1)\ldots(x_0 - x_n)} + \frac{y_1}{(x_1 - x_0)\ldots(x_1 - x_n)} + \cdots$$

$$+ \frac{y_n}{(x_n - x_0)\ldots(x_n - x_{n-1})}. \tag{3.61}$$

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

$$[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0} = \frac{1}{h}\Delta y_0 \tag{3.62}$$

Presented By: http://www.ebooksuit.com

$$[x_0, x_1, x_2] = \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0} = \frac{1}{2h}\left(\frac{\Delta y_1}{h} - \frac{\Delta y_0}{h}\right) = \frac{1}{2h^2}\Delta^2 y_0 = \frac{1}{h^2 2!}\Delta^2 y_0 \tag{3.63}$$

and in general,

$$[x_0, x_1, \ldots, x_n] = \frac{1}{h^n n!}\Delta^n y_0. \tag{3.64}$$

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.

3.10.1 Newton's General Interpolation Formula

We have, from the definition of divided differences,

$$[x, x_0] = \frac{y - y_0}{x - x_0}$$

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 (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

$$\begin{aligned} y = y_0 &+ (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ &+ (x - x_0)(x - x_1)(x - x_2)[x_0, x_1, x_2, x_3] + \cdots \\ &+ (x - x_0)(x - x_1)\ldots(x - x_n)[x, x_0, x_1, \ldots, x_n]. \end{aligned} \tag{3.69}$$

Presented By: http://www.ebooksuit.com

This formula is called *Newton's general interpolation formula with divided differences*, the last term being the remainder term after $(n+1)$ terms.

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

$$f(x) = 3 + (x+1)(-9) + x(x+1)(6) + x(x+1)(x-3)(5) + x(x+1)(x-3)(x-6)$$

$$= x^4 - 3x^3 + 5x^2 - 6.$$

Presented By: http://www.ebooksuit.com

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 $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, 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

$$\Delta_{012\cdots n}(x) = \frac{1}{x_n - x_{n-1}}\begin{vmatrix} \Delta_{012\ldots\overline{n-1}}(x) & x_{n-1} - x \\ \Delta_{012\ldots\overline{n-2}n}(x) & x_n - x \end{vmatrix}. \tag{3.72}$$

The computations may conveniently be arranged as in Table 3.7 below:

Table 3.7 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 3.8. Neville's scheme is particularly suited for iterated inverse interpolation.

Presented By: http://www.ebooksuit.com

Table 3.8 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 3.7) is

y	x		
4	1		
		1.750	
12	3		1.857
		1.600	
19	4		

Presented By: http://www.ebooksuit.com

whereas Neville's scheme (see Table 3.8) 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

$$u_3 = \frac{1}{\Delta y_0}\left[y_u - y_0 - \frac{u_2(u_2-1)}{2}\Delta^2 y_0 - \frac{u_2(u_2-1)(u_2-2)}{6}\Delta^3 y_0\right] \tag{3.77}$$

Presented By: http://www.ebooksuit.com

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 3.15.

Presented By: http://www.ebooksuit.com

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%.

3.13 SPLINE INTERPOLATION

We have so far discussed methods of finding an nth-order polynomial passing through $(n+1)$ given data points. Because of round-off and systematic errors, these polynomials were found to give erroneous results in certain cases. This is particularly so when the function undergoes sudden changes in the vicinity of a point in its range. 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 range. These connecting piecewise polynomials are called *spline functions*, named after the draftman'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 spline functions, viz., linear, quadratic, cubic, quintic, etc. Of all these, the cubic spline (spline of *degree three* or *order four*) has been found to be the most popular

Presented By: http://www.ebooksuit.com

in engineering applications. We shall, however, start with a discussion of linear and quadratic splines, since this would set the theme for the derivation of the governing equations of the cubic spline.

3.13.1 Linear Splines

Let the given data points be

$$(x_i, y_i), \qquad i = 0, 1, 2, \ldots, n, \tag{3.78}$$

where

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b$$

and let

$$h_i = x_i - x_{i-1}, \qquad i = 1, 2, \ldots, n. \tag{3.79}$$

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

$$s_i(x) = y_{i-1} + m_i(x - x_{i-1}), \tag{3.80}$$

where

$$m_i = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}. \tag{3.81}$$

Setting $i = 1, 2, \ldots, n$ successively in (3.80), 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 3.27 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.

Presented By: http://www.ebooksuit.com

3.13.2 Quadratic Splines

With reference to the data points given in (3.78), 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

$$s_i(x_i) = y_i, \qquad i = 0, 1, 2, \ldots, n. \tag{3.82}$$

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

$$s_i'(x) = \frac{1}{h_i}\left[(x_i - x)m_{i-1} + (x - x_{i-1})m_i\right], \tag{3.83}$$

where

$$m_i = s_i'(x_i). \tag{3.84}$$

Integrating (3.83) with respect to x, we obtain

$$s_i(x) = \frac{1}{h_i}\left[-\frac{(x_i - x)^2}{2}m_{i-1} + \frac{(x - x_{i-1})^2}{2}m_i\right] + c_i, \tag{3.85}$$

where c_i are constants to be determined. Putting $x = x_{i-1}$ in (3.85), 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 (3.85) becomes:

$$s_i(x) = \frac{1}{h_i}\left[-\frac{(x_i - x)^2}{2}m_{i-1} + \frac{(x - x_{i-1})^2}{2}m_i\right] + y_{i-1} + \frac{h_i}{2}m_{i-1}. \tag{3.86}$$

In (3.86), 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

$$s_i(x_i -) = s_{i+1}(x_i +) \tag{3.87}$$

From (3.86), we obtain

$$\begin{aligned} s_i(x_i -) &= \frac{h_i}{2}m_i + y_{i-1} + \frac{h_i}{2}m_{i-1} \\ &= \frac{h_i}{2}(m_{i-1} + m_i) + y_{i-1}. \end{aligned} \tag{3.88}$$

Further,

$$s_{i+1}(x) = \frac{1}{h_{i+1}}\left[-\frac{(x_{i+1} - x)^2}{2}m_i + \frac{(x - x_i)^2}{2}m_{i+1}\right] + y_i + \frac{h_{i+1}}{2}m_i,$$

Presented By: http://www.ebooksuit.com

and therefore

$$s_{i+1}(x_i+) = -\frac{h_{i+1}}{2}m_i + y_i + \frac{h_{i+1}}{2}m_i = y_i. \tag{3.89}$$

Equality of (3.88) and (3.89) produces the recurrence relation

$$m_{i-1} + m_i = \frac{2}{h_i}(y_i - y_{i-1}), \qquad i = 1, 2, \ldots, n \tag{3.90}$$

for the spline first derivatives m_i. Equations (3.90) 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 (3.86) 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

$$m_0 = m_1. \tag{3.91}$$

Therefore, Eqs. (3.90) and (3.91) can be solved for m_i, which when substituted in (3.86) gives the required quadratic spline

Example 3.28 Determine the quadratic splines satisfying the data given in Example 3.27. Find also approximate values of $y(2.5)$ and $y'(2.0)$.

We have $n = 2$ and $h = 1$. Equations (3.90) give

$$m_0 + m_1 = 14 \quad \text{and} \quad m_1 + m_2 = 38.$$

Since $m_0 = m_1$, we obtain $m_0 = m_1 = 7$, and $m_2 = 31$.

Hence, Eq. (3.86) 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,

$$y(2.5) \approx s_2(2.5) = 5.5 \quad \text{and} \quad y'(2.0) \approx 7.0.$$

Presented By: http://www.ebooksuit.com

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

$$s_i(x) = a_i + b_i x + c_i x^2, \tag{3.92}$$

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

$$s_i(x_i) = a_i + b_i x_i + c_i x_i^2 \qquad i = 1, 2, \ldots, n-1. \tag{3.93}$$

Next, $s_i(x)$ is continuous at $x = x_i$. This condition requires

$$s_i(x_i-) = s_{i+1}(x_i+). \tag{3.94}$$

Hence, we must have

$$a_i + b_i x_i + c_i x_i^2 = a_{i+1} + b_{i+1} x_i + c_{i+1} x_i^2, \qquad i = 1, 2, \ldots, n-1. \tag{3.95}$$

Again, $s_i'(x)$ is continuous at $x = x_i$. This gives

$$b_i + 2c_i x_i = b_{i+1} + 2c_{i+1} x_i, \qquad i = 1, 2, \ldots, n-1. \tag{3.96}$$

We thus have $3n-3$ conditions and we require three more conditions. Since the spline passes through the end points also, we must have

$$y_0 = a_1 + b_1 x_0 + c_1 x_0^2 \tag{3.97}$$

and

$$y_n = a_n + b_n x_n + c_n x_n^2. \tag{3.98}$$

Finally, for the natural spline, we have

$$s_1''(x_0) = 0, \tag{3.99}$$

and this gives

$$c_1 = 0. \tag{3.100}$$

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.

3.14 CUBIC SPLINES

We consider the same set of data points, viz., the data defined in (3.78), 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

Presented By: http://www.ebooksuit.com

(i) $s_i(x)$ is almost 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 $[x_{i-1}, x_i]$:

$$s_i''(x) = \frac{1}{h_i}\left[(x_i - x)M_{i-1} + (x - x_{i-1})M_i\right], \tag{3.101}$$

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 (3.101) twice with respect to x, we get

$$s_i(x) = \frac{1}{h_i}\left[\frac{(x_i - x)^3}{6}M_{i-1} + \frac{(x - x_{i-1})^3}{6}M_i\right] + c_i(x_i - x) + d_i(x - x_{i-1}), \tag{3.102}$$

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

$$c_i = \frac{1}{h_i}\left(y_{i-1} - \frac{h_i^2}{6}M_{i-1}\right) \quad \text{and} \quad d_i = \frac{1}{h_i}\left(y_i - \frac{h_i^2}{6}M_i\right). \tag{3.103}$$

Substituting for c_i and d_i in (3.102), we obtain

$$s_i(x) = \frac{1}{h_i}\left[\frac{(x_i - x)^3}{6}M_{i-1} + \frac{(x - x_{i-1})^3}{6}M_i + \left(y_{i-1} - \frac{h_i^2}{6}M_{i-1}\right)(x_i - x)\right.$$
$$\left. + \left(y_i - \frac{h_i^2}{6}M_i\right)(x - x_{i-1})\right]. \tag{3.104}$$

In (3.104), the spline second derivatives, M_i, are still not known. To determine them, we use the condition of continuity of $s_i'(x)$. From (3.104), we obtain by differentiation:

$$s_i'(x) = \frac{1}{h_i}\left[\frac{-3(x_i - x)^2}{6}M_{i-1} + \frac{3(x - x_{i-1})^2}{6}M_i\right.$$
$$\left. - \left(y_{i-1} - \frac{h_i^2}{6}M_{i-1}\right) + \left(y_i - \frac{h_i^2}{6}M_i\right)\right]$$

Presented By: http://www.ebooksuit.com

Setting $x = x_i$ in the above, we obtain the left-hand derivative

$$s_i'(x_i-) = \frac{h_i}{2}M_i - \frac{1}{h_i}\left(y_{i-1} - \frac{h_i^2}{6}M_{i-1}\right) + \frac{1}{h_i}\left(y_i - \frac{h_i^2}{6}M_i\right)$$

$$= \frac{1}{h_i}(y_i - y_{i-1}) + \frac{h_i}{6}M_{i-1} + \frac{h_i}{3}M_i \qquad (i = 1, 2, \ldots, n). \tag{3.105}$$

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. (3.104)

$$s_{i+1}(x) = \frac{1}{h_{i+1}}\left[\frac{(x_{i+1} - x)^3}{6}M_i + \frac{(x - x_i)^3}{6}M_{i+1} + \left(y_i - \frac{h_{i+1}^2}{6}M_i\right)(x_{i+1} - x)\right.$$

$$\left. + \left(y_{i+1} - \frac{h_{i+1}^2}{6}M_{i+1}\right)(x - x_i)\right], \tag{3.106}$$

where $h_{i+1} = x_{i+1} - x_i$. Differentiating (3.106) and setting $x = x_i$, we obtain the right-hand derivative at $x = x_i$

$$s_{i+1}'(x_i+) = \frac{1}{h_{i+1}}(y_{i+1} - y_i) - \frac{h_{i+1}}{3}M_i - \frac{h_{i+1}}{6}M_{i+1} \qquad (i = 0, 1, \ldots, n-1). \tag{3.107}$$

Equality of (3.105) and (3.107) produces the recurrence relation

$$\frac{h_i}{6}M_{i-1} + \frac{1}{3}(h_i + h_{i+1})M_i + \frac{h_{i+1}}{6}M_{i+1}$$

$$= \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \qquad (i = 1, 2, \ldots, n-1). \tag{3.108}$$

For equal intervals, we have $h_i = h_{i+1} = h$ and Eq. (3.108) simplifies to

$$M_{i-1} + 4M_i + M_{i+1} = \frac{6}{h^2}(y_{i+1} - 2y_i + y_{i-1}), \qquad (i = 1, 2, \ldots, n-1). \tag{3.109}$$

The system of Eqs. (3.108) has some special significance. If M_0 and M_n are known, then the system can be written as

Presented By: http://www.ebooksuit.com

$$\left.\begin{aligned}
2(h_1+h_2)M_1+h_2M_2 &= 6\left(\frac{y_2-y_1}{h_2}-\frac{y_1-y_0}{h_1}\right)-h_1M_0\\
h_2M_1+2(h_2+h_3)M_2+h_3M_3 &= 6\left(\frac{y_3-y_2}{h_3}-\frac{y_2-y_1}{h_2}\right)\\
h_3M_2+2(h_3+h_4)M_3+h_4M_4 &= 6\left(\frac{y_4-y_3}{h_4}-\frac{y_3-y_2}{h_3}\right)\\
&\vdots\\
h_{n-1}M_{n-2}+2(h_{n-1}+h_n)M_{n-1} &= 6\left(\frac{y_n-y_{n-1}}{h_n}-\frac{y_{n-1}-y_{n-2}}{h_{n-1}}\right)-h_nM_n.
\end{aligned}\right\}\quad(3.110)$$

Equations (3.108) or (3.109) 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 (3.110) are called *tridiagonal* systems and in the Ch. 6, we shall describe an efficient and accurate method for solving them. When the M_i are known, Eq. (3.104) then gives the required cubic spline in the subinterval $[x_{i-1}, x_i]$. Also, the y_i' can be obtained from Eqs. (3.105) and (3.107).

Example 3.29 Determine the cubic splines satisfying the data of Example 3.27. 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 (3.109) 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 3.27, 3.28 and 3.29. 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. (3.105) and (3.107), we can derive the following relationship for the m_i :

Presented By: http://www.ebooksuit.com

$$\frac{1}{h_i}m_{i-1}+2\left(\frac{1}{h_i}+\frac{1}{h_{i+1}}\right)m_i+\frac{1}{h_{i+1}}m_{i+1}$$

$$=\frac{3}{h_{i+1}^2}(y_{i+1}-y_i)+\frac{3}{h_i^2}(y_i-y_{i-1}), \qquad i=1,2,\ldots,n-1. \quad (3.111)$$

The cubic spline in (x_{i-1}, x_i) in terms of the m_i is then given by

$$s_i(x)=\frac{1}{h_i^2}\{m_{i-1}(x_i-x)^2(x-x_{i-1})-m_i(x-x_{i-1})^2(x_i-x)\}$$

$$+\frac{1}{h_i^3}\{y_{i-1}(x_i-x)^2[2(x-x_{i-1})+h_i]+y_i(x-x_{i-1})^2[2(x_i-x)+h_i]\}. \quad (3.112)$$

The above result can easily be derived using the Hermite interpolation formula given in section 3.9.3.

For equally spaced knots, Eqs. (3.111) assume the simpler form:

$$m_{i-1}+4m_i+m_{i+1}=\frac{3}{h}(y_{i+1}-y_{i-1}), \qquad i=1,2,\ldots,n-1. \quad (3.113)$$

Equations (3.109) or (3.113) 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]. The following example demonstrates the improvement in accuracy of the cubic spline interpolates with successive interval halving.

Example 3.30 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 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

$$s_1(x)=\frac{2}{\pi}\left(-\frac{2x^3}{\pi^2}+\frac{3x}{2}\right).$$

Presented By: http://www.ebooksuit.com

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 eqs. (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

3.14.1 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

Presented By: http://www.ebooksuit.com

$$I = \int_a^b [z''(x)]^2 dx \tag{3.114}$$

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

$$\begin{aligned}\int_a^b [z''(x)]^2 dx &= \int_a^b [s''(x) + z''(x) - s''(x)]^2 dx \\ &= \int_a^b [s''(x)^2 dx + 2\int_a^b s''(x)[z''(x) - s''(x)]dx \\ &\quad + \int_a^b [z''(x) - s''(x)]^2 dx. \end{aligned} \tag{3.115}$$

Now,

$$\begin{aligned}\int_a^b s''(x)[z''(x) - s''(x)]dx &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} s''(x)[z''(x) - s''(x)]dx \\ &= \sum_{i=0}^{n-1} \{s''(x)[z'(x) - s'(x)]\}_{x_i}^{x_{i+1}} \\ &\quad - \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} s'''(x)[z'(x) - s'(x)]dx. \end{aligned} \tag{3.116}$$

The first term in (3.116) 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 (3.116) 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, (3.115) becomes

$$\int_a^b [z''(x)]^2 dx = \int_a^b [s''(x)]^2 dx + \int_a^b [z''(x) - s''(x)]^2 dx \tag{3.117}$$

or

$$\int_a^b [z''(x)]^2 dx \geq \int_a^b [s''(x)]^2 dx. \tag{3.118}$$

It follows that the integral

$$I = \int_a^b [z''(x)]^2 dx$$

Presented By: http://www.ebooksuit.com

will be minimum if and only if

$$\int_b^a [z''(x) - s''(x)]^2 dx = 0, \tag{3.119}$$

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

$$z(x) = s(x), \quad a \le x \le b.$$

3.14.2 Error in the Cubic Spline and Its Derivatives

An estimation of error in the cubic spline and its derivatives will be useful in practical applications.

The natural cubic spline yields a good approximation of a smooth function together with several derivatives, which is testified by the following theorem:

Theorem 3.1 If $y \in c^2[a,b]$, $a = x_0 < x_1 < x_2 < \cdots < x_n = b$, and if $s(x)$ is the natural cubic spline for which

$$s(x_i) = y_i, \quad i = 0, 1, 2, \ldots, n$$

then

$$\max_{x_0 \le x \le x_n} |y(x) - s(x)| \le \frac{1}{2} M h^2, \tag{3.120}$$

where

$$h = x_{i+1} - x_i, \qquad i = 0, 1, 2, \ldots, n$$

and

$$M = \max |y''(x)|, \qquad x_0 \le x \le x_n.$$

It is clear that as the interval length h becomes smaller the better approximation the spline gives. This is in contrast to the known peculiarities of Lagrange interpolation. The errors in the spline derivatives can be obtained by using the operator notation. To find the errors in the first derivatives, we start with the recurrence relation (3.113), viz.,

$$m_{i-1} + 4m_i + m_{i+1} = \frac{3}{h}(y_{i+1} - y_{i-1}).$$

That is,

$$s'(x_{i-1}) + 4s'(x_i) + s'(x_{i+1}) = \frac{3}{h}(y_{i+1} - y_{i-1}).$$

Using the operator notation, the above equation can be written as

$$(E^{-1} + 4 + E)\, s'(x_i) = \frac{3}{h}(E - E^{-1}) y_i. \tag{3.121}$$

Presented By: http://www.ebooksuit.com

Since $E \equiv e^{hD}$, where $D \equiv d/dx$, Eq. (3.121) becomes

$$(e^{-hD} + 4 + e^{hD})s'(x_i) = \frac{3}{h}(e^{hD} - e^{-hD})y_i. \tag{3.122}$$

Now,

$$e^{hD} = 1 + hD + \frac{h^2D^2}{2!} + \frac{h^3D^3}{3!} + \frac{h^4D^4}{4!} + \frac{h^5D^5}{5!} + \cdots$$

and

$$e^{-hD} = 1 - hD + \frac{h^2D^2}{2!} - \frac{h^3D^3}{3!} + \frac{h^4D^4}{4!} - \frac{h^5D^5}{5!} + \cdots$$

Hence

$$e^{hD} + e^{-hD} = 2\left(1 + \frac{h^2D^2}{2} + \frac{h^4D^4}{24} + \frac{h^6D^6}{720} + \cdots\right)$$

and

$$e^{hD} - e^{-hD} = 2\left(hD + \frac{h^3D^3}{6} + \frac{h^5D^5}{120} + \cdots\right).$$

Using the above expressions in (3.122), we obtain

$$\left[2\left(1 + \frac{h^2D^2}{2} + \frac{h^4D^4}{24} + \cdots\right) + 4\right]s'(x_i) = \frac{3}{h} \times 2\left(hD + \frac{h^3D^3}{6} + \frac{h^5D^5}{120} + \cdots\right)y_i$$

$$= 6\left(D + \frac{h^2D^3}{6} + \frac{h^4D^5}{120} + \cdots\right)y_i.$$

The above equation simplifies to

$$s'(x_i) = \frac{6\,(D + h^2D^3/6 + h^4D^5/120 + \cdots)}{6 + h^2D^2 + h^4D^4/12 + \cdots}y_i = \frac{D + h^2D^3/6 + h^4D^5/120 + \cdots}{1 + h^2D^2/6 + h^4D^4/72 + \cdots}y_i$$

$$= \left(D + \frac{h^2D^3}{6} + \frac{h^4D^5}{120} + \cdots\right)\left[1 + \left(\frac{h^2D^2}{6} + \frac{h^4D^4}{72} + \cdots\right)\right]^{-1} y_i$$

$$= \left(D + \frac{h^2D^3}{6} + \frac{h^4D^5}{120} + \cdots\right)\left[1 - \left(\frac{h^2D^2}{6} + \frac{h^4D^4}{72} + \cdots\right)\right.$$

$$\left. + \left(\frac{h^2D^2}{6} + \frac{h^4D^4}{72} + \cdots\right)^2 - \cdots\right]y_i$$

Presented By: http://www.ebooksuit.com

$$= \left(D + \frac{h^2 D^3}{6} + \frac{h^4 D^5}{120} + \cdots \right) \left(1 - \frac{h^2 D^2}{6} - \frac{h^4 D^4}{72} - \cdots + \frac{h^4 D^4}{36} + \cdots \right) y_i$$

$$= \left(D + \frac{h^2 D^3}{6} + \frac{h^4 D^5}{120} + \cdots \right) \left(1 - \frac{h^2 D^2}{6} + \frac{h^4 D^4}{72} - \cdots \right) y_i$$

$$= \left(D - \frac{h^2 D^3}{6} + \frac{h^4 D^5}{72} - \cdots + \frac{h^2 D^3}{6} - \frac{h^4 D^5}{36} + \frac{h^4 D^5}{120} + \cdots \right) y_i$$

$$= \left(D - \frac{1}{180} h^4 D^5 + \cdots \right) y_i.$$

Hence

$$s'(x_i) = y_i' - \frac{1}{180} h^4 y_i^{\text{v}} + O(h^6). \tag{3.123}$$

In a similar manner, we can derive the relations:

$$s''(x_i) = y''(x_i) - \frac{1}{12} h^2 y^{\text{iv}}(x_i) + \frac{1}{360} h^4 y^{\text{vi}}(x_i) + O(h^6) \tag{3.124}$$

$$\frac{1}{2}[s'''(x_i+) + s'''(x_i-)] = y'''(x_i) + \frac{1}{12} h^2 y^{\text{v}}(x_i) + O(h^4). \tag{3.125}$$

$$s'''(x_i+) - s'''(x_i-) = h y^{\text{iv}}(x_i) - \frac{1}{720} h^5 y^{\text{viii}}(x_i) + O(h^7). \tag{3.126}$$

From (3.123) to (3.126), we obtain

$$y'(x_i) = s'(x_i) + O(h^4) \tag{3.127}$$

$$y''(x_i) = s''(x_i) + \frac{1}{12} h^2 y^{\text{iv}}(x_i) + O(h^4) \tag{3.128}$$

$$y'''(x_i) = \frac{1}{2}[s'''(x_i+) + s'''(x_i-)] + O(h^2) \tag{3.129}$$

$$y^{\text{iv}}(x_i) = \frac{1}{2}[s'''(x_i+) - s'''(x_i-)] + O(h^4). \tag{3.130}$$

Relations (3.127)–(3.130) demonstrate that we can approximate $y'(x_i), y''(x_i)$ and $y^{\text{iv}}(x_i)$ more accurately than $y'''(x_i)$, and this fact will be useful in the solution of differential equations with given boundary conditions. The above relations are due to Curtis and Powell [1967].

Presented By: http://www.ebooksuit.com

3.15 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

$$\left.\begin{aligned} L_i(x_j) = \delta_{ij} &= 1, \quad j = i \\ &= 0, \quad j \neq i. \end{aligned}\right\} \tag{3.131}$$

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. The cardinal splines are given by

$$L_i(x) = \frac{1}{h}\left[\frac{(x_j - x)^3}{3!} M_{i,j-1} + \frac{(x - x_{j-1})^3}{3!} M_{i,j} + (x_j - x)\left(\delta_{i,j-1} - \frac{h^2}{3!} M_{i,j-1}\right)\right.$$
$$\left. + (x - x_{j-1})\left(\delta_{i,j} - \frac{h^2}{3!} M_{i,j}\right)\right], \tag{3.132}$$

where $M_{i,j} = L_i''(x_j)$. It is easy to verify that (3.132) satisfies conditions (3.131). As in the case of general splines, the condition of continuity of the first derivatives leads to the recurrence relation

$$M_{i,j-1} + 4M_{i,j} + M_{i,j+1} = \frac{6}{h^2}(\delta_{i,j-1} - 2\delta_{i,j} + \delta_{i,j+1}). \tag{3.133}$$

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

$$s(x) = \sum_{i=0}^{n} L_i(x)\, y_i, \tag{3.134}$$

where $L_i(x)$ are given by (3.132).

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

*See, Ichida and Kiyono [1974].

Presented By: http://www.ebooksuit.com

now is to determine the value of z at an arbitrary point in the rectangular region. The cubic spline formula is given by

$$s(x, y) = \sum_{i=0}^{n} \sum_{j=0}^{n} L_i(x)\, L_j(y)\, z_{i,j}, \tag{3.135}$$

where $L_i(x)$ and $L_j(y)$ are given by formulae of the type (3.132). The spline second derivatives, M_{ij}, are calculated from the recurrence relation (3.133) 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 3.31 Using the data of Example 3.27, 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. (3.132) gives:

$$\begin{aligned} L_i(x) &= \frac{(2-x)^3}{6} M_{i,0} + \frac{(x-1)^3}{6} M_{i,1} + (2-x)\left(\delta_{i,0} - \frac{1}{6} M_{i,0}\right) \\ &\quad + (x-1)\left(\delta_{i,1} - \frac{1}{6} M_{i,1}\right) \\ &= \frac{(x-1)^3}{6} M_{i,1} + (2-x)\,\delta_{i,0} + (x-1)\left(\delta_{i,1} - \frac{1}{6} M_{i,1}\right), \end{aligned} \tag{i}$$

since $M_{i,0} = 0$ for the *natural cubic spline*.

Similarly, the recurrence relation (3.133), 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:

$$L_0(x) = \frac{(x-1)^3}{6}\left(\frac{3}{2}\right) + (2-x) + (x-1)\left(-\frac{1}{4}\right) = \frac{1}{4}(x-1)^3 - \frac{5}{4}x + \frac{9}{4}, \tag{ii}$$

$$L_1(x) = -\frac{1}{2}(x-1)^3 + \frac{3}{2}(x-1), \tag{iii}$$

$$L_2(x) = \frac{1}{4}(x-1)^3 - \frac{1}{4}(x-1). \tag{iv}$$

Hence, in $1 \le x \le 2$, the general natural cubic spline is given by

Presented By: http://www.ebooksuit.com

$$s(x)=\sum_{i=0}^{2} y_i L_i(x)$$

$$=\left[\frac{1}{4}(x-1)^3-\frac{5}{4}x+\frac{9}{4}\right](-8)+\left[\frac{3}{2}(x-1)-\frac{1}{2}(x-1)^3\right](-1)$$

$$+\left[\frac{1}{4}(x-1)^3-\frac{1}{4}(x-1)\right](18)$$

$$=3(x-1)^3+4x-12,$$

which is the same as that obtained in Example 3.29. The next example demonstrates the use of *cardinal splines* in surface fitting.

Example 3.32 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

$$L_0(x)=\frac{x^3}{4}-\frac{5x}{4}+1,$$

$$L_1(x)=-\frac{1}{2}x^3+\frac{3}{2}x,$$

Presented By: http://www.ebooksuit.com

$$L_2(x) = \frac{1}{4}x^3 - \frac{1}{4}x.$$

Hence, in $0 \le x,\ y \le 1$, we have

$$\begin{aligned} s(x, y) &= \sum_{i=0}^{2} \sum_{j=0}^{2} L_i(x)\, L_j(y)\, z_{i,j} \\ &= L_0(x)[L_0(y)z_{0,0} + L_1(y)z_{0,1} + L_2(y)z_{0,2}] \\ &\quad + L_1(x)[L_0(y)z_{1,0} + L_1(y)z_{1,1} + L_2(y)z_{1,2}] \\ &\quad + L_2(x)[L_0(y)z_{2,0} + L_1(y)z_{2,1} + L_2(y)z_{2,2}]. \end{aligned}$$

Since $x = y = 0.5$, the above 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%.

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) $\Delta[f(x)/g(x)]$.

3.3. Locate and correct the error in the following table of values:

x	y
2.5	4.32
3.0	4.83
3.5	5.27
4.0	5.47
4.5	6.26
5.0	6.79
5.5	7.23

Presented By: http://www.ebooksuit.com

3.4. Prove the following:

(a) $u_x = u_{x-1} + \Delta u_{x-2} + \Delta^2 u_{x-3} + \cdots + \Delta^{n-1} u_{x-n} + \Delta^n u_{x-n-1}$

(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) $u_1 + u_2 + \cdots + u_n = {}^nC_1 u_0 + {}^nC_2 \Delta u_0 + \cdots + \Delta^{n-1} u_0$.

3.5. From the following table, find the number of students who obtained marks between 60 and 70:

Marks obtained	No. of students
0–40	250
40–60	120
60–80	100
80–100	70
100–120	50

3.6. In the following table, the values of y are consecutive terms of a series of which the number 31 is the 5th term. Find the first and the tenth terms of the series. Find also the polynomial which approximates these values:

x	y
3	13
4	21
5	31
6	43
7	57
8	73
9	91

3.7. From the following table of values of x and $f(x)$, determine (i) $f(0.23)$ and (ii) $f(0.29)$:

x	$f(x)$
0.20	1.6596
0.22	1.6698
0.24	1.6804
0.26	1.6912
0.28	1.7024
0.30	1.7139

3.8. Find the 7th term and the general term of the series 3, 9, 20, 38, 65, ...

Presented By: http://www.ebooksuit.com

3.9. The following values are taken from the table of cubes:

x	$y = x^3$
6.1	226.981
6.2	238.328
6.3	250.047
6.4	262.144
6.5	274.625
6.6	287.496
6.7	300.763

Find $(6.36)^3$ and $(6.61)^3$.

3.10. Define the operators, Δ, ∇, δ and E, E^{-1} and show that

(i) $\Delta \equiv E\nabla$ (ii) $\nabla = E^{-1}\Delta$

(iii) $E \equiv 1 + \Delta$ (iv) $E^{-1} \equiv 1 - \nabla$

(v) $\Delta y_k = \nabla^r y_{k+r} = \delta^r y_{k+r,2}$ (vi) $\Delta\nabla y_k = \nabla\Delta y_k = \delta^2 y_k$

(vii) $\Delta(y_k^2) = (y_k + y_{k+1})\Delta y_k$ (viii) $\Delta(1/y_k) = -\Delta y_k/(y_k y_{k+1})$

3.11. Show that $E \equiv 1 + \Delta$ and $\Delta \equiv \nabla(1-\nabla)^{-1}$. Also, deduce that $1 + \Delta \equiv (E-1)\nabla^{-1}$.

3.12. The population of a town in decennial census were as under. Estimate the population for the year 1955

Year	Population (in thousands)
1921	46
1931	66
1941	81
1951	93
1961	101

3.13. 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 $3^3 = 27$?

Presented By: http://www.ebooksuit.com

3.14. The probability integral

$$p = \sqrt{\frac{2}{\pi}} \int_0^x \exp\left(-\frac{1}{2}t^2\right) dt$$

has the following values:

x	p
1.00	0.682689
1.05	0.706282
1.10	0.728668
1.15	0.749856
1.20	0.769861
1.25	0.788700

Calculate p for $x = 1.235$.

3.15. Prove the following relations where the operators have their usual meanings

(i) $\delta^2 E = \Delta^2$ (ii) $E^{-1/2} = \mu - \delta/2$

(iii) $1 + \delta^2\mu^2 = (1 + \delta^2/2)^2$ (iv) $\mu E = E\mu$

(v) $\nabla = \delta E^{-1/2}$ (vi) $\Delta - \nabla = \delta^2$

(vii) $\mu = \cosh(u/2)$ where $u = hD$

(viii) $f'(x) = \mu\delta f(x) - (1/6)\mu\delta^2 f(x) + (1/30)\mu\delta^5 f(x)$

3.16. The values of the elliptic integral

$$K(m) = \int_0^{\pi/2} (1 - m\sin^2\theta)^{-1/2} d\theta$$

for certain equidistant values of m are given below. Use Everett's or Bessel's formula to determine $K(0.25)$.

m	$K(m)$
0.20	1.659624
0.22	1.669850
0.24	1.680373
0.26	1.691208
0.28	1.702374
0.30	1.713889

3.17. From Bessel's formula, derive the following formula for midway interpolation:

Presented By: http://www.ebooksuit.com

$$y_{1/2} = \frac{1}{2}(y_0 + y_1) - \frac{1}{16}(\Delta^2_{y-1} + \Delta^2_{y0}) + \frac{3}{256}(\Delta^4_{y-2} + \Delta^4_{y-1}) - \cdots$$

Also, deduce this formula from Everett's formula.

3.18. State, without proof, Stirling's formula for central interpolation and mention its limitations.

From the following table of values of x and $y = e^x$, interpolate the value of y when $x = 1.91$

x	$y = e^x$
1.7	5.4739
1.8	6.0496
1.9	6.6859
2.0	7.3891
2.1	8.1662
2.2	9.0250

3.19. Use Stirling's formula to find u_{32} from the following table:

$u_{20} = 14.035$, $u_{25} = 13.674$, $u_{30} = 13.257$,

$u_{35} = 12.734$, $u_{40} = 12.089$, $u_{45} = 11.309$.

3.20. From the following table, find y when $x = 1.45$.

x	y
1.0	0.0
1.2	−0.112
1.4	−0.016
1.6	0.336
1.8	0.992
2.0	2.0

3.21. The following values of x and y are given. Find $y(0.543)$:

x	$y(x)$
0.1	2.631
0.2	3.328
0.3	4.097
0.4	4.944
0.5	5.875
0.6	6.896
0.7	8.013

Presented By: http://www.ebooksuit.com

3.22. Using Gauss's forward formula, find the value of $f(32)$ given that

$$f(25) = 0.2707,\ f(30) = 0.3027,$$
$$f(35) = 0.3386,\ f(40) = 0.3794.$$

3.23. Using Gauss's backward formula, find the value of $\sqrt{12516}$ given that

$$\sqrt{12500} = 111.803399, \qquad \sqrt{12510} = 111.848111,$$
$$\sqrt{12520} = 111.892806, \qquad \sqrt{12530} = 111.937483$$

3.24. Evaluate sin (0.197) from the following table:

x	$\sin x$
0.15	0.14944
0.17	0.16918
0.19	0.18886
0.21	0.20846
0.23	0.22798

3.25. Using Everett's formula, evaluate $f(25)$ from the following table:

x	$f(x)$
20	2854.0
24	3162.0
28	3544.0
32	3992.0

3.26. Given the table of values:

x	$y = \sqrt{x}$
150	12.247
152	12.329
154	12.410
156	12.490

evaluate $\sqrt{155}$ using Lagrange's interpolation formula.

3.27. Show that

$$\sum_{i=1}^{n} \frac{\pi(t)}{(t - t_i)\,\pi'(t_i)} = 1$$

Presented By: http://www.ebooksuit.com

3.28. If $x(t)$ is analytic inside the closed contour C and if $t, t_1, t_2, \ldots, t_n$ lie inside C, show that the remainder term in the error formula for polynomial interpolation can be written as

$$\frac{\pi(t)}{2\pi i} \int_C \frac{x(\tau)}{(\tau - t)\,\pi(\tau)} d\tau.$$

3.29. Show that $\sum_{i=0}^{n} l_i(x) = 1$ for all x.

3.30. Applying Lagrange's formula, find a cubic polynomial which approximates the following data:

x	$y(x)$
-2	-12
-1	-8
2	3
3	5

3.31. Using Lagrange's formula, express the rational function

$$\frac{3x^2 + x + 1}{(x-1)(x-2)(x-3)}$$

as a sum of partial fractions.

[*Hint:* Let $f(x) = 3x^2 + x + 1$. Form a table of values of $f(x)$ for $x = 1, 2, 3$. Obtain the second-order Lagrange polynomial $L_2(x)$]. [Stanton]

3.32. Express the function $(x^2 + x - 3)/(x^3 - 2x^2 - x + 2)$ as a sum of partial fractions.

3.33. Given the data points $(1, -3)$, $(3, 9)$, $(4, 30)$ and $(6, 132)$ satisfying the function $y = f(x)$, compute $f(5)$ using Lagrange polynomials of orders 1 to 3.

3.34. 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.

3.35. If $f(x) = 1/x^2$, find the divided-differences $[a, b]$ and $[a, b, c]$.

3.36. Given the set of tabulated points $(1, -3)$, $(3, 9)$, $(4, 30)$ and $(6, 132)$, obtain the value of y when $x = 2$ using:

(a) Newton's divided-difference formulae of orders 1 to 3, and

(b) Aitken's method.

3.37. Show that the nth divided-difference $[x_0, x_1, \ldots, x_n]$ can be expressed as the quotient of two determinants, shown as follows:

Presented By: http://www.ebooksuit.com

$$[x_0, x_1, \ldots, x_n] = \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_0 & x_1 & x_2 & \cdots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & & \vdots \\ x_0^{n-1} & x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \\ y_0 & y_1 & y_2 & \cdots & y_n \end{vmatrix}$$

$$\div \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_0 & x_1 & x_2 & \cdots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & & \vdots \\ x_0^n & x_1^n & x_2^n & \cdots & x_n^n \end{vmatrix}$$

3.38. If the abscissae $x_i, i = 0, 1, \ldots, n$ are distinct and if $y = f(x)$ is n times continuously differentiable, show that

$$[x_0, x_1, \ldots, x_n] = \int \cdots \int y^{(n)}(t_0 x_0 + t_1 x_1 + \cdots + t_n x_n)\, dt_1 \ldots dt_n,$$

where $t_0 \geq 0$ and $\Sigma_{i=0}^{n} t_i = 1$.

3.39. Tabulate the function $y = \sin x$ for $x = 0$ to 1.0 in steps of $h = 0.01$. Find the error of linear interpolation in this table.

3.40. Find the error of quadratic interpolation in the above example.

3.41. Prove that the third divided difference of the function $f(x) = 1/x$ with arguments p, q, r, s is $-1/(pqrs)$.

3.42. If $f(x) = 1/x$, prove that

$$[x_0, x_1, \ldots, x_r] = \frac{(-1)^r}{x_0 x_1 \ldots x_r}.$$

3.43. Given the table of values

x	$\sqrt[3]{x}$
50	3.684
52	3.732
54	3.779
56	3.825

Use Lagrange's formula to find x when $\sqrt[3]{x} = 3.756$.

Presented By: http://www.ebooksuit.com

3.44. From the table of values

x	y
1.8	2.9422
2.0	3.6269
2.2	4.4571
2.4	5.4662
2.6	6.6947

find x when $y = 5.0$ using the method of successive approximations.

3.45. From the following table of values, find x for which $\sinh x = 5$:

x	$\sinh x$
2.2	4.457
2.4	5.466
2.6	6.695
2.8	8.198
3.0	10.018

3.46. Develop a subprogram, in FORTRAN or C, to implement Lagrange interpolation and test it on the data of Problem 24. Compare it with the result obtained by using the MATLAB 'polyfit function' to fit a fifth-order polynomial.

3.47. *Reciprocal differences:* The concept of reciprocal differences will be useful in determining a continued fraction approximation which agrees with a tabulated function $f(x)$ at the set of points $x_0, x_1, \ldots, x_n$.

We define quantities $\phi_0[x], \phi_1[x_0, x], \phi_2[x_0, x_1, x], \phi_3[x_0, x_1, x_2, x], \ldots$, called the *reciprocal differences* in the following way:

$$\phi_0[x] = f(x)$$

$$\phi_1[x_0, x] = \frac{x - x_0}{\phi_0[x] - \phi_0[x_0]} = \frac{x - x_0}{f(x) - f(x_0)}$$

$$\phi_2[x_0, x_1, x_2] = \frac{x - x_1}{\phi_1[x_0, x] - \phi_1[x_0, x_1]}$$

and so on. Then following the procedure outlined in the derivation of Newton's divided-difference formula, we derive the general formula

$$\phi_0[x] = f(x_0) + \frac{x - x_0}{\phi_1[x_0, x_1]} + \frac{x - x_1}{\phi_2[x_0, x_1, x_2]} + \frac{x - x_2}{\phi_3[x_0, x_1, x_2, x_3]} + \cdots,$$

which is the required *continued fraction approximation* to the given set of tabulated values.

Presented By: http://www.ebooksuit.com

Use the above method to obtain a continued fraction approximation to the set of points (1, 1) (2, 4),(3, 9) and (4, 16).

3.48. Apply reciprocal differences to recover the function $f(x) = 1/(1+x^2)$ from the following data:

x	$f(x)$
0	1
1	1/2
2	1/5
3	1/10
4	1/17
5	1/26

3.49. Using Hermite's interpolation formula, estimate the value of ln 3.2 from the following table:

x	$y = \ln x$	$y' = 1/x$
3.0	1.09861	0.33333
3.5	1.25276	0.28571
4.0	1.38629	0.25000

3.50. Find the Hermite polynomial of the third degree approximating the function $y(x)$ such that $y(x_0) = 1$, $y(x_1) = 0$ and $y'(x_0) = y'(x_1) = 0$.

3.51. Show that the error in Hermite's formula is given by

$$y(x) - H_{2n+1}(x) = \frac{[\pi_{n+1}(x)]^2}{(2n+2)!} y^{(2n+2)}(\xi),$$

where $y(x)$ is assumed to have continuous derivatives of order $(2n+2)$ and $\xi = \xi(x)$ is in the interval determined by the points $x, x_0, \ldots, x_n$.

3.52. The function $y = x^3 + 9$ is tabulated below:

x	y
3	36
4	73
5	134

Predict the value of $y(4.5)$ by using quadratic and cubic splines and state the absolute error in each case.

3.53. Fit a cubic spline to the function defined by the set of points given in the table:

Presented By: http://www.ebooksuit.com

x	$y = e^x$
0.10	1.1052
0.15	1.1618
0.20	1.2214
0.25	1.2840
0.30	1.3499

Use the end conditions:

(i) $M_0 = M_N = 0$

(ii) $s'(0.10) = y'(0.10)$ and $s'(0.30) = y'(0.30)$.

(iii) $s''(0.10) = y''(0.10)$ and $s''(0.30) = y''(0.30)$.

Interpolate in each case for $x = 0.12$ and state which of the end conditions gives the best fit.

3.54. Deduce the expression for the error in the spline second derivative:

$$s''(x_i) = y_i'' - \frac{1}{12} h^2 y_i^{\text{iv}} + O(h^4).$$

3.55. Determine the cubic spline $s(x)$ valid in the interval $[x_{i-1}, x_i]$ for the following data, given that $s''(x_0) = y''(x_0)$ and $s''(x_n) = y''(x_n)$:

(a)

x	$y = x \ln x$
6.2	11.3119
6.4	11.8803
6.6	12.4549

(b)

x	$y = \tan x$
1.3	3.6021
1.4	5.7979
1.5	14.1014

3.56. In the interval $[x_i, x_{i+1}]$, the cubic spline $s_i(x)$ may be expressed as

$$s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \; i = 0, 1, \ldots, n-1$$

Determine the constants a_i, b_i, c_i and d_i, using the conditions for a natural cubic spline.

3.57. Develop a subprogram to implement the natural cubic spline interpolation and test your program on the data of Problem 53. Also, use MATLAB spline function on the same data and compare the results.

3.58. The following table gives the values of $z = f(x, y)$ for different values of x and y. Use the methed of Section 3.12 to find z when $x = 2.5$ and $y = 1.5$. Compare your result with the actual value obtained from $f(x, y) = x^2 + y^2 + y$.

Presented By: http://www.ebooksuit.com

	x		
y	1	2	3
1	3	6	11
2	7	10	15
3	13	16	21

3.59. Repeat problem 58 using cardinal splines.

3.60. Develop a subprogram to evaluate the cardinal splines $L_i(x)$ and the general cubic spline $s(x, y)$. Test your subprogram on the data of problem 58.

Presented By: http://www.ebooksuit.com

CHAPTER

Least Squares, B-splines 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. This is contrary to the case considered in the preceding chapter where we assumed that the data are free of errors. A common strategy for such cases is to derive an approximating function that broadly fits the general trend of the data without necessarily passing through the individual points. The curve drawn is such that the discrepancy between the data points and the curve is least. The method of least squares is most commonly applied in such cases and is described in the earlier sections of this chapter.

Data fiting by means of polynomials has been considered, in the previous chapter, from the viewpoint of cubic splines. The resulting approximation, called the cubic spline approximation, suffers from the disadvantage of being a global approximation, which means that a change in one point affects the entire approximating curve. We describe, in the present chapter, a method based on basis splines, called *B-splines*, which possess a local character, viz., a change in one point introduces a change only in the immediate neighbourhood of that point. The B-spline method finds important applications in computer graphics and smoothing of data. The 'B-spline and its computation' will be discussed in Section 4.5.

In the previous chapter, we concentrated on polynomial interpolation, i.e. interpolation based on a linear combination of the functions $1, x, x^2, \ldots, x^n$. On the other hand, trigonometric interpolation, i.e. interpolation based on trigonometric functions $1, \cos x, \cos 2x, \ldots, \cos nx, \sin x, \sin 2x, \ldots, \sin nx$, plays

Presented By: http://www.ebooksuit.com

an important role in modelling oscillating or vibrating, systems. The Fourier series is a useful tool for dealing with periodic systems but for aperiodic waveforms, the Fourier integral or the Fourier transform is the primary tool available. Numerical methods for the computation of discrete 'Fourier transforms' are discussed in Section 4.6.

We shall finally consider, in the concluding section, the representation of functions by Chebyshev polynomials as also the economization of power series. These are important from the standpoint of digital computation.

4.2 LEAST-SQUARES CURVE FITTING PROCEDURES

Usually a mathematical equation is fitted to experimental data by plotting the data on a graph paper and then passing a straight line through the data points. The method has the obvious drawback in that the straight line drawn may not be unique. The method of least squares is probably the most systematic procedure to fit a unique curve through given data points and is widely used in practical computations. It can also be easily implemented on a digital computer.

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 experimental (or observed) value of the 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 $Y = a_0 + a_1 x$ be the straight line to be fitted to the given data. Then, corresponding to Eq. (4.2) we have

$$S = [y_1 - (a_0 + a_1 x_1)]^2 + [y_2 - (a_0 + a_1 x_2)]^2 + \cdots + [y_m - (a_0 + a_1 x_m)]^2. \tag{4.3}$$

For S to be minimum, we have

$$\frac{\partial S}{\partial a_0} = 0 = -2[y_1 - (a_0 + a_1 x_1)] - 2[y_2 - (a_0 + a_1 x_2)] - \cdots - 2[y_m - (a_0 + a_1 x_m)] \tag{4.4a}$$

Presented By: http://www.ebooksuit.com

and

$$\frac{\partial S}{\partial a_1} = 0 = -2x_1[y_1 - (a_0 + a_1x_1)] - 2x_2[y_2 - (a_0 + a_1x_2)]$$
$$- \cdots - 2x_m[y_m - (a_0 + a_1x_m)]. \tag{4.4b}$$

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

$$a_0(x_1 + x_2 + \cdots + x_m) + a_1(x_1^2 + x_2^2 + \cdots + x_m^2) = x_1y_1 + x_2y_2 + \cdots + x_my_m \tag{4.5b}$$

or, more compactly to

$$ma_0 + a_1\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \tag{4.6a}$$

and

$$a_0 + \sum_{i=1}^{m} x_i + a_1\sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_iy_i. \tag{4.6b}$$

Since the x_i and y_i are known quantities, Eqs. (4.5) or (4.6), called the *normal equations*, can be solved for the two unknown a_0 and a_1.

Differentiating Eqs. (4.4a) and (4.4b) with respect to a_0 to a_1 respectively, we find that $\partial^2 S/\partial a_0^2$ and $\partial^2 S/\partial a_1^2$ will both be positive at the points a_0 and a_1. Hence these values provide a *minimum* of S.

Further, dividing Eqs. (4.6a) throughout by m, we obtain

$$a_0 + a_1\bar{x} = \bar{y},$$

where $(\bar{x}, \bar{y})$ is the centroid of the given data points. It follows that the fitted straight line passes through the centroid of the data points. The following example demonstrates the working of this method.

Example 4.1 The table below gives the temperatures T (in °C) and lengths l (in mm) of a heated rod. If $l = a_0 + a_1T$, find the best values for a_0 and a_1.

T (in °C)	l (in mm)
20	800.3
30	800.4
40	800.6
50	800.7
60	800.9
70	801.0

Presented By: http://www.ebooksuit.com

To use formulae (4.6), we require ΣT, Σl, ΣT^2 and ΣTl, and these are computed as in the following table:

T (in °C)	l (in mm)	T^2	Tl
20	800.3	400	16006
30	800.4	900	24012
40	800.6	1600	32024
50	800.7	2500	40035
60	800.9	3600	48054
70	801.0	4900	56070
270	4803.9	13900	216201

Using formulae (4.6) we then obtain

$$6a_0 + 270a_1 = 4803.9 \quad \text{and} \quad 270a_0 + 13900a_1 = 216201,$$

from which we get $a_0 = 800$ and $a_1 = 0.0146$.

Example 4.2 Certain experimental values of x and y are given below

x	y
0	−1
2	5
5	12
7	20

If $y = a_0 + a_1x$, find approximate values of a_0 and a_1. As in the previous example, we form the following table of values:

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

Formulae (4.6) give the two equations

$$4a_0 + 14a_1 + 36 \quad \text{and} \quad 14a_0 + 78a_1 = 210.$$

Solving the above two equations, we obtain $a_0 = -1.1381$ and $a_1 = 2.8966$. Using these values we obtain $y(5) \approx 13.3449$.

It may be noted that the given table is obtained from the relation $y = -1.0334 + 2.6222x$ so that the correct value of $y(5)$ is 12.0776.

4.2.2 Nonlinear Curve Fitting

In this section, we consider a power function, a polynomial of the nth degree and an exponential function to fit the given data points

$$(x_i, y_i), \quad i = 1, \ldots, m$$

Presented By: http://www.ebooksuit.com

Power function Let $y = ax^c$ be the function to be fitted to the given data. Taking logarithms of both sides, we obtain the relation

$$\log y = \log a + c \log x, \tag{4.7}$$

which is of the form $Y = a_0 + a_1 X$, where $Y = \log y$, $a_0 = \log a$, $a_1 = c$ and $X = \log x$. Hence the procedure outlined in the previous section can be followed to evaluate a_0 and a_1. Then a and c can be calculated from the formulae $a_0 = \log a$ and $c = a_1$.

Polynomial of the nth degree Let the polynomial of the nth degree, viz.,

$$Y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \tag{4.8}$$

be fitted to the data points (x_i, y_i), $i = 1, 2, \ldots, m$. We then have

$$S = [y_1 - (a_0 + a_1 x_1 + \cdots + a_n x_1^n)]^2 + [y_2 - (a_0 + a_1 x_2 + \cdots + a_n x_2^n)]^2$$
$$+ \cdots + [y_m - (a_0 + a_1 x_m + \cdots + a_n x_m^n)]^2. \tag{4.9}$$

Equating, as before, the first partial derivatives to zero and simplifying, we get the following normal equations

$$\left.\begin{aligned}
m a_0 + a_1 \sum_{i=1}^{m} x_i + a_2 \sum_{i=1}^{m} x_i^2 + \cdots + a_n \sum_{i=1}^{m} x_i^n &= \sum_{i=1}^{m} y_i \\
a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 + \cdots + a_n \sum_{i=1}^{m} x_i^{n+1} &= \sum_{i=1}^{m} x_i y_i \\
&\vdots \\
a_0 \sum_{i=1}^{m} x_i^n + a_1 \sum_{i=1}^{m} x_i^{n+1} + \cdots + a_n \sum_{i=1}^{m} x_i^{2n} &= \sum_{i=1}^{m} x_i^n y_i.
\end{aligned}\right\} \tag{4.10}$$

These are $(n+1)$ equations in $(n+1)$ unknowns and hence can be solved for $a_0, a_1, \ldots, a_n$. Equation (4.8) then gives the required polynomial of the nth degree.

It should be noted that for large values of n, the normal equations given by (4.10) are unstable (or ill-conditioned) with the result that roundoff errors in the data may cause large changes in the solution. Such systems occur, quite often, in practice and we shall study their nature in a later chapter. It is sufficient to remark here that 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 4.3 Fit a polynomial of the second degree to the data points given in the following table

x	y
0.0	1.0
1.0	6.0
2.0	17.0

Presented By: http://www.ebooksuit.com

In Eq. (4.10), we require the quantities $\Sigma x_i, \Sigma x_i^2, \Sigma x_i^3, \Sigma x_i^4, \Sigma y_i, \Sigma x_i y_i$ and $\Sigma x_i^2 y_i$. These are computed as in the following table:

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

Using Eqs. (4.10), we now obtain the equations

$$3a_0 + 3a_1 + 5a_2 = 24$$

$$3a_0 + 5a_1 + 9a_2 = 40$$

$$5a_0 + 9a_1 + 17a_2 = 74.$$

the solution to which is $a_0 = 1$, $a_1 = 2$ and $a_2 = 3$.

The required polynomial is then given by $Y = 1 + 2x + 3x^2$. From the given data points, it is seen that this polynomial fitting is 'exact'.

Exponential function Let the curve

$$y = a_0 e^{a_1 x} \tag{4.11}$$

be fitted to the given data. Then, as before, taking logarithms of both sides of (4.11), we get

$$\log y = \log a_0 + a_1 x, \tag{4.12}$$

which can be written in the form

$$Z = A + Bx,$$

where $Z = \log y$, $A = \log a_0$ and $B = a_1$. The problem therefore reduces to finding a least-squares straight line through the given data.

Example 4.4 Determine the constants a and b by the method of least squares such that $y = ae^{bx}$ fits the following data

x	y
2	4.077
4	11.084
6	30.128
8	81.897
10	222.62

The given relation is $y = ae^{bx}$. Taking logarithms of both sides, we obtain

$$\ln y = \ln a + bx.$$

Setting $\ln y = Y, x = X$, $\ln a = a_0$ and $b = a_1$, the above relation takes the form $y = a_0 + a_1 x$, which is a straight line.

Presented By: http://www.ebooksuit.com

The method of procedure is the same as in Section 4.2.1 and we form the following table:

$X = x$	$Y = \ln y$	X^2	XY
2	1.405	4	2.810
4	2.405	16	9.620
6	3.405	36	20.430
8	4.405	64	35.240
10	5.405	100	54.050
30	17.025	220	122.150

Formulae (4.6) give

$$5a_0 + 30a_1 = 17.025, \qquad 30a_0 + 220a_1 = 122.150,$$

which yield the solution:

$$a_0 = 0.405 \quad \text{and} \quad a_1 = 0.5.$$

Hence

$$a = e^{a0} = e^{0.405} = 1.499 \quad \text{and} \quad b = a_1 = 0.5.$$

4.2.3 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

$$y = f(x) = A_1 e^{\lambda_1 x} + A_2 e^{\lambda_2 x} + \cdots + A_n e^{\lambda_n x} \tag{4.13}$$

to a set of data points, say $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

In (4.13), we assume that n is known and $A_1, A_2, \ldots, A_n, \lambda_1, \lambda_2, \ldots, \lambda_n$ are to be determined. It can be seen that $f(x)$ satisfies a differential equation of the type

$$\frac{d^n y}{dx^n} + a_1 \frac{d^{n-1} y}{dx^{n-1}} + a_2 \frac{d^{n-2} y}{dx^{n-2}} + \cdots + a_n y = 0, \tag{4.14}$$

where the coefficients $a_1, a_2, \ldots, a_n$ are presently unknown.

A method suggested by Froberg consists in numerically evaluating the derivatives $d^n y/dx^n$, $d^{n-1} y/dx^{n-1}, \ldots$ at the n data points and substituting them in (4.14), thus obtaining a system of n linear equations for the n unknowns $a_1, a_2, \ldots, a_n$, which can then be solved. Again, it can be verified that $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the roots of the algebraic equation

$$\lambda^n + a_1 \lambda^{n-1} + a_2 \lambda^{n-2} + \cdots + a_n = 0, \tag{4.15}$$

which when solved enables us to compute $A_1, A_2, \ldots, A_n$ from (4.13) by the method of least squares. An obvious disadvantage of the method is the numerical evaluation of the derivatives whose accuracy deteriorates with their increasing order and leading, therefore, to unreliable results.

Presented By: http://www.ebooksuit.com

We describe now a computational technique, due to Moore [1974], which leads to more reliable results. We demonstrate the method for the case $n = 2$.

Let the function to be fitted to a given data be of the form

$$y = A_1 e^{\lambda_1 x} + A_2 e^{\lambda_2 x}, \tag{4.16}$$

which 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. Assuming that 'a' is the initial value of x, we obtain, by integrating (4.17) from 'a' to x, the following equation

$$y'(x) - y'(a) = a_1 y(x) - a_1 y(a) + a_2 \int_a^x y(x)\, dx, \tag{4.18}$$

where $y'(x)$ denotes dy/dx.

Integrating (4.18) again from a to x, we obtain

$$y(x) - y(a) - y'(a)(x-a) = a_1 \int_a^x y(x)\, dx - a_1(x-a)y(a) + a_2 \int_a^x \int_a^x y(x)\, dx\, dx. \tag{4.19}$$

Using the formula

$$\int_a^x \cdots \int_a^x f(x)\, dx \ldots dx = \frac{1}{(n-1)!} \int_a^x (x-t)^{n-1} f(t)\, dt \tag{4.20}$$

Eq. (4.19) simplifies to

$$y(x) - y(a) - (x-a)y'(a) = a_1 \int_a^x y(x)\, dx - a_1(x-a)\, y(a) + a_2 \int_a^x (x-t)\, y(t)\, dt. \tag{4.21}$$

In order to use Eq. (4.21) to set up a linear system for a_1 and a_2, $y'(a)$ should be eliminated and this is done in the following way. Choosing two data points x_1 and x_2 such that $a - x_1 = x_2 - a$, we obtain from (4.21)

$$y(x_1) - y(a) - (x_1 - a)y'(a) = a_1 \int_a^{x_1} y(x)\, dx - a_1(x_1 - a)\, y(a) + a_2 \int_a^{x_1} (x_1 - t)\, y(t)\, dt$$

$$y(x_2) - y(a) - (x_2 - a)y'(a) = a_1 \int_a^{x_2} y(x)\, dx - a_1(x_2 - a)\, y(a) + a_2 \int_a^{x_2} (x_2 - t)\, y(t)\, dt.$$

Presented By: http://www.ebooksuit.com

On adding the above two equations and simplifying, we obtain

$$y(x_1)+y(x_2)-2y(a)=a_1\left[\int_a^{x_1} y(x)\,dx+\int_a^{x_2} y(x)\,dx\right]$$

$$+a_2\left[\int_a^{x_1}(x_1-t)\,y(t)\,dt+\int_a^{x_2}(x_2-t)\,y(t)dt\right]. \quad (4.22)$$

Equation (4.22) 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

$$\lambda^2=a_1\lambda+a_2. \quad (4.23)$$

Finally, A_1 and A_2 can be obtained by the method of least squares. The method of procedure is illustrated by the following example.

Example 4.5 Fit a function of the form

$$y=A_1e^{\lambda_1 x}+A_2e^{\lambda_2 x} \quad \text{(i)}$$

to the data given by

x	y	x	y
1.0	1.54	1.5	2.35
1.1	1.67	1.6	2.58
1.2	1.81	1.7	2.83
1.3	1.97	1.8	3.11
1.4	2.15		

Choosing $x_1=1.0$, $a=1.2$ and $x_2=1.4$, Eq. (4.22) 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

$$1.81a_1+2.180a_2=2.10. \quad \text{(ii)}$$

Again, choosing $x_1=1.4$, $a=1.6$ and $x_2=1.8$, and evaluating the integrals in (4.22), as before, we obtain the equation

$$2.88a_1+3.104a_2=3.00. \quad \text{(iii)}$$

Presented By: http://www.ebooksuit.com

Solving (ii) and (iii), we obtain $a_1 = 0.03204$ and $a_2 = 0.9364$. Equation (4.23) now gives

$$\lambda^2 - 0.03204\lambda - 0.9364 = 0,$$

from which we obtain

$$\lambda_1 = 0.988 = 0.99 \quad \text{and} \quad \lambda_2 = -0.96.$$

Using the method of least squares, we 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 $A_1 = A_2 = 1/2$, $\lambda_1 = 1.0$ and $\lambda_2 = -1.0$.

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

$$\begin{aligned} S &= W_1 [y_1 - f(x_1)]^2 + W_2 [y_2 - f(x_2)]^2 + \cdots + W_m [y_m - f(x_m)]^2 \\ &= W_1 e_1^2 + W_2 e_2^2 + \cdots + W_m e_m^2. \end{aligned} \tag{4.24}$$

In (4.24), 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

$$S(a_0, a_1) = \sum_{i=1}^{m} W_i [y_i - (a_0 + a_1 x_i)]^2. \tag{4.25}$$

For maxima or minima, we have

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = 0, \tag{4.26}$$

which give

$$\frac{\partial S}{\partial a_0} = -2 \sum_{i=1}^{m} W_i [y_i - (a_0 + a_1 x_i)] = 0 \tag{4.27}$$

Presented By: http://www.ebooksuit.com

and

$$\frac{\partial S}{\partial a_1} = -2\sum_{i=1}^{m} W_i\,[y_i - (a_0 + a_1 x_i)]\,x_i = 0. \tag{4.28}$$

Simplification yields the system of equations for a_0 and a_1:

$$a_0\sum_{i=1}^{m} W_i + a_1\sum_{i=1}^{m} W_i x_i = \sum_{i=1}^{m} W_i y_i \tag{4.29}$$

and

$$a_0\sum_{i=1}^{m} W_i x_i + a_1\sum_{i=1}^{m} W_i x_i^2 = \sum_{i=1}^{m} W_i x_i y_i, \tag{4.30}$$

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 4.6 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. (4.29) and (4.30) 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

$$y = -1.349345 + 2.73799x.$$

We obtain

$$y(5.0) \approx 12.34061 = 12.34061,$$

which is a better approximation than that obtained in Example 4.2.

Presented By: http://www.ebooksuit.com

Example 4.7 We consider Example 4.6 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 above 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

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \tag{4.31}$$

be fitted to the given data points. We then have

$$S(a_0, a_1, \ldots, a_n) = \sum_{i=1}^{m} W_i\,[y_i - (a_0 + a_1x_i + \cdots + a_nx_i^n)]^2. \tag{4.32}$$

If a minimum occurs at $(a_0, a_1, \ldots, a_n)$, then we have

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = \frac{\partial S}{\partial a_2} = \cdots = \frac{\partial S}{\partial a_n} = 0. \tag{4.33}$$

Presented By: http://www.ebooksuit.com

These conditions yield the normal equations

$$\left.\begin{aligned}
a_0\sum_{i=1}^{m} W_i + a_1\sum_{i=1}^{m} W_i x_i + \cdots + a_n\sum_{i=1}^{m} W_i x_i^n &= \sum_{i=1}^{m} W_i y_i \\
a_0\sum_{i=1}^{m} W_i x_i + a_1\sum_{i=1}^{m} W_i x_i^2 + \cdots + a_n\sum_{i=1}^{m} W_i x_i^{n+1} &= \sum_{i=1}^{m} W_i x_i y_i \\
&\vdots \\
a_0\sum_{i=1}^{m} W_i x_i^n + a_1\sum_{i=1}^{m} W_i x_i^{n+1} + \cdots + a_n\sum_{i=1}^{m} W_i x_i^{2n} &= \sum_{i=1}^{m} W_i x_i^n y_i.
\end{aligned}\right\} \tag{4.34}$$

Equations (4.34) 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

$$y(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \tag{4.35}$$

be chosen to minimize

$$S(a_0, a_1, \ldots, a_n) = \int_a^b W(x)[y(x) - (a_0 + a_1 x + \cdots + a_n x^n)]^2\, dx. \tag{4.36}$$

The necessary conditions for a minimum are given by

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = \cdots = \frac{\partial S}{\partial a_n} = 0, \tag{4.37}$$

which yield

$$\left.\begin{aligned}
-2\int_a^b W(x)[y(x) - (a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n)]\, dx &= 0 \\
-2\int_a^b W(x)[y(x) - (a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n)]\, x\, dx &= 0 \\
-2\int_a^b W(x)[y(x) - (a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n)]\, x^2\, dx &= 0 \\
&\vdots \\
-2\int_a^b W(x)[y(x) - (a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n)]\, x^n\, dx &= 0.
\end{aligned}\right\} \tag{4.38}$$

Presented By: http://www.ebooksuit.com

Rearrangement of terms in (4.38) gives the system

$$\left.\begin{aligned}
a_0 \int_a^b W(x)dx + a_1 \int_a^b xW(x)\,dx + \cdots + a_n \int_a^b x^n W(x)\,dx &= \int_a^b W(x)y(x)dx \\
a_0 \int_a^b xW(x)dx + a_1 \int_a^b x^2 W(x)\,dx + \cdots + a_n \int_a^b x^{n+1} W(x)\,dx &= \int_a^b x\,W(x)y(x)dx \\
&\vdots \\
a_0 \int_a^b x^n W(x)dx + a_1 \int_a^b x^{n+1} W(x)\,dx + \cdots + a_n \int_a^b x^{2n} W(x)\,dx &= \int_a^b x^n W(x)y(x)dx.
\end{aligned}\right\} \tag{4.39}$$

The system in (4.39) 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 4.8 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 (4.39), 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 (ii), we obtain

$$a_0 \frac{\pi}{2} + a_1 \frac{\pi^2}{8} + a_2 \frac{\pi^3}{24} = 1$$

$$a_0 \frac{\pi^2}{8} + a_1 \frac{\pi^3}{24} + a_2 \frac{\pi^4}{64} = 1$$

$$a_0 \frac{\pi^3}{24} + a_1 \frac{\pi^4}{64} + a_2 \frac{\pi^5}{160} = 2\left(\frac{\pi}{2} - 1\right),$$

Presented By: http://www.ebooksuit.com

whose solution is

$$\left.\begin{aligned} a_0 &= \frac{18}{\pi} + \frac{96}{\pi^2} - \frac{480}{\pi^3} \\ a_1 &= -\frac{144}{\pi^2} - \frac{1344}{\pi^3} + \frac{5760}{\pi^4} \\ a_2 &= \frac{240}{\pi^3} + \frac{2880}{\pi^4} - \frac{11520}{\pi^5}. \end{aligned}\right\} \qquad \text{(iii)}$$

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 above 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:

$$Y(x) = a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x), \qquad (4.40)$$

where $f_j(x)$ is a polynomial in x of degree j. Then, we write

$$S(a_0, a_1, \ldots, a_n) = \int_a^b W(x)\{y(x) - [a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x)]\}^2 dx. \qquad (4.41)$$

For S to be minimum, we must have

Presented By: http://www.ebooksuit.com

$$\left.\begin{aligned}
\frac{\partial S}{\partial a_0} &= 0 = -2\int_a^b W(x)\{y(x)-[a_0 f_0(x)+a_1 f_1(x)+\cdots+a_n f_n(x)]\} f_0(x)\,dx \\
\frac{\partial S}{\partial a_1} &= 0 = -2\int_a^b W(x)\{y(x)-[a_0 f_0(x)+a_1 f_1(x)+\cdots+a_n f_n(x)]\} f_1(x)\,dx \\
&\vdots \\
\frac{\partial S}{\partial a_n} &= 0 = -2\int_a^b W(x)\{y(x)-[a_0 f_0(x)+a_1 f_1(x)+\cdots+a_n f_n(x)]\} f_n(x)\,dx
\end{aligned}\right\} \quad (4.42)$$

The normal equations are now given by

$$\left.\begin{aligned}
&a_0\int_a^b W(x) f_0^2(x)\,dx + a_1\int_a^b W(x) f_0(x) f_1(x)\,dx + \cdots + a_n\int_a^b W(x) f_0(x) f_n(x)\,dx \\
&\quad = \int_a^b W(x)\, y(x) f_0(x)\,dx \\
&a_0\int_a^b W(x) f_1(x) f_0(x)\,dx + a_1\int_a^b W(x) f_1^2(x)\,dx + \cdots + a_n\int_a^b W(x) f_1(x) f_n(x)\,dx \\
&\quad = \int_a^b W(x)\, y(x) f_1(x)\,dx \\
&\quad \vdots \\
&a_0\int_a^b W(x) f_n(x) f_0(x)\,dx + a_1\int_a^b W(x) f_n(x) f_1(x)\,dx + \cdots + a_n\int_a^b W(x) f_n^2(x)\,dx \\
&\quad = \int_a^b W(x)\, y(x) f_n(x)\,dx.
\end{aligned}\right\} \quad (4.43)$$

The above system can be written more simply as

$$a_0\int_a^b W(x) f_0(x) f_j(x)\,dx + a_1\int_a^b W(x) f_1(x) f_j(x)\,dx + \cdots$$
$$+ a_n\int_a^b W(x) f_n(x) f_j(x)\,dx = \int_a^b W(x)\, y(x) f_j(x)\,dx, \qquad j = 0, 1, 2, \ldots, n. \quad (4.44)$$

Presented By: http://www.ebooksuit.com

In (4.43), we find products of the type $f_p(x)\,f_q(x)$ in the integrands, and if we assume that

$$\int_a^b W(x) f_p(x) f_q(x)\,dx = \begin{cases} 0, & p \neq q \\ \displaystyle\int_a^b W(x) f_p^2(x)\,dx, & p = q, \end{cases} \tag{4.45}$$

then the system (4.43) reduces to

$$a_0 \int_a^b W(x) f_0^2(x)\,dx = \int_a^b W(x)\,y(x)\,f_0(x)\,dx$$

$$\vdots$$

$$a_n \int_a^b W(x) f_n^2(x)\,dx = \int_a^b W(x)\,y(x)\,f_n(x)\,dx.$$

From the above, we obtain

$$a_j = \frac{\displaystyle\int_a^b W(x)\,y(x)\,f_j(x)\,dx}{\displaystyle\int_a^b W(x)\,f_j^2(x)\,dx}, \qquad j = 0, 1, 2, \ldots, n. \tag{4.46}$$

Substitution of $a_0, a_1, \ldots, a_n$ in (4.40) then yields the required least squares approximation, but the functions $f_0(x), f_1(x), \ldots, f_n(x)$ are still not known. The $f_j(x)$, which are polynomials in x satisfying the conditions (4.45), 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}

*For more details concerning orthogonal polynomials, see Abramovitz and Stegun [1965].

Presented By: http://www.ebooksuit.com

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

$$f_0(x) = 1 \tag{4.47}$$

we find that the linear polynomial $f_1(x)$, with leading term x, can be written as

$$f_1(x) = x + k_{1,0}\ f_0(x), \tag{4.48}$$

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 (4.45) and (4.48). From the above, we obtain

$$k_{1,0} = -\frac{\int_a^b x\, W(x) f_0(x)\, dx}{\int_a^b W(x) f_0^2(x)\, dx} \tag{4.49}$$

and Eq. (4.48) 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

$$f_2(x) = x^2 + k_{2,0} f_0(x) + k_{2,1} f_1(x), \tag{4.50}$$

where the constants $k_{2,0}$ and $k_{2,1}$ are to be determined by using the orthogonality conditions in (4.45). Since $f_2(x)$ is orthogonal to $f_0(x)$, we have

Presented By: http://www.ebooksuit.com

$$\int_a^b W(x) f_0(x)[x^2 + k_{2,0} f_0(x) + k_{2,1} f_1(x)]\,dx = 0.$$

Since $\int_a^b W(x) f_0(x) f_1(x)\,dx = 0$, the above equation gives

$$k_{2,0} = -\frac{\int_a^b x^2 W(x) f_0(x)\,dx}{\int_a^b W(x) f_0^2(x)\,dx} = -\frac{\int_a^b x^2 W(x)\,dx}{\int_a^b W(x)\,dx} \tag{4.51}$$

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

$$k_{2,1} = -\frac{\int_a^b x^2 W(x) f_1(x)\,dx}{\int_a^b W(x) f_1^2(x)\,dx}. \tag{4.52}$$

Since $k_{2,0}$ and $k_{2,1}$ are known, Eq. (4.50) determines $f_2(x)$. Proceeding in this way, the method can be generalized and we write

$$f_j(x) = x^j + k_{j,0} f_0(x) + k_{j,1} f_1(x) + \cdots + k_{j,j-1} f_{j-1}(x), \tag{4.53}$$

where the constants $k_{j,i}$ are so chosen that $f_j(x)$ is orthogonal to $f_0(x), f_1(x), \ldots, f_{j-1}(x)$. These conditions yield

$$k_{j,i} = -\frac{\int_a^b x^j W(x) f_i(x)\,dx}{\int_a^b W(x) f_i^2(x)\,dx}. \tag{4.54}$$

Since the a_i and $f_i(x)$ in (4.40) are known, the approximation $Y(x)$ can now be determined. The following example illustrates the method of procedure.

Example 4.9 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. (4.49) gives

Presented By: http://www.ebooksuit.com

$$k_{1,0} = -\frac{\int_{-1}^{1} x\,dx}{\int_{-1}^{1} dx} = 0.$$

We then obtain from Eq. (4.48), $f_1(x) = x$. Equations (4.51) and (4.52) 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. (4.50) yields $f_2(x) = x^2 - 1/3$.

In a similar manner, we obtain

$$k_{3,0} = -\frac{\int_{-1}^{1} x^3 dx}{\int_{-1}^{1} dx} = 0,$$

$$k_{3,1} = -\frac{\int_{-1}^{1} x^3 x\,dx}{\int_{-1}^{1} x^2\,dx} = -\frac{3}{5},$$

$$k_{3,2} = -\frac{\int_{-1}^{1} x^3(x^2 - 1/3)\,dx}{\int_{-1}^{1} (x^2 - 1/3)^2 dx} = 0.$$

Presented By: http://www.ebooksuit.com

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 (4.45). An important application of Legendre polynomials occurs in numerical quadrature (*see* Chapter 5).

4.5 CUBIC B-SPLINES

We have seen that a curve passing through a given set of data points must be dependent on some interpolation formula or an approximating function to establish its relationship with the given data. The interpolation formulae (including the cubic spline formula) discussed so far are 'global in nature', since they do not permit local changes in the data or curve.

The B-splines are 'non-global'. These are *basis* functions. This basis allows the degree of the resulting curve to be changed without any change in the data. The B-splines can be of any degree but, in computer graphics and other applications, B-splines of degree 2 or 3 are generally found to be sufficient. We therefore restrict our study to a discussion of cubic B-splines only. The cubic B-spline resembles the ordinary cubic spline, discussed in the previous chapter, in that a separate cubic is derived for each interval. Specifically, a cubic B-spline (or a *B-spline of order four*), denoted by $B_{4i}(x)$, is a cubic spline with knots $k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}$ and k_i, which is zero everywhere except in the range $k_{i-4} < x < k_i$. In such a case, $B_{4i}(x)$ is said to have a support $[k_{i-4}, k_i]$. It may be noted that a B-spline need not necessarily pass through any or all of the data points. Similarly, a B-spline of order n (degree $n-1$), denoted by $B_{ni}(x)$, is nonzero only in the range $k_{i-n} < x < k_i$. 'The theory for B-splines' was first suggested by Schoenberg [1946] and a recurrence formula for its numerical computation was independently discovered by Cox [1972] and de Boor [1972]. The B-splines may be defined in several ways. A useful representation is that based on divided differences and this will be given in the next section.

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 $a < k_1 < k_2 < \cdots < k_p < b$. Then the cubic spline $B_{4,5}(x)$ with knots k_1, k_2, k_3, k_4 and k_5 must satisfy the following *properties*:

(i) On each interval, the B-spline must be a polynomial of degree 3 or less,

(ii) The B-spline and its first-two derivatives must be continuous over the entire curve,

Presented By: http://www.ebooksuit.com

(iii) $B_{4,5}(x) > 0$ inside $[k_1, k_5]$, i.e., the B-spline is non-zero only over four successive intervals,

(iv) $B_{4,5}(x)$ is identically zero outside $[k_1, k_5]$.

For computational purposes, it would be convenient to use the *normalized* B-splines, $N_{4i}(x)$, defined by

$$N_{4i}(x) = (k_i - k_{i-4})\, B_{4i}(x). \tag{4.55}$$

The sum of all the normalized B-splines in the given range is equal to 1.

Figure 4.1 shows the graph of a cubic spline $B_{42}(x)$ with knots $-2, -1, 0, 1$ and 2.

Figure 4.1 A typical cubic B-spline.

In Fig. 4.1., $B_{4,2}(x)$ has the following properties:

$$\left.\begin{aligned} &\text{(i)}\ B_{4,2}(-2) = B_{4,2}(2) = 0 \text{ and } B_{4,2}^{(0)} = 1 \\ &\text{(ii)}\ B'_{4,2}(-2) = B'_{4,2}(2) = 0 \\ &\text{(iii)}\ B''_{4,2}(-2) = B''_{4,2}(2) = 0. \end{aligned}\right\} \tag{4.56}$$

Suppose now we have p knots denoted by $k_1, k_2, \ldots, k_p$. To define the full set of B-splines, it is necessary to introduce 'eight additional knots,' viz. k_{-3}, k_{-2}, k_{-1}, k_0, k_{p+1}, k_{p+2}, k_{p+3} and k_{p+4}. These are chosen such that

$$k_{-3} < k_{-2} < k_{-1} < k_0 = a \quad \text{and} \quad b = k_{p+1} < k_{p+2} < k_{p+3} < k_{p+4}. \tag{4.57}$$

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

$$s(x) = \sum_{i=1}^{p+4} \alpha_i N_{4,i}(x), \qquad a \le x \le b, \tag{4.58}$$

where $N_{4,i}(x)$ are the normalized B-splines of order 4 and α_i are constants to be determined.

Presented By: http://www.ebooksuit.com

4.5.1 Least-squares Solution

To determine the coefficients α_i in (4.58), we substitute $x = x_r$ and obtain

$$s(x_r) = y_r = \sum_{i=1}^{p+4} \alpha_i N_{4,i}(x_r), \qquad r = 1, 2, \ldots, m. \tag{4.59}$$

where p is chosen such that $m >> p+4$. In matrix form, Eqs. (4.59) can be written as

$$N\boldsymbol{\alpha} = \mathbf{y}, \tag{4.60}$$

where N is an $m \times (p+4)$ band matrix and $\boldsymbol{\alpha}, \mathbf{y}$ are column vectors. The solution of Eqs. (4.59) may be obtained by solving the normal equations:

$$N^{\mathrm{T}} N\boldsymbol{\alpha} = N^{\mathrm{T}}\mathbf{y}. \qquad \text{i.e.} \quad N\boldsymbol{\alpha} = \mathbf{y} \tag{4.61}$$

4.5.2 Representations of B-splines

To define the cubic B-spline at $x = k_i$, we first consider the five knots $k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}$, and k_i, where $a < k_{i-4}$ and $k_i < b$. We also define the function

$$P_+^3 = \begin{cases} P^3 & \text{when } P \geq 0 \\ 0 & \text{when } P \leq 0. \end{cases} \tag{4.62}$$

Then a unique representation of the cubic B-spline with knots $k_{i-4}, \ldots, k_i$ is given by (Greville [1968])

$$s(x) = B_{4,i}(x) = \sum_{j=0}^{3} \alpha_j x^j + \sum_{m=i-4}^{i} \beta_m (x - k_m)_+^3. \tag{4.63}$$

Unfortunately, the representation of the cubic spline, as given by (4.63), is computationally inefficient because of loss of accuracy through cancellation. Another representation of the B-spline, a traditional one, is through divided differences. The divided difference of fourth order of the function $(k_p - x)_+^3$ with respect to the knots $k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}$ and k_i as arguments is denoted by $[k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}, k_i]$. We then have (*see* Section 3.10)

$$\begin{aligned} B_{4,i}(x) &= [k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}, k_i]. \\ &= \frac{(k_{i-4} - x)_+^3}{(k_{i-4} - k_{i-3})(k_{i-4} - k_{i-2})(k_{i-4} - k_{i-1})(k_{i-4} - k_i)} \\ &\quad + \frac{(k_{i-3} - x)_+^3}{(k_{i-3} - k_{i-4})(k_{i-3} - k_{i-2})(k_{i-3} - k_{i-1})(k_{i-3} - k_i)} \\ &\quad + \cdots + \frac{(k_i - x)_+^3}{(k_i - k_{i-4})(k_i - k_{i-3})(k_i - k_{i-2})(k_i - k_{i-1})}. \end{aligned} \tag{4.64}$$

Presented By: http://www.ebooksuit.com

Setting

$$\pi_{4,i}(x) = (x - k_{i-4})(x - k_{i-3})(x - k_{i-2})(x - k_{i-1})(x - k_i) \tag{4.65}$$

Eq. (4.64) can be expressed in the more compact form

$$B_{4,i}(x) = \sum_{m=i-4}^{i} \frac{(k_m - x)_+^3}{\pi'_{4,i}(k_m)}. \tag{4.66}$$

More generally, a B-spline of order n (degree $n-1$) is defined by

$$B_{n,i}(x) = [k_{i-n}, k_{i-n+1}, \ldots, k_i] = \sum_{m=i-n}^{i} \frac{(k_m - x)_+^{n-1}}{\pi'_{n,i}(k_m)}, \tag{4.67}$$

where

$$\pi_{n,i}(x) = (x - k_{i-n})(x - k_{i-n+1}) \ldots (x - k_i). \tag{4.68}$$

Recalling that

$$[k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}, k_i] = \frac{[k_{i-3}, k_{i-2}, k_{i-1}, k_i] - [k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}]}{k_i - k_{i-4}}, \tag{4.69}$$

we obtain the relation

$$B_{4,i}(x) = \frac{B_{3,i}(x) - B_{3,i-1}(x)}{k_i - k_{i-4}}, \tag{4.70}$$

which is a recurrence relation. Similarly, for B-splines of order n, we obtain the relation

$$B_{n,i}(x) = \frac{B_{n-1,i}(x) - B_{n-1,i-1}(x)}{k_i - k_{i-n}} \tag{4.71}$$

for a recursive computation of the B-splines $B_{n,i}(x)$. Unfortunately, *computational algorithms based on formula* (4.70) or (4.71) have been found to be numerically unstable even for simple examples. However, algorithms based on a recurrence relation discovered independently by de Boor [1972] and Cox [1972] have been found to be both stable and efficient. This recurrence relation will be stated and illustrated in the next section. We conclude the present section with an example on the computation of cubic B-splines represented by (4.63).

Example 4.10 Using the relation (4.63), determine the cubic B-spline $s(x)$ with support $[0, 4]$ on the knots 0,1,2,3,4. Show further that such a representation will be unique if $s(1)$ is specified.

Since $s(x)$ is a cubic B-spline over $[0, 4]$, we have

Presented By: http://www.ebooksuit.com

$$\left.\begin{aligned} s(0) = s(4) = 0 \\ s'(0) = s'(4) = 0 \\ s''(0) = s''(4) = 0. \end{aligned}\right\} \tag{i}$$

Because of symmetry, we also have

$$s'(2) = 0 \tag{ii}$$

$$s(1) = s(3). \tag{iii}$$

Now, on the interval [0, 1], let $s(x)$ be given by

$$s(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3. \tag{iv}$$

Since $s(0) = 0$, we immediately have $c_0 = 0$.

Also, the conditions $s'(0) = s''(0) = 0$ given $c_1 = 0$ and $c_2 = 0$. Hence (iv) becomes

$$s(x) = c_3 x^3, \tag{v}$$

which is the cubic spline on [0, 1]. Obviously, $c_3 = s(1)$. Again, on the interval [0, 2], let $s(x)$ be represented by

$$s(x) = c_3 x^3 + \beta_1 (x-1)_+^3, \tag{vi}$$

where β_1 is to be determined. Now,

$$s'(x) = 3c_3 x^2 + 3\beta_1 (x-1)_+^2.$$

But $s'(2) = 0$. Hence we obtain

$$12c_3 + 3\beta_1 = 0,$$

which gives

$$\beta_1 = -4c_3.$$

Substituting for β_1 in (vi), we obtain

$$s(x) = c_3 x^3 - 4c_3 (x-1)_+^3, \tag{vii}$$

which is the cubic B-spline valid in the interval [0, 2]. Further, let $s(x)$ be represented on [0, 3] as

$$s(x) = c_3 x^3 - 4c_3 (x-1)_+^3 + \beta_2 (x-2)_+^3. \tag{viii}$$

But $s(3) = s(1) = c_3$. Substitution in (viii) gives

$$c_3 = 27c_3 - 32c_3 + \beta_2,$$

from which, we obtain

$$\beta_2 = 6c_3.$$

Hence, (viii) becomes:

$$s(x) = c_3 [x^3 - 4(x-1)_+^3 + 6(x-2)_+^3]. \tag{ix}$$

Presented By: http://www.ebooksuit.com

Finally, let $s(x)$ be represented on [0, 4] as:

$$s(x) = c_3[x^3 - 4(x-1)_+^3 + 6(x-2)_+^3] + \beta_3(x-3)_+^3. \tag{x}$$

Since $s(4) = 0$, eq. (x) gives

$$0 = c_3(64 - 108 + 48) + \beta_3,$$

from which, we obtain

$$\beta_3 = -4c_3.$$

Substitution in (x) gives the required B-spline as

$$s(x) = c_3x^3 - 4c_3(x-1)_+^3 + 6c_3(x-2)_+^3 - 4c_3(x-3)_+^3,$$

which will be *unique* if $c_3 = s(1)$ is specified.

4.5.3 Computation of B-splines

B-splines are most conveniently computed by the Cox–de Boor recurrence formula which is both stable and efficient. For B-splines of order n (degree $n-1$), the formula is given by

$$B_{n,i}(x) = \frac{(x - k_{i-n})B_{n-1,i-1}(x) + (k_i - x)B_{n-1,i}(x)}{k_i - k_{i-n}} \tag{4.72}$$

and holds for all values of x. For proof of this formula, see Cox [1972]. It is seen from (4.72) that the computation of $B_{n,i}(x)$ for any value of x depends on the values of $B_{n-1,i-1}(x)$ and $B_{n-1,i}(x)$. Thus, to compute the cubic B-spline based on the knots, k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}, k_i, we need to compute, from left to right, the elements in the array:

$B_{1,i-3}$			
	$B_{2,i-2}$		
$B_{1,i-2}$		$B_{3,i-1}$	
	$B_{2,i-1}$		$B_{4,i}$
$B_{1,i-1}$		$B_{3,i}$	
	$B_{2,i}$		
$B_{1,i}$			

Further, advantage may be taken of the fact that some of the elements in the above array may vanish because of the properties of the B-spline. For example, if $k_{i-4} \le x < k_{i-3}$, then using the relation

$$\left.\begin{aligned} B_{1,j} &= \frac{1}{k_j - k_{j-1}}, && \text{if } k_{j-1} \le x < k_j \\ &= 0, && \text{otherwise.} \end{aligned}\right\} \tag{4.73}$$

Presented By: http://www.ebooksuit.com

the above array takes the form:

$B_{1,i-3}$			
	$B_{2,i-2}$		
0		$B_{3,i-1}$	
	0		$B_{4,i}$
0		0	
	0		
0			

The numerical computation of the B-splines will now, become more simpler. This is illustrated in the following example considered by Cox [1972].

Example 4.11 We consider knots 0, 1, 2, 3, 4, 5, 6 and compute B-splines of order 6 (degree 5) at $x=1$ and $x=2$, i.e. at the interior knots 1 and 2.

Corresponding to $x=1$, we have $k_2 \le x < k_3$.
Then

$$B_{1,3} = \frac{1}{k_3 - k_2} = 1$$

and we need to compute $B_{6,7}$. Consequently, we need to compute only the elements in the following array:

0					
	$B_{2,3}$				
$B_{1,3}$		$B_{3,4}$			
	0		$B_{4,5}$		
0		0		$B_{5,6}$	
	0		0		$B_{6,7}$
0		0		0	
	0		0		
0		0			
	0				
0					

Using the Cox–de Boor formula, the values at $x=1$ of the above B-splines are given by:

$$B_{1,3}=1, \quad B_{2,3}=1/2, \quad B_{3,4}=1/6, \quad B_{4,5}=1/24, \quad B_{4,6}=0,$$

$$B_{5,6}=1/120, \quad B_{5,7}=0, \quad B_{6,7}=1/720=0.0013888\ldots,$$

which is the same as that value obtained by Cox.

Again, corresponding to $x=2$, we have $k_3 \le x < k_4$. Hence,

$$B_{1,4} = \frac{1}{k_4 - k_3} = 1,$$

and we require the value of $B_{6,7}$ at $x=2$.

Presented By: http://www.ebooksuit.com

The elements to be computed are given in the following array:

0					
	0				
0		$B_{3,4}$			
	$B_{2,4}$		$B_{4,5}$		
$B_{1,4}$		$B_{3,5}$		$B_{5,6}$	
	$B_{2,5}$		$B_{4,6}$		$B_{6,7}$
0		$B_{3,6}$		$B_{5,7}$	
	0		$B_{4,7}$		
0		0			
	0				
0					

The values obtained at $x = 2$ using the Cox–de Boor formula are given by:

$$B_{2,4} = 1/2, \qquad B_{2,5} = 0,$$

$$B_{3,4} = 1/6, \qquad B_{3,5} = 1/6, \qquad B_{3,6} = 0,$$

$$B_{4,5} = 1/6, \qquad B_{4,6} = 1/24, \qquad B_{4,7} = 0,$$

$$B_{5,6} = 11/120, \qquad B_{5,7} = 1/120,$$

$$B_{6,7} = 26/720 = 0.0361111\ldots,$$

which is the same as that obtained by Cox.

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

$$f(t+T) = f(t), \tag{4.74}$$

where T is the smallest value satisfying Eq. (4.74). Then the Fourier series for $f(t)$ is written as

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n \cos\frac{2\pi nt}{T} + b_n \sin\frac{2\pi nt}{T}\right), \tag{4.75}$$

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 (4.75) from 0 to T, we obtain

$$\int_0^T f(t)\,dt = \frac{a_0}{2}\int_0^T dt + \int_0^T\left(a_n \cos\frac{2\pi nt}{T} + b_n \sin\frac{2\pi nt}{T}\right)dt = \frac{a_0}{2}T,$$

Presented By: http://www.ebooksuit.com

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

$$a_0 = \frac{2}{T}\int_0^T f(t)\,dt. \tag{4.76}$$

Again, multiplying both the sides of (4.75) by $\cos(2\pi nt/T)$ and then integrating from 0 to T, we get

$$a_n = \frac{2}{T}\int_0^T f(t)\cos\left(\frac{2\pi nt}{T}\right)dt, \tag{4.77}$$

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 side of (4.75) by $\sin(2\pi nt/T)$ and then integrating from 0 to T, we obtain

$$b_n = \frac{2}{T}\int_0^T f(t)\sin\left(\frac{2\pi nt}{T}\right)dt. \tag{4.78}$$

Thus the coefficients a_0, a_n and b_n in the representation (4.75) are evaluated. If $T = 2\pi$, i.e. if $f(t)$ is of period 2π, the formula (4.76)–(4.78) become:

$$\left.\begin{aligned} a_0 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)dt, \\ a_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos nt\,dt, \\ b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin nt\,dt. \end{aligned}\right\} \tag{4.79}$$

The Fourier series becomes further simplified if $f(t)$ is an even or odd function. If $f(t)$ is even, then we have

$$\left.\begin{aligned} f(t) &= \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt, \\ \text{where} \quad & \\ a_n &= \frac{2}{\pi}\int_0^{\pi} f(t)\cos nt\,dt, \end{aligned}\right\} \tag{4.80}$$

since $b_n = 0$.

Presented By: http://www.ebooksuit.com

Similarly, if $f(t)$ is an odd function, then we have

$$\left.\begin{aligned} f(t) &= \sum_{n=1}^{\infty} b_n \sin nt, \\ \text{where} \quad & \\ b_n &= \frac{2}{\pi}\int_0^{\pi} f(t)\sin nt\, dt. \end{aligned}\right\} \tag{4.81}$$

since $a_0 = a_n = 0$.

The formulae (4.75)–(4.78) can be expressed in a different way. For this, the well-known relations are used:

$$\cos nt = \frac{e^{int} + e^{-int}}{2} \quad \text{and} \quad \sin nt = \frac{e^{int} - e^{-int}}{2i}. \tag{4.82}$$

Using (4.82), Eqs. (4.75)–(4.78) can be expressed as

$$f(t) = \sum_{p=-\infty}^{\infty} A_p e^{2\pi ipt/T}, \tag{4.83}$$

where

$$A_p = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\, e^{-2\pi ipt/T}\, dt, \qquad p = 0, 1, 2, \ldots \tag{4.84}$$

These formulae directly lead us to the discussion of Fourier transforms but, before this, we consider an illustrative example on Fourier series.

Example 4.12 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

Fig. 4.2

Presented By: http://www.ebooksuit.com

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 The 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 story 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. (4.83) and (4.84), by making T approach infinity so that the function becomes aperiodic. When $T \to \infty$, Eq. (4.84) can be written in the form

$$F(i\omega_0) = \int_{-\infty}^{\infty} f(t)\, e^{-i\omega_0 t}\, dt, \tag{4.85}$$

and is called the *Fourier transform* of $f(t)$. Similarly, Eq. (4.83) is written as

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(i\omega_0) e^{i\omega_0 t}\, d\omega_0, \tag{4.86}$$

and is called the *inverse Fourier transform* of $f(t)$. Equations (4.85) and (4.86) enable us to transform from time domain to frequency domain and from

Presented By: http://www.ebooksuit.com

frequency to time domain, respectively. Physically, $F(i\omega_0)$ represents the frequency content of the signal. In Eq. (4.85), 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, i.e. $f(t)$ is specified only at the points $t_i, i = 0, 1, 2, \cdots, N-1$, and $\Delta t = T/N$. Thus f_k denotes the value of $f(t)$ at t_k. Then, corresponding to Eq. (4.85) and (4.86), we have:

$$F_p = \sum_{k=0}^{N-1} f_k \, e^{-2\pi i k p/N}, \qquad p = 0, 1, 2, \ldots, N-1 \tag{4.87}$$

and

$$f_k = \frac{1}{N} \sum_{p=0}^{N-1} F_p \, e^{2\pi i \omega_0 k/N}, \qquad k = 0, 1, 2, \ldots, N-1 \tag{4.88}$$

Denoting

$$W_N = e^{-2\pi i/N}, \tag{4.89}$$

Eqs. (4.87) and (4.88) become

$$F_p = \sum_{k=0}^{N-1} f_k W_N^{kp}, \qquad p = 0, 1, 2, \cdots, N-1 \tag{4.90}$$

and

$$f_k = \frac{1}{N} \sum_{p=0}^{N-1} F_p W_N^{-kp}, \qquad k = 0, 1, 2, \cdots, N-1 \tag{4.91}$$

The above equations are, respectively, called the *discrete Fourier transform* (DFT) and the inverse DFT. They are the discrete analogues of Eqs. (4.85) and (4.86), respectively. The coefficients $|F_p|$ form a periodic sequence when extended outside of the range $p = 0, 1, 2, \ldots, N-1$, and we have

$$F_{p+N} = F_p \tag{4.92}$$

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 relations

$$P = \frac{1}{N} \sum_{k=0}^{N-1} |F_k|^2 = \sum_{k=0}^{N-1} |F_k|^2. \tag{4.93}$$

Therefore, the sequence

$$p_k = |F_k|^2, \qquad k = 0, 1, 2, \cdots, N-1 \tag{4.94}$$

Presented By: http://www.ebooksuit.com

is the distribution of power as a function of frequency and is called the *power density spectrum* of the periodic signal. The power spectrum is the plot of p_k as a function of frequency $\omega_0 k$. Since $F_k (k = 0, 1, 2, \ldots, N-1)$ is also a periodic sequence with period N, it follows that the spectrum of $F_k (k = 0, 1, 2, \ldots, N-1)$ is also a periodic sequence with period N. Hence, any N consecutive samples of the signal or its spectrum provide a complete description of the signal in the time or frequency domains.

Example 4.13 Find the DFT of the sequence $\{1, 1, 1, \ldots, 1\}$ for $k = 0, 1, 2, \ldots, N-1$.

We have

$$F_p = \sum_{k=0}^{N-1} f_k W_N^{pk}, \tag{i}$$

where

$$W_N = e^{-2\pi i/N}, \qquad i = \sqrt{-1} \tag{ii}$$

since $f_k = 1$ for all $k = 0, 1, \ldots, N-1$, it follows that

$$F_p = \sum_{k=0}^{N-1} W_N^{pk} = \sum_{k=0}^{N-1} (W_N^p)^k,$$

which is a geometric series of N terms with a common ratio of W_N^p. We therefore have

$$F_p = \frac{1-(W_N^p)^N}{1-W_N^p} = \frac{1-W_N^{pN}}{1-W_N^p}, \qquad p = 0, 1, 2, \ldots, N-1. \tag{iii}$$

for $p = 0$, it is seen that the ratio on the right side of (iii) is of the form 0/0. Hence we obtain its limiting value as N by using L'Hospital's rule. Similarly, for $p = 1, 2, \ldots, N-1$, the limiting value of F_p is calculated and is found to be zero. We thus have

$$F_p = \begin{cases} N, & \text{when } p = 0 \\ 0, & \text{when } p = 1, 2, \ldots, N-1. \end{cases}$$

4.6.2 The Fast Fourier Transform

The computation of DFT using Eq. (4.90) is inefficient because it does not make use of the symmetric and periodic properties of the factor W_N, viz.,

$$W_N^{k(N-p)} = (W_N^{kp})^* \quad \text{and} \quad W_N^{k(N+p)} = W_N^{kp} = W_N^{(k+N)p} \tag{4.94}$$

The direct use of Eq. (4.90) requires N^2 complex operations and also memory to store the values of $f(t)$ and W_N^{kp}. As N increases, the computation of DFT

Presented By: http://www.ebooksuit.com

demands very high memory requirements and becomes a complex and time –consuming process.

A class of algorithms, called the *fast Fourier transforms* (FFT), computes the DFT in an economic fashion using properties (4.94) and thereby reducing the number of operations to $N \log_2 N$. This means that, in terms of computing time and memory requirements, the FFT is far superior to the DFT. For example, for $N = 50$, the FFT requires about 250 complex operations compared to about 2500 complex operations required by the direct use of Eq. (4.90). This contrast, therefore, points to the importance of FFT algorithms. There exist several FFT algorithms and the basic idea behind all these is that a DFT of length N is *decimated* (or decomposed) into successive smaller DFTs. One class of FFT algorithms, called *radix*-2 algorithms, is based on the assumption 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, viz., (a) decimation-in-time (DIT), and (b) decimation-in-frequency (DIF). The Cooley–Tukey algorithm belongs to the type (a), whereas the Sandey–Tukey algorithm to the type (b). Both the algorithms require $N \log_2 N$ operations but differ in organization. The Cooley–Tukey algorithm is discussed in the next section.

4.6.3 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 DFTs, then decompose each of the $N/2$-point DFTs into two $N/4$-point DFTs and continuing this process until we obtain $N/2$ two-point DFTs. The number of steps required to achieve this is clearly m. For easy understanding, we present this algorithm for $N = 8$ and it will be seen that it is easily generalized. The first step (or stage) of the algorithm is described below:

Let $f_0, f_1, f_2, \ldots, f_7$ be a sequence of values of $f(t)$. The DFT for f_k is given by

$$F_p = \sum_{k=0}^{7} f_k W_8^{pk}, \qquad p = 0, 1, 2, \ldots, 7 \tag{4.90}$$

where

$$W_8 = e^{-2\pi i/8}. \tag{4.89}$$

We split the summation on the right side of (4.90) into two equal parts of length 4, one containing the even-indexed values of $f(t)$ and the other of the odd-indexed values. We therefore write

$$F_p = \sum_{k(\text{even})} f_k W_8^{kp} + \sum_{k(\text{odd})} f_k W_8^{kp}. \tag{4.95}$$

Presented By: http://www.ebooksuit.com

Putting $k = 2r$ in the first sum and $k = 2r+1$ in the second sum of (4.95), we obtain

$$F_p = \sum_{r=0}^{3} f_{2r} W_8^{2pr} + \sum_{r=0}^{3} f_{2r+1} W_8^{(2r+1)p}. \tag{4.96}$$

But

$$\left.\begin{aligned} W_8^{2pr} &= e^{-2\pi i(2pr)/8} = e^{-2\pi i pr/4} = W_4^{pr} && (4.97\text{a}) \\ \text{and} \quad & \\ W_8^{(2r+1)p} &= W_8^{2rp} W_8^{p} = W_8^{p} W_4^{pr}. && (4.97\text{b}) \end{aligned}\right\}$$

Using (4.97) in (4.96), we get

$$F_p = \sum_{r=0}^{3} f_{2r} W_4^{rp} + W_8^{p} \sum_{r=0}^{3} f_{2r+1} W_4^{rp}. \tag{4.98}$$

It is easily seen that the two sums on the right side of (4.98) represent 4-point DFTs. Setting

$$\left.\begin{aligned} F_p^e &= \sum_{r=0}^{3} f_{2r} W_4^{pr} && (4.99\text{a}) \\ \text{and} \quad & \\ F_p^o &= \sum_{r=0}^{3} f_{2r+1} W_4^{pr}, && (4.99\text{b}) \end{aligned}\right\}$$

Eq. (4.98) becomes:

$$F_p = F_p^e + W_8^p F_p^o, \qquad p = 0, 1, 2, 3, \tag{4.100}$$

where F_p^e and F_p^o are the 4-point DFTs of the even and odd-indexed sequences defined by (4.99). This completes the first stage of decomposing the 8-point DFT into two 4-point DFTs. Further, to compute (4.100) for $p = 4, 5, 6, 7$ we use the formula:

$$F_p = F_{p-4}^e + W_8^p F_{p-4}^o, \qquad p = 4, 5, 6, 7 \tag{4.101}$$

The computations involving equations (4.100) and (4.101) for the first stage of the 8-point DIT–FFT are shown in the flow-graph in Fig. 4.3.

Presented By: http://www.ebooksuit.com

Figure 4.3 First stage of the 8-point DIT–FFT.

In the second stage, each of the 4-point transforms in (4.100) is decomposed into two 2-point transforms. We then write

$$\begin{aligned} F_p^e &= \sum_{r=0}^{3} f_{2r} W_4^{pr} \\ &= \sum_{s=0}^{1} f_{4s} W_2^{sp} + W_4^p \sum_{s=0}^{1} f_{4s+2} W_2^{sp} \\ &= F_p^{ee} + W_4^p F_p^{eo}, \end{aligned} \tag{4.102}$$

where

$$F_p^{ee} = \sum_{s=0}^{1} f_{4s} W_2^{sp} \quad \text{and} \quad F_p^{eo} = \sum_{s=0}^{1} f_{4s+2} W_2^{sp}. \tag{4.103}$$

Similarly, we obtain

$$F_p^o = F_p^{oe} + W_4^p F_p^{oo}, \tag{4.104}$$

where

$$F_p^{oe} = \sum_{l=0}^{1} f_{4l+1} W_2^{lp} \quad \text{and} \quad F_p^{oo} = \sum_{l=0}^{1} f_{4l+3} W_2^{lp}. \tag{4.105}$$

Presented By: http://www.ebooksuit.com

This completes the second stage of decomposition 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 4.4.

Figure 4.4 Second stage of the decomposition.

From Eq. (4.102), we have

$$F_p^{ee} = \sum_{s=0}^{1} f_{4s} W_2^{sp} = f_0 W_2^0 + f_4 W_2^p \tag{4.106a}$$

and

$$F_p^{eo} = \sum_{s=0}^{1} f_{4s+2} W_2^{sp} = f_2 W_2^0 + f_6 W_2^p. \tag{4.106b}$$

Equations (4.106a) and (4.106b) show that at the third stage (which is the final stage, since $N = 8$), we obtain

$$F_p^{eee} = f_0, \qquad F_p^{eeo} = f_4, \qquad F_p^{eoe} = f_2, \qquad F_p^{eoo} = f_6. \tag{4.107}$$

It follows that, for the 8-point computation, we start with the input sequence f_0, f_4, f_2, f_6, f_1, f_5, f_3, and f_7, and then compute the various Fourier coefficients. These computations can conveniently be depicted in a flow graph (Fig. 4.5).

Presented By: http://www.ebooksuit.com

Figure 4.5 Flowgraph of an 8-point DIT–FFT.

A close inspection of Fig. 4.5 enables us to make the following observations:

(i) The input data is shuffled and are in the order $f_0, f_4, f_2, f_6, f_1, f_5, f_3$, and f_7. They are in the *bit-reversed* order, as shown in Table 4.2.

Table 4.2 Input Data in the Reversed Bits

Input position	Binary equivalent	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 data for the Fourier coefficients F_k is in the natural order.

(iii) The computations are carried out in terms of a fundamental molecule called *butterfly*. A typical butterfly is shown in Fig. 4.6, where i and j represent the position numbers in the stage and m represents the stage of the computation.

Figure 4.6 A typical butterfly.

Presented By: http://www.ebooksuit.com

The outputs g_i^{m+1} and g_j^{m+1} are given by

$$g_i^{m+1} = g_i^m + W_N^r g_j^m, \qquad g_j^{m+1} = g_i^m - W_N^r g_j^m \tag{4.108}$$

where r is a variable depending on the position of the butterfly.

The method of computation is illustrated in the following numerical example.

Example 4.14 Using the Cooley–Tukey algorithm, find the DFT of the sequence

$$f_k = \{1, 2, 3, 4, 4, 3, 2, 1\}.$$

We have

$$W_8^0 = 1, \quad W_8^1 = e^{-2\pi i/8} = (1-i)/\sqrt{2}, \quad W_8^2 = (e^{-2\pi i/8})^2 = -i,$$

$$W_8^3 = -(1+i)/\sqrt{2}, \quad W_8^4 = -1, \quad W_8^5 = -(1-i)/\sqrt{2},$$

$$W_8^6 = i, \quad W_8^7 = (1+i)/\sqrt{2}$$

The DIT-FFT flowgraph for DFT computation is given below in Fig. 4.7:

Figure 4.7 Flowgraph for Example 4.14.

Presented By: http://www.ebooksuit.com

4.6.4 Sande–Tukey Algorithm

This is an alternative approach and is a member of the class called *decimation-in-frequency* techniques. Effectively, it is the reverse of the Cooley–Tukey algorithm described in the previous section. However, in this case, the final results are scrambled.

We start again with Eq. (4.90), viz.,

$$F_p = \sum_{k=0}^{7} f_k W_8^{pk}, \qquad p = 0, 1, 2, \ldots, 7$$

and divide the sum in terms of the first and last four points as:

$$F_p = \sum_{k=0}^{3} f_k W_8^{pk} + \sum_{k=4}^{7} f_k W_8^{pk}. \tag{4.109}$$

In the second sum on the right side of Eq. (4.109), we make a change of variable and write the equation as

$$F_p = \sum_{k=0}^{3} f_k W_8^{pk} + \sum_{k=0}^{3} f_{k+4} W_8^{p(k+4)}. \tag{4.110}$$

But

$$W_8^{p(k+4)} = W_8^{pk} W_8^{4p} = (-1)^p W_8^{pk},$$

which is positive if p is even and negative, otherwise.

Accordingly, (4.110) may be written as

$$\begin{aligned} F_{2r} &= \sum_{k=0}^{3} (f_k + f_{k+4}) W_8^{2rk} \\ &= \sum_{k=0}^{3} (f_k + f_{k+4}) W_4^{rk}, \qquad r = 0, 1, 2, 3 \end{aligned} \tag{4.111}$$

for the even components, and

$$\begin{aligned} F_{2r+1} &= \sum_{k=0}^{3} (f_k - f_{k+4}) W_8^{(2r+1)k} \\ &= \sum_{k=0}^{3} (f_k - f_{k+4}) W_8^{2rk} W_8^{k}, \\ &= \sum_{k=0}^{3} (f_k - f_{k+4}) W_8^{k} W_4^{rk}, \qquad r = 0, 1, 2, 3 \end{aligned} \tag{4.112}$$

Presented By: http://www.ebooksuit.com

for odd components. From Eqs. (4.111) and (4.112), it can be seen that F_{2r} and F_{2r+1} are the transforms of the functions $(f_k + f_{k+4})$ and $(f_k - f_{k+4})W_8^k$, respectively. We therefore set

$$f_k + f_{k+4} = s_k \quad \text{and} \quad (f_k - f_{k+4})W_8^k = t_k, \qquad k = 0, 1, 2, 3. \tag{4.113}$$

In view of (4.113), Eqs. (4.111) and (4.112) assume the form

$$F_{2r} = S_r \quad \text{and} \quad F_{2r+1} = T_r, \qquad r = 0, 1, 2, 3. \tag{4.114}$$

It is clear that this approach can be repeated at the second stage to break each of the 4-point transforms into two 2-point transforms. In the general case, the final result is obtained after $\log_2 N$ stages. Figure 4.8 shows the flowgraph for the 8-point decimation in frequency–FFT.

Figure 4.8 Flowgraph for an 8-point DIF–FFT.

Figure 4.8 shows that:

(i) The input is in the natural order whereas the output for the frequency components is in the bit-reversed order.

(ii) the computations can be carried out by using the butterfly structure.

4.6.5 Computation of the Inverse DFT

The inverse DFT is given by Eq. (4.91), viz.

$$f_k = \frac{1}{N}\sum_{p=0}^{N-1} F_p W_N^{-kp}, \qquad k = 0, 1, 2, \ldots, N-1.$$

Presented By: http://www.ebooksuit.com

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 flowgraph for the calculation of DFT can also be adopted for the computation of inverse DFT after making the above changes.

The Fast Fourier transform is extensively employed in many areas of electrical engineering such as signal processing. Hence, excellent software packages have been developed by MATLAB, IMSL, etc. Numerical Recipes discusses a number of FORTRAN programs. Any of these programs can conveniently be used as subprograms in the solution of research problems.

4.7 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.*

4.7.1 Chebyshev Polynomials

The Chebyshev polynomial of degree n over the interval $[-1, 1]$ is defined by the relation

$$T_n(x) = \cos(n \cos^{-1} x), \tag{4.115}$$

from which follows immediately the relation

$$T_n(x) = T_{-n}(x). \tag{4.116}$$

Let $\cos^{-1} x = \theta$ so that $x = \cos\theta$ and (4.115) gives

$$T_n(x) = \cos n\theta.$$

*Refer to Fox and Parker [1968] for further details and other applications of Chebyshev polynomials.

Presented By: http://www.ebooksuit.com

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

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{4.117}$$

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:

$$\left.\begin{aligned}
T_0(x) &= 1 \\
T_1(x) &= x \\
T_2(x) &= 2x^2 - 1 \\
T_3(x) &= 4x^3 - 3x \\
T_4(x) &= 8x^4 - 8x^2 + 1 \\
T_5(x) &= 16x^5 - 20x^3 + 5x \\
T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1.
\end{aligned}\right\} \tag{4.118}$$

The graph of the first four Chebyshev polynomials are shown in Fig. 4.9

Figure 4.9 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

$$\frac{dy}{dx} = \frac{n\sin n\theta}{\sin\theta}$$

Presented By: http://www.ebooksuit.com

and

$$\frac{d^2y}{dx^2} = \frac{-n^2\cos n\theta + n\sin n\theta\cot\theta}{\sin^2\theta} = \frac{-n^2y + x(dy/dx)}{1-x^2}$$

so that

$$(1-x^2)\frac{d^2y}{dx^2} - x\frac{dy}{dx} + n^2y = 0, \tag{4.119}$$

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

$$\left.\begin{aligned}
1 &= T_0(x)\\
x &= T_1(x)\\
x^2 &= \frac{1}{2}[T_0(x)+T_2(x)]\\
x^3 &= \frac{1}{4}[3T_1(x)+T_3(x)]\\
x^4 &= \frac{1}{8}[3T_0(x)+4T_2(x)+T_4(x)]\\
x^5 &= \frac{1}{16}[10T_1(x)+5T_3(x)+T_5(x)]\\
x^6 &= \frac{1}{32}[10T_0(x)+15T_2(x)+6T_4(x)+T_6(x)].
\end{aligned}\right\} \tag{4.120}$$

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

$$\int_{-1}^{1} \frac{T_m(x)\,T_n(x)\,dx}{\sqrt{1-x^2}} = \begin{cases} 0, & m \neq n \\ \pi/2, & m = n \neq 0 \\ \pi, & m = n = 0 \end{cases} \tag{4.121}$$

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 (4.121).

Presented By: http://www.ebooksuit.com

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

$$P_n(x) = 2^{1-n} T_n(x), \qquad (n \geq 1). \tag{4.122}$$

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 rang* $(-1, 1)$. Since $|T_n(x)| \leq 1$, 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 (4.122) 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 4.15 Find the best lower-order approximation to the cubic $2x^3 + 3x^2$. Using the relations given in (4.120), 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 $\pm 1/2$ in the range $(-1, 1)$.

A similar application of Chebyshev series in the *economization* of power series which is discussed next.

4.7.2 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

$$f(x) = A_0 + A_1 x + A_2 x^2 + \cdots + A_n x^n, \qquad (-1 \leq x \leq 1). \tag{4.123}$$

Using the relations given in (4.120), we convert the above series into an expansion in Chebyshev polynomials. We obtain

$$f(x) = B_0 + B_1 T_1(x) + B_2 T_2(x) + \cdots + B_n T_n(x). \tag{4.124}$$

For a large number of functions, an expansion as in (4.124) above, converges more rapidly than the power series given by (4.123). This is known as *economization of the power series* and is illustrated in Example 4.16.

Presented By: http://www.ebooksuit.com

Example 4.16 Economize the power series

$$\sin x \approx x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040}.$$

Since $1/5040 = 0.000198\ldots$, 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 (i) into Chebyshev polynomials by using the relations given in (4.120). 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 $1/1920 = 0.00052\ldots$, 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 (4.118), the economized series is therefore given by

$$\sin x \approx \frac{169}{192}x - \frac{5}{128}(4x^3 - 3x) = \frac{383}{384}x - \frac{5}{32}x^3.$$

EXERCISES

4.1. Fit a straight line of the form $Y = a_0 + a_1 x$ to the data:

x	y	x	y
1	2.4	4	4.2
2	3.1	6	5.0
3	3.5	8	6.0

4.2. Find the values of a_0 and a_1 so that $Y = a_0 + a_1 x$ fits the data given in the table:

x	y
0	1.0
1	2.9
2	4.8
3	6.7
4	8.6

Presented By: http://www.ebooksuit.com

4.3. If the straight line $y = a_0 + a_1 x$ is the best fit to the set of points (x_1, y_1), $(x_2, y_2) \ldots (x_n, y_n)$ then show that

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma y_i^2 & \Sigma x_i \end{vmatrix} = 0, \qquad (i = 1, 2, \ldots, n).$$

4.4. Use the method of least squares to fit the straight line $Y = a + bx$ to the data

x	y	w
0	2	1
1	5	1
2	8	1
3	11	1

4.5. Find the values of a, b and c so that $Y = a + bx + cx^2$ is the best fit to the data

x	y
0	1
1	0
2	3
3	10
4	21

4.6. The table below gives the temperatures T (in 0 °C) and lengths l (in mm) of a heated rod. If $l = a_0 + a_1 T$, find the values of a_0 and a_1 using linear least squares

T	l
40	600.5
50	600.6
60	600.8
70	600.9
80	601.0

4.7. Determine the normal equations if the cubic polynomial $Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ is fitted to the data (x_i, y_i), $i = 0, 1, 2, \ldots, m$.

4.8. Find best values of a_0, a_1 and a_2 so that the parabola $y = a_0 + a_1 x + a_2 x^2$ fits the data:

Presented By: http://www.ebooksuit.com

x	y	x	y
1.0	1.1	3.0	2.8
1.5	1.2	3.5	3.3
2.0	1.5	4.0	4.1
2.5	2.6		

4.9. Determine the constants a and b by the least-squares method such that $y = ae^{bx}$, fits the following data:

x	y
1.0	40.170
1.2	73.196
1.4	133.372
1.6	243.02

4.10. Fit a function of the form $y = ax^b$ to the following data:

x	y	x	y
2	43	20	8
4	25	40	5
7	18	60	3
10	13	80	2

4.11. Fit an exponential function of the type $y = ae^{bx}$ to the following data:

x	y	x	y
0	0.10	1.5	9.15
0.5	0.45	2.0	40.35
1.0	2.15	2.5	180.75

4.12. The curve $y = ce^{bx}$ is fitted to the data:

x	y	x	y
1	1.5	4	40.1
2	4.6	5	125.1
3	13.9	6	299.5

Find the best values of c and b.

4.13. Fit a function of the form $y = A_1e^{\lambda_1 x} + A_2e^{\lambda_2 x}$ to the data given in the following table:

Presented By: http://www.ebooksuit.com

x	y	x	y
1.0	1.175	1.5	2.129
1.1	1.336	1.6	2.376
1.2	1.510	1.7	2.646
1.3	1.698	1.8	2.942
1.4	1.904		

4.14. Fit a natural cubic B-spline, s, to the data $-2, -1, 0, 1, 2$. Show also that s is unique if $s(-1)$ is prescribed.

4.15. Given the set of knots 0, 1, 2, 3, 4, 5, 6, evaluate the B-spline of order 6 at each of the interval knots by the divided difference method.

4.16. Prove the Cox–de Boor recurrence formula given by (4.72).

4.17. Repeat Problem No.15 using the Cox–de Boor recurrence formula.

4.18. Find a Fourier series approximation to the function defined by the graph in Fig. 4.10.

Figure 4.10.

4.19. Compute the FFT of the sequence $\{1, 2, 3, 4, 4, 3, 2, 1\}$ using Sande–Tukey algorithm.

4.20. Compute the FFT of the function $f(t)$ defined by the points $f_0 = 1$, $f_1 = -1$, $f_2 = -1$, $f_3 = -1$, $f_4 = 1$, $f_5 = 1$, $f_6 = 1$, $f_7 = -1$, using the Cooley–Tukey algorithm.

4.21. If $P_n(x)$ is any polynomial of degree n with leading coefficient unity, then prove that

$$\max_{-1 \le x \le 1} \left| \frac{T_n(x)}{2^{n-1}} \right| \le \max_{-1 \le x \le 1} |P_n(x)|.$$

4.22. Prove that $x^2 = 1/2[T_0(x) + T_2(x)]$.

4.23. Prove that $T_n(x)$ is a polynomial in x of degree n.

Presented By: http://www.ebooksuit.com

4.24. Prove that the coefficient of x^n in $T_n(x)$ is 2^{n-1}.

4.25. Express the following as polynomials in x:

(a) $T_0(x)+2T_1(x)+T_2(x)$

(b) $2T_0(x)-\frac{1}{4}T_2(x)+\frac{1}{8}T_4(x)$.

4.26. Express the following polynomials as sums of Chebyshev polynomials:

(a) $1+x-x^2+x^3$

(b) $1-x^2+2x^4$.

4.27. Obtain the best lower-degree approximation to the cubic x^3+2x^2.

4.28. For x nearer 1, the sum

$$S=1-x+\frac{x^2}{2}-\frac{x^3}{6}+\frac{x^4}{24}-\frac{x^5}{120}+\frac{x^6}{720}-\frac{x^7}{5040}$$

gives a result which is correct to five decimal places. Economize the above series if the fourth decimal place is not to be affected.

4.29. Economize the series

$$\sinh x=x+\frac{x^3}{6}+\frac{x^5}{120}+\frac{x^7}{5040},$$

on the interval $[-1, 1]$, allowing for a tolerance of 0.0005.

4.30. Economize the series given by

$$f(x)=1-\frac{1}{2}x-\frac{1}{8}x^2-\frac{1}{16}x^3.$$

Presented By: http://www.ebooksuit.com

CHAPTER

Numerical Differentiation and Integration

5.1 INTRODUCTION

In Chapter 3, we were concerned with the general problem of interpolation, viz., given the set of values $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ 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) $\dfrac{dy}{dx}, \dfrac{d^2y}{dx^2}, \ldots$ for any value of x in $[x_0, x_n]$, and

(ii) $\displaystyle\int_{x_0}^{x_n} y\,dx.$

5.2 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.

Presented By: http://www.ebooksuit.com

Consider Newton's forward difference formula:

$$y = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + \cdots, \tag{5.1}$$

where

$$x = x_0 + uh. \tag{5.2}$$

Then

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = \frac{1}{h}\left(\Delta y_0 + \frac{2u-1}{2}\Delta^2 y_0 + \frac{3u^2-6u+2}{6}\Delta^3 y_0 + \cdots\right). \tag{5.3}$$

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 (5.2), and hence (5.3) gives

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{1}{h}\left(\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \cdots\right). \tag{5.4}$$

Differentiating (5.3) once again, we obtain

$$\frac{d^2y}{dx^2} = \frac{1}{h^2}\left(\Delta^2 y_0 + \frac{6u-6}{6}\Delta^3 y_0 + \frac{12u^2-36u+22}{24}\Delta^4 y_0 + \cdots\right), \tag{5.5}$$

from which we obtain

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{1}{h^2}\left(\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 + \cdots\right). \tag{5.6}$$

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

$$\left[\frac{dy}{dx}\right]_{x=x_n} = \frac{1}{h}\left(\nabla y_n + \frac{1}{2}\nabla^2 y_n + \frac{1}{3}\nabla^3 y_n + \cdots\right) \tag{5.7}$$

and

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_n} = \frac{1}{h^2}\left(\nabla^2 y_n + \nabla^3 y_n + \frac{11}{12}\nabla^4 y_n + \frac{5}{6}\nabla^5 y_n + \cdots\right). \tag{5.8}$$

(b) Stirling's formula gives

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{1}{h}\left(\frac{\Delta y_{-1} + \Delta y_0}{2} - \frac{1}{6}\,\frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{2} + \frac{1}{30}\,\frac{\Delta^5 y_{-3} + \Delta^5 y_{-2}}{2} + \cdots\right) \tag{5.9}$$

Presented By: http://www.ebooksuit.com

and

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{1}{h^2}\left(\Delta^2 y_{-1} - \frac{1}{12}\Delta^4 y_{-2} + \frac{1}{90}\Delta^6 y_{-3} - \cdots\right). \tag{5.10}$$

If a derivative is required near the end of a table, one of the following formulae may be used to obtain better accuracy

$$hy_0' = \left(\Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \frac{1}{4}\Delta^4 + \frac{1}{5}\Delta^5 - \frac{1}{6}\Delta^6 + \cdots\right)y_0 \tag{5.11}$$

$$= \left(\Delta + \frac{1}{2}\Delta^2 - \frac{1}{6}\Delta^3 + \frac{1}{12}\Delta^4 - \frac{1}{20}\Delta^5 + \frac{1}{30}\Delta^6 - \cdots\right)y_{-1} \tag{5.12}$$

$$h^2 y_0'' = \left(\Delta^2 - \Delta^3 + \frac{11}{12}\Delta^4 - \frac{5}{6}\Delta^5 + \frac{137}{180}\Delta^6 - \frac{7}{10}\Delta^7 + \frac{363}{560}\Delta^8 - \cdots\right)y_0 \tag{5.13}$$

$$= \left(\Delta^2 - \frac{1}{12}\Delta^4 + \frac{1}{12}\Delta^5 - \frac{13}{180}\Delta^6 + \frac{11}{180}\Delta^7 - \frac{29}{560}\Delta^8 + \cdots\right)y_{-1} \tag{5.14}$$

$$hy_n' = \left(\nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \frac{1}{4}\nabla^4 + \frac{1}{5}\nabla^5 + \frac{1}{6}\nabla^6 + \frac{1}{7}\nabla^7 + \frac{1}{8}\nabla^8 + \cdots\right)y_n \tag{5.15}$$

$$= \left(\nabla - \frac{1}{2}\nabla^2 - \frac{1}{6}\nabla^3 - \frac{1}{12}\nabla^4 - \frac{1}{20}\nabla^5 - \frac{1}{30}\nabla^6 - \frac{1}{42}\nabla^7 - \frac{1}{56}\nabla^8 - \cdots\right)y_{n+1} \tag{5.16}$$

$$h^2 y_n'' = \left(\nabla^2 + \nabla^3 + \frac{11}{12}\nabla^4 + \frac{5}{6}\nabla^5 + \frac{137}{180}\nabla^6 + \frac{7}{10}\nabla^7 + \frac{363}{560}\nabla^8 + \cdots\right)y_n \tag{5.17}$$

$$= \left(\nabla^2 - \frac{1}{12}\nabla^4 - \frac{1}{12}\nabla^5 - \frac{13}{180}\nabla^6 - \frac{11}{180}\nabla^7 - \frac{29}{560}\nabla^8 - \cdots\right)y_{n+1}. \tag{5.18}$$

For more details, the reader is referred to Interpolation and Allied Tables. The following examples illustrate the use of the formulae stated above.

Example 5.1 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		

Presented By: http://www.ebooksuit.com

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 (5.11) 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 (5.12), 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 (5.13) 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 (5.14), we obtain

$$\left[\frac{d^2y}{dx^2}\right]_{x=1.2} = \frac{1}{0.04}\left[0.1333 - \frac{1}{12}(0.0067) + \frac{1}{12}(0.0013)\right] = 3.32.$$

Presented By: http://www.ebooksuit.com

Example 5.2 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 5.1. Here $x_n = 2.2$, $y_n = 9.0250$ and $h = 0.2$. Hence formula (5.15) 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 (5.15) or (5.16). Formula (5.15) gives

$$\left[\frac{dy}{dx}\right]_{x=2.0} = \frac{1}{0.2}\left[1.3395 + \frac{1}{2}(0.2429) + \frac{1}{3}(0.0441) + \frac{1}{4}(0.0080) + \frac{1}{5}(0.0013) + \frac{1}{6}(0.0001)\right]$$

$$= 7.3896.$$

whereas from formula (5.16), 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 5.3 Find dy/dx and d^2y/dx^2 at $x = 1.6$ for the tabulated function of Example 5.1.

Choosing $x_0 = 1.6$, formula (5.9) 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 (5.10) yields

$$\left[\frac{d^2y}{dx^2}\right]_{x=1.6} = \frac{1}{0.04}\left[0.1988 - \frac{1}{12}(0.0080) + \frac{1}{90}(0.0001)\right] = 4.9525.$$

Presented By: http://www.ebooksuit.com

In the above 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.

5.2.1 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.) From Table 3.4 (p. 71), it is then clear 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, Stirlings formula (5.9). This can be written in the form

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{\Delta y_{-1} + \Delta y_0}{2h} + T_1 = \frac{y_1 - y_{-1}}{2h} + T_1, \tag{5.19}$$

where T_1, the truncation error, is given by

$$T_1 = \frac{1}{6h}\left|\frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{2}\right|. \tag{5.20}$$

Similarly, formula (5.10) can be written as

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{1}{h^2}\Delta^2 y_{-1} + T_2, \tag{5.21}$$

where

$$T_2 = \frac{1}{12h^2}|\Delta^4 y_{-2}|. \tag{5.22}$$

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

Presented By: http://www.ebooksuit.com

second derivatives, and so on. Thus *rounding error* increases as h decreases. Considering again Stirling's formula in the form of (5.19), 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

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{\Delta y_{-1} + \Delta y_0}{2h} - \frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{12h} + \cdots$$

$$= \frac{y_{-2} - 8y_{-1} + 8y_1 - y_2}{12h} + \cdots \tag{5.23}$$

has the maximum rounding error

$$\frac{18\varepsilon}{12h} = \frac{3\varepsilon}{2h}.$$

Finally, the formula

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{\Delta^2 y_{-1}}{h^2} + \cdots = \frac{y_{-1} - 2y_0 + y_1}{h^2} + \cdots \tag{5.24}$$

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 5.4 Assuming that the function values given in the table of Example 5.1 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 $x = 1.6$:

$$\textit{Truncation error} = \frac{1}{6h}\left|\frac{\Delta^3 y_{-1} + \Delta^3 y_0}{2}\right|, \quad \text{from (5.20)}$$

$$= \frac{1}{6(0.2)}\,\frac{0.0361 + 0.0441}{2}$$

$$= 0.03342$$

and

$$\textit{Rounding error} = \frac{3\varepsilon}{2h}, \quad \text{from (5.23)}$$

$$= \frac{3(0.5)10^{-4}}{0.4}$$

$$= 0.00038.$$

Presented By: http://www.ebooksuit.com

Hence,

$$\text{Total error} = 0.03342 + 0.00038 = 0.0338.$$

Using Stirling's formula form (5.19), 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 (5.24), 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

$$\text{Total error in } \left[\frac{d^2y}{dx^2}\right]_{x=1.6} = 0.0167 + 0.0050 = 0.0217.$$

5.2.2 The Cubic Spline Method

The cubic spline derived in Section 3.14 can conveniently be used to compute the *first* and *second* derivatives of a function. For a natural cubic spline, the recurrence formulae (3.108) or (3.109) may be used to compute the spline second derivatives depending upon the choice of the subdivisions. Then Eq. (3.106) 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 (3.105) and (3.107) directly. If, on the other hand, end conditions involving the first derivatives are given, then recurrence formulae (3.111) or (3.113) may be used to compute the remaining first derivatives.

The following example illustrates the use of the spline formulae in numerical differentiation.

Example 5.5 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

$$y_0 = y_2 = 0, \quad y_1 = 1 \quad \text{and} \quad M_0 = M_2 = 0.$$

Presented By: http://www.ebooksuit.com

Using formulae (3.109), 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 (3.106) 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

$$s'(x) = \frac{2}{\pi}\left[-\frac{2}{\pi^2}(3x^2) + \frac{3}{2}\right]. \tag{i}$$

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

Presented By: http://www.ebooksuit.com

It is seen that there is a greater inaccuracy in the values of the spline second derivatives.

5.3 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

$$\frac{dy}{dp} = \Delta y_0 + \frac{2p-1}{2}\Delta^2 y_0 + \frac{3p^2-3p+2}{6}\Delta^3 y_0 + \cdots \tag{5.25}$$

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

$$c_0 + c_1 p + c_2 p^2 = 0, \tag{5.26}$$

where

$$\left.\begin{aligned} c_0 &= \Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 \\ c_1 &= \Delta^2 y_0 - \Delta^3 y_0 \\ \text{and} \quad c_2 &= \frac{1}{2}\Delta^3 y_0. \end{aligned}\right\} \tag{5.27}$$

Values of x can then be found from the relation $x = x_0 + ph$.

Example 5.6 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

Presented By: http://www.ebooksuit.com

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 (5.25), 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

$$y(1.58) = 0.9996 - 0.2(0.0021) + \frac{-0.2(-0.2+1)}{2}(-0.0099)$$

$$= 0.9996 - 0.0004 + 0.0008$$

$$= 1.0.$$

5.4 NUMERICAL INTEGRATION

The general problem of numerical integration may be stated as follows. Given a set of data points $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ of a function $y = f(x)$, where $f(x)$ is not known explicitly, it is required to compute the value of the definite integral

$$I = \int_a^b y\,dx. \tag{5.28}$$

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

Presented By: http://www.ebooksuit.com

$$I = \int_{x_0}^{x_n} y\,dx.$$

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

$$\int_{x_0}^{x_n} y\,dx = nh\left[y_0 + \frac{n}{2}\Delta y_0 + \frac{n(2n-3)}{12}\Delta^2 y_0 + \frac{n(n-2)^2}{24}\Delta^3 y_0 + \cdots \right]. \quad (5.29)$$

From this *general formula*, we can obtain different integration formulae by putting $n = 1, 2, 3, \ldots$, 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.

5.4.1 Trapezoidal Rule

Setting $n = 1$ in the general formula (5.29), all differences higher than the first will become zero and we obtain

$$\int_{x_0}^{x_1} y\,dx = h\left(y_0 + \frac{1}{2}\Delta y_0 \right) = h\left[y_0 + \frac{1}{2}(y_1 - y_0) \right] = \frac{h}{2}(y_0 + y_1). \quad (5.30)$$

For the next interval $[x_1, x_2]$, we deduce similarly

$$\int_{x_1}^{x_2} y\,dx = \frac{h}{2}(y_1 + y_2) \quad (5.31)$$

and so on. For the last interval $[x_{n-1}, x_n]$, we have

$$\int_{x_{n-1}}^{x_n} y\,dx = \frac{h}{2}(y_{n-1} + y_n). \quad (5.32)$$

Presented By: http://www.ebooksuit.com

Combining all these expressions, we obtain the rule

$$\int_{x_0}^{x_n} y\,dx = \frac{h}{2}[y_0 + 2(y_1 + y_2 + \cdots + y_{n-1}) + y_n], \tag{5.33}$$

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

$$\int_{x_0}^{x_1} y\,dx = \int_{x_0}^{x_1} \left[y_0 + (x - x_0)y_0' + \frac{(x - x_0)^2}{2} y_0'' + \cdots \right] dx$$

$$= hy_0 + \frac{h^2}{2} y_0' + \frac{h^3}{6} y_0'' + \cdots \tag{5.34}$$

Similarly,

$$\frac{h}{2}(y_0 + y_1) = \frac{h}{2}\left(y_0 + y_0 + hy_0' + \frac{h^2}{2} y_0'' + \frac{h^3}{6} y_0''' + \cdots \right)$$

$$= hy_0 + \frac{h^2}{2} y_0' + \frac{h^3}{4} y_0'' + \cdots \tag{5.35}$$

From (5.34) and (5.35), we obtain

$$\int_{x_0}^{x_1} y\,dx - \frac{h}{2}(y_0 + y_1) = -\frac{1}{12} h^3 y_0'' + \cdots, \tag{5.36}$$

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., $[x_1, x_2], [x_2, x_3], \ldots$ and $[x_{n-1}, x_n]$. We thus have

$$E = -\frac{1}{12} h^3 (y_0'' + y_1'' + \cdots + y_{n-1}''), \tag{5.37}$$

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 (5.37), we obtain

Presented By: http://www.ebooksuit.com

$$E = -\frac{1}{12}h^3 n y''(\bar{x}) = -\frac{b-a}{12}h^2 y''(\bar{x}) \tag{5.38}$$

since $nh = b - a$.

5.4.2 Simpson's 1/3-Rule

This rule is obtained by putting $n = 2$ in Eq. (5.29), 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

$$\int_{x_0}^{x_n} y\,dx = \frac{h}{3}[y_0 + 4(y_1 + y_3 + y_5 + \cdots + y_{n-1}) + 2(y_2 + y_4 + y_6 + \cdots + y_{n-2}) + y_n], \tag{5.39}$$

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 5.4.1, it can be shown that the error in Simpson's rule is given by

$$\int_a^b y\,dx = \frac{h}{3}[y_0 + 4(y_1 + y_3 + y_5 + \cdots + y_{n-1}) + 2(y_2 + y_4 + y_6 + \cdots + y_{n-2}) + y_n]$$

$$= -\frac{b-a}{180}h^4 y^{\text{iv}}(\bar{x}), \tag{5.40}$$

where $y^{\text{iv}}(\bar{x})$ is the largest value of the fourth derivatives.

Presented By: http://www.ebooksuit.com

5.4.3 Simpson's 3/8-Rule

Setting $n = 3$ in (5.29) 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

$$\int_{x_0}^{x_n} y\,dx = \frac{3h}{8}[(y_0 + 3y_1 + 3y_2 + y_3) + (y_3 + 3y_4 + 3y_5 + y_6) + \cdots$$

$$+ (y_{n-3} + 3y_{n-2} + 3y_{n-1} + y_n)]$$

$$= \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + 2y_3 + 3y_4 + 3y_5 + 2y_6 + \cdots$$

$$+ 2y_{n-3} + 3y_{n-2} + 3y_{n-1} + y_n). \tag{5.41}$$

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 $-(3/80)\ h^5 y^{iv}(\bar{x})$.

5.4.4 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

$$\int_{x_0}^{x_4} y\,dx = \frac{2h}{45}(7y_0 + 32y_1 + 12y_2 + 32y_3 + 7y_4). \tag{5.42}$$

The leading term in the error of this formula can be shown to be

$$-\frac{8h^7}{945} y^{vi}(\bar{x}).$$

Presented By: http://www.ebooksuit.com

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

$$\int_{x_0}^{x_6} y\,dx = \frac{3h}{10}(y_0 + 5y_1 + y_2 + 6y_3 + y_4 + 5y_5 + y_6), \tag{5.43}$$

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.

5.4.5 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 (3.104). On carrying out the integration and simplifying, we obtain

$$I = \sum_{i=1}^{n} \left[\frac{h}{2}(y_{i-1} + y_i) - \frac{h^3}{24}(M_{i-1} + M_i) \right], \tag{5.44}$$

where M_i, the spline second-derivatives, are calculated from the recurrence relation

$$M_{i-1} + 4M_i + M_{i+1} = \frac{6}{h^2}(y_{i-1} - 2y_i + y_{i+1}), \qquad i = 1, 2, \ldots, n-1. \tag{3.109}$$

The use of the cubic spline method is demonstrated in Example 5.12.

5.4.6 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

Presented By: http://www.ebooksuit.com

$$I = \int_a^b y\,dx$$

and evaluate it by the trapezoidal rule (5.33) with two different subintervals of widths h_1 and h_2 to obtain the approximate values I_1 and I_2, respectively. Then Eq. (5.38) gives the errors E_1 and E_2 as

$$E_1 = -\frac{1}{12}(b-a)h_1^2 y''(\bar{x}) \tag{5.45}$$

and

$$E_2 = -\frac{1}{12}(b-a)h_2^2 y''(\bar{\bar{x}}). \tag{5.46}$$

Since the term $y''(\bar{\bar{x}})$ in (5.46) 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

$$E_2 = \frac{h_2^2}{h_2^2 - h_1^2}(I_2 - I_1). \tag{5.47}$$

We therefore obtain a new approximation I_3 defined by

$$I_3 = I_2 - E_2 = \frac{I_1 h_2^2 - I_2 h_1^2}{h_2^2 - h_1^2}, \tag{5.48}$$

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

$$h_2 = \frac{1}{2}h_1 = \frac{1}{2}h$$

Eq. (5.48) can be written in the more convenient form

$$I\left(h, \frac{1}{2}h\right) = \frac{1}{3}\left[4I\left(\frac{1}{2}h\right) - I(h)\right], \tag{5.49}$$

Presented By: http://www.ebooksuit.com

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*.

5.4.7 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

$$I = \int_a^b y\,dx \tag{5.50}$$

is evaluated by an integration formula of the type

$$I_n = \sum_{i=0}^{n} C_i y_i, \tag{5.51}$$

Presented By: http://www.ebooksuit.com

where the coefficients C_i are determined completely by the abscissae x_i. Integration formulae of the type (5.51) 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 Eqs. (5.47)–(5.50) 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

$$\text{(a)} \quad \int_{x_0}^{x_2} y\,dx = 2hy_1 + \frac{h^3}{3} y''(\bar{x}), \ (x_0 < \bar{x} < x_2) \tag{5.52}$$

$$\text{(b)} \quad \int_{x_0}^{x_3} y\,dx = \frac{3h}{2}(y_1 + y_2) + \frac{3h^3}{4} y''(\bar{x}), \ (x_0 < \bar{x} < x_3) \tag{5.53}$$

$$\text{(c)} \quad \int_{x_0}^{x_4} y\,dx = \frac{4h}{3}(2y_1 - y_2 + 2y_3) + \frac{14}{45} h^5 y^{\text{iv}}(\bar{x}), \ (x_0 < \bar{x} < x_4) \tag{5.54}$$

$$\text{(d)} \quad \int_{x_0}^{x_5} y\,dx = \frac{5h}{24}(11y_1 + y_2 + y_3 + 11y_4) + \frac{95}{144} h^5 y^{\text{iv}}(\bar{x}), \ (x_0 < \bar{x} < x_5) \tag{5.55}$$

$$\text{(e)} \quad \int_{x_0}^{x_6} y\,dx = \frac{6h}{20}(11y_1 - 14y_2 + 26y_3 - 14y_4 + 11y_5) + \frac{41}{140} h^7 y^{\text{vi}}(\bar{x}), \quad (x_0 < \bar{x} < x_6). \tag{5.56}$$

A convenient method for determining the coefficients in the Newton–Cotes formulae is the method of undetermined coefficients. This is demonstrated in Example 5.13.

Example 5.7 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

Presented By: http://www.ebooksuit.com

We know that

$$\text{Area} = \int_{7.47}^{7.52} f(x)\,dx$$

with $h = 0.01$, the trapezoidal rule (5.32) gives

$$\text{Area} = \frac{0.01}{2}[1.93 + 2(1.95 + 1.98 + 2.01 + 2.03) + 2.06] = 0.0996.$$

Example 5.8 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

$$V = \frac{\pi(0.25)}{3}[1.0000 + 4(0.9793 + 0.8261) + 2(0.9195) + 0.7081]$$

$$= 2.8192.$$

Presented By: http://www.ebooksuit.com

Example 5.9 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

$$I = \frac{1}{12}[1.0 + 4(0.8000 + 0.5714) + 2(0.6667) + 0.5] = 0.6932.$$

Presented By: http://www.ebooksuit.com

(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 5.10 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 (5.49), we obtain

$$I\left(h, \frac{1}{2}h\right) = 0.6970 + \frac{1}{3}(0.6970 - 0.7084) = 0.6932.$$

Presented By: http://www.ebooksuit.com

$$I\left(\frac{1}{2}h, \frac{1}{4}h\right) = 0.6941 + \frac{1}{3}(0.6941 - 0.6970) = 0.6931$$

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 5.11 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 5.12 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 $y = \sin \pi x$ is

Presented By: http://www.ebooksuit.com

x	y
0	0
0.5	1.0
1.0	0.0

Using (3.109) with $M_0 = M_2 = 0$, we obtain $M_1 = -12$. Then formula (5.44) 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 5.13 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

$$a_{-1} = \frac{2}{3} = a_1 \quad \text{and} \quad a_0 = \frac{4h}{3}.$$

Presented By: http://www.ebooksuit.com

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 5.4.2.

5.5 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

$$\frac{1}{e^x - 1} = \frac{1}{x} - \frac{1}{2} + B_1 x + B_3 x^3 + B_5 x^5 + \cdots, \tag{5.57}$$

where

$$B_{2r} = 0, \qquad B_1 = \frac{1}{12}, \qquad B_3 = -\frac{1}{720}, \qquad B_5 = \frac{1}{30,240}, \text{ etc.}$$

In (5.57), if we set $x = hD$ and use the relation $E \equiv e^{hD}$ (see Section 3.3.4), we obtain the identity

$$\frac{1}{E-1} \equiv \frac{1}{hD} - \frac{1}{2} + B_1 hD + B_3 h^3 D^3 + B_5 h^5 D^5 + \cdots \tag{5.58}$$

or equivalently

$$\frac{E^n - 1}{E-1} = \frac{1}{hD}(E^n - 1) - \frac{1}{2}(E^n - 1) + B_1 hD(E^n - 1) + B_3 h^3 D^3 (E^n - 1) + \cdots \tag{5.58}$$

Operating this identity on y_0, we obtain

$$\begin{aligned}\frac{E^n - 1}{E-1} y_0 &= \frac{1}{hD}(E^n - 1)y_0 - \frac{1}{2}(E^n - 1)y_0 + B_1 hD(E^n - 1)y_0 + \cdots \\ &= \frac{1}{hD}(y_n - y_0) - \frac{1}{2}(y_n - y_0) + B_1 h(y_n' - y_0') + B_3 h^3 (y_n''' - y_0''') \\ &\quad + B_5 h^5 (y_n^{\text{v}} - y_0^{\text{v}}) + \cdots \end{aligned} \tag{5.59}$$

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.

Presented By: http://www.ebooksuit.com

Hence, Eq. (5.59) becomes

$$\int_{x_0}^{x_n} y\,dx = \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n) - \frac{h^2}{12}(y_n' - y_0')$$

$$+ \frac{h^4}{720}(y_n''' - y_0''') - \frac{h^6}{30,240}(y_0^{v} - y_0^{v}) + \cdots \tag{5.60}$$

which is called the *Euler–Maclaurin's formula* for integration. The first expression on the right-hand side of (5.60) 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 5.14 Evaluate

$$I = \int_0^{\pi/2} \sin x\,dx$$

using the Euler–Maclaurin's formula.

In this case, formula (5.60) 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

$$\int_0^{\pi/2} \sin x\,dx = \frac{\pi}{16}[(0 + 2(0.382683) + .707117 + 0.923879 + 1.000000)$$

$$= 0.987119 + 0.012851 + 0.000033$$

$$= 1.000003.$$

Presented By: http://www.ebooksuit.com

Example 5.15 Use the Euler–Maclaurin formula to prove

$$\sum_{1}^{n} x^2 = \frac{n(n+1)(2n+1)}{6}.$$

In this case, rewrite Eq. (5.60) as

$$\frac{1}{2}y_0 + y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2}y_n = \frac{1}{h}\int_{x_0}^{x_n} y\,dx + \frac{h}{12}(y_n' - y_0') - \frac{h^3}{720}(y_n''' - y_0''')$$

$$+ \frac{h^5}{30,240}(y_n^{v} - y_0^{v}) - \cdots \qquad \text{(i)}$$

Here $y(x) = x^2$, $y'(x) = 2x$ and $h = 1$.

Hence eq. (i) gives

$$\text{Sum} = \int_{1}^{n} x^2\,dx + \frac{1}{2}(n^2+1) + \frac{1}{12}(2n-2)$$

$$= \frac{1}{3}(n^3-1) + \frac{1}{2}(n^2+1) + \frac{1}{6}(n-1)$$

$$= \frac{1}{6}(2n^3 + 3n^2 + n)$$

$$= \frac{n(n+1)(2n+1)}{6}.$$

5.6 ADAPTIVE QUADRATURE METHODS

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 and this can easily be modified to the other integration formulae:

Suppose that we wish to approximate the integral

$$I = \int_{a}^{b} y(x)\,dx \qquad (5.61)$$

Presented By: http://www.ebooksuit.com

to within an accuracy $\varepsilon > 0$. Using Simpson's (1/3)-rule with $h = (b-a)/2$, we obtain

$$I = \int_a^b y(x)\,dx \approx \frac{h}{3}\left[y(a) + 4y\frac{a+b}{2} + y(b)\right] - \frac{h^5}{90}y^{\text{iv}}(\xi_1), \quad a < \xi_1 < b$$

$$= I(a,b) - \frac{(b-a)h^4}{180}y^{\text{iv}}(\xi_1), \tag{5.62}$$

where

$$I(a,b) = \frac{h}{3}\left[y(a) + 4y\frac{a+b}{2} + y(b)\right]. \tag{5.63}$$

Now, we subdivide the interval and set $h = (b-a)/4$. Simpson's (1/3)-rule then gives

$$I = \int_a^b y(x)\,dx = \frac{h}{6}\left[y(a) + 4y\frac{3a+b}{4} + 2y\frac{a+b}{2} + 4y\frac{a+3b}{4} + y(b)\right]$$

$$-\frac{h^4(b-a)}{180\times 16}y^{\text{iv}}(\xi_2)$$

$$= \frac{h}{6}\left[y(a) + 4y\frac{3a+b}{4} + y\frac{a+b}{2}\right]$$

$$+\frac{h}{6}\left[y\frac{a+b}{2} + 4y\frac{a+3b}{4} + y(b)\right] - \frac{(b-a)h^4}{180\times 16}y^{\text{iv}}(\xi_2)$$

$$= I\left(a, \frac{a+b}{2}\right) + I\left(\frac{a+b}{2}, b\right) - \frac{(b-a)h^4}{180\times 16}y^{\text{iv}}(\xi_2), \tag{5.64}$$

where

$$I\left(a, \frac{a+b}{2}\right) = \frac{h}{6}\left[y(a) + 4y\frac{3a+b}{4} + y\frac{a+b}{2}\right] \tag{5.65a}$$

and

$$I\left(\frac{a+b}{2}, b\right) = \frac{h}{6}\left[y\frac{a+b}{2} + 4y\frac{a+3b}{4} + y(b)\right]. \tag{5.65b}$$

Assuming

$$y^{\text{iv}}(\xi_1) = y^{\text{iv}}(\xi_2)$$

Presented By: http://www.ebooksuit.com

Eqs. (5.62) and (5.64) give on simplification

$$\frac{1}{15}\left[I(a,b)-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right]=\frac{(b-a)h^4}{180\times 16}y^{\text{iv}}(\xi_2). \tag{5.66}$$

Substituting (5.66) in (5.64), we obtain an estimate for the error, viz.

$$\left|\int_a^b y(x)\,dx-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right|$$

$$=\frac{1}{15}\left|I(a,b)-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right|. \tag{5.67}$$

If we suppose

$$\frac{1}{15}\left|I(a,b)-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right|<\varepsilon \tag{5.68}$$

for some $\varepsilon>0$ in the interval $[a,b]$, then Eq. (5.67) means that

$$\left|\int_a^b y(x)\,dx-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right|<\varepsilon \tag{5.69}$$

and that

$$\int_a^b y(x)\,dx\approx I\left(a,\frac{a+b}{2}\right)+I\left(\frac{a+b}{2},b\right). \tag{5.70}$$

to within an accuracy of $\varepsilon>0$.

If the inequality (5.68) 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 5.16 Test the error estimate given by (5.67) in the evaluation of the integral

$$I=\int_0^{\pi/2}\cos x\,dx.$$

Presented By: http://www.ebooksuit.com

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.

5.7 GAUSSIAN INTEGRATION

We consider the numerical evaluation of the integral

$$I = \int_a^b f(x)\, dx. \tag{5.71}$$

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

$$\int_{-1}^{1} F(u)\, du = W_1 F(u_1) + W_2 F(u_2) + \cdots + W_n F(u_n) = \sum_{i=1}^{n} W_i F(u_i), \tag{5.72}$$

where the W_i and u_i are called the *weights* and *abscissae*, respectively. An advantage of this formula is that the 'abscissae and weights' are symmetrical with respect to the middle point of the interval.

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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

$$W_0=\int_{-1}^{1}\frac{u-u_1}{u_0-u_1}\,du=\frac{1}{u_0-u_1}\left[\frac{u^2}{2}-u_1u\right]_{-1}^{1}=1$$

and

$$W_1=\int_{-1}^{1}\frac{u-u_0}{u_1-u_0}\,du=\frac{1}{u_1-u_0}\left[\frac{u^2}{2}-u_0u\right]_{-1}^{1}=1.$$

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 found from (5.82). 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 5.1, the abscissae and weights for values of n up to $n=6$.

Table 5.1 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

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

which is singular at $t = x$. The principal value, $P(I)$, of the integral is defined by

$$P(I) = \lim_{\varepsilon \to 0} \left[\int_a^{t-\varepsilon} \frac{f(x)}{x-t} dx + \int_{t+\varepsilon}^{b} \frac{f(x)}{x-t} dx \right], \qquad (a < t < b)$$
$$= I(f), \qquad \text{for } t < a \quad \text{or} \quad t > b. \tag{5.85}$$

Setting $x = a + uh$ and $t = a + kh$ in (5.84), we obtain

$$P(I) = P \int_0^p \frac{f(a+uh)}{u-k} du.$$

Replacing $f(a+uh)$ by *Newton's forward difference formula*. Table 3.1 (p.66) (*see* Section 3.6) at $x = a$ and simplifying, we have

$$I(f) = \sum_{j=0}^{\infty} \frac{\Delta^j f(a)}{j!} c_j, \tag{5.86}$$

where the constants c_j are given by

$$c_j = P \int_0^p \frac{(u)_j}{u-k} du. \tag{5.87}$$

In (5.87), $(u)_0 = 1$, $(u)_1 = u$, $(u)_2 = u(u-1)$, etc. Various approximate formulae can be obtained by truncating the series on the right side of (5.86). Thus, by writing (5.86) in the form

$$I_n(f) = \sum_{j=0}^{n} \frac{\Delta^j f(a)}{j!} c_j \tag{5.88}$$

we obtain rules of orders 1, 2, 3,... etc., by setting $n = 1, 2, 3, \ldots$ respectively.

(a) *Two-point rule*, $n = 1$:

$$I_1(f) = \sum_{j=0}^{1} \frac{\Delta^j f(a)}{j!} c_j$$
$$= c_0 f(a) + c_1 \Delta f(a) \tag{5.89}$$
$$= (c_0 - c_1) f(a) + c_1 f(a+h).$$

Presented By: http://www.ebooksuit.com

(b) *Three-point rule*, $n = 2$:

$$I_2(f) = \sum_{j=0}^{2} \frac{\Delta^j f(a)}{j!} c_j$$

$$= c_0 f(a) + c_1 \Delta f(a) + c_2 \Delta^2 f(a)$$

$$= \left(c_0 - c_1 + \frac{1}{2}c_2\right) f(a) + (c_1 - c_2) f(a+h) + \frac{1}{2} c_2 f(a+2h). \quad (5.90)$$

In the above relations, the values of c_j are given by

$$\left.\begin{aligned} c_0 &= \log_e \left| \frac{p-k}{k} \right| \\ c_1 &= p + c_0 k \\ c_2 &= \frac{p^2}{2} + p(k-1) + c_0 k(k-1). \end{aligned}\right\} \quad (5.91)$$

A discussion of errors in these formulae may be found in the paper by Delves [1968].

5.8.2 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

$$I(s) = \int_a^b f(t)\, \phi(t-s)\, dt, \quad (5.92)$$

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 (5.92) can be written as

$$I(s) = \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} f(t)\, \phi(t-s)\, dt. \quad (5.93)$$

The method to be followed here is to approximate $f(t)$ in (5.93) by the linear interpolating function $f_n(t)$, where

$$f_n(t) = \frac{1}{h}[(t_{j+1} - t) f(t_j) + (t - t_j) f(t_{j+1})]. \quad (5.94)$$

Presented By: http://www.ebooksuit.com

Substituting $f_n(t)$ for $f(t)$ in (5.93), 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

$$I(s)=h\sum_{j=0}^{n-1}[\alpha_j f(t_j)+\beta_j f(t_{j+1})], \tag{5.95}$$

where

$$\alpha_j=h\int_0^1(1-p)\,\phi(t_j+ph-s)\,dp \tag{5.96a}$$

and

$$\beta_j=h\int_0^1 p\phi(t_j+ph-s)\,dp. \tag{5.96b}$$

It is clear from (5.96a) and (5.96b) that if $\phi(u)\equiv 1$, then $\alpha_j=\beta_j=h/2$, and hence Eq. (5.95) 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 5.4.1. Hence the rule defined by (5.95), (5.96a) and (5.96b) 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.[1] 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}).[2]

The error in generalized quadrature can also be estimated by the method outlined in Section 5.4.1. 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 $[a, b]$.

1. See, for example, Sastry [1973; 1976].

2. See, Noble [1964], p. 241.

Presented By: http://www.ebooksuit.com

5.9 NUMERICAL CALCULATION OF FOURIER INTEGRALS

We consider, in this section, the problem of computing definite integrals which involve oscillatory functions, i.e. integrals of the form

$$I_c = \int_a^b f(x) \cos \omega x \, dx \tag{5.97}$$

and

$$I_s = \int_a^b f(x) \sin \omega x \, dx. \tag{5.98}$$

Such integrals, called the *Fourier integrals*, occur in practical applications, e.g. *spectral analysis*. We describe below three methods for the numerical integration of such integrals and compare their accuracies through a numerical example. Only the outlines of the methods will be indicated here. For details, the reader is referred to the research papers cited.

For definiteness, we outline the methods with reference to the particular example

$$I = \int_0^\infty e^{-x} \cos \omega x \, dx = \frac{1}{1+\omega^2}, \tag{5.99}$$

but these methods also hold good for equations of the type (5.98). In all the formulae below, a step-length h is used.

5.9.1 Trapezoidal Rule

Using this rule, the integral in (5.99) approximates to

$$\begin{aligned}
I \approx I_1 &= \frac{h}{2} + h \sum_{n=1}^{\infty} e^{-nh} \cos \omega nh \\
&= h\left(\frac{1}{2} + \text{Re} \sum_{n=1}^{\infty} e^{-nh} e^{i\omega nh}\right) \\
&= h\left[\frac{1}{2} + \text{Re} \sum_{n=1}^{\infty} e^{(-1+i\omega)nh}\right] \\
&= h\left\{\frac{1}{2} + \text{Re}\left[\frac{e^{(-1+i\omega)h}}{1-e^{(-1+i\omega)h}}\right]\right\} \\
&= h\left[\frac{1}{2} + \text{Re}\left(\frac{e^{i\omega h}}{e^h - e^{i\omega h}}\right)\right] \\
&= h\left[\frac{1}{2} + \text{Re}\left(\frac{\cos \omega h + i \sin \omega h}{\cosh h + \sinh h - \cos \omega h - i \sin \omega h}\right)\right],
\end{aligned}$$

Presented By: http://www.ebooksuit.com

which gives on simplification

$$I_1 = \frac{h}{2} \frac{\sinh h}{\cosh h - \cos \omega h}. \tag{5.100}$$

Formula (5.100) is due to Einarsson [1972].

5.9.2 Filon's Formula

In his original paper, Filon [1928] derived formulae for integrals of the type (5.97) and (5.98). In his method, the interval $[a, b]$ is divided into $2N$ subintervals and in each *double interval*, $f(x)$ is approximated by a *quadratic*. Thus this rule is similar to Simpson's rule except that there is an extra factor $\cos \omega x$ in the present case. Since the derivation of the formula is quite involved, only the relevant details are given below.

With $h = (b-a)/(2N)$, let

$$C_1 = \frac{1}{2} f(a) \cos \omega a + f(a+2h) \cos \omega (a+2h)$$
$$+ f(a+4h) \cos \omega (a+4h) + \cdots + \frac{1}{2} f(b) \cos \omega b \tag{5.101}$$

and

$$C_2 = f(a+h) \cos \omega (a+h) + f(a+3h) \cos \omega (a+3h) + \cdots$$
$$+ f(b-h) \cos \omega (b-h). \tag{5.102}$$

Then,

$$\int_a^b f(x) \cos \omega x \, dx = h \{\alpha [f(b) \sin \omega b - f(a) \sin \omega a] + \beta C_1 + \delta C_2\}, \tag{5.103}$$

where

$$\left.\begin{aligned} \alpha &= \frac{\omega^2 h^2 + \omega h \sin \omega h \cos \omega h - 2 \sin^2 (\omega h)}{\omega^3 h^3} \\ \beta &= \frac{2 \{\omega h [1 + \cos^2 (\omega h)] - \sin 2\omega h\}}{\omega^3 h^3} \\ \delta &= \frac{4 (\sin \omega h - \omega h \cos \omega h)}{\omega^3 h^3}. \end{aligned}\right\} \tag{5.104}$$

A similar formula for the integral (5.98) is given by

$$\int_a^b f(x) \sin \omega x \, dx = h \{-\alpha [f(b) \cos \omega b - f(a) \cos \omega a] + \beta S_1 + \delta S_2\}, \tag{5.105}$$

Presented By: http://www.ebooksuit.com

where S_1 and S_2 are sums similar to C_1 and C_2 for $f(x)\sin \omega x$.

For the integral in (5.99), the Filon formula is given by

$$I \approx I_2 = h\left[p\left(\frac{1}{2} + e^{-2h}\cos 2\omega h + e^{-4h}\cos 4\omega h + \cdots\right)\right.$$
$$\left. + q\,(e^{-h}\cos \omega h + e^{-3h}\cos 3\omega h + \cdots)\right], \tag{5.106}$$

where

$$p = 2\left[\frac{1+\cos^2(\omega h)}{\omega^2 h^2} - \frac{\sin 2\omega h}{\omega^3 h^3}\right] \tag{5.107a}$$

and

$$q = 4\left(\frac{\sin \omega h}{\omega^3 h^3} - \frac{\cos \omega h}{\omega^2 h^2}\right). \tag{5.107b}$$

For computational purposes, however, the right side of (5.106) can be put into a more convenient form (see Einarsson [1972]). We have

$$\begin{aligned}
\frac{1}{2} + e^{-2h}\cos 2\omega h + e^{-4h}\cos 4\omega h + \cdots &= \frac{1}{2} + \sum_{n=1}^{\infty} e^{-2nh}\cos 2\omega nh \\
&= \frac{1}{2} + \mathrm{Re}\sum_{n=1}^{\infty} e^{-2nh}\, e^{2i\omega nh} \\
&= \frac{1}{2} + \mathrm{Re}\sum_{n=1}^{\infty} e^{(-1+i\omega)2nh} \\
&= \frac{1}{2} + \mathrm{Re}\left(\frac{e^{2i\omega h}}{e^{2h} - e^{2i\omega h}}\right) \\
&= \frac{1}{2}\,\frac{\sinh 2h}{\cosh 2h - \cos 2\omega h},
\end{aligned}$$

on simplification.

In a similar manner, we obtain

$$e^{-h}\cos \omega h + e^{-3h}\cos 3\omega h + \cdots = \frac{\sinh h - \cos \omega h}{\cosh 2h - \cos 2\omega h}.$$

Presented By: http://www.ebooksuit.com

Hence (5.106) becomes

$$I_2 = h\left(\frac{p}{2}\frac{\sinh 2h}{\cosh 2h - \cos 2\omega h} + q\frac{\sinh h.\cos \omega h}{\cosh 2h - \cos 2\omega h}\right)$$

$$= \frac{h \sinh h}{\cosh 2h - \cos 2\omega h}(p \cosh h + q \cos \omega h), \tag{5.108}$$

where p and q are given by (5.107).

5.9.3 The Cubic Spline Method

We first consider the integral (5.97) and let the interval $[a, b]$ be divided into n equal subintervals, each of which is of length $h = (b-a)/n$. Let $f(x_i) = y_i$, $i = 0, 1, \ldots, n$. If $s(x)$ is the cubic spline interpolating to the data values (x_i, y_i), then we have

$$s(x) = M_{i-1}\frac{(x_i - x)^3}{6h} + M_i\frac{(x - x_{i-1})^3}{6h} + \left(y_{i-1} - \frac{M_{i-1}}{6}h^2\right)\frac{x_i - x}{h}$$

$$+\left(y_i - \frac{M_i}{6}h^2\right)\frac{x - x_{i-1}}{h}, \tag{5.109}$$

where the M_i $[= s''(x_i)]$ satisfy the recurrence relation

$$M_{i-1} + 4M_i + M_{i+1} = \frac{6}{h^2}(y_{i-1} - 2y_i + y_{i+1}), \qquad i = 1, 2, \ldots, n-1. \tag{5.110}$$

Hence we have

$$I_c = \int_a^b f(x)\cos \omega x\, dx \approx \int_a^b s(x)\cos \omega x\, dx.$$

Since $s(x) \in c^2[a, b]$, we can integrate the above integral three times and obtain

$$I_c = \left[s(x)\frac{\sin \omega x}{\omega}\right]_a^b - \int_a^b s'(x)\frac{\sin \omega x}{\omega}dx$$

$$= \frac{1}{\omega}[s(b)\sin b\omega - s(a)\sin a\omega] + \frac{1}{\omega^2}[\cos b\omega\, s'(b) - \cos a\omega\, s'(a)]$$

$$-\frac{1}{\omega^3}[s''(b)\sin b\omega - s''(a)\sin a\omega] + \frac{1}{\omega^3}\int_a^b \sin \omega x\, s'''(x)dx. \tag{5.111}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Similarly, applying Simpson's rule to the integral

$$I = \int_{y_{j-1}}^{y_{j+1}} \int_{x_{i-1}}^{x_{i+1}} f(x,y)\,dx\,dy, \tag{5.118}$$

we obtain

$$I = \frac{h}{3} \int_{y_{j-1}}^{y_{j+1}} [f(x_{i-1},y) + 4f(x_i,y) + f(x_{i+1},y)]\,dy$$

$$= \frac{hk}{9}[f(x_{i-1},y_{j-1}) + 4f(x_{i-1},y_j) + f(x_{i-1},y_{j+1})$$

$$+ 4\{f(x_i,y_{j-1}) + 4f(x_i,y_j) + f(x_i,y_{j+1})\}$$

$$+ f(x_{i+1},y_{j-1}) + 4f(x_{i+1},y_j) + f(x_{i+1},y_{j+1})]$$

$$= \frac{hk}{9}[f_{i-1,j-1} + f_{i-1,j+1} + f_{i+1,j-1} + f_{i+1,j+1}$$

$$+ 4(f_{i-1,j} + f_{i,j-1} + f_{i,j+1} + f_{i+1,j}) + 16f_{i,j}]. \tag{5.119}$$

A numerical example is given below.

Example 5.18 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' (5.117) repeatedly, we obtain

$$I = \frac{0.25}{4}[1.0 + 4(1.6487) + 6(2.7183) + 4(4.4817) + 7.3891]$$

$$= \frac{12.3050}{4}$$

$$= 3.0762.$$

Presented By: http://www.ebooksuit.com

Using 'Simpson's rule' (5.119) 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

5.1. Find $\frac{d}{dx}(J_0)$ at $x = 0.1$ from the following table:

x	$J_0(x)$
0.0	1.0000
0.1	0.9975
0.2	0.9900
0.3	0.9776
0.4	0.9604

5.2. The following table gives the angular displacements θ (radians at different intervals of time t (seconds)

θ	t	θ	t
0.052	0	0.327	0.08
0.105	0.02	0.408	0.10
0.168	0.04	0.489	0.12
0.242	0.06		

Calculate the angular velocity at the instant $t = 0.06$.

5.3. From the following values of x and y, find dy/dx when $x = 6$:

x	y	x	y
4.5	9.69	6.5	26.37
5.0	12.90	7.0	32.34
5.5	16.71	7.5	39.15
6.0	21.18		

Presented By: http://www.ebooksuit.com

5.4. A rod is rotating in a plane. The following table gives the angle θ (radians) through which the rod has turned for various values of the time t in seconds. Find the angular velocity of the rod when $t = 0.6$.

x	y	x	y
0	0	0.8	2.022
0.2	0.122	1.0	3.200
0.4	0.493	1.2	4.666
0.6	1.123		

5.5. The following table of values of x and y is given:

x	y	x	y
0	6.9897	4	8.4510
1	7.4036	5	8.7506
2	7.7815	6	9.0309
3	8.1291		

Find dy/dx when (i) $x = 1$, (ii) $x = 3$, and (iii) $x = 6$. Also find d^2y/dx^2 when $x = 3$.

5.6. A function $y = f(x)$ is defined as follows:

x	$y = f(x)$	x	$y = f(x)$
1.0	1.0	1.20	1.095
1.05	1.025	1.25	1.118
1.10	1.049	1.30	1.140
1.15	1.072		

Compute the values of dy/dx and d^2y/dx^2 at $x = 1.05$.

5.7. Tabulate the function $f(x) = 5x^4 - 3x^3 + 10x - 6$ at $x_0 = -0.50$, $x_1 = 1.00$ and $x_2 = 2.00$. Compute its first and second derivatives as accurately as possible. Compare your results with the true values.

5.8. The distances travelled by a rocket at different times are as given below:

t	s	t	s
0	0	4	38
1	3	5	50
2	7		
3	15		

Estimate the rocket's velocity and acceleration for each value of t.

Presented By: http://www.ebooksuit.com

5.9. A cubic function $y = f(x)$ satisfies the following data:

x	$f(x)$
0	1
1	4
3	40
4	85

Determine the function $f(x)$ and hence find $f'(2)$ and $f''(2)$.

5.10. The temperature T of a cooling body drops at a rate which is proportional to the difference $T - T_s$ where T_s is the constant temperature of the surrounding medium. A metal ball which is initially at 150°C is dropped into water that is held at constant temperature $T_s = 40°C$. The temperature of the ball at time t is given as follows:

t (in min.)	T (in °C)
0	150
5	74.8
10	68.5
15	50.7
20	44.4

Determine dT/dt at each value of t. If $dT/dt = -k(T - T_s)$, estimate the value of k by linear least squares.

5.11. The function $y = 3xe^{-x}$ is tabulated below:

x	y
3	0.4481
3.2	0.3913
5	0.1010

Develop a subprogram to find the first derivative values of y, test it with the above data and compare your results with the actual values.

5.12. From the following table of values of x and y, find dy/dx at $x = 2$ using the cubic spline method.

x	y
2	11
3	49
4	123

Presented By: http://www.ebooksuit.com

5.13. From the following table of values of x and y, determine the value of dy/dx at each of the points by fitting a cubic spline through them.

x	y
1	1
2	3
4	4
5	2

5.14. If $y = A + Bx + Cx^2$ and y_0, y_1, y_2 are the values of y corresponding to $x = a,\ a+h,\ a+2h$ respectively, prove that

$$\int_a^{a+2h} y\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2).$$

5.15. Evaluate

(a) $\displaystyle\int_0^{\pi} t \sin t\,dt$ (b) $\displaystyle\int_{-2}^{2} \frac{t}{5+2t}\,dt$

using the trapezoidal rule.

5.16. Discuss a method for finding an approximate area under a given curve. A curve is given by the points of the table given below:

x	y	x	y
0	23	2.5	16
0.5	19	3.0	19
1.0	14	3.5	20
1.5	11	4.0	20
2.0	12.5		

Estimate the area bounded by the curve, the x-axis and the extreme ordinates.

5.17. 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 by direct integration.

5.18. Evaluate

$$\int_0^{\pi/2} \sqrt{\sin\theta}\,d\theta,$$

Using Simpson's rule with $h = \pi/12$.

Presented By: http://www.ebooksuit.com

5.19. Find the value of

$$\int_3^7 x^2 \log x \, dx$$

by taking 4 strips.

5.20. The velocities of a car (running on a straight road) at intervals of 2 minutes are given below.

Time (in min.)	*Velocity* (in km/hr)
0	0
2	22
4	30
6	27
8	18
10	7
12	0

Apply Simpson's rule to find the distance covered by the car.

5.21. Compute the values of

$$I = \int_0^1 \frac{dx}{1+x^2}$$

by using the trapezoidal rule with $h = 0.5$, 0.25 and 0.125. Then obtain a better estimate by using Romberg's method. Compare your result with the true value.

5.22. A reservoir discharging through sluices at a depth h below the water surface has a surface area A for various values of h as given below.

h (in ft)	A (in sq. ft)
10	950
11	1070
12	1200
13	1350
14	1530

If t denotes the time in minutes, the rate of fall of the surface is given by

$$\frac{dh}{dt} = -\frac{48}{A}\sqrt{h}.$$

Presented By: http://www.ebooksuit.com

Estimate the time taken for the water level to fall from 14 ft to 10 ft above the sluices.

5.23. Find the approximate value of

$$\int_0^{\pi/2} \sqrt{\cos\theta}\, d\theta$$

by dividing the interval into six parts.

5.24. Evaluate

$$\int_0^1 \cos x\, dx$$

using $h = 0.2$.

5.25. Determine the maximum error in evaluating the integral

$$\int_0^{\pi/2} \cos x\, dx$$

by both the trapezoidal and Simpson's rules using four sub-intervals.

5.26. Estimate the value of

$$\int_0^{\pi} \frac{\sin t}{t}\, dt.$$

5.27. Derive Simpson's (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

$$I = \int_0^1 \frac{1}{1+x}\, dx$$

with $h = 1/6$. Evaluate the integral by using Simpson's (1/3)-rule and compare the results.

5.28. Deduce Weddle's rule

$$\int_{x_0}^{x_6} y\, dx = \frac{3h}{10} (y_0 + 5y_1 + y_2 + 6y_3 + y_4 + 5y_5 + y_6)$$

and use it to obtain an approximate value of π from the formula

$$\frac{\pi}{4} = \int_0^1 \frac{1}{1+x^2}\, dx$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

5.36. Use the three-point Gauss-Legendre formula to evaluate the integral

$$\int_0^{\pi/2} \sin x \, dx.$$

Compare this result with that obtained by Simpson's rule using seven points.

5.37. Use the method of undetermined coefficients to derive the formula

$$\int_0^{2\pi} f(x) \sin x \, dx = f(0) - f(2\pi).$$

5.38. Apply Filon's formula to obtain the value of

$$\int_0^{2\pi} \log(1+u) \sin 10u \, du.$$

5.39. Using Filon's method, evaluate the integral

$$\int_0^{2\pi} e^{-t} \sin 10t \, dt.$$

Compare your result with the analytical solution given by

$$\frac{10}{101}(1 - e^{-10}).$$

5.40. If $I = \int_0^{\pi/4} \cos^2 x \, dx$, compute

$$I\left(0, \frac{\pi}{4}\right), \qquad I\left(0, \frac{\pi}{8}\right), \qquad I\left(\frac{\pi}{8}, \frac{\pi}{4}\right).$$

5.41. Verify the error estimate (5.67) for problem 40.

5.42. Use *adaptive quadrature* to evaluate the integral

$$\int_{1/10}^{2.0} \sin \frac{1}{x} \, dx$$

to within an accuracy $\varepsilon = 0.001$.

5.43. Evaluate the following double integral

$$\int_{-2}^{2} \int_0^4 (x^2 - xy + y^2) \, dx \, dy$$

by using Simpson's (1/3)-rule.

Presented By: http://www.ebooksuit.com

CHAPTER

Matrices and Linear Systems of Equations

6.1 INTRODUCTION

Matrices occur in a variety of problems of interest; for example, in the solution of linear algebraic systems, solution of ordinary and partial differential equations, and eigenvalue problems. The matrix notation is convenient and powerful in expressing basic relationships in fields like elasticity and electrical engineering. In this chapter, we introduce the matrices independently although they can be treated, more conveniently, through the theory of linear transformations. We assume that the reader is familiar with the concept of a determinant and its properties and we describe briefly some simple properties of matrices which will be used in the solution of linear algebraic systems to which some considerable attention will be given in the later sections. The eigenvalue problem will be discussed in Section 6.5, whereas Section 6.6 will be devoted to a discussion of the singular value decomposition of matrices. The theorems will be stated without proof.

6.2 BASIC DEFINITIONS

A *matrix* is an array of mn elements arranged in m rows and n columns. Such a matrix A is usually denoted by

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = [a_{ij}], \tag{6.1}$$

Presented By: http://www.ebooksuit.com

where $a_{11}, a_{12}, \ldots$ are called its *elements* and may be either real or complex. The matrix A is said to be of size $(m \times n)$.

If $m = n$, the matrix is said to be a *square* matrix of *order* n. Thus,

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

is a square matrix of order 3. We may also have *single-row* or *single-column* matrices. These are called *vectors*. Thus, $[a_{11}, a_{12}, a_{13}, \ldots, a_{1n}]$ is a single-row matrix or a *row vector*, and

$$\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}$$

is a single-column matrix or a *column vector*.

The elements a_{ii} in a square matrix form the *principal diagonal* (or *main diagonal*). Their sum $a_{11} + a_{22} + \cdots + a_{nn}$ is called the *trace* of A. If all the elements of a square matrix are zero, then the matrix is said to be a *null* matrix. Thus, if $a_{ij} = 0$ for $i, j = 1, 2, \ldots, n$, then A is a null matrix of order n. On the other hand, if only the elements on the main diagonal are nonzero, then the matrix is said to be a *diagonal* matrix. For example,

$$C = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

is a diagonal matrix.

In particular, the diagonal matrix

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

in which all of the diagonal elements are equal to one, is called a *unit* matrix of order 3. Unit matrices are usually denoted by I.

A square matrix is said to be an *upper-triangular* matrix if $a_{ij} = 0$ for $i > j$, and a *lower-triangular* matrix if $a_{ij} = 0$ for $i < j$. For example,

$$A = \begin{bmatrix} 2 & 3 & 4 \\ 0 & 5 & 6 \\ 0 & 0 & 7 \end{bmatrix}$$

Presented By: http://www.ebooksuit.com

is an upper-triangular matrix, and

$$B = \begin{bmatrix} 2 & 0 & 0 \\ 3 & 4 & 0 \\ 5 & 6 & 7 \end{bmatrix}$$

is a lower-triangular matrix. Matrices of the type

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix}$$

are called *tridiagonal* matrices. A square matrix A in which $a_{ij} = a_{ji}$ is said to be *symmetric*; if $a_{ij} = -a_{ji}$, it is said to be *skew-symmetric*. For example,

$$A = \begin{bmatrix} 2 & 5 & 6 \\ 5 & 8 & 7 \\ 6 & 7 & 4 \end{bmatrix}$$

is a 'symmetric matrix,' and

$$B = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$$

is a 'skew-symmetric matrix.'

Every square matrix A is associated with a number called its *determinant*, which is written as

$$|A| = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}.$$

The minor M_{ij} of the element a_{ij} of $|A|$ is that determinant of order $(n-1)$ obtained by deleting the row and column containing a_{ij}. The *cofactor* A_{ij} of a_{ij} is given by

$$A_{ij} = (-1)^{i+j} M_{ij}. \tag{6.2}$$

If $|A| \neq 0$, then A is said to be a *nonsingular* matrix; otherwise, it is said to be *singular*. Thus,

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 3 & 4 & 0 \\ 5 & 6 & 0 \end{bmatrix}$$

is singular since $|A| = 0$.

Presented By: http://www.ebooksuit.com

6.2.1 Matrix Operations

Equality of two matrices Two matrices are said to be *equal* if they are of the same size and if their corresponding elements are equal.

Addition and subtraction of matrices Two matrices of the same size can be added or subtracted by adding or subtracting their corresponding elements. Thus, if

$$A = \begin{bmatrix} 4 & 3 \\ 5 & 2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 2 & 1 \\ 3 & 1 \end{bmatrix}$$

then

$$A + B = \begin{bmatrix} 6 & 4 \\ 8 & 3 \end{bmatrix} \quad \text{and} \quad A - B = \begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix}.$$

Multiplication of a matrix by a scalar If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

then

$$kA = \begin{bmatrix} ka_{11} & ka_{12} & ka_{13} \\ ka_{21} & ka_{22} & ka_{23} \\ ka_{31} & ka_{32} & ka_{33} \end{bmatrix},$$

where k is scalar.

The following properties of matrices easily follow from the definitions:

(i) $A + (B + C) = (A + B) + C$

(ii) $A + B = B + A$

(iii) $k(A + B) = kA + kB$, k being a scalar.

(iv) $(k_1 + k_2)A = k_1A + k_2A$, k_1 and k_2 being scalars.

Multiplication of a matrix by another matrix Two matrices A and B can be *multiplied* only if the number of columns of A is equal to the number of rows of B. Thus, if A and B are of sizes (2×3) and (3×2), respectively, then their product C, given by

$$C = AB,$$

is defined, and will be of size (2×2). The elements of C are obtained by the rule that the element C_{ij} of C is equal to the sum of the products of the corresponding elements of the ith row of A by those of the jth column of B. In general, if A is of size $(l \times m)$ and B is of size $(m \times n)$, i.e. if

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

We find that

$$A' = \begin{bmatrix} 0 & -2 & -3 \\ 2 & 0 & -5 \\ 3 & 5 & 0 \end{bmatrix} = -\begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 5 \\ -3 & -5 & 0 \end{bmatrix} = -A.$$

It should be noted that this is a 'skew-symmetric matrix.'

Example 6.4 Express the matrix

$$A = \begin{bmatrix} 1 & 7 & 8 \\ 6 & 2 & 9 \\ 5 & 4 & 3 \end{bmatrix}$$

as the 'sum of a symmetric' and a 'skew-symmetric matrix.'

In general, we can write A as

$$A = \frac{A + A'}{2} + \frac{A - A'}{2} = C + D, \text{ say.}$$

Now

$$C' = \frac{A + A'}{2} = \frac{A' + A}{2} = C,$$

which shows that C is a symmetric matrix. Again,

$$D' = \frac{A - A'}{2} = \frac{A' - A}{2} = -\frac{A - A'}{2} = -D,$$

which shows that D is skew-symmetric.

For the example given above, we have

$$A' = \begin{bmatrix} 1 & 6 & 5 \\ 7 & 2 & 4 \\ 8 & 9 & 3 \end{bmatrix}.$$

Hence

$$\frac{A + A'}{2} = \begin{bmatrix} 1 & 6.5 & 6.5 \\ 6.5 & 2 & 6.5 \\ 6.5 & 6.5 & 3 \end{bmatrix}, \text{ which is a symmetric matrix,}$$

and

$$\frac{A - A'}{2} = \begin{bmatrix} 0 & 0.5 & 1.5 \\ -0.5 & 0 & 2.5 \\ -1.5 & -2.5 & 0 \end{bmatrix}, \text{ which is a skew-symmetric matrix.}$$

The reader may verify that their sum is equal to A itself.

Presented By: http://www.ebooksuit.com

6.2.3 The Inverse of a Matrix

Let A be a nonsingular square matrix of order n. Let B be another square matrix of the same order such that

$$BA = I,$$

where I is the unit matrix of order n. Then B is said to be the *inverse* of A which is written as A^{-1} so that

$$AA^{-1} = A^{-1}A = I. \tag{6.4}$$

The following *ten properties* can be shown to hold on the inverse of a square matrix:

(i) A^{-1} exists if and only if $|A| \neq 0$. If $|A| = 0$, A is said to be a *singular* matrix.

(ii) If A^{-1} exists, it is *unique*.

(iii) If A^{-1} exists, $|A^{-1}| = |A|^{-1} = 1/|A|$.

(iv) $(A^{-1})^{-1} = A$.

(v) $(A')^{-1} = (A^{-1})'$.

(vi) $(AB)^{-1} = B^{-1}A^{-1}$.

(vii) If A is a diagonal matrix with diagonal elements a_{ii}, then A^{-1} is also a diagonal matrix with diagonal elements $1/a_{ii}$. For example, if

$$A = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix},$$

then

$$A^{-1} = \begin{bmatrix} 1/a_{11} & 0 & 0 \\ 0 & 1/a_{22} & 0 \\ 0 & 0 & 1/a_{33} \end{bmatrix}.$$

(viii) $I^{-1} = I$.

(ix) The inverse of an upper-triangular matrix is also an upper-triangular matrix. For example, if

$$A = \begin{bmatrix} 2 & 3 & 4 \\ 0 & 5 & 6 \\ 0 & 0 & 7 \end{bmatrix},$$

then

$$A^{-1} = \frac{1}{70}\begin{bmatrix} 35 & -21 & -2 \\ 0 & 14 & -12 \\ 0 & 0 & 10 \end{bmatrix}.$$

Presented By: http://www.ebooksuit.com

Similarly, the inverse of a lower-triangular matrix is also a lower-triangular matrix.

(x) The inverse A^{-1}, when it exists, can be computed, as follows

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{21} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix}, \tag{6.5}$$

where $A_{11}, A_{12}, \ldots$ are the cofactors of $a_{11}, a_{12}, \ldots$ in the determinant of the transpose A' of A. The matrix on the right side of (6.5) is called the *adjoint* of A. This is not an efficient method for the computation of the *inverse*. A better method is the 'Gaussian elimination method' (*see* Section 6.3.2).

For a nonsquare matrix also, it is possible to define an inverse, called the *generalized inverse*. However, in this book, we shall consider inverses for square matrices only (*see* Section 6.6).

Example 6.5 Find the inverse of the matrix

$$A = \begin{bmatrix} 5 & -2 & 4 \\ -2 & 1 & 1 \\ 4 & 1 & 0 \end{bmatrix}.$$

We have $|A| = -37$, and

$$A' = \begin{bmatrix} 5 & -2 & 4 \\ -2 & 1 & 1 \\ 4 & 1 & 0 \end{bmatrix} = A.$$

Hence

$$A^{-1} = -\frac{1}{37}\begin{bmatrix} -1 & 4 & -6 \\ 4 & -16 & -13 \\ -6 & -13 & 1 \end{bmatrix}.$$

The reader should verify that $AA^{-1} = I$.

6.2.4 Rank of a Matrix

Consider a square matrix of order n. Of the n rows and n columns, if there are at least k rows and k columns which must be deleted in order to obtain a nonvanishing determinant, then the order of the highest ordered nonvanishing determinant in A is given by $r = n - k$, and this number is defined as the *rank* of A and is written $r(A)$. Hence, *the rank of a matrix is equal to the order of the highest ordered nonvanishing determinant in A*. It follows,

Presented By: http://www.ebooksuit.com

therefore, that for a nonsingular square matrix of order n, the rank is equal to n. To determine the rank of a matrix, we have to find the order of the highest ordered nonvanishing determinant. This method, although general, would be tedious when applied to matrices of higher order, for which the 'Gaussian elimination method,' to be described in Section 6.3.2, would be particularly suitable.

Example 6.6

(a) $A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. $r(A) = 0$, since all the elements are zero.

(b) $A = \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix}$. $r(A) = 2$, since $|A| \neq 0$.

(c) $A = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}$. $|A| = 0$; hence $r(A) = 1$.

(d) $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & 2 & 1 \end{bmatrix}$. $|A| = 0$; and $\begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix} \neq 0$; hence $r(A) = 2$.

(e) $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 8 \\ 3 & 2 & 1 \end{bmatrix}$. $|A| \neq 0$; hence $r(A) = 3$.

6.2.5 Consistency of a Linear System of Equations

Consider the system of m linear equations in n unknowns:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m. \end{aligned}\right\} \tag{6.6}$$

The matrix

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

is called the *coefficient matrix*, and the matrix defined by

Presented By: http://www.ebooksuit.com

$$(A,\ b) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}b_1 \\ a_{21} & a_{22} & \cdots & a_{2n}b_2 \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn}b_m \end{bmatrix}$$

is called the *augmented matrix*. If $r(A)$ is the rank of A and $r(A,\ b)$ that of $(A,\ b)$ then the following theorem is proved in books on linear algebra (for example, *see* W.L. Ferrar's *Algebra*).

Theorem 6.3 If $r(A) < r(A,\ b)$, then the equations defined by (6.6) are inconsistent and there will be no solution; if $r(A) = r(A,\ b)$, the equations are consistent and there exists at least one solution to the system (6.6).

Example 6.7 Examine for consistency the equations

$$2x - 3y + 5z = 1$$
$$3x + y - z = 2$$
$$x + 4y - 6z = 1.$$

We have

$$A = \begin{bmatrix} 2 & -3 & 5 \\ 3 & 1 & -1 \\ 1 & 4 & -6 \end{bmatrix}.$$

Then

$$|A| = 0, \quad \text{but} \begin{vmatrix} 2 & -3 \\ 3 & 1 \end{vmatrix} \neq 0; \quad \text{hence} \quad r(A) = 2.$$

Further

$$(A, b) = \begin{bmatrix} 2 & -3 & 5 & 1 \\ 3 & 1 & -1 & 2 \\ 1 & 4 & -6 & 1 \end{bmatrix},$$

and it can be seen that all determinants of the third order formed from $(A,\ b)$ are zero and that $r(A, b) = 2$. It follows, therefore, that the equations are consistent.

Example 6.8 Find whether the following system is consistent

$$x - 4y + 5z = 8$$
$$3x + 7y - z = 3$$
$$x + 15y - 11z = -14.$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

$$\|\mathbf{x}\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2} = \left[\sum_{i=1}^{n} |x_i|^2\right]^{1/2} = \|\mathbf{x}\|_e \tag{6.12}$$

$$\|\mathbf{x}\|_\infty = \max_i |x_i|. \tag{6.13}$$

The norm $\|\cdot\|_2$ is called the *Euclidean* norm since it is just the formula for distance in the three-dimensional Euclidean space. The norm $\|\cdot\|_\infty$ is called the *maximum* norm or the *uniform* norm.

It is easy to show that the three norms $\|\mathbf{x}\|_1, \|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$ satisfy the conditions (6.7) to (6.9), given above. Conditions (6.7) and (6.8) are trivially satisfied. Only condition (6.9), the triangle inequality, needs to be shown to be *true*. For the norm $\|\mathbf{x}\|_1$ we observe that

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\| &= \sum_{i=1}^{n} |x_i + y_i| \\ &\le \sum_{i=1}^{n} (|x_i| + |y_i|) \\ &= \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| \\ &= \|\mathbf{x}\|_1 + \|\mathbf{y}\|_1\end{aligned} \tag{6.14}$$

Similarly, for $\|\mathbf{x}\|_\infty$, we have

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_\infty &= \max_i |x_i + y_i| \\ &\le \max_i (|x_i| + |y_i|) \\ &= \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty .\end{aligned} \tag{6.15}$$

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,

$$|A+B| \le |A| + |B| \tag{6.16}$$

$$|AB| \le |A|\,|B| \tag{6.17}$$

$$|\alpha A| = |\alpha|\,|A| \quad (\alpha \text{ a scalar}). \tag{6.18}$$

From (6.17) it follows that

$$|A^p| \le |A|^p, \tag{6.19}$$

Presented By: http://www.ebooksuit.com

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

$$\|A\| \geq 0 \quad \text{and} \quad \|A\| = 0 \quad \text{if and only if } A = 0 \tag{6.20}$$

$$\|\alpha A\| = |\alpha| \, \|A\| \quad (\alpha \text{ a scalar}) \tag{6.21}$$

$$\|A + B\| \leq \|A\| + \|B\| \tag{6.22}$$

$$\|AB\| \leq \|A\| \, \|B\|. \tag{6.23}$$

From (6.23), it easily follows that

$$\|A^p\| \leq \|A\|^p, \tag{6.24}$$

where p is a natural number.

Corresponding to the vector norms given in (6.11)–(6.13), we have the three matrix norms

$$\|A\|_1 = \max_j \sum_i |a_{ij}| \quad \text{(the column norm)} \tag{6.25}$$

$$\|A\|_e = \left[\sum_{i,j} |a_{ij}|^2\right]^{1/2} \quad \text{(the Euclidean norm)} \tag{6.26}$$

$$\|A\|_\infty = \max_i \sum_j |a_{ij}| \quad \text{(the row norm).} \tag{6.27}$$

In addition to the above, we have $\|A\|_2$ defined by

$$\|A\|_2 = (\text{Maximum eigenvalue of } A^{\mathrm{T}}A)^{1/2}. \tag{6.28}$$

The eigenvalues of a matrix will be discussed in Section 6.5.

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 6.9 Given the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

find $\|A\|_1, \|A\|_e$ and $\|A\|_\infty$.

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

This method is obviously unsuitable for solving large systems, since the computation of A^{-1} by cofactors will then become exceedingly difficult, and one should therefore adopt methods which do not require the computation of the cofactors. We will describe such methods in the subsections below and these can be applied to *any* number of equations.

6.3.2 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*.

We consider the system given in (6.29), viz., the system of n linear equations in n unknowns

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= b_n. \end{aligned}\right\} \tag{6.29}$$

There are two steps in the solution of the system (6.29), 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 (6.29), we obtain

$$\left(a_{22} - a_{12}\frac{a_{21}}{a_{11}}\right)x_2 + \left(a_{23} - a_{13}\frac{a_{21}}{a_{11}}\right)x_3 + \cdots + \left(a_{2n} - a_{1n}\frac{a_{21}}{a_{11}}\right)x_n = b_2 - b_1\frac{a_{21}}{a_{11}}, \tag{6.32}$$

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 (6.29). This eliminates the unknown x_1 from the third equation of (6.29) and we obtain

$$a'_{32}x_2 + a'_{33}x_3 + \cdots + a'_{3n}x_n = b'_3. \tag{6.33}$$

In a similar fashion, we can eliminate x_1 from the remaining equations and

Presented By: http://www.ebooksuit.com

after eliminating x_1 from the last equation of (6.29), we obtain the system

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2 \\ a'_{32}x_2 + a'_{33}x_3 + \cdots + a'_{3n}x_n &= b'_3 \\ &\vdots \\ a'_{n2}x_2 + a'_{n3}x_3 + \cdots + a'_{nn}x_n &= b'_n. \end{aligned}\right\} \tag{6.34}$$

We next eliminate x_2 from the last $(n-2)$ equations of (6.34). 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 (6.34) 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 (6.34), 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

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2 \\ a''_{33}x_3 + \cdots + a''_{3n}x_n &= b''_3 \\ &\vdots \\ a''_{n3}x_3 + \cdots + a''_{nn}x_n &= b''_n. \end{aligned}\right\} \tag{6.35}$$

In (6.35), the 'double points' 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:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2 \\ a''_{33}x_3 + \cdots + a''_{3n}x_n &= b''_3 \\ &\vdots \\ a^{(n-1)}_{nn}x_n &= b^{(n-1)}_n, \end{aligned}\right\} \tag{6.36}$$

where $a^{(n-1)}_{nn}$ 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 (6.36). From the last equation of this system, we obtain

Presented By: http://www.ebooksuit.com

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}.$$

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*.

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 6.11 Use Gauss elimination to solve

$$2x + y + z = 10$$

$$3x + 2y + 3z = 18$$

$$x + 4y + 9z = 16.$$

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

$$y + 3z = 6. \qquad \text{(i)}$$

Similarly, we multiply the first equation by $(-1/2)$ and add it to the third to get

$$7y + 17z = 22. \qquad \text{(ii)}$$

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

$$2x + y + z = 10$$

$$y + 3z = 6$$

$$z = 5.$$

Presented By: http://www.ebooksuit.com

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 6.12 Solve the system

$$0.0002x + 0.3003y = 0.1002$$

$$2.0000x + 3.0000y = 2.0000.$$

The exact solution of the system is easily seen to be $x = 1/2$ and $y = 1/3$.

We first solve the system without pivoting. Multiplying the first equation by (–2/0.0002) and adding it to the second, we obtain

$$\left(3.0000 - \frac{0.3003 \times 2}{0.0002}\right)y = 2.0000 - \frac{0.1002 \times 2}{0.0002},$$

which simplifies to

$$1498.5y = 499.$$

Hence the triangular system is

$$0.0002x + 0.3003y = 0.1002$$

$$1498.5y = 499.$$

The solution to the system is given by $y = 0.3330$ and $x = 0.5005$; the errors in the solution being due to the large multiplier.

We next interchange the two rows so that the system is written as

$$2.0000x + 3.0000y = 2.0000$$

$$0.0002x + 0.3003y = 0.1002$$

Multiplying the first equation by (–0.0002/2) and adding it to the second, we obtain

$$\left(0.3003 - \frac{3.0000 \times 0.0002}{2}\right)y = 0.1002 - \frac{2 \times 0.0002}{2},$$

which simplifies to

$$0.3000y = 0.1000.$$

Hence the solution is

$$y = \frac{1}{3} \quad \text{and} \quad x = \frac{1}{2}.$$

6.3.3 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:

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Example 6.14 We shall consider again the system given in Example 6.13. We have here

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 3 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix}.$$

The augmented system is

$$\left[\begin{array}{ccc|ccc} 2 & 1 & 1 & 1 & 0 & 0 \\ 3 & 2 & 3 & 0 & 1 & 0 \\ 1 & 4 & 9 & 0 & 0 & 1 \end{array}\right].$$

After the first stage, this becomes

$$\left[\begin{array}{ccc|ccc} 2 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1/2 & 3/2 & -3/2 & 1 & 0 \\ 0 & 7/2 & 17/2 & -1/2 & 0 & 1 \end{array}\right].$$

Finally, at the end of the second stage, the system becomes:

$$\left[\begin{array}{ccc|ccc} 2 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1/2 & 3/2 & -3/2 & 1 & 0 \\ 0 & 0 & -2 & 10 & -7 & 1 \end{array}\right].$$

This is equivalent to the three systems:

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 1 \\ 0 & 1/2 & 3/2 & -3/2 \\ 0 & 0 & -2 & 10 \end{array}\right],$$

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 0 \\ 0 & 1/2 & 3/2 & 1 \\ 0 & 0 & -2 & -7 \end{array}\right]$$

and

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 0 \\ 0 & 1/2 & 3/2 & 0 \\ 0 & 0 & -2 & 1 \end{array}\right],$$

whose solution by back substitution yields the three columns of the matrix:

Presented By: http://www.ebooksuit.com

$$\begin{bmatrix} -3 & 5/2 & -1/2 \\ 12 & -17/2 & 3/2 \\ -5 & 7/2 & -1/2 \end{bmatrix},$$

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.

6.3.5 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. (6.32), 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

Presented By: http://www.ebooksuit.com

$$\sum_{p=1}^{n-1} (n-p)(n+2-p) = \sum [(n-p)^2 + 2(n-p)]$$

$$= \sum_{p=1}^{n-1} (n^2 + p^2 - 2np + 2n - 2p)$$

$$= n^2(n-1) + \frac{(n-1)(n)(2n-2+1)}{6} - 2n\frac{(n-1)n}{2}$$

$$+ 2n(n-1) - 2\frac{(n-1)n}{2}$$

$$\approx \frac{n^3}{3},$$

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.

6.3.6 *LU* Decomposition

This method is based on the fact that a square matrix A can be factorized into the form LU, where L is unit lower triangular and U is upper triangular, if all the principal minors of A are nonsingular, i.e. if

$$a_{11} \neq 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \neq 0, \text{ etc.}$$

It is a standard result of linear algebra that such a factorization, when it exists, is unique.

We consider, for definiteness, the linear system

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3,$$

which can be written in the form

$$AX = B \cdot \tag{6.38}$$

Presented By: http://www.ebooksuit.com

Let

$$A = LU, \tag{6.39}$$

where

$$L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \tag{6.40}$$

and

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}. \tag{6.41}$$

Hence (6.38) becomes

$$LUX = B. \tag{6.42}$$

If we set

$$UX = Y, \tag{6.43}$$

then (6.42) may be written as

$$LY = B, \tag{6.44}$$

which is equivalent to the system

$$\begin{aligned} y_1 &= b_1 \\ l_{21}y_1 + y_2 &= b_2 \\ l_{31}y_1 + l_{32}y_2 + y_3 &= b_3 \end{aligned}$$

and can therefore be solved for y_1, y_2, y_3 by the forward substitution. When Y is known, the system (6.43) becomes

$$\begin{aligned} u_{11}x_1 + u_{12}x_2 + u_{13}x_3 &= y_1 \\ u_{22}x_2 + u_{23}x_3 &= y_2, \\ u_{33}x_3 &= y_3, \end{aligned}$$

which can be solved for x_1, x_2, x_3 by the backward substitution.

We shall now describe a scheme for computing the matrices L and U, and illustrate the procedure with a matrix of order 3. From the relation (6.39), we obtain

$$\begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Presented By: http://www.ebooksuit.com

Multiplying the matrices on the left and equating the corresponding elements of both sides, we get

$$\left.\begin{aligned} &u_{11} = a_{11}, \quad u_{12} = a_{12}, \quad u_{13} = a_{13} \\ &l_{21}u_{11} = a_{21}, \quad l_{21}u_{12} + u_{22} = a_{22}, \quad l_{21}u_{13} + u_{23} = a_{23} \\ &l_{31}u_{11} = a_{31}, \quad l_{31}u_{12} + l_{32}u_{22} = a_{32}, \quad l_{31}u_{13} + l_{32}u_{23} + u_{33} = a_{33} \end{aligned}\right\} \quad (6.45a)$$

Solving them, we get

$$\left.\begin{aligned} &l_{21} = \frac{a_{21}}{a_{11}}; \qquad l_{31} = \frac{a_{31}}{a_{11}}; \\ &u_{22} = a_{22} - \frac{a_{21}}{a_{11}}a_{12}; \qquad u_{23} = a_{23} - \frac{a_{21}}{a_{11}}a_{13} \\ &l_{32} = \frac{a_{32} - (a_{31}/a_{11})a_{12}}{u_{22}} \end{aligned}\right\} \quad (6.45b)$$

from which u_{33} can be computed.

We thus have a systematic procedure to evaluate the elements of L and U. 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. The procedure can be obviously generalized.

When the factorization is effected, the inverse of A can be computed from the formula

$$A^{-1} = (LU)^{-1} = U^{-1}L^{-1}. \quad (6.46)$$

Example 6.15 Solve the equations

$$2x + 3y + z = 9$$
$$x + 2y + 3z = 6$$
$$3x + y + 2z = 8$$

by LU decomposition.

We have

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}. \quad \text{(i)}$$

Let

$$\begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix}\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}. \quad \text{(ii)}$$

Presented By: http://www.ebooksuit.com

Clearly,

$$u_{11} = 2, \quad u_{12} = 3, \quad u_{13} = 1;$$

also

$$l_{21}u_{11} = 1 \quad \text{so that} \quad l_{21} = \frac{1}{2}$$

and

$$l_{31}u_{11} = 3 \quad \text{so that} \quad l_{31} = \frac{3}{2}.$$

For u_{22} and u_{23}, we have the equations

$$l_{21}u_{12} + u_{22} = 2 \quad \text{and} \quad l_{21}u_{13} + u_{23} = 3,$$

from which we obtain

$$u_{22} = \frac{1}{2} \quad \text{and} \quad u_{23} = \frac{5}{2}.$$

Finally, l_{32} and u_{33} are obtained from

$$l_{31}u_{12} + l_{32}u_{22} = 1 \quad \text{and} \quad l_{31}u_{13} + l_{32}u_{23} + u_{33} = 2,$$

and hence

$$l_{32} = -7 \quad \text{and} \quad u_{33} = 18.$$

It follows that

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 3/2 & -7 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 \\ 0 & 1/2 & 5/2 \\ 0 & 0 & 18 \end{bmatrix} \tag{iii}$$

and hence the given system of equations can be written as

$$\begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 3/2 & -7 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 \\ 0 & 1/2 & 5/2 \\ 0 & 0 & 18 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 9 \\ 6 \\ 8 \end{bmatrix} \tag{iv}$$

or, as

$$\begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 3/2 & -7 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 6 \\ 8 \end{bmatrix}, \tag{v}$$

where

$$\begin{bmatrix} 2 & 3 & 1 \\ 0 & 1/2 & 5/2 \\ 0 & 0 & 18 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}. \tag{vi}$$

Presented By: http://www.ebooksuit.com

Solving the system in (v) by forward substitution, we get

$$y_1 = 9, \quad y_2 = \frac{3}{2}, \quad y_3 = 5.$$

With these values of y_1, y_2, y_3, eq. (vi) can now be solved by the back substitution process and we obtain

$$x = \frac{35}{18}, \qquad y = \frac{29}{18}, \qquad z = \frac{5}{18}.$$

6.3.7 *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

$$A\mathbf{X} = \mathbf{b}, \tag{6.47}$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

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

$$a'_{22}x_2 + a'_{23}x_3 = b'_2. \tag{6.48}$$

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

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 &= b'_2 \\ a'_{32}x_2 + a'_{33}x_3 &= b'_3. \end{aligned}\right\} \tag{6.49}$$

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

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

and

$$\begin{vmatrix} -73 & 78 & 24 \\ 92 & 66 & 25 \\ -80 & 37 & 10.01 \end{vmatrix} = -118.94.$$

Example 6.18 The system

$$\left.\begin{aligned} 10x_1 + 7x_2 + 8x_3 + 7x_4 &= 32 \\ 7x_1 + 5x_2 + 6x_3 + 5x_4 &= 23 \\ 8x_1 + 6x_2 + 10x_3 + 9x_4 &= 33 \\ 7x_1 + 5x_2 + 9x_3 + 10x_4 &= 31 \end{aligned}\right\} \qquad \text{(i)}$$

is ill-conditioned since the system

$$\begin{aligned} 10x_1 + 7x_2 + 8x_3 + 7x_4 &= 32.1 \\ 7x_1 + 5x_2 + 6x_3 + 5x_4 &= 22.9 \\ 8x_1 + 6x_2 + 10x_3 + 9x_4 &= 32.9 \\ 7x_1 + 5x_2 + 9x_3 + 10x_4 &= 31.1 \end{aligned}$$

has the solution [6, −7.2, 2.9, −0.1] whereas the system (i) has the solution [1, 1, 1, 1].

Ill-conditioning can usually be expected when $|A|$, in the system $A\mathbf{x} = \mathbf{b}$, is small. The quantity $\nu(A)$ defined by

$$\nu(A) = \| A \| \, \| A^{-1} \|, \qquad (6.65)$$

where $\| A \|$ is any matrix norm, gives a *measure of the condition of the matrix*. It is called the *condition number* of the matrix. Large condition numbers, as a rule, indicate ill-conditioning of a matrix. We give below examples of ill-conditioned and well-conditioned matrices.

Example 6.19 Let

$$A = \begin{bmatrix} 2 & 1 \\ 2 & 1.01 \end{bmatrix}$$

Taking the Euclidean norms, we obtain

$$\| A \|_e = 3.165 \quad \text{and} \quad \| A^{-1} \|_e = 158.273.$$

Hence $\nu(A) = 500.974$. It follows that A is 'ill-conditioned.'

Example 6.20 Let

$$B = \begin{bmatrix} -0.6 & 0.6 \\ 0.4 & 0.2 \end{bmatrix}.$$

Presented By: http://www.ebooksuit.com

We have

$$\| B \|_e = 0.959 \quad \text{and} \quad \| B^{-1} \|_e = 2.664.$$

Hence

$$\nu(B) = \| B \|_e \| B^{-1} \|_e = 2.555.$$

It follows that B is a well-conditioned matrix.

Another indicator of ill-conditioning is the following. If $A = [a_{ij}]$ and

$$s_i = (a_{i1}^2 + a_{i2}^2 + \cdots + a_{in}^2)^{1/2} \tag{6.66}$$

then the quantity

$$k = \frac{|A|}{s_1 \, s_2 \ldots s_n} \tag{6.67}$$

indicates, in some sense, the smallness of the determinant of A. If k is very small compared to 1, then the matrix A is 'ill-conditioned.' Otherwise, it is well-conditioned. For the matrix A in Example 6.19 above, we obtain $|A| = 0.02$, $s_1 = \sqrt{5} = 2.2360679$ $s_2 = 2.240$ and $k = 3.993 \times 10^{-3}$. Similarly, for the matrix B in Example 6.20, we obtain $|B| = -0.36$, $s_1 = \sqrt{0.72} = 0.848$, $s_2 = 0.447$ and $k = 0.950$.

6.3.11 Method for Ill-conditioned Matrices

One method of improving the accuracy for an ill-conditioned system is by means of working all the calculations to more number of significant digits. But multilength arithmetic is time-consuming and therefore uneconomical. One possible alternative is to improve upon the accuracy of the approximate solution by an iterative procedure. This is described below.

Let the system be

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3. \end{aligned}\right\} \tag{6.68}$$

Let $\bar{x}_1, \bar{x}_2, \bar{x}_3$ be an approximate solution. Substituting these values on the left-hand side, we get new values of b_1, b_2, b_3, say, $\bar{b}_1, \bar{b}_2, \bar{b}_3$. Thus the new system is

$$\left.\begin{aligned} a_{11}\bar{x}_1 + a_{12}\bar{x}_2 + a_{13}\bar{x}_3 &= \bar{b}_1 \\ a_{21}\bar{x}_1 + a_{22}\bar{x}_2 + a_{23}\bar{x}_3 &= \bar{b}_2 \\ a_{31}\bar{x}_1 + a_{32}\bar{x}_2 + a_{33}\bar{x}_3 &= \bar{b}_3. \end{aligned}\right\} \tag{6.69}$$

Presented By: http://www.ebooksuit.com

Subtracting each equation in (6.69) from the corresponding equation in (6.68), we get

$$\left.\begin{aligned} a_{11}e_1 + a_{12}e_2 + a_{13}e_3 &= d_1 \\ a_{21}e_1 + a_{22}e_2 + a_{23}e_3 &= d_2 \\ a_{31}e_1 + a_{32}e_2 + a_{33}e_3 &= d_3, \end{aligned}\right\} \tag{6.70}$$

where $e_i = x_i - \tilde{x}_i$ and $d_i = b_i - \tilde{b}_i$. We now solve the system (6.70) for e_1, e_2 and e_3. Since $e_i = x_i - \tilde{x}_i$, we obtain

$$x_i = e_i + \tilde{x}_i, \tag{6.71}$$

which is a better approximation for x_i. The procedure can be repeated to improve upon the accuracy.

6.4 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

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n &= b_3 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= b_n \end{aligned}\right\} \tag{6.72}$$

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 (6.72) as

Presented By: http://www.ebooksuit.com

$$\left.\begin{aligned}
x_1 &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2 - \frac{a_{13}}{a_{11}}x_3 - \cdots - \frac{a_{1n}}{a_{11}}x_n \\
x_2 &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1 - \frac{a_{23}}{a_{22}}x_2 - \cdots - \frac{a_{2n}}{a_{22}}x_n \\
x_3 &= \frac{b_3}{a_{33}} - \frac{a_{31}}{a_{33}}x_1 - \frac{a_{32}}{a_{33}}x_2 - \cdots - \frac{a_{3n}}{a_{33}}x_n \\
&\vdots \\
x_n &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1 - \frac{a_{n2}}{a_{nn}}x_2 - \cdots - \frac{a_{n,n-1}}{a_{nn}}x_{n-1}.
\end{aligned}\right\} \tag{6.73}$$

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 (6.73), we find a system of second approximations

$$\left.\begin{aligned}
x_1^{(2)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2^{(1)} - \cdots - \frac{a_{1n}}{a_{11}}x_n^{(1)}, \\
x_2^{(2)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1^{(1)} - \cdots - \frac{a_{2n}}{a_{22}}x_n^{(1)}, \\
x_3^{(2)} &= \frac{b_3}{a_{33}} - \frac{a_{31}}{a_{33}}x_1^{(1)} - \cdots - \frac{a_{3n}}{a_{33}}x_n^{(1)}, \\
&\vdots \\
x_n^{(2)} &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1^{(1)} - \cdots - \frac{a_{n,n-1}}{a_{nn}}x_{n-1}^{(1)}.
\end{aligned}\right\} \tag{6.74}$$

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

$$\left.\begin{aligned}
x_1^{(n+1)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2^{(n)} - \cdots - \frac{a_{1n}}{a_{11}}x_n^{(n)}, \\
x_2^{(n+1)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1^{(n)} - \cdots - \frac{a_{2n}}{a_{22}}x_n^{(n)}, \\
&\vdots \\
x_n^{(n+1)} &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1^{(n)} - \cdots - \frac{a_{n,n-1}}{a_{nn}}x_{n-1}^{(n)}.
\end{aligned}\right\} \tag{6.75}$$

If we write (6.73) in the matrix form

$$X = BX + C \tag{6.76}$$

Presented By: http://www.ebooksuit.com

then the iteration formula (6.75) may be written as

$$X^{(n+1)} = BX^{(n)} + C. \tag{6.77}$$

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

$$\| B \| < 1. \tag{6.78}$$

A simple modification in this method sometimes yields faster convergence and is described below:

In the first equation of (6.73), 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 (6.73) 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

$$\sum_{j=1,\, j \neq i}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \leq 1, \qquad (i = 1, 2, \ldots, n), \tag{6.79}$$

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 example:

Example 6.21 We consider the equations:

$$10x_1 - 2x_2 - x_3 - x_4 = 3$$
$$-2x_1 + 10x_2 - x_3 - x_4 = 15$$
$$-x_1 - x_2 + 10x_3 - 2x_4 = 27$$
$$-x_1 - x_2 - 2x_3 + 10x_4 = -9.$$

To solve these equations by the iterative methods, we re-write them as follows:

$$x_1 = 0.3 + 0.2x_2 + 0.1x_3 + 0.1x_4$$
$$x_2 = 1.5 + 0.2x_1 + 0.1x_3 + 0.1x_4$$

Presented By: http://www.ebooksuit.com

$$x_3 = 2.7 + 0.1x_1 + 0.1x_2 + 0.2x_4$$

$$x_4 = -0.9 + 0.1x_1 + 0.1x_2 + 0.2x_3.$$

It can be verified that these equations satisfy the condition given in (6.79). The results are given in Tables 6.1 and 6.2:

Table 6.1 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 6.2 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 6.1 and 6.2, it is clear that twelve iterations are required by Jacobi's method to achieve the same accuracy as seven Gauss–Seidel iterations.

6.5 THE 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

$$AX = \lambda X. \tag{6.80}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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

$$X_2 = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}.$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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

$$\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}.$$

6.5.1 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

$$A_1 = \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{12} & a_{22} & a_{23} \\ 0 & a_{23} & a_{33} \end{bmatrix}. \tag{6.84}$$

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

$$\phi_3(\lambda) = 0. \tag{6.85}$$

Expanding the determinant in terms of the third row, we obtain

$$\begin{aligned} \phi_3(\lambda) &= (a_{33}-\lambda)\begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{12} & a_{22}-\lambda \end{vmatrix} - a_{23}\begin{vmatrix} a_{11}-\lambda & 0 \\ a_{12} & a_{23} \end{vmatrix} \\ &= (a_{33}-\lambda)\phi_2(\lambda) - a_{23}(a_{11}-\lambda)a_{23} \\ &= (a_{33}-\lambda)\phi_2(\lambda) - a_{23}^2\phi_1(\lambda) \\ &= 0. \end{aligned} \tag{6.86}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}. \tag{6.93}$$

Householder's method consists in finding a real symmetric orthogonal matrix P_1 such that

$$P_1 A P_1 = \begin{bmatrix} a_{11} & a'_{12} & 0 \\ a'_{12} & a'_{22} & a'_{23} \\ 0 & a'_{23} & a'_{33} \end{bmatrix}, \tag{6.94}$$

where the primes denote that the elements have changed. Householder suggested that P_1 should be of the form

$$P_1 = I - 2VV^{\mathrm{T}}, \tag{6.95}$$

where

$$V = [0 \quad v_2 \quad v_3]^{\mathrm{T}} \quad \text{and} \quad V^{\mathrm{T}}V = 1. \tag{6.96}$$

It is easily verified that

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-2v_2^2 & -2v_2v_3 \\ 0 & -2v_2v_3 & 1-2v_3^2 \end{bmatrix}, \tag{6.97}$$

where

$$v_2^2 + v_3^2 = 1. \tag{6.98}$$

Further,

$$\begin{aligned} P_1P_1^{\mathrm{T}} &= (I - 2VV^{\mathrm{T}})(I - 2VV^{\mathrm{T}})^{\mathrm{T}} \\ &= (I - 2VV^{\mathrm{T}})(I - 2VV^{\mathrm{T}}) \\ &= I - 4VV^{\mathrm{T}} + 4VV^{\mathrm{T}}VV^{\mathrm{T}} \\ &= I. \end{aligned} \tag{6.99}$$

We thus see that P_1 is both symmetric and orthogonal. By direct multiplication, we find that

$$AP_1 = \begin{bmatrix} a_{11} & a_{12}(1-2v_2^2) - 2a_{13}v_2v_3 & -2a_{12}v_2v_3 + a_{13}(1-2v_3^2) \\ a_{12} & a_{22}(1-2v_2^2) - 2a_{23}v_2v_3 & -2a_{22}v_2v_3 + a_{23}(1-2v_3^2) \\ a_{13} & a_{23}(1-2v_2^2) - 2a_{33}v_2v_3 & -2a_{23}v_2v_3 + a_{33}(1-2v_3^2) \end{bmatrix}$$

Presented By: http://www.ebooksuit.com

and therefore

$$P_1AP_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-2v_2^2 & -2v_2v_3 \\ 0 & -2v_2v_3 & 1-2v_3^2 \end{bmatrix}$$

$$\times \begin{bmatrix} a_{11} & a_{12}(1-2v_2^2)-2a_{13}v_2v_3 & -2a_{12}v_2v_3+a_{13}(1-2v_3^2) \\ a_{12} & a_{22}(1-2v_2^2)-2a_{23}v_2v_3 & -2a_{22}v_2v_3+a_{23}(1-2v_3^2) \\ a_{13} & a_{23}(1-2v_2^2)-2a_{33}v_2v_3 & -2a_{23}v_2v_3+a_{33}(1-2v_3^2) \end{bmatrix}. \quad (6.100)$$

Comparing (6.100) with (6.94), we find that

$$\begin{aligned} 0 &= -2a_{12}v_2v_3 + a_{13}(1-2v_3^2) \\ &= a_{13} - 2v_3(a_{12}v_2 + a_{13}v_3) \\ &= a_{13} - 2rv_3, \end{aligned} \quad (6.101)$$

where

$$r = a_{12}v_2 + a_{13}v_3. \quad (6.102)$$

Also,

$$\begin{aligned} a'_{12} &= a_{12}(1-2v_2^2) - 2a_{13}v_2v_3 \\ &= a_{12} - 2v_2(a_{12}v_2 + a_{13}v_3) \\ &= a_{12} - 2rv_2 \end{aligned} \quad (6.103)$$

using (6.102). From (6.101) and (6.103), it follows that

$$\begin{aligned} (a'_{12})^2 &= (a_{12} - 2rv_2)^2 + (a_{13} - 2rv_3)^2 \\ &= a_{12}^2 + a_{13}^2 + 4r^2(v_2^2 + v_3^2) - 4r(a_{12}v_2 + a_{13}v_3) \\ &= a_{12}^2 + a_{13}^2, \end{aligned} \quad (6.104)$$

using (6.98) and (6.102). Hence,

$$a'_{12} = \pm\sqrt{a_{12}^2 + a_{13}^2} = a_{12} - 2rv_2 = \pm S, \text{ say}. \quad (6.105)$$

We therefore have two equations, viz.,

$$a_{13} - 2rv_3 = 0 \quad (6.106)$$

and

$$a_{12} - 2rv_2 = \pm S, \quad (6.107)$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

$$R \equiv P_{n-1} P_{n-2} \dots P_2 P_1 A_1. \tag{6.114}$$

To complete the construction, we define the orthogonal matrix

$$Q^{\mathrm{T}} = P_{n-1} P_{n-2} \dots P_2 P_1 \tag{6.115}$$

so that $A_1 = QR$ as required. The sequence $\{A_k\}$ converges either to a triangular matrix with the eigenvalues of A on its diagonal, or to a near-triangular matrix from which also the eigenvalues can be easily calculated. Since the sequence converges slowly, a technique, known as *shifting*, is used to accelerate the convergence. This technique will not be discussed here, and the interested reader may refer to advanced texts for this.

6.6 SINGULAR VALUE DECOMPOSITION

We have so far considered square matrices only and in Section 6.3.6 we obtained the LU decomposition of a square matrix. A somewhat similar decomposition is also possible of a rectangular matrix and this is called the *singular value decomposition* (SVD). The SVD is of great importance in matrix theory since it is useful in finding the generalized inverse of a singular matrix and has several image processing applications.

Let A be an $(m \times n)$ real matrix with $m \geq n$. Then the matrices $A^{\mathrm{T}}A$ and AA^{T} are non-negative, symmetric and have identical eigenvalues, say λ_n. We can then obtain the n orthonormalized eigenvectors, say $\mathbf{x}_n$, of $A^{\mathrm{T}}A$ such that

$$A^{\mathrm{T}} A \mathbf{x}_n = \lambda_n \mathbf{x}_n. \tag{6.116}$$

If we assume $\mathbf{y}_n$ to be the n orthonormalized eigenvectors of AA^{T}, then we have

$$AA^{\mathrm{T}} \mathbf{y}_n = \lambda_n \mathbf{y}_n. \tag{6.117}$$

Then A can be decomposed into the form

$$A = UDV^{\mathrm{T}}, \tag{6.118}$$

where

$$U^{\mathrm{T}}U = V^{\mathrm{T}}V = VV^{\mathrm{T}} = I_n \tag{6.119}$$

and

$$D = \mathrm{diag}\ (\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_n}). \tag{6.120}$$

The decomposition defined by (6.118) is called the singular value decomposition of A. The matrix $V(n \times n)$ consists of $\mathbf{x}_n$ which are the n orthonormalized eigenvectors of $A^{\mathrm{T}}A$. It follows then that the eigenvectors of AA^{T}, i.e. $\mathbf{y}_n$ are given by

$$\mathbf{y}_n = \frac{1}{\sqrt{\lambda_n}} A\mathbf{x}_n. \tag{6.121}$$

Presented By: http://www.ebooksuit.com

The matrix D is a diagonal matrix given by

$$D = \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\lambda_n} \end{bmatrix} \tag{6.122}$$

where $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_n}$ are called the *singular values* of A and are such that

$$\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \dots \geq \sqrt{\lambda_n} \geq 0. \tag{6.123}$$

If the rank of A is $r < n$, then

$$\sqrt{\lambda_{r+1}} = \sqrt{\lambda_{r+2}} = \dots = \sqrt{\lambda_n} = 0. \tag{6.124}$$

It can be shown that the singular value decomposition of A is *unique* if the λ_i are distinct and (6.123) is satisfied. In case, A is a square matrix of order n, then the matrices U, D and V are also square matrices of the same size and the inverse of A can be trivially computed, since

$$A^{-1} = VD^{-1}U^{\mathrm{T}} \tag{6.125}$$

and

$$D^{-1} = \operatorname{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \dots, \frac{1}{\sqrt{\lambda_n}}\right). \tag{6.126}$$

If any of the λ_i's are zero, then the matrix A is singular. Similarly, if the λ_i are very small, then the matrix A is very nearly singular. Thus the singular-value decomposition of a matrix gives a clear indication whether the matrix is singular or very nearly singular. The following example demonstrates the method.

Example 6.25 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

Presented By: http://www.ebooksuit.com

$$\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 16.64\begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

which gives the system:

$$3x_1 + 6x_2 = 16.64x_1$$
$$6x_1 + 14x_2 = 16.64x_2.$$

The solution is given by

$$\mathbf{x}_1 = \begin{bmatrix} 0.4033 \\ 0.9166 \end{bmatrix}.$$

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

$$\mathbf{x}_2 = \begin{bmatrix} 0.9166 \\ -0.4033 \end{bmatrix}.$$

We also have $\sqrt{\lambda_1} = 4.080$ and $\sqrt{\lambda_2} = 0.60$.

The eigenvectors of AA^T can then be obtained from (6.121): 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

6.1. Obtain AB, when

$$A = \begin{bmatrix} 2 & 5 & -2 \\ -1 & 0 & 0 \\ 2 & 3 & 4 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 3 & 5 \\ 1 & 0 \\ 2 & 0 \end{bmatrix}.$$

Presented By: http://www.ebooksuit.com

6.2. Compute A^2, A^3 and A^4, when

$$A = \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix}.$$

6.3. Form AB and BA:

$$A = \begin{bmatrix} 2 & 3 & 2 \\ 1 & 0 & 0 \\ 2 & 0 & 2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -1 & 2 & 1 \\ 2 & 3 & 4 \\ 1 & -2 & 3 \end{bmatrix}.$$

6.4. Compute A^{-1} and check your result by direct multiplication with A, where

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

6.5. Find the inverse in the following cases:

$$\text{(a)} \begin{bmatrix} 2 & 4 & 3 \\ 0 & 1 & 1 \\ 2 & 2 & -1 \end{bmatrix} \qquad \text{(b)} \begin{bmatrix} 1 & 6 & 4 \\ 0 & 2 & 3 \\ 0 & 1 & 2 \end{bmatrix}.$$

6.6. Compute the inverse of the matrix

$$\begin{bmatrix} 3 & 2 & 4 \\ 2 & 1 & 1 \\ 1 & 3 & 5 \end{bmatrix}$$

and use the result to solve the system of equations:

$$3x + 2y + 4z = 7$$
$$2x + y + z = 7$$
$$x + 3y + 5z = 2.$$

6.7. Find whether the following systems are consistent, or not:

(a)
$$2x - y + z = 4$$
$$3x - y + z = 6$$
$$4x - y + 2z = 7$$
$$-x + y - z = 9$$

(b)
$$5x - 3y + 7z = 4$$
$$3x + 26y + 2z = 9$$
$$7x + 2y + 10z = 5.$$

6.8. Show that the equations

$$x_1 + 2x_2 - x_3 = 3$$
$$x_1 - x_2 + 2x_3 = 1$$
$$2x_1 - 2x_2 + 3x_3 = 2$$
$$x_1 - x_2 + x_3 = -1$$

are consistent and solve them.

Presented By: http://www.ebooksuit.com

6.9. Use Gaussian 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 substituting into the original equations.

6.10. Use Gauss–Jordan method to solve the system in Problem 9.

6.11. Find the inverse of the matrix:

$$A = \begin{bmatrix} 1 & -1 & 1 \\ 1 & -2 & 4 \\ 1 & 2 & 2 \end{bmatrix}$$

using Gauss elimination.

6.12. Solve the system

$$5x - 2y + z = 4$$
$$7x + y - 5z = 8$$
$$3x + 7y + 4z = 10$$

by (a) Gauss elimination (b) Gauss–Jordan method. In both the cases, check your answers by substituting them into the original equations.

6.13. Decompose the matrix

$$A = \begin{bmatrix} 5 & -2 & 1 \\ 7 & 1 & -5 \\ 3 & 7 & 4 \end{bmatrix}$$

into the form LU and hence solve the system $A\mathbf{x} = \mathbf{b}$ where $\mathbf{b} = [4 \quad 8 \quad 10]^T$. Determine also L^{-1} and U^{-1} and hence find A^{-1}.

6.14. Develop a subprogram in a language of your choice to solve a system of equations using Gauss elimination with partial pivoting. Test your subprogram using the system given in Problem 12.

6.15. Develop a subprogram in a language of your Choice to decompose a matrix A into the form LU using partial pivoting. Include the possibility of computing the inverse also. Test your program with the matrix given in Problem 13.

6.16. Solve the system of equations:

$$2x - y = 0$$
$$-x + 2y - z = 0$$
$$-y + 2z - u = 0$$
$$-z + 2u = 1$$

by the procedure described in Section 6.3.8.

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

(a) $\begin{bmatrix} 1 & 6 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & \sqrt{2} \\ \sqrt{2} & 1 \end{bmatrix}$ (c) $\begin{bmatrix} -9 & 2 & 6 \\ 5 & 0 & -3 \\ -16 & 4 & 11 \end{bmatrix}$.

6.23. Determine the largest eigenvalue and the corresponding eigenvector of the matrices

(a) $\begin{bmatrix} 10 & -2 & 1 \\ -2 & 10 & -2 \\ 1 & -2 & 10 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 3 & -1 \\ 3 & 2 & 4 \\ -1 & 4 & 10 \end{bmatrix}$.

6.24. Use the iterative method to find the largest eigenvalue and the corresponding eigenvector of the matrix

$$A = \begin{bmatrix} 5 & 2 & 1 & -2 \\ 2 & 6 & 3 & -4 \\ 1 & 3 & 19 & 2 \\ -2 & -4 & 2 & 1 \end{bmatrix}.$$

6.25. Reduce the following matrices to the tridiagonal form by Householder's method

(a) $\begin{bmatrix} 1 & 3 & 4 \\ 3 & 1 & 2 \\ 4 & 2 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$.

6.26. Use the QR algorithm to find the eigenvalues of the matrix

$$A = \begin{bmatrix} 0 & 1 & 4 \\ 1 & 3 & 1 \\ 2 & 1 & 0 \end{bmatrix}.$$

6.27. Compute the SVD of the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 3 \end{bmatrix}.$$

6.28. Find the eigenvalues and the corresponding orthonormalized eigenvectors of the matrix $A^{\mathrm{T}}A$ if

$$A = \begin{bmatrix} 20 & 4 \\ 10 & 14 \\ 5 & 5 \end{bmatrix}.$$

Hence determine the SVD of A.

Presented By: http://www.ebooksuit.com

CHAPTER 7

Numerical Solution of Ordinary Differential Equations

7.1 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

$$\frac{dy}{dx} = f(x, y) \tag{7.1a}$$

with the initial condition,

$$y(x_0) = y_0 \tag{7.1b}$$

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

Presented By: http://www.ebooksuit.com

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. (7.1) 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 (7.1), and at the end of the chapter we will outline methods for solving boundary value problems for second-order differential equations.

7.2 SOLUTION BY TAYLOR'S SERIES

We consider the differential equation

$$y' = f(x, y) \tag{7.1a}$$

with the initial condition

$$y(x_0) = y_0. \tag{7.1b}$$

If $y(x)$ is the exact solution of (7.1), then the Taylor's series for $y(x)$ around $x = x_0$ is given by

$$y(x) = y_0 + (x - x_0)y_0' + \frac{(x - x_0)^2}{2!} y_0'' + \cdots \tag{7.2}$$

If the values of $y_0', y_0'', \ldots$ are known, then (7.2) 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

$$y''' = f'' = f_{xx} + f_{xy}f + f(f_{yx} + f_{yy}f) + f_y(f_x + f_y f)$$

$$= f_{xx} + 2ff_{xy} + f^2 f_{yy} + f_x f_y + ff_y^2$$

and other higher derivatives of y. The method can easily be extended to simultaneous and higher-order differential equations.

Presented By: http://www.ebooksuit.com

Example 7.1 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 7.2 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

$$y'''(x) = xy''(x) + y'(x) + y'(x) = xy''(x) + 2y'(x), \tag{ii}$$

$$y^{\text{iv}}(x) = xy'''(x) + y''(x) + 2y''(x) = xy'''(x) + 3y''(x), \tag{iii}$$

$$y^{\text{v}}(x) = xy^{\text{iv}}(x) + y'''(x) + 3y'''(x) = xy^{\text{iv}}(x) + 4y'''(x), \tag{iv}$$

$$y^{\text{vi}}(x) = xy^{\text{v}}(x) + y^{\text{iv}}(x) + 4y^{\text{iv}}(x) = xy^{\text{v}}(x) + 5y^{\text{iv}}(x), \tag{v}$$

Presented By: http://www.ebooksuit.com

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

$$y(0.1)=1+\frac{(0.1)^2}{2}+\frac{(0.1)^4}{24}(3)+\frac{(0.1)^6}{720}(5)+\cdots$$

$$=1+0.005+0.0000125, \text{ neglecting the last term}$$

$$=1.0050125, \text{ correct to seven decimal places.}$$

7.3 PICARD'S METHOD OF SUCCESSIVE APPROXIMATIONS

Integrating the differential equation in (7.1), we obtain

$$y=y_0+\int_{x_0}^{x} f(x,y)\,dx. \tag{7.3}$$

Equation (7.3), 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 (7.3), 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 (7.3) 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

$$y^{(n)}=y_0+\int_{x_0}^{x} f(x,y^{(n-1)})\,dx \qquad \text{with } y^{(0)}=y_0 \tag{7.4}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

If we assume the continuity of $\partial f/\partial y$, then the expression in the brackets can be simplified by using the mean value theorem [see Theorem 1.5, Chapter 1]

$$f(x_n, y_n) - f(x_n, y(x_n)) = [y_n - y(x_n)]\, f_y(x_n, \xi_n) = e_n f_y(x_n, \xi_n),$$

where ξ_n lies between $y(x_n)$ and y_n. We thus have obtained the recurrence formula

$$\begin{aligned} e_{n+1} &= e_n + h e_n f_y(x_n, \xi_n) + R_{n+1} + L_{n+1} \\ &= e_n[1 + h f_y(x_n, \xi_n)] + R_{n+1} + L_{n+1}. \end{aligned} \tag{7.12}$$

The first term on the right side of (7.12) is the *propagated error*, i.e. the error in y_{n+1} resulting from the error in the previous approximation y_n.

Expressions for e_{n+1} can be obtained by successive substitutions into (7.12). Thus we obtain

$$e_0 = 0,$$

$$e_1 = R_1 + L_1,$$

$$e_2 = [1 + h f_y(x_1, \xi_1)](R_1 + L_1) + R_2 + L_2$$

$$e_3 = [1 + h f_y(x_2, \xi_2)]\{[1 + h f_y(x_1, \xi_1)](R_1 + L_1) + R_2 + L_2\} + R_3 + L_3$$

and so on.

An upper bound for the total solution error can be obtained analytically, but this will not be attempted here. The interested reader is referred to the book by Isaacson and Keller for more details. The step-by-step calculation of the solution error using relation (7.12) is demonstrated in the following illustrative example.

Example 7.6 We consider, again, the differential equation $y' = -y$ with the condition $y(0) = 1$, which we have solved by Euler's method in Example 7.5.

Choosing $h = 0.01$, we have

$$1 + h f_y(x_n, \xi_n) = 1 + 0.01(-1) = 0.99.$$

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

$$|L_1| \le 0.00005 = 5 \times 10^{-5},$$

$$|L_2| \le (0.00005)(0.99) < 5 \times 10^{-5},$$

$$|L_3| \le (0.00005)(0.9801) < 5 \times 10^{-5},$$

Presented By: http://www.ebooksuit.com

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 (7.12) the estimates

$$e_0 = 0,$$

$$|e_1| \le 5 \times 10^{-5}$$

$$|e_2| \le 0.99e_1 + 5 \times 10^{-5} < 10^{-4}$$

$$|e_3| \le 0.99e_2 + 5 \times 10^{-5} < 10^{-4} + 5 \times 10^{-5}$$

$$|e_4| \le 0.99e_3 + 5 \times 10^{-5} < 10^{-4} + 10^{-4} = 2 \times 10^{-4} = 0.0002$$

$$\vdots$$

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 7.5.

7.4.2 Modified Euler's Method

Instead of approximating $f(x, y)$ by $f(x_0, y_0)$ in (7.6), we now approximate the integral in (7.6) by means of trapezoidal rule to obtain

$$y_1 = y_0 + \frac{h}{2}[f(x_0, y_0) + f(x_1, y_1)] \tag{7.13}$$

We thus obtain the iteration formula

$$y_1^{(n+1)} = y_0 + \frac{h}{2}[f(x_0, y_0) + f(x_1, y_1^{(n)})], \quad n = 0, 1, 2, \ldots \tag{7.14}$$

where $y_1^{(n)}$ is the nth approximation to y_1. The iteration formula (7.14) can be started by choosing $y_1^{(0)}$ from Euler's formula:

$$y_1^{(0)} = y_0 + hf(x_0, y_0).$$

Example 7.7 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 (7.14) and we obtain

$$y_1^{(1)} = 1.0513.$$

Presented By: http://www.ebooksuit.com

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.

7.5 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. (7.13), 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

$$y_1 = y_0 + \frac{1}{2}(k_1 + k_2), \tag{7.15}$$

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

$$k_2 = hf\,(x_0 + h, y_0 + hf_0) = h\left[f_0 + h\frac{\partial f}{\partial x_0} + hf_0\frac{\partial f}{\partial y_0} + O(h^2)\right].$$

Since

$$\frac{df(x, y)}{dx} = \frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y},$$

we obtain

$$k_2 = h[f_0 + hf_0' + O(h^2)] = hf_0 + h^2 f_0' + O(h^3),$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

where the parameters have to be determined by expanding both sides of the first equation of (7.18a) 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

$$\left.\begin{array}{lll} \alpha_0 = \beta_0 = \dfrac{1}{2}, & \alpha_1 = \dfrac{1}{2}, & \alpha_2 = 1, \\ \beta_1 = \dfrac{1}{2}(\sqrt{2}-1), & \beta_2 = 0 & \\ \nu_1 = 1 - \dfrac{1}{\sqrt{2}}, & \nu_2 = -\dfrac{1}{\sqrt{2}}, & \delta_1 = 1 + \dfrac{1}{\sqrt{2}}, \\ W_1 = W_4 = \dfrac{1}{6}, & W_2 = \dfrac{1}{3}\left(1 - \dfrac{1}{\sqrt{2}}\right), & W_3 = \dfrac{1}{3}\left(1 + \dfrac{1}{\sqrt{2}}\right), \end{array}\right\} \tag{7.19}$$

we obtain the method of Gill, whereas the choice

$$\left.\begin{array}{ll} \alpha_0 = \alpha_1 = \dfrac{1}{2}, & \beta_0 = \nu_1 = \dfrac{1}{2} \\ \beta_1 = \beta_2 = \nu_2 = 0, & \alpha_2 = \delta_1 = 1 \\ W_1 = W_4 = \dfrac{1}{6}, & W_2 = W_3 = \dfrac{2}{6} \end{array}\right\} \tag{7.20}$$

leads to the fourth-order Runge–Kutta formula, the most commonly used one in practice:

$$y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{7.21a}$$

where

$$\left.\begin{array}{l} k_1 = hf(x_0, y_0) \\ k_2 = hf\left(x_0 + \dfrac{1}{2}h, y_0 + \dfrac{1}{2}k_1\right) \\ k_3 = hf\left(x_0 + \dfrac{1}{2}h, y_0 + \dfrac{1}{2}k_2\right) \\ k_4 = hf(x_0 + h, y_0 + k_3) \end{array}\right\} \tag{7.21b}$$

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.

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

We take $h = 0.2$. With $x_0 = y_0 = 0$, we obtain from (7.21a) and (7.21b),

$$k_1 = 0.2,$$

$$k_2 = 0.2\,(1.01) = 0.202,$$

$$k_3 = 0.2\,(1 + 0.010201) = 0.20204,$$

$$k_4 = 0.2\,(1 + 0.040820) = 0.20816,$$

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 (7.21a) and (7.21b) gives

$$k_1 = 0.2\,[1 + (0.2027)^2] = 0.2082,$$

$$k_2 = 0.2\,[1 + (0.3068)^2] = 0.2188,$$

$$k_3 = 0.2\,[1 + (0.3121)^2] = 0.2195,$$

$$k_4 = 0.2\,[1 + (0.4222)^2] = 0.2356,$$

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 7.10 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

Presented By: http://www.ebooksuit.com

7.6 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 $x_n, x_{n-1}, x_{n-2}, \ldots$ 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 7.6.1 we derive Predictor–corrector formulae which use backward differences and in Section 7.6.2 we describe Milne's method which uses forward differences.

7.6.1 Adams–Moulton Method

Newton's backward difference interpolation formula can be written as

$$f(x, y) = f_0 + n\nabla f_0 + \frac{n(n+1)}{2}\nabla^2 f_0 + \frac{n(n+1)(n+2)}{6}\nabla^3 f_0 + \cdots \tag{7.22}$$

where

$$n = \frac{x - x_0}{h} \quad \text{and} \quad f_0 = f(x_0, y_0).$$

If this formula is substituted in

$$y_1 = y_0 + \int_{x_0}^{x_1} f(x, y)\,dx, \tag{7.23}$$

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, $y_{-1}, y_{-2}, \ldots$; all of which are known. Hence this formula can be used to compute y_1. We therefore write it as

$$y_1^{\mathrm{p}} = y_0 + \left(1 + \frac{1}{2}\nabla + \frac{5}{12}\nabla^2 + \frac{3}{8}\nabla^3 + \frac{251}{720}\nabla^4 + \cdots\right) f_0 \tag{7.24}$$

Presented By: http://www.ebooksuit.com

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:

$$f(x, y) = f_1 + n\nabla f_1 + \frac{n(n+1)}{2}\nabla^2 f_1 + \frac{n(n+1)(n+1)}{6}\nabla^3 f_1 + \cdots \tag{7.25}$$

Substituting (7.25) in (7.23), we obtain

$$y_1 = y_0 + \int_{x_0}^{x_1} \left[f_1 + n\nabla f_1 + \frac{n(n+1)}{2}\nabla^2 f_1 + \cdots \right] dx$$

$$= y_0 + h \int_1^0 \left[f_1 + n\nabla f_1 + \frac{n(n+1)}{2}\nabla^2 f_1 + \cdots \right] dn$$

$$= y_0 + h\left(1 - \frac{1}{2}\nabla - \frac{1}{12}\nabla^2 - \frac{1}{24}\nabla^3 - \frac{19}{720}\nabla^4 - \cdots\right) f_1 \tag{7.26}$$

The right side of (7.26) depends on $y_1, y_0, y_{-1}, \ldots$ where for y_1 we use y_1^{p}, the predicted value obtained from (7.24). The new value of y_1 thus obtained from (7.26) is called the *corrected* value, and hence we rewrite the formula as

$$y_1^{\mathrm{c}} = y_0 + h\left(1 - \frac{1}{2}\nabla - \frac{1}{12}\nabla^2 - \frac{1}{24}\nabla^3 - \frac{19}{720}\nabla^4 - \cdots\right) f_1^{\mathrm{p}} \tag{7.27}$$

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 (7.24) and (7.27) 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 (7.24) and (7.27) can be written as

$$y_1^{\mathrm{p}} = y_0 + \frac{h}{24}(55f_0 - 59f_{-1} + 37f_{-2} - 9f_{-3}) \tag{7.28}$$

and

$$y_1^{\mathrm{c}} = y_0 + \frac{h}{24}(9f_1^{\mathrm{p}} + 19f_0 - 5f_{-1} + f_{-2}) \tag{7.29}$$

in which the errors are approximately

$$\frac{251}{720}h^5 f_0^{(4)} \quad \text{and} \quad -\frac{19}{720}h^5 f_0^{(4)} \quad \text{respectively.}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

and

$$y_{n+1}^{c} = y_{n-1} + \frac{h}{3}(f_{n-1} + 4f_n + f_{n+1}) \tag{7.34a}$$

The application of this method is illustrated by the following example.

Example 7.12 We consider again the differential equation discussed in Examples 7.9 and 7.10, 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 7.9 and these 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 (7.32) 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 (7.34) 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 (7.34).

Example 7.13 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 $y(0.3)$.

Presented By: http://www.ebooksuit.com

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 (7.32), 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 (7.34), we need to compute $y'(0.3)$. We have

$$y'(0.3) = (0.3)^2 + (0.614616)^2 - 2 = -1.532247.$$

Now, (7.34) gives the corrected value of $y(0.3)$:

$$y(0.3) = 0.89 + \frac{0.1}{3}[-1.197900 + 4(-1.38164) + (-1.532247)] = 0.614776.$$

7.7 THE CUBIC SPLINE METHOD

The governing equations of a cubic spline have been discussed in detail in Section 3.14, 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 $s'(x_i) = m_i$:

$$s(x) = m_{i-1}\frac{(x_i - x)^2(x - x_{i-1})}{h^2} - m_i\frac{(x - x_{i-1})^2(x_i - x)}{h^2}$$
$$+ y_{i-1}\frac{(x_i - x)^2[2(x - x_{i-1}) + h]}{h^3} + y_i\frac{(x - x_{i-1})^2[2(x_i - x) + h]}{h^3}, \quad (7.35)$$

where $h = x_i - x_{i-1}$. Differentiating (7.35) with respect to x and simplifying, we obtain

$$s'(x) = \frac{m_{i-1}}{h^2}(x_i - x)(2x_{i-1} + x_i - 3x) - \frac{m_i}{h^2}(x - x_{i-1})(x_{i-1} + 2x_i - 3x)$$
$$+ \frac{6(y_i - y_{i-1})}{h^3}(x - x_{i-1})(x_i - x). \quad (7.36)$$

Again,

$$s''(x) = -\frac{2m_{i-1}}{h^2}(x_{i-1} + 2x_i - 3x) - \frac{2m_i}{h^2}(2x_{i-1} + x_i - 3x)$$

Presented By: http://www.ebooksuit.com

$$+\frac{6(y_i - y_{i-1})}{h^3}(x_{i-1} + x_i - 2x), \tag{7.37}$$

which gives

$$s''(x_i) = \frac{2m_{i-1}}{h} + \frac{4m_i}{h} - \frac{6}{h^2}(y_i - y_{i-1})$$

$$= \frac{2m_{i-1}}{h} + \frac{4m_i}{h} - \frac{6}{h^2}(s_i - s_{i-1}). \tag{7.38}$$

If we now consider the initial-value problem

$$\frac{dy}{dx} = f(x, y) \tag{7.39a}$$

and

$$y(x_0) = y_0 \tag{7.39b}$$

then from (7.39a), we obtain

$$\frac{d^2y}{dx^2} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx},$$

or

$$y''(x_i) = f'_x(x_i, y_i) + f'_y(x_i, y_i)\, f(x_i, y_i)$$

$$= f'_x(x_i, s_i) + f'_y(x_i, s_i)\, f(x_i, s_i). \tag{7.40}$$

Equating (7.38) and (7.40), we obtain

$$\frac{2m_{i-1}}{h} + \frac{4m_i}{h} - \frac{6}{h^2}(s_i - s_{i-1}) = f'_x(x_i, s_i) + f'_y(x_i, s_i) f(x_i, s_i) \tag{7.41}$$

from which s_i can be computed. Substitution in (7.35) gives the required solution.

The following example demonstrates the usefulness of the spline method.

Example 7.14 We consider again the boundary-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,

$$f(x_i, s_i) = 3x_i + \frac{1}{2}s_i.$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

In a similar manner, one can extend the Taylor series method or Picard's method to the system (7.42). The extension of the Runge–Kutta method to a system of n equations is quite straightforward.

We now consider the second-order differential equation

$$y'' = F(x, y, y') \tag{7.44a}$$

with the initial conditions

$$y(x_0) = y_0 \quad \text{and} \quad y'(x_0) = y'_0. \tag{7.45a}$$

By setting $z = y'$, the problem (7.44a) and (7.45a) can be reduced to the problem of solving the system

$$y' = z \quad \text{and} \quad z' = F(x, y, z) \tag{7.44b}$$

with the conditions

$$y(x_0) = y_0 \quad \text{and} \quad z(x_0) = y'_0 \tag{7.45b}$$

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.

7.9 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. 7.1) 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 (7.14) 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.

Presented By: http://www.ebooksuit.com

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.

7.10 BOUNDARY-VALUE PROBLEMS

Some simple examples of two-point linear boundary-value problems are:

(a) $$y''(x) + f(x)y'(x) + g(x)y(x) = r(x) \tag{7.46}$$

with the boundary conditions

$$y(x_0) = a \quad \text{and} \quad y(x_n) = b \tag{7.47}$$

(b) $$y^{iv}(x) = p(x)\, y(x) = q(x) \tag{7.48}$$

with

$$y(x_0) = y'(x_0) = A \quad \text{and} \quad y(x_n) = y'(x_n) = B. \tag{7.49}$$

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 7.10.1. An alternate method, called the *shooting method*, will be described next. Finally, in Section 7.10.3, we, discuss a method based on the application of cubic splines.

7.10.1 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

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2}y''(x) + \frac{h^3}{6}y'''(x) + \cdots \tag{7.50}$$

from which we obtain

$$y'(x) = \frac{y(x+h) - y(x)}{h} - \frac{h}{2}y''(x) - \cdots$$

Thus we have

$$y'(x) = \frac{y(x+h) - y(x)}{h} + O(h) \tag{7.51}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

We have explained the method with simple boundary conditions (7.47) 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 7.15 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:

$$y_0 - \frac{31}{16} y_1 + y_2 = -\frac{1}{16}$$

$$y_1 - \frac{31}{16} y_2 + y_3 = -\frac{1}{16}$$

$$y_2 - \frac{31}{16} y_3 + y_4 = -\frac{1}{16},$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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 (7.61) 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.

7.10.2 The Shooting Method

This method consists in transforming the boundary value problem into an initial-value problem. Its main advantage is that it is easy to apply. The method requires good initial guesses for the first derivative and can be applied to both linear and nonlinear problems. To describe the method, we consider the boundary-value problem defined by

$$y''(x) = f(x); \qquad y(0) = 0, \quad y(1) = 1. \tag{7.62}$$

The main steps involved in this method are:

(i) transformation of the boundary-value problem into an initial-value problem

(ii) solution of the initial-value problem by any of the known methods, and finally

(iii) solution of the given boundary-value problem.

To apply any initial value method, we must know $y'(0)$. Let the true value of $y'(0)$ be m. We start with two initial guesses for m and then determine the corresponding values of $y(1)$ using any initial value method. Let the two guesses for m be m_0 and m_1 and also let the corresponding values of $y(1)$ be Y_0 and Y_1, obtained by the initial value method. Using linear interpolation, we can then obtain a better approximation m_2 for m. This is given by

$$\frac{m_2 - m_0}{y(1) - Y_0} = \frac{m_1 - m_0}{Y_1 - Y_0}, \tag{7.63}$$

which gives

$$m_2 = m_0 + (m_1 - m_0).\frac{y(1) - Y(0)}{Y_1 - Y_0} \tag{7.64}$$

Presented By: http://www.ebooksuit.com

With this value of m_2, we solve the initial value problem

$$y''(x) = f(x); \qquad y(0) = 0, \quad y'(0) = m_2, \tag{7.65}$$

and obtain Y_2. If this agrees with $y(1)$ to the desired accuracy, the solution to the boundary-value problem is obtained. Otherwise, linear interpolation is carried out with (m_1, Y_1) and (m_2, Y_2) to obtain m_3. The process is repeated until convergence is obtained, i.e. until the value of Y_i agrees with $y(1)$ to the desired accuracy. The speed of convergence depends on how good the initial guesses were. The method will be tedious to apply to higher-order boundary-value problems and in the case of nonlinear problems, linear interpolation yields unsatisfactory results.

The method is illustrated with a simple linear second-order boundary-value problem.

Example 7.17 Solve the boundary value problem

$$y''(x) = y(x); \qquad y(0) = 0, \quad y(1) = 1.1752.$$

by the shooting method, taking $m_0 = 0.7$ and $m_1 = 0.8$. By Taylor's series, we have

$$y(x) = y(0) + xy'(0) + \frac{x^2}{2}y''(0) + \frac{x^3}{6}y'''(0) + \frac{x^4}{24}y^{\text{iv}}(0)$$
$$+ \frac{x^5}{120}y^{\text{v}}(0) + \frac{x^6}{720}y^{\text{vi}}(0) + \cdots \tag{i}$$

Since $y''(x) = y(x)$, we have

$$y'''(x) = y'(x), \quad y^{\text{iv}}(x) = y''(x) = y(x),$$
$$y^{\text{v}}(x) = y'(x), \quad y^{\text{vi}}(x) = y''(x) = y(x), \ldots$$

Putting $x = 0$ in the above, we obtain

$$y''(0) = y(0) = 0, \quad y'''(0) = y'(0),$$
$$y^{\text{iv}}(0) = 0, \qquad y^{\text{v}}(0) = y'(0), \ldots$$

Substitution in (i) gives

$$y(x) = y'(0)\left(x + \frac{x^3}{6} + \frac{x^5}{120} + \frac{x^7}{5040} + \frac{x^9}{362800} + \cdots\right), \qquad \text{since } y(0) = 0.$$

Hence

$$y(1) = y'(0)\left(1 + \frac{1}{6} + \frac{1}{120} + \frac{1}{5040} + \cdots\right) = y'(0)\,(1.1752) \tag{ii}$$

With $y'(0) \approx m_0 = 0.7$, (ii) gives

$$y(1) \approx Y_0 = 0.8226.$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

If we divide the interval [0, 1] into two equal subintervals, then from Eq. (7.62) and the recurrence relations for M_i, we obtain

$$y(0.5) = \frac{3}{22} = 0.13636, \tag{ii}$$

Then

$$M_0 = -1, \quad M_1 = -\frac{25}{22}, \quad M_2 = -1$$

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 of the problem has an error of 2.24% (see Bickley [1968]).

Example 7.19 Given the boundary-value problem

$$x^2y'' + xy' - y = 0; \qquad y(1) = 1, \quad y(2) = 0.5$$

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 in (7.67) and (7.68), 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

$$10M_0 - M_1 + 24y_1 = 36$$

$$16M_1 - M_2 - 32y_1 = -12$$

$$M_0 + 20M_1 + 16y_1 = 24$$

$$M_1 + 26M_2 - 24y_1 = -9$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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. (7.68) 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.

EXERCISES

7.1. Given

$$\frac{dy}{dx} - 1 = xy \quad \text{and} \quad y(0) = 1,$$

obtain the Taylor series for $y(x)$ and compute $y(0.1)$ correct to four decimal places.

7.2. Using the Taylor's series method, prove that the solution of

$$\frac{d^2y}{dx^2} + xy = 0$$

is given by

$$y = d\left(1 - \frac{x^3}{3!} + \frac{1\times 4}{6!}x^6 - \frac{1\times 4\times 7}{9!}x^9 + \cdots\right),$$

when the conditions are $x = 0$, $y = d$ and $dy/dx = 0$.

7.3. If

$$\frac{dy}{dx} = \frac{1}{x^2 + y},$$

where $y(4) = 4$, compute the values of $y(4.1)$ and $y(4.2)$ by Taylor's series method.

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

7.13. Given the problem

$$\frac{dy}{dx} = f(x, y) \quad \text{and} \quad y(x_0) = y_0$$

an approximate value y_1 at $x = x_1$ is given by the formula

$$y_1 = y_0 + \frac{1}{6}(k_1 + 4k_2 + k_3) + R_4,$$

where

$$k_1 = hf(x_0, y_0),$$

$$k_2 = hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1\right)$$

$$k_3 = hf(x_0 + h,\ y_0 + 2k_2 - k_1).$$

Show that R_4 is of order h^4.

7.14. Tabulate the solution of the equation

$$\frac{dy}{dx} - x = 0.1y^2, \qquad y(0) = 0$$

for the range $0 < x < 0.5$ at intervals of 0.1. Obtain the solution correct to four decimal places, and compare it with the Taylor's series solution.

7.15. Use the Runge–Kutta method to solve

$$10\frac{dy}{dx} = x^2 + y^2, \qquad y(0) = 1$$

for the interval $0 < x \le 0.4$ with $h = 0.1$.

7.16. Use the predictor–corrector formulae for tabulating a solution of

$$10\frac{dy}{dx} = x^2 + y^2, \qquad y(0) = 1$$

for the range $0.5 \le x \le 1.0$.

7.17. Tabulate the solution of

$$\frac{dy}{dx} = x + y, \qquad y(0) = 0$$

for $0.4 < x \le 1.0$ with $h = 0.1$ using the predictor–corrector formulae.

7.18. Using Milne's method, find $y(0.8)$ given that

$$\frac{dy}{dx} = x - y^2, \quad y(0) = 0, \quad y(0.2) = 0.02,$$

$$y(0.4) = 0.0795, \quad y(0.6) = 0.1762.$$

Presented By: http://www.ebooksuit.com

7.19. Using Milne's method, solve the differential equation

$$(1+x)\frac{dy}{dx}+y=0, \qquad \text{with } y(0)=2,$$

for $x=1.5$ to $x=2.5$. Obtain the starting values by using the fourth-order Runge–Kutta method with $h=0.5$.

7.20. Solve the system of differential equations

$$\frac{dx}{dt}=y-t, \qquad \frac{dy}{dt}=x+t$$

with $x=1$, $y=1$ when $t=0$, taking $\Delta t=h=0.1$.

7.21. Solve the differential equation

$$\frac{d^2y}{dx^2}-x\left(\frac{dy}{dx}\right)^2+y^2=0,$$

with $y(0)=1$ and $y'(0)=0$, using the fourth-order Runge–Kutta method with $h=0.2$.

7.22. Given that

$$\frac{dv}{du}=uvw, \qquad \frac{dw}{du}=\frac{uv}{w},$$

and

$$v(1)=\frac{1}{3}, \qquad w(1)=1.$$

compute $v(1.1)$ and $w(1.1)$.

7.23. Solve the equation

$$\frac{d^2y}{dx^2}+y=0$$

with the conditions $y(0)=1$ and $y'(0)=0$. Compute $y(0.2)$ and $y(0.4)$.

7.24. Solve the boundary-value problem

$$y''-64y+10=0; \qquad y(0)=y(1)=0$$

by the finite-difference method. Compute the value of $y(0.5)$ and compare it with the true value.

7.25. Use the spline method to solve the initial value problem

$$x\,dy+2y\,dx=0 \quad \text{and} \quad y(2)=1$$

7.26. Using the cubic spline technique, solve the following boundary-value problems

(i) $y''-y=0$, $y(0)=0$ and $y(1)=1$

(ii) $y''+2y'+y=30x$, $y(0)=0$ and $y(1)=0$

Presented By: http://www.ebooksuit.com

(iii) $y'' - 64y + 10 = 0,\ y(0) = y(1) = 0.$

In each case, divide the given interval into two equal subintervals, and compare your solution with the analytical solution at the midpoint of the interval.

7.27. Solve the boundary value problem

$$y'' = y(x); \qquad y(0) = 0, \quad y(1) = 1,$$

by the shooting method.

7.28. Solve the boundary-value problem

$$y'' - 64y + 10 = 0; \qquad y(0) = y(1) = 0$$

by the shooting method.

7.29. Solve the boundary-value problem

$$\frac{d^2y}{dx^2} + \frac{4x}{1+x^2}\frac{dy}{dx} + \frac{2}{1+x^2}y = 0.$$

with the boundary conditions $y(0) = 1$ and $y(2) = 0.2$. Use the cubic spline method first with $h = 1$ and then with $h = 1/2$ to determine the value of $y(1)$. Compare your answers with the exact value obtained from the analytical solution $y = 1/(1+x^2)$ [Albasiny and Hoskins].

7.30. Solve the boundary-value problem

$$(1+x)^2\frac{d^2y}{dx^2} + (1+x)\frac{dy}{dx} - y = 0,$$

with

$$y(0) = 1 \text{ and } y(1) = 0.5.$$

Use the cubic spline method to determine the value of $y(0.5)$ and compare it with that obtained from the exact solution $y = 1/(1+x)$.

Presented By: http://www.ebooksuit.com

CHAPTER

Numerical Solution of Partial Differential Equations

8.1 INTRODUCTION

Partial differential equations occur in many branches of applied mathematics, for example, in hydrodynamics, elasticity, quantum mechanics and electromagnetic theory. The analytical treatment of these equations is a rather involved process and requires application of advanced mathematical methods. On the other hand, it is generally easier to produce sufficiently approximate solutions by simple and efficient numerical methods. Several numerical methods have been proposed for the solution of partial differential equations, but only the *finite-difference methods* have become popular and are more gainfully employed than others. We will therefore restrict ourselves to a treatment of the finite-difference methods and, in the sequel, we will discuss, very briefly, some of the numerical procedures with simple illustrative examples. Only the rudiments of the method will be given here and the interested reader is referred to the text by G.D. Smith (see Bibliography) for further details.

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

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G, \tag{8.1}$$

where A, B, C, ..., G, are all functions of x and y.

Presented By: http://www.ebooksuit.com

Equations of the form (8.1) can be classified with respect to the sign of the discriminant.

$$\Delta_s = B^2 - 4AC, \tag{8.2}$$

in the following way. If $\Delta_s < 0$ at a point in the (x, y) plane, the equation is said to be of *elliptic* type, to be of *hyperbolic* type when $\Delta_s > 0$ at that point, and to be of *parabolic* type when $\Delta_s = 0$.

In the following, we will restrict ourselves to three simple particular cases of Eq. (8.1), namely

$$u_{xx} + u_{yy} = 0 \text{ (the Laplace equation)} \tag{8.3}$$

$$u_{xx} - \frac{1}{c^2} u_{tt} = 0 \text{ (the wave equation)} \tag{8.4}$$

$$u_{xx} - u_t = 0 \text{ (the heat conduction equation)}, \tag{8.5}$$

where (x, y) are space coordinates and t is the time coordinate. It is easy to see that the Laplace equation is of elliptic type, that the wave equation is of hyperbolic type and that the heat equation is of parabolic type.

In a similar way, we conclude that the partial differential equation

$$xu_{xx} + u_{yy} = 0$$

is

(i) parabolic if $x = 0$

(ii) elliptic if $x > 0$

(iii) hyperbolic if $x < 0$.

It is clear that the region plays an important role in the classification of partial differential equations. 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, to find a function u satisfying the Laplace equation in R, i.e. to find u such that

$$\left.\begin{aligned} u_{xx} + u_{yy} &= 0 \text{ in } R \\ u &= f \text{ on } C \end{aligned}\right\} \tag{8.6}$$

(ii) *Cauchy's problem:*

$$\left.\begin{aligned} u_{tt} - u_{xx} &= 0 \quad \text{for } t > 0 \\ u(x, 0) &= f(x) \\ u_t(x, 0) &= g(x) \end{aligned}\right\} \tag{8.7}$$

$f(x)$ and $g(x)$ being arbitrary.

Presented By: http://www.ebooksuit.com

$$\text{(iii)} \quad \left.\begin{aligned} u_t - u_{xx} &= 0 \quad \text{for } t > 0 \\ u(x, 0) &= f(x) \end{aligned}\right\} \tag{8.8}$$

These problems are all *well-defined* (or *well-posed*) and it is proved in textbooks of partial differential equations that they possess unique solutions. At this juncture, it is, however, important to mention a point of difference between ordinary and partial differential equations. In contrast with ordinary differential equations, the form of a partial differential equation is always connected with a particular type of associated conditions. Thus, the problem of Laplace's equation with Cauchy boundary conditions, viz., the problem defined by

$$\left.\begin{aligned} u_{xx} + u_{yy} &= 0 \\ u(x, 0) &= f(x) \\ u_y(x, 0) &= g(x) \end{aligned}\right\} \tag{8.9}$$

is an *ill-posed* problem.

8.2 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 (see Section 7.10 of Chapter 7)

$$u_x = \frac{u_{i+1,j} - u_{i,j}}{h} + O(h) \tag{8.10}$$

$$= \frac{u_{i,j} - u_{i-1,j}}{h} + O(h) \tag{8.11}$$

$$= \frac{u_{i+1,j} - u_{i-1,j}}{2h} + O(h^2) \tag{8.12}$$

and

$$u_{xx} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + O(h^2) \tag{8.13}$$

where

$$u_{i,j} = u(ih, jk) = u(x, y)$$

Presented By: http://www.ebooksuit.com

Similarly, we have the approximations

$$u_y = \frac{u_{i,\,j+1} - u_{i,\,j}}{k} + O(k) \tag{8.14}$$

$$= \frac{u_{i,\,j} - u_{i,\,j-1}}{k} + O(k) \tag{8.15}$$

$$= \frac{u_{i,\,j+1} - u_{i,\,j-1}}{2k} + O(k^2) \tag{8.16}$$

and

$$u_{yy} = \frac{u_{i,\,j-1} - 2u_{i,\,j} + u_{i,\,j+1}}{k^2} + O(k^2) \tag{8.17}$$

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

$$\frac{1}{h^2}(u_{i+1,\,j} - 2u_{i,\,j} + u_{i-1,\,j}) + \frac{1}{k^2}(u_{i,\,j+1} - 2u_{i,\,j} + u_{i,\,j-1}) = 0. \tag{8.18}$$

If $h = k$, this gives

$$u_{i,j} = \frac{1}{4}(u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1}), \tag{8.19}$$

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. 8.1a), and is written

$$u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1} - 4u_{i,\,j} = 0 \tag{8.20}$$

By expanding the terms on the right side of (8.19) by Taylor's series, it can be shown that

$$u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1} - 4u_{i,\,j} = h^2(u_{xx} + u_{yy}) - \frac{1}{6}h^4 u_{xxyy} + O(h^6)$$

$$= -\frac{1}{6}h^4 u_{xxyy} + O(h^6) \tag{8.21}$$

Instead of formula (8.19), we may also use the formula

$$u_{i,\,j} = \frac{1}{4}(u_{i-1,\,j-1} + u_{i+1,\,j-1} + u_{i+1,\,j+1} + u_{i-1,\,j+1}) \tag{8.22}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

$$u_6 = \frac{1}{4}(u_3 + c_7 + u_9 + u_5); \quad u_2 = \frac{1}{4}(c_3 + u_3 + u_5 + u_1).$$

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.

8.3.1 Jacobi's Method

Let $u_{i,j}^{(n)}$ denotes the nth iterative value of $u_{i,j}$. An iterative procedure to solve (8.20) is

$$u_{i,j}^{(n+1)} = \frac{1}{4}[u_{i-1,j}^{(n)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n)} + u_{i,j+1}^{(n)}] \tag{8.26}$$

for the interior mesh points. This is called the *point Jacobi method.*

8.3.2 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:

$$u_{i,j}^{(n+1)} = \frac{1}{4}[u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)}] \tag{8.27}$$

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.*

8.3.3 Successive Over-relaxation (or SOR Method)

Equation (8.27) can be written as

$$\begin{aligned} u_{i,j}^{(n+1)} &= u_{i,j}^{(n)} + \frac{1}{4}[u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)} - 4u_{i,j}^{(n)}] \\ &= u_{i,j}^{(n)} + \frac{1}{4}R_{i,j} \end{aligned}$$

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

$$\begin{aligned} u_{i,j}^{(n+1)} &= u_{i,j}^{(n)} + \frac{1}{4}\omega R_{i,j} \\ &= \frac{1}{4}\omega[u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)}] + (1-\omega)u_{i,j}^{(n)} \end{aligned} \tag{8.28}$$

The rate of convergence of (8.28) depends on the choice of ω, which is called the *accelerating factor* and lies between 1 and 2.

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

$$u_3^{(1)} = \frac{1}{4}(1+1+0+0) = 0.5;$$

$$u_4^{(1)} = \frac{1}{4}(1+1+0+0) = 0.5.$$

The iterations have been continued using the formula (8.26), 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 the formula (8.28) 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 8.3 Solve Laplace's equation for Fig. 8.5 given below:

Figure 8.5

Presented By: http://www.ebooksuit.com

We first compute the quantities u_5, u_7, u_9, u_1 and u_3 by using the diagonal five-point formula (8.22). 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 (8.20)

$$u_8^{(1)} = 53.12; \qquad u_4^{(1)} = 18.75;$$

$$u_6^{(1)} = 18.75; \qquad u_2^{(1)} = 9.38.$$

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 8.3. We give below the first-four iterates obtained by using the Gauss–Seidel formula (8.27).

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 8.4 Solve the Poisson equation

$$u_{xx} + u_{yy} = -10\,(x^2 + y^2 + 10).$$

in the domain of Fig. 8.6.

Figure 8.6

Let the values of u at the four grid points, A, B, C, D bet u_1, u_2, u_3, u_4, respectively. Let the grid points be defined by $x = ih$, $y = jh$, where $h = 1$,

Presented By: http://www.ebooksuit.com

$i, j = 0, 1, 2, 3$. 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

$$u_1^{(n+1)} = \frac{1}{2} u_2^{(n)} + 30$$

$$u_2^{(n+1)} = \frac{1}{4}[u_1^{(n+1)} + u_4^{(n)} + 150]$$

$$u_4^{(n+1)} = \frac{1}{2} u_2^{(n+1)} + 45.$$

For the first iteration, we assume that $u_2^{(0)} = u_4^{(0)} = 0$. Hence we obtain

$$u_1^{(1)} = 30,$$

$$u_2^{(1)} = \frac{1}{4}(30 + 0 + 150) = 45$$

$$u_4^{(1)} = \frac{1}{2}(45) + 45 = 67.5.$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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.,

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0 \tag{8.29}$$

and the standard five-point formula

$$u_{i-1,\,j}+u_{i+1,\,j}+u_{i,\,j-1}+u_{i,\,j+1}-4u_{i,\,j}=0 \tag{8.20}$$

The use of formula (8.20) involves the solution of a system of algebraic equations, whose coefficient matrix, for $n=6$, is of the form

$$A=\begin{bmatrix} -4 & 1 & 0 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 1 & -4 & 0 & 0 & 1 \\ 1 & 0 & 0 & -4 & 1 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & 0 & 1 & -4 \end{bmatrix} \tag{8.30}$$

The general form of such a system is given by

$$B=\begin{bmatrix} T & I & & & 0 & \\ I & T & I & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & & I & T & I \\ & & & & I & T \end{bmatrix}, \tag{8.31}$$

where T is a tridiagonal matrix of the form

$$T=\begin{bmatrix} -4 & 1 & & & & \\ 1 & -4 & 1 & & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & & 1 & -4 & 1 \\ & & & & 1 & -4 \end{bmatrix} \tag{8.32}$$

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 system of the type (8.32) 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.

Presented By: http://www.ebooksuit.com

We rearrange formula (8.20) in either of two ways:

$$u_{i-1,j} - 4u_{i,j} + u_{i+1,j} = -u_{i,j-1} - u_{i,j+1} \tag{8.33}$$

or

$$u_{i,j-1} - 4u_{i,j} + u_{i,j+1} = -u_{i-1,j} - u_{i+1,j} \tag{8.34}$$

The ADI is an *iteration* method and formulae (8.33) and (8.34) are used as iteration formulae

$$u_{i-1,j}^{(r+1)} - 4u_{i,j}^{(r+1)} + u_{i+1,j}^{(r+1)} = -u_{i,j-1}^{(r)} - u_{i,j+1}^{(r)} \tag{8.35}$$

and

$$u_{i\,j-1}^{(r+2)} - 4u_{i,j}^{(r+2)} + u_{i\,j+1}^{(r+2)} = -u_{i-1,j}^{(r+1)} - u_{i+1,j}^{(r+1)} \tag{8.36}$$

Formula (8.35) is used to compute function values at all internal mesh points along rows and formula (8.36) those along columns. For $j = 1, 2, 3, \ldots, n-1$, Eq. (8.35) yields a tridiagonal system of equations and can easily be solved. Similarly, for $i = 1, 2, 3, \ldots, n-1$, Eq. (8.36) also yields a tridiagonal system of equations.

In the ADI method, formulae (8.35) and (8.36) are used alternately. For example, for the first row $j = 1$, and formula (8.35) gives

$$u_{i-1,1}^{(r+1)} - 4u_{i,1}^{(r+1)} + u_{i+1,1}^{(r+1)} = -u_{i,0}^{(r)} - u_{i,2}^{(r)}, \qquad (i = 1, 2, 3, \ldots, n-1) \tag{8.37}$$

Together with the boundary conditions, Eqs. (8.37) represent 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 formula (8.36) 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 8.5 demonstrates the method of solution.

Example 8.5 Solve Laplace's equation, $u_{xx} + u_{yy} = 0$, in the domain of Fig. 8.7 (see Example 8.2).

Figure 8.7

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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 (8.36)

$$u_{2,j-1}^{(2)} - 4u_{2,j}^{(2)} + u_{2,j+1}^{(2)} = -u_{1,j}^{(1)} - u_{3,j}^{(1)} \tag{iv}$$

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.

8.4 PARABOLIC EQUATIONS

We consider the heat conduction equation:

$$C\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad C \text{ being a constant.} \tag{8.38}$$

Let the (x, t) plane be divided into smaller rectangles by means of the sets of lines

$$x = ih, \qquad i = 0, 1, 2, \ldots$$

$$t = jk, \qquad j = 0, 1, 2, \ldots$$

Using the approximations

$$\frac{\partial u}{\partial t} = \frac{u_{i,j+1} - u_{i,j}}{k} \tag{8.39a}$$

and

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) \tag{8.39b}$$

Eq. (8.38) can be replaced by the finite-difference analogue

$$\frac{C}{k}(u_{i,j+1} - u_{i,j}) = \frac{1}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}),$$

Presented By: http://www.ebooksuit.com

which can be written as

$$u_{i,j+1} = u_{i,j} + \lambda(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}), \tag{8.40a}$$

where $\lambda = k/(Ch^2)$.

This formula expresses the unknown function value at the $(i, j+1)$th interior point in terms of the known function values and hence it is called the *explicit formula*. It can be shown that this formula is valid only for $0 < \lambda \leq 1/2$.

For $\lambda = 1/2$, Eq. (8.40) reduces to

$$u_{i,j+1} = \frac{1}{2}(u_{i-1,j} + u_{i+1,j}), \tag{8.40b}$$

which is called *Bender–Schmidt recurrence relation.*

In formula (8.40a), we have used the function values along the jth row only in the approximation of $\partial^2 u/\partial x^2$.

Crank and Nicolson proposed a method in 1947 according to which $\partial^2 u/\partial x^2$ is replaced by the average of its finite-difference approximations on the jth and $(j+1)$th rows. Thus,

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{2}\left(\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1}}{h^2}\right)$$

and hence Eq. (8.38) is replaced by

$$\frac{C}{k}(u_{i,j+1} - u_{i,j}) = \frac{1}{2h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j} + u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1}),$$

which gives on rearranging

$$-\lambda u_{i-1,j+1} + (2+2\lambda)u_{i,j+1} - \lambda u_{i+1,j+1} = \lambda u_{i-1,j} + (2-2\lambda)u_{i,j} + \lambda u_{i+1,j}, \tag{8.41}$$

where $\lambda = k/(Ch^2)$.

On the left side of (8.41) we have three unknowns and on the right side all the three quantities are known. Equation (8.41), which is an *implicit* scheme, is called *Crank–Nicolson formula* and is convergent for all finite values of λ.

If there are N internal mesh points on each row, then formula (8.41) gives N simultaneous equations for the N unknowns in terms of the given boundary values. Similarly, the internal mesh points on all rows can be calculated.

Example 8.6 Use the Bender–Schmidt recurrence relation to solve the equation

$$\frac{\partial^2 u}{\partial x^2} = 2\frac{\partial u}{\partial t}$$

with the conditions

$$u(x, 0) = 4x - x^2, \qquad u(0, t) = u(4, t) = 0.$$

Presented By: http://www.ebooksuit.com

Taking $h = 1$, we obtain

$$k = \frac{1}{2}h^2 C = 1.$$

Also, $u(0, 0) = 0$, $u(1, 0) = 3$, $u(2, 0) = 4$, $u(3, 0) = 3$, and $u(4, 0) = 0$. For the first time step, $k = 1$. Using the Bender–Schmidt recurrence relation, we obtain

$$u_{1,1} = \frac{1}{2}(0+4) = 2, \quad u_{2,1} = \frac{1}{2}(3+3) = 3, \quad u_{3,1} = \frac{1}{2}(4+0) = 2.$$

For $k = 2$, we have

$$u_{1,2} = \frac{1}{2}(0+3) = 1.5, \quad u_{2,2} = \frac{1}{2}(2+2) = 2, \quad u_{3,2} = \frac{1}{2}(3+0) = 1.5,$$

For $k = 3$, we obtain

$$u_{1,3} = \frac{1}{2}(0+2) = 1, \quad u_{2,3} = \frac{1}{2}(1.5+1.5) = 1.5, \quad u_{3,3} = \frac{1}{2}(2+0) = 1$$

With $k = 4$, we have

$$u_{1,4} = \frac{1}{2}(0+1.5) = 0.75, \quad u_{2,4} = \frac{1}{2}(1+1) = 1, \quad u_{3,4} = \frac{1}{2}(1.5+0) = 0.75$$

Similarly with $k = 5$, we obtain

$$u_{1,5} = \frac{1}{2}(0+1.0) = 0.5, \quad u_{2,5} = \frac{1}{2}(0.75+0.75) = 0.75, \quad u_{3,5} = \frac{1}{2}(1+0) = 0.5.$$

The computations can be continued to any number of time steps.

Example 8.7 Solve

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

subject to the initials condition $u = \sin \pi x$ at $t = 0$ for $0 \le x \le 1$ and the boundary conditions $u = 0$ at $x = 0$ and $x = 1$ for $t > 0$. Take $h = 0.2$, $\lambda = 1$ and compute the values of u at the internal mesh points up to two time steps.

We have $h = 0.2$ and $\lambda = 1$. Hence $k = h^2 = 0.04$.

The Crank–Nicolson formula corresponding to $\lambda = 1$ is

$$u_{i-1,\,j+1} - 4u_{i,\,j+1} + u_{i+1,\,j+1} = -u_{i-1,\,j} - u_{i+1,\,j}. \tag{i}$$

Applying (i) at the mesh point u_1, we obtain

$$0 - 4u_1 + u_2 = -0.9511. \tag{ii}$$

Again, applying (i) at u_2, we get

$$u_1 - 4u_2 + u_3 = -0.5878 - 0.9511 = -1.5389. \tag{iii}$$

Similarly, application of (i) at the mesh points u_3 and u_4 gives, respectively:

$$u_2 - 4u_3 + u_4 = -0.9511 - 0.5878 = -1.5389 \tag{iv}$$

Presented By: http://www.ebooksuit.com

and

$$u_3 - 4u_4 = -0.9511. \tag{v}$$

By symmetry, we have $u_1 = u_4$ and $u_2 = u_3$.

Hence, eqs. (ii) to (v) reduce to the two equations:

$$4u_1 - u_2 = 0.9511, \qquad u_1 - 3u_2 = -1.5389,$$

the solution of which is

$$u_1 = 0.3993 \quad \text{and} \quad u_2 = 0.6461.$$

For the second time step, let u_5, u_6, u_7 and u_8 be the values of u at the *internal mesh* points. Then, applying formula (i) at these mesh points, we obtain

$$-4u_5 + u_6 = -0.6461$$

$$u_5 - 4u_6 + u_7 = -0.3993 - 0.6461 = -1.0454$$

$$u_6 - 4u_7 + u_8 = -1.0454$$

$$u_7 - 4u_8 = -0.6461.$$

By symmetry, $u_5 = u_8$ and $u_6 = u_7$. Hence the above equations reduce to the two equations:

$$-4u_5 + u_6 = -0.6461 \quad \text{and} \quad u_5 - 3u_6 = -1.0454,$$

the solution of which is

$$u_5 = 0.2712 \quad \text{and} \quad u_6 = 0.4387.$$

Example 8.8 Solve the heat equation:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

subject to the conditions

$$u(x, 0) = 0, \qquad u(0, t) = 0, \qquad u(1, t) = t$$

(i) We first choose $k = 1/8$ and $h = 1/2$ so that $\lambda = k/h^2 = 1/2$. The Crank–Nicolson scheme (8.41) now becomes

$$-u_{i-1,\,j+1} + 6u_{i,\,j+1} - u_{i+1,\,j+1} = u_{i-1,\,j} + 2u_{i,\,j} + u_{i+1,\,j} \tag{i}$$

Let the value of u corresponding to $t = 1/8$ and $x = 1/2$, i.e. at the mesh point P be u_1 (see Fig. 8.8). Applying the Crank–Nicolson scheme (i) given above at this point, we obtain

$$0 + 6u_1 - \frac{1}{8} = 0 \qquad \text{which gives } u_1 = 0.02083.$$

Presented By: http://www.ebooksuit.com

Figure 8.8

(ii) We now choose $k = 1/8, h = 1/4$ so that $\lambda = 2$ (see Fig. 8.9). The Crank–Nicolson scheme corresponding to this value of λ is given by

$$-u_{i-1,\,j+1} + 3u_{i,\,j+1} - u_{i+1,\,j+1} = u_{i-1,\,j} - u_{i,\,j} + u_{i+1,\,j}. \qquad \text{(ii)}$$

Applying the above equation at the mesh point P, we obtain

$$0 + 3u_1 - u_2 = 0 \quad \text{or} \quad 3u_1 = u_2$$

Figure 8.9

Similarly, applying the same equation at the mesh points Q and R, we obtain the two equations

$$-u_1 + 3u_2 - u_3 = 0 \quad \text{and} \quad -u_2 + 3u_3 - \frac{1}{8} = 0.$$

We have thus three equations in the three unknowns u_1, u_2, u_3 and the solution is

$$u_1 = 0.00595, \qquad u_2 = 0.01785, \qquad u_3 = 0.04760$$

(iii) As our final choice, we choose $k = 1/16$, $h = 1/4$ so that $\lambda = 1$. This means that we propose to find our solution for $t = 1/8$ in two steps instead of one as in (i) and (ii) above.

Presented By: http://www.ebooksuit.com

The Crank–Nicolson scheme corresponding to this value of λ is now

$$-u_{i-1,j+1} + 4u_{i,j+1} - u_{i+1,j+1} = u_{i-1,j} + u_{i+1,j}. \qquad \text{(iii)}$$

Applying the scheme (iii) above at the mesh points P, Q and R, we obtain the three equations:

$$4u_1 - u_2 = 0, \quad -u_1 + 4u_2 - u_3 = 0, \quad -u_2 + 4u_3 - \frac{1}{16} = 0$$

whose solution is

$$u_1 = \frac{1}{56 \times 16}, \quad u_2 = \frac{1}{56 \times 4}, \quad u_3 = \frac{15}{56 \times 16}$$

Again, applying the scheme (iii) at each of the mesh points X, Y, Z in Fig. 8.10, we obtain the three equations:

$$4u_4 - u_5 = \frac{1}{4 \times 56}$$

$$-u_4 + 4u_5 - u_6 = \frac{1}{56}$$

$$-u_5 + 4u_6 - \frac{1}{8} = \frac{1}{4 \times 56} + \frac{1}{16}$$

Figure 8.10

The solution is

$$u_4 = 0.005899, \quad u_5 = 0.019132, \quad u_6 = 0.052771$$

The exact solution of the problem is given by Froberg: *Introduction to Numerical Analysis*, p. 269.

$$u(x,t) = \frac{1}{6}(x^3 - x + 6xt) + \frac{2}{\pi^3} \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^3} e^{-n^2\pi^2 t} \sin n\pi x$$

Presented By: http://www.ebooksuit.com

which gives

$$u\left(\frac{1}{4},\frac{1}{8}\right)=0.00541, \qquad u\left(\frac{1}{2},\frac{1}{8}\right)=0.01878, \qquad u\left(\frac{3}{4},\frac{1}{8}\right)=0.05240.$$

8.5 ITERATIVE METHODS FOR THE SOLUTION OF EQUATIONS

The iterative methods discussed in Section 8.3 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

$$(1+r)u_{i,j+1}=u_{i,j}+\frac{1}{2}r(u_{i-1,j+1}+u_{i+1,j}+u_{i+1,j+1}+u_{i-1,j}-2u_{i,j}) \tag{8.42}$$

where $r=k/h^2$.

In Eq. (8.42), 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

$$c_i=u_{i,j}+\frac{1}{2}r(u_{i-1,j}-2u_{i,j}+u_{i+1,j}) \tag{8.43}$$

Eq. (8.42) can be written as

$$u_i=\frac{r}{2(1+r)}(u_{i-1}+u_{i+1})+\frac{c_i}{1+r} \tag{8.44}$$

From Eq. (8.44), we obtain the iteration formula

$$u_i^{(n+1)}=\frac{r}{2(1+r)}[u_{i-1}^{(n)}+u_{i+1}^{(n)}]+\frac{c_i}{1+r}, \tag{8.45}$$

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. (8.45) 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 (8.45) by its latest value available, namely by $u_{i-1}^{(n+1)}$. Accordingly, we obtain the formula

$$u_i^{(n+1)}=\frac{r}{2(1+r)}[u_{i-1}^{(n+1)}+u_{i+1}^{(n)}]+\frac{c_i}{1+r} \tag{8.46}$$

Presented By: http://www.ebooksuit.com

which is called the *Gauss–Seidel iteration formula*. It can be shown that the scheme (8.46) converges for all finite values of r and that it converges twice as fast as Jacobi's scheme.

Equation (8.46) 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

$$u_i^{(n+1)} = u_i^{(n)} + \omega \left\{ \frac{r}{2(1+r)} [u_{i-1}^{(n+1)} + u_{i+1}^{(n)}] + \frac{c_i}{1+r} - u_i^{(n)} \right\} \tag{8.47}$$

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 8.9 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 $r = k/h^2 = 1/2$. The formula (8.46) 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. 8.11.

Figure 8.11

Presented By: http://www.ebooksuit.com

Applying the formula (i) at the four interior mesh points, we obtain successively

$$u_1^{(n+1)} = \frac{1}{6}[0 + u_2^{(n)}] + \frac{2}{3}\left[0.5878 + \frac{1}{4}(0 - 2\times 0.5878 + 0.9511)\right]$$

$$= \frac{1}{6}u_2^{(n)} + 0.3544 \qquad \text{(ii)}$$

$$u_2^{(n+1)} = \frac{1}{6}[u_1^{(n+1)} + u_3^{(n)}] + \frac{2}{3}\left[0.9511 + \frac{1}{4}(0.5878 - 2\times 0.9511 + 0.9511)\right]$$

$$= \frac{1}{6}[u_1^{(n+1)} + u_3^{(n)}] + 0.5736 \qquad \text{(iii)}$$

$$u_3^{(n+1)} = \frac{1}{6}[u_2^{(n+1)} + u_4^{(n)}] + \frac{2}{3}\left[0.9511 + \frac{1}{4}(0.9511 - 2\times 0.9511 + 0.5878)\right]$$

$$= \frac{1}{6}[u_2^{(n+1)} + u_4^{(n)}] + 0.5736 \qquad \text{(iv)}$$

$$u_4^{(n+1)} = \frac{1}{6}[u_3^{(n+1)} + 0] + \frac{2}{3}\left[0.5878 + \frac{1}{4}(0.9511 - 2\times 0.5878 + 0.0)\right]$$

$$= \frac{1}{6}u_3^{(n+1)} + 0.3544 \qquad \text{(v)}$$

Formulae (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.*

*For the derivation of a more general finite-difference representation of the parabolic equation, see the paper by Sastry [1976].

Presented By: http://www.ebooksuit.com

8.6 HYPERBOLIC EQUATIONS

We consider the boundary-value problem defined by

$$u_{tt} = c^2 u_{xx} \tag{8.48}$$

$$u(x, 0) = f(x) \tag{8.49}$$

$$u_t(x, 0) = \phi(x) \tag{8.50}$$

$$u(0, t) = \psi_1(t) \tag{8.51}$$

$$u(1, t) = \psi_2(t) \tag{8.52}$$

for $0 \le t \le T$, which models the transverse vibrations of a stretched string. As in the previous cases, we use the following difference approximations for the derivatives

$$u_{xx} = \frac{1}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) + O(h^2) \tag{8.53}$$

and

$$u_{tt} = \frac{1}{k^2}(u_{i,j-1} - 2u_{i,j} + u_{i,j+1}) + O(k^2), \tag{8.54}$$

where $x = ih$, $i = 0, 1, 2, \ldots$, and $t = jk$, $j = 0, 1, 2, \ldots$

Further, $u_t(x, t)$ is approximated as follows

$$u_t(x, t) = \frac{u_{i,j+1} - u_{i,j-1}}{2k} + O(k^2) \tag{8.55}$$

Substituting (8.53) and (8.54) in (8.48), we obtain

$$\frac{1}{k^2}(u_{i,j-1} - 2u_{i,j} + u_{i,j+1}) = \frac{c^2}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j})$$

Putting $\alpha = ck/h$ in the above and rearranging the terms, we obtain

$$u_{i,j+1} = -u_{i,j-1} + \alpha^2(u_{i-1,j} + u_{i+1,j}) + 2(1 - \alpha^2)\, u_{i,j} \tag{8.56}$$

Formula (8.56) shows that the function values at the jth and $(j-1)$th time levels are required in order to determine those at the $(j+1)$th time level. Such difference schemes are called *three level* difference schemes compared to the two level schemes derived in the parabolic case.

By expanding the terms in (8.56) as Taylor's series and simplifying, it can be shown that the truncation error in (8.56) is $O(k^2 + h^2)$. Further, formula (8.56) holds good if $\alpha < 1$, which is the condition for stability.

There exist implicit finite difference schemes for the equation given by (8.48). Two such schemes are

Presented By: http://www.ebooksuit.com

$$\frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{k^2}=\frac{c^2}{2h^2}[(u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1})$$

$$+(u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1})] \tag{8.57}$$

and

$$\frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{k^2}=\frac{c^2}{4h^2}[(u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1})$$

$$+2(u_{i+1,j}-2u_{i,j}+u_{i-1,j})$$

$$+(u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1})] \tag{8.58}$$

Formulae (8.57) and (8.58) hold good for all values of ck/h. The use of formula (8.56) is demonstrated in the following examples.

Example 8.10 Solve the equation

$$\frac{\partial^2 u}{\partial t^2}=\frac{\partial^2 u}{\partial x^2}$$

subject to the following conditions

$$u(0,t)=0, \quad u(1,t)=0 \qquad t>0$$

and

$$\frac{\partial u}{\partial t}(x,0)=0, \quad u(x,0)=\sin^3(\pi x) \qquad \text{for all } x \text{ in } 0\le x\le 1.$$

This problem admits an exact solution which is given by

$$u(x,t)=\frac{3}{4}\sin \pi x\cos \pi t-\frac{1}{4}\sin 3\pi x\cos 3\pi t \tag{i}$$

We use the explicit formula given by (8.56), viz.,

$$u_{i,j+1}=-u_{i,j-1}+\alpha^2(u_{i-1,j}+u_{i+1,j})+2(1-\alpha^2)u_{i,j} \qquad \text{where } \alpha=\frac{k}{h}<1 \tag{ii}$$

Let $h=0.25$ and $k=0.2$. Hence $\alpha=0.2/0.25=0.8$, so that the stability condition is satisfied. Let $u_{ij}=u(ih, jk)$, so that the boundary conditions become

$$u_{0,j}=0 \tag{iii}$$

$$u_{4,j}=0 \tag{iv}$$

$$u_{i,0}=\sin^3(\pi ih), \quad i=1,\ 2,\ 3,\ 4 \tag{v}$$

Presented By: http://www.ebooksuit.com

and

$$u_{i,1} - u_{i,-1} = 0 \quad \text{so that} \quad u_{i,-1} = u_{i,1}. \tag{vi}$$

Substituting the value of $\alpha = 0.8$, eq. (ii) becomes

$$u_{i,j+1} = -u_{i,j-1} + 0.64\,(u_{i-1,j} + u_{i+1,j}) + 2(0.36)\,u_{i,j} \tag{vii}$$

At the first step, $j = 0$ and the above equation becomes

$$u_{i,1} = -u_{i,-1} + 0.64\,(u_{i-1,j} + u_{i+1,j}) + 2(0.36)u_{i,j}$$

or,

$$u_{i,1} = 0.32\,(u_{i-1,j} + u_{i+1,j}) + 0.36\,u_{i,j}, \tag{viii}$$

using (vi)

Hence
$$\begin{aligned} u_{1,1} &= 0.32\,(u_{0,j} + u_{2,0}) + 0.36u_{1,0} \\ &= 0.32\,(0+1) + 0.36\,(0.3537) \\ &= 0.4473. \end{aligned}$$

The exact value $u(0.25, 0.2) = 0.4838$.

Again,

$$u_{2,1} = 0.32\,(0.3537 + 0.3537) + 0.36\,(1.0) = 0.5867$$

Exact value = 0.5296.

Finally,

$$u_{3,1} = 0.32\,(1.0 + 0) + 0.36\,(0.3537) = 0.4473$$

Exact value = 0.4838.

The computations can be continued for $j = 1, 2, \ldots$

Example 8.11 Solve the boundary-value problem $u_{tt} = 4u_{xx}$ subject to the conditions:

$$u(0,t) = 0 = u(4,t), \qquad u_t(x,0) = 0, \qquad u(x,0) = 4x - x^2.$$

We take $h = 1$ and $\alpha = 1$ so that $k = 1/2 = 0.5$.

Since $u(0,t) = u(4,t) = 0$, u vanishes from $x = 0$ to $x = 4$, i.e.

$$u_{0,j} = u_{4,j} = 0 \qquad \text{for all } j.$$

Again, since $u_t(x,0) = 0$, we have

$$\frac{u_{i,j+1} - u_{i,j-1}}{2k} = 0,$$

or

$$u_{i,1} - u_{i,-1} = 0 \qquad \text{for } j = 0.$$

The above relation shows that the values of u are the same for $j = 1$ and $j = -1$.

Presented By: http://www.ebooksuit.com

Finally, $u(x, 0) = 4x - x^2$ gives

$$u_{i,0} = 4i - i^2, \qquad \text{since } h = 1.$$

Then

$$u_{0,0} = 0, \quad u_{1,0} = 3, \quad u_{2,0} = 4, \quad u_{3,0} = 3, \quad u_{4,0} = 0.$$

Now, for $\alpha = 1$, Eq. (8.56) becomes

$$u_{i,j+1} = -u_{i,j-1} + u_{i-1,j} + u_{i+1,j} \tag{i}$$

For $j = 0$, the above relation gives

$$u_{i,1} = -u_{i,-1} + u_{i-1,0} + u_{i+1,0}$$

or

$$u_{i,1} = \frac{1}{2}(u_{i-1,0} + u_{i+1,0}), \qquad \text{since } u_{i,1} = u_{i,-1}.$$

Figure 8.12

From Fig. 8.12, we obtain

$$u_{1,1} = \frac{1}{2}(u_{0,0} + u_{2,0}) = \frac{1}{2}(0 + 4) = 2,$$

$$u_{2,1} = \frac{1}{2}(3 + 3) = 3,$$

$$u_{3,1} = \frac{1}{2}(4 + 0) = 2.$$

For $k = 1$, we use eq. (i) with $j = 1$:

$$u_{i,2} = -u_{i,0} + u_{i-1,1} + u_{i+1,1}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

CHAPTER

Numerical Solution of Integral Equations

9.1 INTRODUCTION

Any equation in which the unknown function appears under the integral sign is known as an *integral equation*. When the limits of the integrals are constants, we have a *Fredholm equation*. For example, equations of the form

$$\int_a^b K(x,t)\ f(t)\ dt = \phi(x) \tag{9.1}$$

and

$$\lambda \int_a^b K(x,t)\ f(t)\ dt = f(x) + \phi(x) \tag{9.2}$$

are called *linear Fredholm integral equations of the first and second kinds* respectively. In each case the unknown function is $f(x)$ and occurs to the first degree, $\phi(x)$ is a known function and the *kernel* $K(x, t)$ is also known. If the constant b in Eqs. (9.1) and (9.2) is replaced by x, the variable of integration, the equations are called *Volterra integral equations*. For example,

$$\lambda \int_a^x K(x,t)\ f(t)\ dt = f(x) + \phi(x) \tag{9.3}$$

is the Volterra integral equation of the second kind.

If $\phi(x) = 0$ in Eq. (9.2), then the equation is called *homogeneous*, otherwise *nonhomogeneous*. For nonhomogeneous equations, λ is a numerical parameter

Presented By: http://www.ebooksuit.com

whereas for homogeneous equations it is an *eigenvalue parameter* because in such a case the integral equation presents an eigenvalue problem in which the objective is to determine those values of λ, called the *eigenvalues* for which the integral equation possesses nontrivial solutions called *eigenfunctions*.

If the kernel $K(x, t)$ 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 equation is said to be *singular*.

To solve an integral equation of any type is to find the unknown function satisfying that equation. In practical cases, however, the solution of an integral equation by analytical techniques is out of the question, and hence it would be necessary to adopt a numerical method of solution.

Fredholm integral equations, particularly those of the second kind, occur quite frequently in practice and hence, we restrict ourselves, in this chapter, to a brief treatment of numerical methods for the solution of nonhomogeneous linear Fredholm integral equations of the second kind. Before presenting these methods, it would be instructive to demonstrate the relationship between integral equations and initial-value problems. This is shown in Example 9.1

Example 9.1 We consider the initial-value problem

$$y'' + y = 0, \qquad \text{with } y(0) = 0 \text{ and } y'(0) = 1 \tag{9.4}$$

Let

$$\frac{d^2y}{dx^2} = u(x) \tag{9.5}$$

Integrating both sides of (9.5) 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

$$y(x) = \int_0^x (x - t)\,u(t)\,dt + x \tag{9.6}$$

Substituting (9.5) and (9.6) in (9.4), 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.

Presented By: http://www.ebooksuit.com

9.2 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 [1971].

9.2.1 Method of Degenerate Kernels

We consider the integral equation

$$f(x) - \int_a^b K(x,t) f(t)\, dt = \phi x \tag{9.7}$$

A kernel $K(x, t)$ is said to be *degenerate* if it can be expressed in the form

$$K(x,t) = \sum_{i=1}^{n} u_i(x)\, v_i(t) \tag{9.8}$$

Substituting this in (9.7), We obtain

$$f(x) - \sum_{i=1}^{n} \int_a^b u_i(x)\, v_i(t)\, f(t)\, dt = \phi(x) \tag{9.9}$$

Setting

$$\int_a^b v_i(t)\, f(t)\, dt = A_i \tag{9.10}$$

Eq. (9.9) gives

$$f(x) = \sum_{i=1}^{n} A_i\, u_i(x) + \phi(x) \tag{9.11}$$

The constants A_i are still to be detemined, but substituting from (9.11) in (9.10), 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

$$\sum^{n} A_j \int_a^b v_i(t)\, u_j(t)\, dt + \int_a^b v_i(t)\, \phi(t)\, dt = A_i, \tag{9.12}$$

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. (9.11) then gives $f(x)$.

Presented By: http://www.ebooksuit.com

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 (9.48) 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 9.3.

Example 9.2 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

$$f(x) = \frac{2}{2-\lambda} \sin x \quad (\lambda \neq 2).$$

By direct substitution the reader should verify that this is the solution of the given integral equation.

Example 9.3 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.

Presented By: http://www.ebooksuit.com

Hence the given integral equation becomes

$$f(x) = \frac{1}{2}(e^{-x} + 3x - 1) + \int_0^1 (-x^2t^2 + \frac{1}{2}x^3t^4) f(t)\,dt$$

$$= \frac{1}{2}(e^{-x} + 3x - 1) - x^2 \int_0^1 t^2 f(t)\,dt + \frac{1}{2}x^3 \int_0^1 t^4 f(t)\,dt$$

$$= \frac{1}{2}(e^{-x} + 3x - 1) - k_1x^2 + \frac{1}{2}k_2x^3, \tag{i}$$

where

$$K_1 = \int_0^1 t^2 f(t)\,dt \tag{ii}$$

and

$$K_2 = \int_0^1 t^4 f(t)\,dt. \tag{iii}$$

Substituting for $f(t)$ from (i) in (ii), we obtain

$$K_1 = \int_0^1 t^2 \left[\frac{1}{2}(e^{-t} + 3t - 1) - K_1t^2 + \frac{1}{2}K_2t^3\right] dt. \tag{iv}$$

Since

$$\int_0^1 t^2e^{-t}\,dt = 2 - \frac{5}{e},$$

eq. (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

$$K_1 = 0.2522 \text{ and } K_2 = 0.1685.$$

Hence the solution of the given integral equation is

$$f(x) = \frac{1}{2}(e^{-x} + 3x - 1) - 0.2522x^2 + \frac{1}{2}(0.1685)x^3.$$

Presented By: http://www.ebooksuit.com

9.2.2 Quadrature Methods

We consider the integral equation in the form

$$f(x) - \int_a^b K(x,t)\, f(t)\, dt = \phi(x). \tag{9.13}$$

Since a definite integral can be closely approximated by a quadrature formula, we approximate the integral term in (9.13) by a formula of the form

$$\int_a^b F(x)\, dx = \sum_{m=1}^{n} A_m F(x_m), \tag{9.14}$$

where A_m and x_m are the weights and abscissae, respectively. Consequently, (9.13) can be written as

$$f(x) - \sum_{m=1}^{n} A_m K(x, t_m) f(t_m) = \phi(x), \tag{9.15}$$

where $t_1, t_2, \ldots, t_n$ are points in which the interval (a, b) is subdivided. Further, Eq. (9.15) 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

$$f(t_i) - \sum_{m=1}^{n} A_m K(t_i, t_m) f(t_m) = \phi(t_i), \quad i = 1, 2, \ldots, n. \tag{9.16}$$

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. (9.15) 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 9.4 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

$$f(x) - \frac{1}{4}\left[x f_0 + 2\left(x + \frac{1}{2}\right) f_1 + (x+1) f_2\right] = \frac{3}{2}x - \frac{5}{6}, \qquad \text{where } f_i = f(x_i).$$

Presented By: http://www.ebooksuit.com

Setting $x = t_i$, where $t_0 = 0$, $t_1 = 1/2$ and $t_2 = 1$, this gives the system of equations

$$12f_0 - 3f_1 - 3f_2 = -10$$

$$-3f_0 + 12f_1 - 9f_2 = -2$$

$$-3f_0 - 9f_1 + 6f_2 = 8$$

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

$$6f_0 - 2f_1 - f_2 = -5$$

$$-f_0 + 4f_1 - 3f_2 = -1$$

$$-f_0 - 6f_1 + 4f_2 = 4.$$

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 9.5 The integral equation

$$y(x) + \int_{-1}^{1} K(x, s)\, y(s)\, ds = 1, \tag{i}$$

where

$$K(x, s) = \frac{1}{\pi} \frac{1}{1 + (x - s)^2} \tag{ii}$$

Presented By: http://www.ebooksuit.com

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:

$$y_i + \frac{h}{2}K(x_i, s_0)y_0 + \frac{h}{2}K(x_i, s_n)y_n + h\sum_{j=1}^{n-1} K(x_i, s_j)y_j = 1 \qquad \text{(iii)}$$

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

9.2.3 Use of Chebyshev Series

We consider the Fredholm integral equation in the form

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

9.2.4 The Cubic Spline Method

We know that in the interval $x_{j-1} \le x \le x_j$, $s(x)$ is given by

$$s(x) = M_{j-1}\frac{(x_j - x)^3}{6h} + M_j\frac{(x - x_{j-1})^3}{6h}$$

$$+\left(y_{j-1} - \frac{h^2}{6}M_{j-1}\right)\frac{x_j - x}{h} + \left(y_j - \frac{h^2}{6}M_j\right)\frac{x - x_{j-1}}{h} \qquad (9.33)$$

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 (9.17) by using (9.33), we obtain

$$y(x_i) + \sum_{j=1}^{N}\int_{s_{j-1}}^{s_j} K(x, s)\left[M_{j-1}\frac{(s_j - s)^3}{6h} + M_j\frac{(s - s_{j-1})^3}{6h}\right.$$

$$\left.+\left(y_{j-1} - \frac{h^2}{6}M_{j-1}\right)\frac{s_j - s}{h} + \left(y_j - \frac{h^2}{6}M_j\right)\frac{(s - s_{j-1})}{h}\right]ds$$

$$= f(x_i), \quad i = 0, 1, 2, \ldots, N \qquad (9.34)$$

Putting $s = s_{j-1} + ph$, the above equation simplifies to

$$y(x_i) + h\sum_{j=1}^{N}\int_0^1 K(x_i, s_{j-1} + ph)\left[M_{j-1}\frac{(1-p)^3h^2}{6} + M_j\frac{p^3h^2}{6}\right.$$

$$\left.+\left(y_{j-1} - \frac{h^2}{6}M_{j-1}\right)(1-p) + \left(y_j - \frac{h^2}{6}M_j\right)p\right]dp$$

$$= f(x_i), \quad i = 0, 1, 2, \ldots, N \qquad (9.35)$$

In (9.35), the integrals

$$\int_0^1 K\left(x_i, s_{j-1} + ph\right)p^m\, dp, \qquad m = 0, 1, 2 \text{ and } 3, \qquad (9.36)$$

have to be evaluated. This can be done either analytically (wherever possible) or alternatively, by numerical techniques. When these integrals are evaluated, Eqs. (9.35) together with the relations

$$\left.\begin{aligned} \frac{h}{6}M_{j-1} + \frac{2h}{3}M_j + \frac{h}{6}M_{j+1} &= \frac{y_{j-1} - 2y_j + y_{j+1}}{h} \\ j &= 1, 2, \ldots, N-1 \\ \text{and} \qquad M_0 &= M_N = 0 \end{aligned}\right\} \qquad (9.37)$$

Presented By: http://www.ebooksuit.com

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 9.7 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 (9.36) 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

Presented By: http://www.ebooksuit.com

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 to error in the method. It seems probable that the condition of continuity of the kernel may be relaxed, and the advantage to be achieved by using unequal intervals may also be explored. Finally, the solution obtained by the spline method can be improved upon by regarding it as the initial iterate in an iterative method of higher order convergence.

9.3 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

$$f(x) + \int_a^b K(x,t)\, f(t)\, dt = \phi(x), \qquad a \le x \le b. \tag{9.38}$$

Let $b - a = nh$ and $t_j = a + jh$, $j = 0, 1, \ldots, n$ so that $t_0 = a$ and $t_n = b$. Then (9.38) can be written as

$$f(x) + \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} K(x,t)\, f(t)\, dt = \phi(x). \tag{9.39}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

$$= \frac{1}{h\pi} \tan^{-1}\left[\frac{h}{1+h^2(i-j)(i-j-1)}\right]$$

and

$$X1(i,j) = \int_0^1 pK(x_i, t_j + ph)\, dp$$

$$= \frac{1}{\pi}\int_0^1 \frac{p\, dp}{1+h^2(i-j-p)^2}$$

$$= \frac{1}{2h^2\pi}\log\left[\frac{1+h^2(i-j-1)^2}{1+h^2(i-j)^2}\right] + (i-j)X0(i,j)$$

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 (9.41) the equations

$$1.076f_0 + 0.126f_1 + 0.081f_2 + 0.050f_3 + 0.018f_4 = 1.0$$

$$0.071f_0 + 1.153f_1 + 0.126f_2 + 0.081f_3 + 0.029f_4 = 1.0$$

$$0.047f_0 + 0.126f_1 + 1.153f_2 + 0.126f_3 + 0.047f_4 = 1.0$$

$$0.029f_0 + 0.081f_1 + 0.126f_2 + 0.153f_3 + 0.071f_4 = 1.0$$

$$0.018f_0 + 0.050f_1 + 0.081f_2 + 0.126f_3 + 1.076f_4 = 1.0$$

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 9.5) shows that this rule gives better accuracy than the ordinary trapezoidal rule. The order of convergence is h^2 as in the latter rule.

Presented By: http://www.ebooksuit.com

The next example demonstrates the use of generalized quadrature in dealing with kernels having a logarithmic singularity.

Example 9.9 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

$$\left.\begin{aligned} K(s,\sigma) &= \frac{1}{\sqrt{(x-\xi)^2+(y+\eta)^2}}\left\{\frac{x'y-y'(x-\xi)}{y}K(k)\right. \\ &\quad \left. -E(k)\left[\frac{x'y-y'(x-\xi)}{y}+2\eta\frac{x'(y-\eta)-y'(x-\xi)}{(x-\xi)^2+(y-\eta)^2}\right]\right\} \\ K_2 &= \frac{4y\eta}{(x-\xi)^2+(y+\eta)^2},\ x' = \frac{dx}{ds} \end{aligned}\right\} \tag{ii}$$

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:

$$K(s,\sigma) = P(s,\sigma)\log|s-\sigma| + Q(s,\sigma) \tag{iii}$$

where

$$\left.\begin{aligned} P(s,\sigma) &= -\frac{1}{\sqrt{(x-\xi)^2+(y+\eta)^2}}\left\{\frac{x'y-y'(x-\xi)}{y}\frac{2}{\pi}E(k_1)\right. \\ &\quad \left. -2\eta\frac{x'(y-\eta)-y'(x-\xi)}{(x-\xi)^2+(y-\eta)^2}\frac{2}{\pi}[K(k_1)-E(k_1)]\right\} \\ Q(s,\sigma) &= K(s,\sigma) - P(s,\sigma)\log|s,\sigma| \end{aligned}\right\} \tag{iv}$$

Presented By: http://www.ebooksuit.com

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 5 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].

9.4 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

$$y(x) = g(x) + \int_0^a K(x, s)\, y(s)\, ds \tag{9.43a}$$

where

$$K(x, s) = \int_0^\infty f(xz)\, f(sz)\, w(z)\, dz \tag{9.43b}$$

Presented By: http://www.ebooksuit.com

In the method of invariant imbedding, Eq. (9.43) is first rewritten as a Volterra integral equation in the form

$$y(x,t) = g(x) + \int_0^t K(x,s)\, y(s,t)\, ds; \qquad 0 \le x \le t; \quad 0 \le t \le a. \quad (9.44)$$

An essential feature of the method is to convert the Volterra integral equation (9.44) 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 9.10 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

$$y(x) + \int_0^1 K(x,s)\, y(s)\, ds = -\frac{4}{\pi}, \qquad \text{(i)}$$

where

$$K(x,s) = -\frac{2}{\pi} \int_0^\infty e^{-2\xi H} \cos \xi x \cos \xi s \, d\xi, \qquad \text{(ii)}$$

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:

Presented By: http://www.ebooksuit.com

$$\left.\begin{aligned}\frac{dR_{ik}(t)}{dt} &= \left[\cos(tA_k) + \sum_{m=1}^{N} \frac{2}{(1+a_m)^2} F_m W(A_m) \cos(tA_m)\, R_{mk}(t)\right] \\ &\quad \times \left[\cos(tA_i) + \sum_{m=1}^{N} \frac{2}{(1+a_m)^2} F_m W(A_m) \cos(tA_m)\, R_{im}(t)\right] \\ R_{ik}(0) &= 0\end{aligned}\right\} \text{(iii)}$$

and

$$\left.\begin{aligned}\frac{de_i(t)}{dt} &= \left[g(t) + \sum_{m=1}^{N} \frac{2}{(1+a_m)^2} F_m W(A_m) \cos(tA_m)\, e_m(t)\right] \\ &\quad \times \left[\cos tA_i + \sum_{m=1}^{N} \frac{2}{(1+a_m)^2} F_m W(A_m) \cos(tA_m)\, R_{im}(t)\right] \\ e_i(0) &= 0, \qquad 1 \le i \le N, \quad 0 \le t \le 1.\end{aligned}\right\} \text{(iv)}$$

where

$$e_i(t) = e(A_i, t)$$

and finally,

$$y(x,t) = g(x) + \sum_{m=1}^{N} \frac{2F_m}{(1+a_m)^2} W(A_m) \cos xA_m\, e_m(t), \qquad 0 \le x \le t \le 1. \text{(v)}$$

In (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:

Presented By: http://www.ebooksuit.com

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

9.1. Verify whether the functions given below are solutions of the integral equations indicated against them:

(a) $f(x) = 1 : f(x) + \int_0^1 x(e^{xt} - 1)\, f(t)\, dt = e^x - x$

(b) $u(t) = e^t : u(t) + \lambda \int_0^1 \sin(t\,x)\, u(x)\, dx = 1$

(c) $f(x) = \sin\frac{\pi x}{2} : f(x) - \frac{\pi^2}{4} \int_0^1 K(x, t)\, f(t)\, dt = \frac{x}{2}$

where

$$K(x, t) = \begin{cases} x\left(1 - \frac{t}{2}\right), & 0 \le x \le t \\ t\left(1 - \frac{x}{2}\right), & t \le x \le 1 \end{cases}$$

Presented By: http://www.ebooksuit.com

(d) $\phi(x) = x$: $\phi(x) = \dfrac{15x-2}{18} + \dfrac{1}{3}\displaystyle\int_0^1 (x+t)\,\phi(t)\,dt$

(e) $f(x) = x - 1$: $f(x) - \displaystyle\int_0^1 (x+t)\,f(t)\,dt = \dfrac{3}{2}x - \dfrac{5}{6}$.

9.2. Solve the following integral equations with degenerate kernels:

(a) $f(x) - \lambda \displaystyle\int_{-\pi/4}^{\pi/4} \tan s\, f(s)\,ds = \cot x$.

(b) $f(x) - \lambda \displaystyle\int_0^{\pi/2} \sin x \cos t\, f(t)\,dt = \sin x$.

(c) $f(x) - \lambda \displaystyle\int_0^{\pi} \sin(x-u)\, f(u)\,du = \cos x$.

(d) $f(x) = \sin x + \displaystyle\int_0^1 [1 - x\cos(xt)]\, f(t)\,dt$

9.3. Solve the integral equations given in problem 1(d) and (e) by

(i) the trapezoidal method

(ii) the cubic spline method.

In each case, divide the range into two equal subintervals and approximate to the solution. Compare your results with the exact solution.

Presented By: http://www.ebooksuit.com

CHAPTER

The Finite Element Method

10.1 INTRODUCTION

In Chapters 7 and 8 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 and partial 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 10.1.2. 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 10.2, we discuss two important methods of approximation, viz., the Rayleigh–Ritz method and the Galerkin

Presented By: http://www.ebooksuit.com

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 10.1.1), 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 our discussion to the use of triangular elements in the solution of simple two-dimensional problems (see Section 10.4.2).

Examples of typical finite elements are shown in Fig. 10.1.

(a) Line element (b) Triangular and Quadrilateral elements

Figure 10.1

10.1.1 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

Presented By: http://www.ebooksuit.com

integration is performed over a region. Consider, for example, the integral defined by

$$I(y) = \int_a^b f(x, y, y')\, dx, \tag{10.1}$$

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 (10.1). 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

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0. \tag{10.2}*$$

Similarly, for functionals of the form

$$I(y) = \int_a^b f(x, y, y', y'')\, dx \tag{10.3}$$

the Euler–Lagrange equation takes the form

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial f}{\partial y''}\right) = 0. \tag{10.4}$$

The Euler–Lagrange equation (10.2) 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 [1997, a].

Presented By: http://www.ebooksuit.com

Example 10.1 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

$$\delta \int_a^b f\, y\, x = \int_a^b f\, \delta y\, dx$$

$$= \int_a^b \frac{d^2y}{dx^2} \delta y\, dx, \text{ since } f(x) = \frac{d^2y}{dx^2}.$$

$$= \left[\frac{dy}{dx}\delta y\right]_a^b - \int_a^b \frac{dy}{dx}\frac{d}{dx}(\delta y)dx, \text{ on integrating by parts}$$

$$= -\int_a^b \frac{dy}{dx}\frac{d}{dx}(\delta y)\, dx, \text{ since } \delta y(a) = \delta y(b) = 0$$

$$= -\int_a^b \frac{dy}{dx}\delta\left(\frac{dy}{dx}\right)dx, \text{ since } \frac{d}{dx}(\delta y) = \delta\left(\frac{dy}{dx}\right)$$

$$= -\int_a^b \frac{1}{2}\delta\left(\frac{dy}{dx}\right)^2 dx$$

$$= -\delta \int_a^b \frac{1}{2}\left(\frac{dy}{dx}\right)^2 dx.$$

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

$$I(v) = \int_a^b \left[fv + \frac{1}{2}\left(\frac{dv}{dx}\right)^2\right] dx. \tag{iii}$$

Presented By: http://www.ebooksuit.com

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) $$\frac{d^2y}{dx^2} = f(x),\ y(a) = y(b) = 0 \tag{10.5}$$

$$I(v) = \int_a^b v(2f - v'')\,dx. \tag{10.6}$$

(ii) $$\frac{d^2y}{dx^2} + ky = x^2,\ 0 < x < 1;\ y(0) = 0,\ \left(\frac{dy}{dx}\right)_{x=1} = 1 \tag{10.7}$$

$$I(v) = \frac{1}{2}\int_0^1 \left[\left(\frac{dv}{dx}\right)^2 - kv^2 + 2vx^2\right] dx - v(1). \tag{10.8}$$

(iii) $$x^2y'' + 2xy' = f(x),\ y(a) = y(b) = 0 \tag{10.9}$$

$$I(v) = \int_a^b v\left[2f - \frac{d}{dx}(x^2y')\right] dx. \tag{10.10}$$

(iv) $\nabla^2 u = 0,\ u = 0$ on the boundary C of R. (10.11)

$$I(v) = \iint_R \frac{1}{2}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dx\,dy. \tag{10.12}$$

(v) $\nabla^2 u = -f,\ u = 0$ on the boundary C of R. (10.13)

$$I(v) = \iint_R \left\{\frac{1}{2}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] - uf\right\} dx\,dy. \tag{10.14}$$

(vi) $$\left.\begin{aligned} EI\frac{d^4y}{dx^4} + ky &= f(x),\ 0 < x < l \\ y = 0 = \frac{d^2y}{dx^2} &\text{ at } x = 0,\ l \end{aligned}\right\} \tag{10.15}$$

$$I(v) = \frac{1}{2}\int_0^l \left[EI\left(\frac{d^2v}{dx^2}\right)^2 + kv^2 - 2vf\right] dx. \tag{10.16}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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 10.3 Solve the boundary-value problem defined by

$$y'' + y = -x, \qquad 0 < x < 1 \tag{i}$$

Presented By: http://www.ebooksuit.com

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

$$I(v) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j q_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j r_{ij} + 2\sum_{i=1}^{n} \alpha_i p_i \tag{xi}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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.

10.2.2 The Galerkin Method

The Rayleigh–Ritz method discussed in Section 10.2.1 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.

To explain Galerkin's method, we consider the boundary value problem defined by

$$y'' + p(x)y' + q(x)y = f(x), \qquad a < x < b \tag{10.27}$$

with the boundary conditions

$$\left.\begin{aligned} p_0 y(a) + q_0 y'(a) &= l_0 \\ p_1 y(b) + q_1 y'(b) &= l_1 \end{aligned}\right\} \tag{10.28}$$

To find an approximate solution of the problems (10.27) and (10.28), we choose base functions $\phi_i(x)$ as in the Rayleigh–Ritz method.

Then an approximate solution $v(x)$ is assumed to be a linear combination of the ϕ_i, i.e. $v(x)$ is written as

$$v(x) = \sum_{i=1}^{n} \alpha_i \, \phi_i(x). \tag{10.29}$$

Now, $v(x)$ will not, in general, satisfy (10.27), but produces a *residual* or *discrepancy*. This is equal to the difference between the left-hand and right-hand sides of Eq. (10.27) when on the left side $y(x)$ is replaced by $v(x)$. If $R(v)$ is the residual, we then write

$$R(v) = v'' + p(x)v' + q(x)v - f(x). \tag{10.30}$$

Presented By: http://www.ebooksuit.com

Taking the weight function as $\psi_i(x)$, we write

$$\int_a^b \psi_i(x)\, R(v)\, dx = 0, \tag{10.31}$$

which yields a system of equations for the unknown parameters α_i and can be solved. In Galerkin's method, we usually take $\psi_i(x) = \phi_i(x)$. The method is illustrated with the following example:

Example 10.5 We consider again the problem of Example 10.3, viz.,

$$y'' + y = -x, \quad 0 < x < 1 \tag{i}$$

$$y(0) = y(1) = 0 \tag{ii}$$

As our first approximation, we choose

$$v(x) = \alpha_1 \phi_1(x) = \alpha_1 x(1-x), \tag{iii}$$

where $\phi_1(0) = \phi_1(1) = 0$.

Substituting for v in (i), we obtain

$$R(v) = v'' + v + x \tag{iv}$$

Hence, using (10.31), we write

$$\int_0^1 (v'' + v + x)\, \phi_1(x)\, dx = 0$$

$$\int_0^1 (v'' + v + x)\, x(1-x)\, dx = 0 \tag{v}$$

Now,

$$\int_0^1 v''\, x(1-x)\, dx = [v' x(1-x)]_0^1 - \int_0^1 v'(1-2x)\, dx = -\int_0^1 v'(1-2x)\, dx,$$

since the first expression on the right vanishes. Now,

$$\int_0^1 v''\, x(1-x)\, dx = -[v(1-2x)]_0^1 - \int_0^1 -2v\, dx = -2\int_0^1 v\, dx, \quad \text{since } v(0)=v(1)=0.$$

Hence (v) becomes

$$\int_0^1 [-2v + vx(1-x) + x^2(1-x)]\, dx = 0, \tag{vi}$$

which gives on simplification $\alpha_1 = 5/18 = 0.2778$. (vii)

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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 (ii), we obtain

$$I(v) = \int_0^1 \int_0^1 \alpha xy(x-1)(y-1)[2k - 2\alpha y(y-1) - 2\alpha x(x-1)]\, dx\, dy. \tag{v}$$

Let

$$\left.\begin{aligned} a &= \int_0^1 \int_0^1 xy(x-1)(y-1)\, dx\, dy = \frac{1}{36} \\ b &= \int_0^1 \int_0^1 xy^2(x-1)(y-1)^2\, dx\, dy = -\frac{1}{180} \\ c &= \int_0^1 \int_0^1 x^2 y(x-1)^2(y-1)\, dx\, dy = -\frac{1}{180}. \end{aligned}\right\} \tag{vi}$$

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.

10.4 THE 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

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Instead of Eq. (10.40), we now have

$$K_{ij}^{(e)} y_j^{(e)} = F_i^{(e)}, \tag{10.50}$$

where $K_{ij}^{(e)}$ and $F_i^{(e)}$ are given by (10.41) and (10.42).

With the choice of $\phi_i^{(e)}(x)$ as in Eq. (10.48), we now demonstrate the computation of $K^{(e)}$ and $F^{(e)}$. In particular, we choose $a(x) = 1$ and $f = 2$. With $h_e = x_e - x_{e-1}$, we obtain

$$\frac{d\phi_1^{(e)}}{dx} = -\frac{1}{h_e} \quad \text{and} \quad \frac{d\phi_2^{(e)}}{dx} = \frac{1}{h_e} \tag{10.51}$$

where

$$\left.\begin{aligned} K_{11} &= \int_{x_{e-1}}^{x_e} \left(-\frac{1}{h_e}\right)^2 dx = \frac{1}{h_e} \\ K_{12} &= \int_{x_{e-1}}^{x_e} -\frac{1}{h_e^2}\, dx = -\frac{1}{h_e} = K_{21} \\ K_{22} &= \int_{x_{e-1}}^{x_e} \frac{1}{h_e^2}\, dx = \frac{1}{h_e} \end{aligned}\right\} \tag{10.52}$$

and

$$\left.\begin{aligned} F_1^{(e)} &= 2 \int_{x_{e-1}}^{x_e} \frac{x_e - x}{h_e}\, dx + D_1^{(e)} = h_e + D_1^{(e)} \\ F_2^{(e)} &= 2 \int_{x_{e-1}}^{x_e} \frac{x - x_{e-1}}{h_e}\, dx + D_2^{(e)} = h_e + D_2^{(e)}. \end{aligned}\right\} \tag{10.53}$$

As a particular case, we consider the following example.

Example 10.7 We consider the following problem defined by

$$\frac{d^2y}{dx^2} = -2, \quad 0 < x < 1, \qquad y(0) = 0,\ y'(1) = 0. \tag{i}$$

The exact solution of the above problem is given by

$$y(x) = 2x - x^2 \tag{ii}$$

Comparison with (10.32) shows that $a(x) = 1$ and $f(x) = 2$.

Presented By: http://www.ebooksuit.com

(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 (10.48) and (10.49), 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 10.7. 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 10.7, the correspondence can be expressed mathematically as follows:

$$y_1^{(1)} = Y_1, \quad y_2^{(1)} = Y_2 = y_1^{(2)}, \quad y_2^{(2)} = Y_3.$$

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

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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 $Y_1 = 0$.

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

To find the variational form of Eq. (10.54), we multiply it with the test function $v(x, y)$ and integrate the result over a typical element R_e, to obtain

$$0 = \iint_{R_e} \left[\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right) - vf \right] dx\,dy - \int_{c_e} v q_n \, ds \tag{10.56}$$

where

$$q_n = \eta_x \frac{\partial u}{\partial x} + \eta_y \frac{\partial u}{\partial y},$$

η_x and η_y being the direction cosines of a unit normal $\hat{n}$ on the boundary c_e and ds is an arc length of an infinitesimal element along the boundary.

The next step in the finite element solution of this problem is to set up a finite element model of the given equation. To do this, we approximate u by the expression

$$u = \sum_{j=1}^{n} u_j \phi_j, \tag{10.57}$$

where $u_j = u(x_j, y_i)$ and the ϕ_j have the property

$$\phi_i(x_j, y_j) = \delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j. \end{cases} \tag{10.58}$$

Substituting (10.57) in (10.56) and putting $v = \phi_i$, we obtain

$$0 = \sum_{j=1}^{n} \iint_{R_e} \left(\frac{\partial \phi_i}{\partial x}\frac{\partial \phi_j}{\partial x} + \frac{\partial \phi_i}{\partial y}\frac{\partial \phi_j}{\partial y} \right) u_j \, dx\,dy - \iint_{R_e} f\phi_i \, dx\,dy - \int_{C_e} \phi_i q_n \, ds \tag{10.59}$$

for $i = 1, 2, \ldots, n$.

Equation (10.59) can be written in the form

$$\sum_{j=1}^{n} K_{ij}^{(e)} u_j^{(e)} = F_i^{(e)} \tag{10.60}$$

where

$$K_{ij}^{(e)} = \iint_{R_e} \left(\frac{\partial \phi_i}{\partial x}\frac{\partial \phi_j}{\partial x} + \frac{\partial \phi_i}{\partial y}\frac{\partial \phi_j}{\partial y} \right) dx\,dy \tag{10.61}$$

and

$$F_i^{(e)} = \iint_{R_e} f\phi_i \, dx\,dy + \int_{C_e} q_n \phi_i \, ds. \tag{10.62}$$

Equation (10.60) represents the finite element model of the Poisson equation.

Presented By: http://www.ebooksuit.com

We next consider a triangular element (see Fig. 10.3) in which the nodes are numbered in the counter-clockwise direction and derive the interpolation functions for it. We assume the interpolating polynomial in such a way that the number of terms in it equals the number of nodes in the triangular element. Accordingly, we assume

$$u(x, y) = a_1 + a_2 x + a_3 y \tag{10.63}$$

Figure 10.3

as the required approximation. We also set

$$u(x_i, y_i) = u_i, \quad i = 1, 2, 3 \tag{10.64}$$

where (x_i, y_i), $i = 1, 2, 3$ denote the three vertices of the triangle. Substituting (10.64) in (10.63), we obtain

$$\left.\begin{aligned} u_1 &= a_1 + a_2 x_1 + a_3 y_1 \\ u_2 &= a_1 + a_2 x_2 + a_3 y_2 \\ u_3 &= a_1 + a_2 x_3 + a_3 y_3. \end{aligned}\right\} \tag{10.65}$$

Solving Eqs. (10.65), we obtain

$$\left.\begin{aligned} a_1 &= \frac{1}{2\Delta_e}\begin{vmatrix} u_1 & u_2 & u_3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} \\ a_2 &= \frac{1}{2\Delta_e}\begin{vmatrix} u_1 & u_2 & u_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{vmatrix} \\ a_3 &= \frac{1}{2\Delta_e}\begin{vmatrix} u_1 & u_2 & u_3 \\ 1 & 1 & 1 \\ x_1 & x_2 & x_3 \end{vmatrix}, \end{aligned}\right\} \tag{10.66}$$

where

$$\Delta_e = \text{Area of the triangle} = \frac{1}{2}\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}. \tag{10.67}$$

Presented By: http://www.ebooksuit.com

Substituting for a_1, a_2, a_3 in (10.63) and simplifying, we obtain

$$u(x, y) = \frac{1}{2\Delta_e}[u_1(x_2y_3 - x_3y_2) + u_2(x_3y_1 - x_1y_3) + u_3(x_1y_2 - x_2y_1)]$$
$$+ \frac{1}{2\Delta_e}[u_1(y_2 - y_3) + u_2(y_3 - y_1) + u_3(y_1 - y_2)]x$$
$$+ \frac{1}{2\Delta_e}[u_1(x_3 - x_2) + u_2(x_1 - x_3) + u_3(x_2 - x_1)]y. \tag{10.68}$$

Collecting the coefficients of u_1, u_2 and u_3 in the above, Eq. (10.68) can be written in the form

$$u(x, y) = \sum_{i=1}^{3} u_i \phi_i^{(e)}(x, y), \tag{10.69}$$

where the $\phi_i^{(e)}$ are the linear interpolating functions for the triangular elements under consideration, and are given by

$$\left.\begin{aligned}
\phi_1^{(e)}(x, y) &= \frac{1}{2\Delta_e}\begin{vmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} \\
\phi_2^{(e)}(x, y) &= \frac{1}{2\Delta_e}\begin{vmatrix} 1 & x & y \\ 1 & x_3 & y_3 \\ 1 & x_1 & y_1 \end{vmatrix} \\
\phi_3^{(e)}(x, y) &= \frac{1}{2\Delta_e}\begin{vmatrix} 1 & x & y \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{vmatrix},
\end{aligned}\right\} \tag{10.70}$$

From formulae (10.70), it is easily verified that

$$\phi_i^{(e)}(x_j, y_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \quad \text{and} \quad \sum_{i=1}^{3} \phi_i^{(e)}(x, y) = 1 \tag{10.71}$$

We also have

$$\left.\begin{aligned}
\frac{\partial \phi_1^{(e)}}{\partial x} &= \frac{y_2 - y_3}{2\Delta_e}, \quad \frac{\partial \phi_1^{(e)}}{\partial y} = \frac{x_3 - x_2}{2\Delta_e} \\
\frac{\partial \phi_2^{(e)}}{\partial x} &= \frac{y_3 - y_1}{2\Delta_e}, \quad \frac{\partial \phi_2^{(e)}}{\partial y} = \frac{x_1 - x_3}{2\Delta_e} \\
\frac{\partial \phi_3^{(e)}}{\partial x} &= \frac{y_1 - y_2}{2\Delta_e}, \quad \frac{\partial \phi_3^{(e)}}{\partial y} = \frac{x_2 - x_1}{2\Delta_e}.
\end{aligned}\right\} \tag{10.72}$$

Presented By: http://www.ebooksuit.com

Using Eqs. (10.72) the element matrices $K_{ij}^{(e)}$ and $F_i^{(e)}$ in (10.60) can then be computed easily. These computations will be demonstrated through a simple example.

Example 10.8 We consider a particular case of the problem defined by Eqs. (10.54) and (10.55), viz., the Poisson equation

$$-\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)=2, \quad 0 \le x,\ y \le 1 \tag{i}$$

with the condition

$$u=0 \tag{ii}$$

on the boundary of the square $0 \le x \le 1$, $0 \le y \le 1$.

We divide the square region along the line of symmetry $x = y$ and then consider only the lower-triangular part. We again subdivide the lower triangular part into four triangular elements, as shown in Fig. 10.4. Let the elements be numbered, as shown in the figure, and it is seen that element ①, ② and ④ are symmetrical. Hence the element matrices for these elements will all be of the same type.

Figure 10.4

Now, the vertices 1, 2 and 3 of the element ① are given by (0, 0), (1/2, 0) and (1/2, 1/2) respectively. For this element, we obtain $\Delta_1 = 1/8$ and Eqs. (10.70) give

$$\left.\begin{aligned}
\phi_1^{(1)} &= 4\left[\frac{1}{4}-0+\left(-\frac{1}{2}\right)x\right]=1-2x \\
\phi_2^{(1)} &= 4\left[0+\frac{1}{2}x-\frac{1}{2}y\right]=2(x-y) \\
\phi_3^{(1)} &= 4\left[0+\frac{1}{2}y\right]=2y.
\end{aligned}\right\} \tag{iii}$$

Presented By: http://www.ebooksuit.com

It is easy to see that $\phi_1^{(1)} + \phi_2^{(1)} + \phi_3^{(1)} = 1$, thus verifying (10.71). The element matrices K and F can now be computed easily, using (10.61) and (10.62).

We first obtain the derivatives

$$\left.\begin{aligned} \frac{\partial \phi_1^{(1)}}{\partial x} &= -2, & \frac{\partial \phi_1^{(1)}}{\partial y} &= 0 \\ \frac{\partial \phi_2^{(1)}}{\partial x} &= 2, & \frac{\partial \phi_2^{(1)}}{\partial y} &= -2 \\ \frac{\partial \phi_3^{(1)}}{\partial x} &= 0, & \frac{\partial \phi_3^{(1)}}{\partial y} &= 2 \end{aligned}\right\} \tag{iv}$$

Equation (10.61) now gives

$$\left.\begin{aligned} & K_{11}^{(1)} = \iint_{\Delta_{123}} 4\,dx\,dy = \frac{1}{2},\ K_{12}^{(1)} = \iint -4\,dx\,dy = -\frac{1}{2},\ K_{13}^{(1)} = 0 \\ & K_{21}^{(1)} = -\frac{1}{2},\ K_{22}^{(1)} = 1,\ K_{23}^{(1)} = -\frac{1}{2} \\ & K_{31}^{(1)} = 0,\ K_{32}^{(1)} = -\frac{1}{2},\ K_{33}^{(1)} = \frac{1}{2} \end{aligned}\right\} \tag{v}$$

Similarly Eq. (10.62) yields

$$\left.\begin{aligned} F_1^{(1)} &= \iint_{\Delta_{123}} 2(1-2x)\,dx\,dy + \int_{C_{123}} q_n(1-2x)\,ds \\ &= \int_0^{1/2}\int_0^{1/2} 2(1-2x)\,dx\,dy + I_1^{(1)}, \text{say} \\ &= \frac{1}{12} + I_1^{(1)}, \qquad \text{where } I_1^{(1)} = \int_{C_{123}} q_n(1-2x)\,ds \\ F_2^{(1)} &= \iint_{\Delta_{123}} 2(2x-2y)\,dx\,dy + \int_{C_{123}} q_n(2x-2y)\,ds \\ &= \frac{1}{12} + I_2^{(1)}, \qquad \text{where } I_2^{(1)} = \int_{C_{123}} q_n(2x-2y)\,ds \\ F_3^{(1)} &= \iint_{\Delta_{123}} 4y\,dx\,dy + \int_{C_{123}} q_n(2y)\,ds \\ &= \frac{1}{12} + I_3^{(1)}, \qquad \text{where } I_3^{(1)} = \int_{C_{123}} q_n(2y)\,ds \end{aligned}\right\} \tag{vi}$$

Presented By: http://www.ebooksuit.com

Let the global nodes be U_1, U_2, U_3, U_4, U_5, and U_6 corresponding to the local nodes u_1, u_2, u_3, u_4, u_5 and u_6 at the respective vertices. As there are six nodes, the corresponding matrices will be of order 6. Hence, we obtain for element ①:

$$K^{(1)} = \begin{bmatrix} 1/2 & -1/2 & 0 & 0 & 0 & 0 \\ -1/2 & 1 & -1/2 & 0 & 0 & 0 \\ 0 & -1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } F^{(1)} = \frac{1}{12}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} l_1^{(1)} \\ l_2^{(1)} \\ l_3^{(1)} \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{vii}$$

Since the elements ② and ④ are similar to ①, their element matrices will be of the same type as those of ① given in (vii). Thus, for element ②,

$$K^{(2)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & -1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1/2 & 0 & 1 & -1/2 & 0 \\ 0 & 0 & 0 & -1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } F^{(2)} = \frac{1}{12}\begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ l_1^{(2)} \\ 0 \\ l_2^{(2)} \\ l_3^{(2)} \\ 0 \end{bmatrix}. \tag{viii}$$

Similarly, for element ④,

$$K^{(4)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & -1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1/2 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 0 & -1/2 & 1/2 \end{bmatrix}, \text{ and } F^{(4)} = \frac{1}{12}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ l_1^{(4)} \\ 0 \\ l_2^{(4)} \\ l_3^{(4)} \end{bmatrix}. \tag{ix}$$

Finally, for element ③, we note that the correspondence between its vertices and those of ① is given by $5 \to 1, 3 \to 2$, and $2 \to 3$. Hence, we have

Presented By: http://www.ebooksuit.com

$$K^{(3)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & -1/2 & 0 & 0 & 0 \\ 0 & -1/2 & 1 & 0 & -1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } F^{(3)} = \frac{1}{12}\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ I_3^{(3)} \\ I_2^{(3)} \\ 0 \\ I_1^{(3)} \\ 0 \end{bmatrix}. \tag{x}$$

Assembling the element matrices in (vii), (viii), (ix) and (x) and simplifying, we obtain the matrix equation

$$\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 4 & -2 & -1 & 0 & 0 \\ 0 & -2 & 4 & 0 & -2 & 0 \\ 0 & -1 & 0 & 2 & -1 & 0 \\ 0 & 0 & -2 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \\ U_6 \end{bmatrix} = \frac{1}{12}\begin{bmatrix} 1 \\ 3 \\ 3 \\ 1 \\ 3 \\ 1 \end{bmatrix} + \begin{bmatrix} I_1^{(1)} \\ I_2^{(1)} + I_1^{(2)} + I_3^{(3)} \\ I_3^{(1)} + I_2^{(3)} + I_1^{(4)} \\ I_2^{(2)} \\ I_3^{(2)} + I_1^{(3)} + I_2^{(4)} \\ I_3^{(4)} \end{bmatrix} \tag{xi}$$

From the boundary conditions, we have (see Fig. 10.4)

$$U_1 = U_2 = U_4 = U_5 = U_6 = 0. \tag{xii}$$

Hence, eq. (xi) gives

$$-U_3 = \frac{1}{4} + I_1^{(2)} + I_2^{(1)} + I_3^{(3)} \tag{xiii}$$

$$2U_3 = \frac{1}{4} + I_3^{(1)} + I_2^{(3)} + I_1^{(4)} \tag{xiv}$$

$$-U_3 = \frac{1}{4} + I_3^{(2)} + I_1^{(3)} + I_2^{(4)} \tag{xv}$$

From (xiv), we obtain

$$U_3 = \frac{1}{8} + \frac{1}{2}(I_3^{(1)} + I_1^{(4)} + I_2^{(3)}) \tag{xvi}$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

10.3. $x^2 \dfrac{d^2y}{dx^2} + 2x\dfrac{dy}{dx} = g(x),\ y(0) = y(1) = 0.$

10.4. $\dfrac{d^2y}{dx^2} + p(x)y + q(x) = 0,\ y(a) = y(b) = 0.$

10.5. $\dfrac{d^4y}{dx^4} + ky = f(x),\ 0 < x < 1,\ y = \dfrac{d^2y}{dx^2} = 0$ at $x = 0, 1.$

10.6. $\nabla^2 u = 0,\ u = 0$ on the boundary C of R.

10.7. $\nabla^2 u = -f,\ u = 0$ on the boundary C of R.

Use the Rayleigh–Ritz method to solve the following boundary-value problems (Problems 8–11):

10.8. $\dfrac{d^2y}{dx^2} + 2x = 0,\ y(0) = y(1) = 0.$

10.9. $\dfrac{d^2y}{dx^2} + y = x^2,\ y(0) = y(1) = 0.$ (Use a two-parameter approximate solution).

10.10. $\dfrac{d^2y}{dx^2} + x\dfrac{dy}{dx} + y = 2x,\ y(0) = 1$ and $y(1) = 0.$

10.11. $\dfrac{d^2y}{dx^2} + x\dfrac{dy}{dx} - 2y = 0,\ y'(0) + y(0) = 1$ and $y(1) = 2.$

Compare the two-parameter approximate solution with the analytical solution given by $y = x^2 + 1$. Apply the Galerkin technique to solve the following boundary-value problems (Problems 12–14):

10.12. Exercise 9 above.

10.13. Exercise 11 above.

10.14. $\dfrac{d^2y}{dx^2} - x = 0,\ y(0) = 0,\ y'(1) = -\dfrac{1}{2}.$

10.15. Using Galerkin technique, solve Poisson's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = k, \qquad 0 < x,\ y < 1$$

with $u = 0$ on the boundary C of the region R.

10.16. Use the Galerkin technique to approximate Eq. (10.36) and hence obtain the solution of the boundary-value problem defined by

$$\frac{d^2y}{dx^2} = -2, \qquad 0 < x < 1; \qquad y(0) = 0,\ y'(1) = 0,$$

taking two equal subintervals.

Presented By: http://www.ebooksuit.com

10.17. In the notation of Example 10.8, prove that (a) $l_2^{(3)} = 0$ (b) $l_1^{(4)} = 0$.

10.18. (Reddy). Solve Poisson's equation

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = 1, \qquad 0 < x,\ y < 1,$$

where

$$\left[\frac{\partial u}{\partial x}\right]_{x=0} = \left[\frac{\partial u}{\partial y}\right]_{y=0} = 0; \qquad u(1,\ y) = u(x,\ 1) = 0.$$

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

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.

Presented By: http://www.ebooksuit.com

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., *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.

Presented By: http://www.ebooksuit.com

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.

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*, India, Vol. 8, p. 279, 1968.

Presented By: http://www.ebooksuit.com

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.

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*, India, 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.

Rutishauser, H., *Z. Angew, Math. Phys.*, Vol. 5, p. 233, 1954.

Sastry, S.S., A numerical solution of an integral equationof the second kind occurring in aerodynamics, *Ind. J. Pure and App. Math.*, Vol. 4, p. 838, 1973.

Presented By: http://www.ebooksuit.com

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.

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.

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Presented By: http://www.ebooksuit.com

Hidden page

Presented By: http://www.ebooksuit.com

Introductory Methods of Numerical Analysis

FOURTH EDITION

S.S. Sastry

This completely revised **fourth edition** of the book, appropriate for all engineering undergraduate students, continues to provide a rigorous introduction to the fundamentals of numerical methods required in scientific and technological applications. The book focuses clearly on teaching students numerical methods and in helping them to develop problem-solving skills.

A distinguishing feature of the present edition is that it provides references to MATLAB, IMSL and Numerical Recipes program libraries for implementing the numerical methods described in the book. Several exercises are included to illustrate the use of these libraries.

Additional worked examples and exercises have been added for better appreciation and understanding of the material. Answers to some selected exercises have been provided.

NEW TO THIS EDITION

- Expanded section on cubic splines with the inclusion of linear and quadratic splines.
- A new section on surface fitting by cubic splines.
- A new section on Fourier transforms including a discussion on fast Fourier transforms.
- Enlarged range of about 500 worked examples and exercises.

THE AUTHOR

S.S. SASTRY, Ph.D., Formerly, Applied Mathematics Division, Vikram Sarabhai Space Centre of the Indian Space Research Organization, Thiruvananthapuram. Earlier he taught at BIT, Ranchi.

Rs. 195.00

Prentice-Hall of India
New Delhi
www.phindia.com

Presented By: http://www.ebooksuit.com