CHAPTER ONE

NUMERICAL COMPUTATIONS

1.0 INTRODUCTION

Numerical methods are methods for solving problems on computers by numerical calculations, often giving a table of numbers and/or graphical representations or figures. Numerical methods tend to emphasize the implementation of algorithms. The aim of numerical methods is therefore to provide systematic methods for solving problems in a numerical form. The process of solving problems generally involves starting from an initial data, using high precision digital computers, following the steps in the algorithms, and finally obtaining the results. Often the numerical data and the methods used are approximate ones. Hence, the error in a computed result may be caused by the errors in the data, or the errors in the method or both.

In this chapter, we will describe Taylor's theorem, a few basic ideas and concepts regarding numerical computations, errors considerations, absolute and relative errors, inherent errors, round-off errors and truncation errors, error estimation, general error formulae including approximation of a function, stability and condition.

1.1 TAYLOR'S THEOREM

Taylor's theorem allows us to represent, exactly, and fairly general functions in terms of polynomials with a known, specified and boundable error. Taylor's theorem is stated as follows:

Let $f(x)$ have $n+1$ continuous derivatives on $[a,b]$ for some $n \geq 0$, and $x, x_0 \in [a,b]$.

Then $f(x) = P_n(x) + R_n(x)$ **(1.1)**

for $p_n(x) = \sum_{k=0}^{n} \frac{(x-x_0)^k}{k!} f^{(k)}(x_0)$

and $R_n(x) = \frac{1}{n!}\int_{x_0}^{x}(x-t)^n f^{n+1}(t)dt$ **(1.2)**

Also, there exists a point ξ_x, between x and x_0 such that

$$R_n(x) = \frac{(x-x_0)^{n+1}}{(n+1)!} f^{n+1}(\xi_x) \quad \textbf{(1.3)}$$

where $R_n(x)$ is the remainder.

Taylor's series is an associated formula of Taylor's theorem. Taylor's series gives us a means to predict a function value at one point in terms of the function value and its derivatives at another point.

The Taylor's series expansion is defined by

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \frac{f'''(x_0)}{3!}(x-x_0)^3 + \ldots + \frac{f^n(x_0)}{n!}(x-x_0)^n + R_n \quad \textbf{(1.4)}$$

We note that **Eq. (1.4)** represents an infinite series. The remainder term R_n is included to account for all terms from (n + 1) to infinity:

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)}(x-x_0)^{n+1} \quad \textbf{(1.5)}$$

where the subscript n connotes that this is the remainder for the n^{th} order approximation and ξ is a value of x_i that lies somewhere between x_0 and x.

We can rewrite the Taylor's series in **Eq. (1.5)** by defining a step size $h = x - x_0$ as

$$f(x) = f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2!}h^2 + \frac{f'''(x_0)}{3!}h^3 + \ldots + \frac{f^n(x_0)}{n!}h^n + R_n \quad \textbf{(1.6)}$$

where the remainder term R_n is given by

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)}h^{n+1} \quad \textbf{(1.7)}$$

NB: when $x_0 = 0$ then **Eq. (1.6)** forms a M*aclaurin's* series.

Exercise 1.1

Give the *Maclaurin's* series expansion of the following:

a. e^x b. $\sin x$ c. $\cos x$

(Hint: put $x_i = 0$ *)*

Solution: (@ Lectures)

1.2 ERROR CONSIDERATIONS

Sources of Errors: When a computational procedure is involved in solving a scientific-mathematical problem, errors often will be involved in the process. A rough classification of the kinds of original errors that might occur is as follows:

- *Modeling Errors*: *Mathematical modeling* is a process when mathematical equations are used to represent a physical system. This modeling introduces errors and are called *modeling errors*.

- *Blunders and Mistakes*: Blunders occur at any stage of the mathematical modeling process and consist to all other components of error. Blunders can be avoided by sound knowledge of fundamental principles and with taking proper care in approach and design to a solution. Mistakes are due to the programming errors.

- *Machine Representation and Arithmetic Errors*: These errors are inevitable when using floating-point arithmetic when using computers or calculators. Examples are rounding and chopping errors.

- *Mathematical Approximation Errors*: This error is also known as a *truncation error* or *discretization error*. These errors arise when an approximate formulation is made to a problem that otherwise cannot be solved exactly.

Accuracy and Precision: *Accuracy* refers to how closely a computed or measured value agrees with the true value. *Precision* refers to how closely individual computed or measured values agree with each other. *Inaccuracy* (also known as *bias*) is the systematic deviation from the truth. *Imprecision* (uncertainty) refers to the magnitude of the scatter.

Errors are introduced by the computational process itself. Computers perform mathematical operations with only a finite number of digits. If the number x_a is an approximation to the exact result x_e, then the difference $x_e - x_a$ is called *error*. Hence,

$$\text{Exact value} = \text{approximate value} + \text{error}$$

In numerical computations, we come across the following types of errors:
(*a*) Absolute and relative errors
(*b*) Inherent errors
(*c*) Round-off errors
(*d*) Truncation errors

1.2.1 Absolute and Relative Errors

If X_E is the exact or true value of a quantity and X_A is its approximate value, then $|X_E - X_A|$ is called the *absolute error* E_a. Therefore absolute error

$$E_a = |X_E - X_A| \tag{1.8}$$

and *relative error* is defined by

$$E_r = \left| \frac{X_E - X_A}{X_E} \right|, X_E \neq 0 \tag{1.9}$$

The percentage relative error is

$$E_p = \left| \frac{X_E - X_A}{X_E} \right| \times 100, X_E \neq 0 \qquad \textbf{(1.10)}$$

1.2.2 Inherent Errors

Inherent errors are the errors that pre - exist in the problem statement itself before its solution is obtained. Inherent errors exist because the data being approximate or due to the limitations of the calculations using digital computers. Inherent errors cannot be completely eliminated but can be minimized if we select better data or by employing high precision computer computations.

1.2.3 Round – Off Errors

Round-off error is due to the inaccuracies that arise due to a finite number of digits of precision used to represent numbers. All computers represent numbers, except for integer and some fractions, with imprecision. Digital computers use floating-point numbers of fixed word length. This type of representation will not express the exact or true values correctly. Error introduced by the omission of significant figures due to computer imperfection is called the *round-off error*.

Round-off errors are avoidable in most of the computations. When n digits are used to represent a real number, then one method is keep the first n digits and *chop off* all remaining digits. Another method is to *round* to the n^{th} digit by examining the values of the remaining digits.

Exercise 1.2

Given the number π is approximated using $n = 5$ decimal digits.

a. Determine the relative error due to *chopping* and express it as a per cent.
b. Determine the relative error due to *rounding* and express it as a per cent.

Solution: (@ Lectures)

1.2.4 Truncation Error

Truncation errors are defined as those errors that result from using an approximation in place of an exact mathematical procedure. Truncation error results from terminating after a finite number of terms known as *formula truncation error* or *simply truncation error*. Suppose a function $f(x)$ is infinitely differentiable in an interval which includes the point $x = a$. Then the Taylor series expansion of $f(x)$ about $x = a$ is given by

$$f(x) = \sum_{k=0}^{\infty} \frac{(x-a)^k}{k!} f^k(a) \qquad \textbf{(1.11)}$$

where, $f^k(a)$ denotes the k^{th} derivative of $f(x)$ evaluated at $x = a$.

The error in approximating $E_n(x)$ is equal to the sum of the neglected higher order terms and is often called the *tail* of the series. The tail is given by

$$E_n(x) = \frac{f^{(n+1)}(\xi)}{n!}(x-a)^n \qquad \textbf{(1.12)}$$

The *total numerical error* is the summation of the truncation and round-off errors. The best way to minimize round-off errors is to increase the number of significant figures of the computer. It should be noted here that round-off error increases due to subtractive cancellation or due to an increase in the number of computations in an analysis. The truncation error can be reduced by decreasing the step size. In general, the truncation errors are *decreased* as the round-off errors are *increased* in numerical differentiation.

There exist no systematic and general approaches in evaluating numerical errors for all problems. In most cases, error estimates are based on experience and judgment of the engineer or scientist.

Exercise 1.3

Given the trigonometric function $f(x) = \sin x$,

a. expand $f(x)$ about $x = 0$ using Taylor series
b. truncate the series to $n = 6$ terms
c. find the relative error at $x = \pi/4$ due to truncation in (b)
d. determine the upper bound on the magnitude of the relative error at $x = \pi/4$ and express it as a percent.

Solution: (@ Lectures)

1.2.5 Stability and Condition

A numerical computation is said to be *numerically unstable* if the uncertainty of the input values is grossly magnified by numerical method employed.

Consider the first-order Taylor's series of a function given by

$$f(x) = f(a) + f'(a)(x - a) \tag{1.13}$$

The *relative error* of $f(x)$ then becomes

$$\frac{f(x) - f(a)}{f(x)} \cong \frac{f'(a)(x - a)}{f(a)} \tag{1.14}$$

The *relative error* of x becomes

$$\frac{x - a}{a} \tag{1.15}$$

A *condition number* is often defined as the ratio of the relative errors given by **Eqs. (1.14)** and **(1.15)** as

$$Condition\ number = \frac{af'(a)}{f(a)} \tag{1.16}$$

The condition number given below indicates the extent to which an uncertainty in x is magnified by $f(x)$.

Condition number = 1 (function's relative error = relative error in x)

Condition number > 1 (relative error is amplified)

Condition number < 1 (relative error is attenuated)

Condition number > very large number (the function is ill-conditioned)

Exercise 1.4

Compute and interpret the condition number for

a. $f(x) = \sin x$ for $a = 0.51\pi$

b. $f(x) = \tan x$ for $a = 1.7$

Solution: (@ Lectures)

EXERCISE 1

1. Determine the following hyperbolic trigonometric functions to $o(0.9)^4$.

   a. $\sinh(0.9)$ b. $\cosh(0.9)$

2. Determine the **maximum relative error** for the function

   $$F = 3x^2y^2 + 5y^2z^2 - 7x^2z^2 + 38$$

   for $x = y = z = 1$ and $\Delta x = -0.005$, $\Delta y = 0.001$ and $\Delta z = 0.02$

3. Evaluate and interpret the **condition numbers** for

   a. $f(x) = \left(x^2 - 1\right)^{1/2} - x \quad for\ x = 200$

   b. $f(x) = \dfrac{e^x + 1}{x} \quad for\ x = 0.01$

4. Consider the trigonometric function $f(x) = \cos(2x + 1)$.

   a. find the Taylor series expansion of $f(x)$ about 0.

b. assuming the Taylor series is truncated to $n = 6$ terms. Determine the relative error at $x = \pi/4$ due to truncation. Express it as a percentage.

c. determine an upper bound on the magnitude of the relative error at $x = \pi/4$ expressed as a percentage.

5. Consider the trigonometric function $f(x) = \ln(1+x)$.

   i. find the Taylor series expansion of $f(x)$ about 0.

   ii. assuming the Taylor series is truncated to $n = 7$ terms. Determine the relative error at $x = 5/3$ due to truncation. Express it as a percentage.

   iii. determine an upper bound on the magnitude of the relative error at $x = 5/3$ expressed as a percentage. **(IA: 2012/ 2013)**

CHAPTER TWO

LINEAR SYSTEM OF EQUATION

2.1 INTRODUCTION

In this chapter we present the solution of n linear simultaneous algebraic equations in n unknowns. Linear systems of equations are associated with many problems in engineering and science, as well as with applications of mathematics to the social sciences and quantitative study of business and economic problems. Their most important application in engineering is in the analysis of linear systems (any system whose response is proportional to the input is deemed to be linear). Linear systems include structures, elastic solids, heat flow, seepage of fluids, electromagnetic fields and electric circuits i.e., most topics taught in an engineering curriculum. If the system is discrete, such as a truss or an electric circuit, then its analysis leads directly to linear algebraic equations.

A system of algebraic equations has the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned} \qquad \textbf{(2.1)}$$

where the coefficients a_{ij} and the constants b_j are known and x_j represents the unknowns. In matrix notation, the equations are written as

$$\begin{bmatrix} a_{11} & a_{12} & a_{14} \\ a_{21} & a_{22} & a_{2n} \\ & & \\ a_{n1} & a_{n2} & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \\ b_n \end{bmatrix} \qquad \textbf{(2.2)}$$

or simply $Ax = b$

A system of linear equations in n unknowns has a unique solution, provided that the determinant of the coefficient matrix is *non-singular* i.e., if $|A| \neq 0$. The rows and columns of a non-singular matrix are *linearly independent* in the sense that no row (or column) is a linear combination of the other rows (or columns):

If the coefficient matrix is *singular*, the equations may have infinite number of solutions, or no solutions at all, depending on the constant vector.

With a set of n unknowns, checking the rank of the coefficient matrix A and that of the augmented matrix A_b enables us to see whether;

a. a unique solution exists

   $rank\ A = rank\ A_b = n$

b. an infinite number of solutions exist

   $rank\ A = rank\ A_b = m < n$

c. no solution exists

   $rank\ A < rank\ A_b$

Exercise 2.1

Determine the uniqueness of the following linear system of equations:

$$\begin{bmatrix} 2 & -1 & 7 \\ 4 & 2 & 2 \\ 3 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ 1 \end{bmatrix}$$

Solution: (@ Lectures)

2.2 METHODS OF SOLUTION

There are two classes of methods for solving system of linear, algebraic equations: direct and iterative methods.

The common characteristics of *direct methods* are that they transform the original equation into *equivalent equations* (equations that have the same solution) that can be solved more easily. The transformation is carried out by applying certain operations.
The solution does not contain any truncation errors but the round off errors is introduced due to floating point operations.

Iterative or *indirect methods*, start with a guess of the solution **x**, and then repeatedly refine the solution until a certain convergence criterion is reached. Iterative methods are generally less efficient than direct methods due to the large number of operations or iterations required.
Iterative procedures are self-correcting, meaning that round off errors (or even arithmetic mistakes) in one iteration cycle are corrected in subsequent cycles. The solution contains truncation error. A serious drawback of iterative methods is that they do not always converge to the solution. The initial guess affects only the number of iterations that are required for convergence. The indirect solution technique (iterative) is more useful to solve a set of ill-conditioned equations.

In this chapter, we will present six direct methods and two indirect (iterative) methods.

Direct Methods:

1. Matrix Inverse Method
2. Cramer's Method
3. Gauss Elimination Method
4. Gauss-Jordan Method
5. Cholesky's Triangularisation Method
6. Eigenvalues and Eigenvectors

Indirect or Iterative Methods:

1. Jacobi's Iteration Method
2. Gauss-Seidal Iteration Method

2.2.1 Matrix Inverse Method

Consider a set of three simultaneous linear algebraic equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \qquad \textbf{(2.3)}$$

Equation (2.3) can be expressed in the matrix form:

$$Ax = b \qquad \textbf{(2.4)}$$

Pre-multiplying by the inverse A^{-1}, we obtain the solution of x as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = A^{-1} \begin{pmatrix} 5 \\ 0 \\ 5 \end{pmatrix} \qquad \textbf{(2.5)}$$

If the matrix **A** is non-singular, that is, if $\det(\mathbf{A})$ is not equal to zero, then Eq. (2.3) has a unique solution.

The solution for x_1 is

$$x_1 = \frac{1}{|A|}\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix} = \frac{1}{|A|}\left\{ b_1 \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - b_2 \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + b_3 \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \right\} \qquad \textbf{(2.6)}$$

$$x_1 = \frac{1}{|A|}\{b_1c_{11} - b_2c_{21} + b_3c_{31}\}$$

where $\boldsymbol{A}$ is the determinant of the coefficient matrix $\boldsymbol{A}$, and c_{11}, c_{21} and c_{31} are the cofactors of $\boldsymbol{A}$ corresponding to element 11, 21 and 31. We can also write similar expressions for x_2 and x_3 by replacing the second and third columns by the y column respectively. Hence, the complete solution can be written in matrix form as follows:

$$\begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \frac{1}{|A|} \begin{vmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{vmatrix} \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \end{Bmatrix} \quad \textbf{(2.7)}$$

Although this method is quite general but it is not quite suitable for large systems because evaluation of A^{-1} by co-factors becomes very cumbersome.

Exercise 2.2

Obtain the solution of the following linear simultaneous equations by the **matrix inversion method**:

$$\begin{pmatrix} 1 & -1 & 3 \\ 4 & 2 & -1 \\ 1 & 3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \\ 5 \end{pmatrix}$$

Solution: (@ Lectures)

2.2.2 Cramer's Rule

Consider a set of three simultaneous linear algebraic equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \quad \textbf{(2.8)}$$

and let

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix},\ D_1 = \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}, D_2 = \begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix} and\ D_3 = \begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix} \quad \textbf{(2.9)}$$

Cramer's rule states that the unknowns x_1, x_2 and x_3 can be found from the relations

$$x_1 = \frac{D_1}{\Delta}, x_2 = \frac{D_2}{\Delta},\ x_3 = \frac{D_3}{\Delta} \quad \textbf{(2.10)}$$

provided that the determinant Δ (delta) is not zero.

We observe that the numerators of equation (2.10) are determinants that are formed from Δ by the substitution of the known values b_1, b_2 and b_3, for the coefficients of the desired unknown.

Note that Cramer's rule applies to systems of two or more equations. If equation (2.8) is a homogeneous set of equations, that is, if $b_1 = b_2 = b_3 = 0$ then, D_1, D_2 and D_3 are all zero. Then, $x_1 = x_2 = x_3 = 0$ also.

Exercise 2.3

Use Cramer's rule to find v_1, v_2 and v_3 if

$$2v_1 - 5 - v_2 + 3v_3 = 0$$
$$-2v_3 - 3v_2 - 4v_1 = 8$$
$$v_2 + 3v_1 - 4 - v_3 = 0$$

Solution: (@ Lectures)

2.2.3 Gaussian Elimination Method

Consider a set of three simultaneous linear algebraic equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \qquad \textbf{(2.11)}$$

Gauss elimination is a popular technique for solving simultaneous linear algebraic equations. It reduces the coefficient matrix into an upper triangular matrix through a sequence of operations carried out on the matrix. The vector *b* is also modified in the process. The solution vector $\{x\}$ is obtained from a backward substitution procedure.

In Gauss elimination method, the unknowns are eliminated such that the elimination process leads to an upper triangular system and the unknowns are obtained by back substitution. It is assumed. The method can be described by the following steps:

Step 1: Eliminate x_1 from the 2nd and 3rd equations to obtain:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \quad ------(*)$$
$$a'_{22}x_2 + a'_{23}x_3 = b'_2 \quad -----(2.12)$$
$$a'_{32}x_2 + a'_{33}x_3 = b'_3 \quad ------(2.13)$$

Equation (*) is called the *pivotal equation* and the coefficient a_{11} is the *pivot*.

Step 2: Eliminate x_2 from the Eq. (2.12) using Eq. (2.13) by assuming $a_{22} \neq 0$. We perform the following operation:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \quad ----------(**)$$
$$a'_{22}x_2 + a'_{23}x_3 = b'_2 \quad ----------(2.14)$$
$$a''_{33}x_3 = b''_3 \quad ----------(2.15)$$

Here Eq. (2.14) is called the *pivotal equation* and the coefficient a'_{22} is the *pivot*.

Step 3: To find x_1, x_2 and x_3, we apply back substitution starting from Eq. (2.15) giving x_3, then x_2 from Eq. (2.14) and x_1 from Eq. (**).

Exercise 2.4

Solve the following equations by Gauss elimination method:

$$2x + 4y - 6z = -4$$
$$x + 5y + 3z = 10$$
$$x + 3y + 2z = 5$$

Solution: (@ Lectures)

2.2.4 Gauss-Jordan Method

Gauss-Jordan method is an extension of the Gauss elimination method. The set of equations $Ax = b$ is reduced to a diagonal set $Ix = b'$, where ***I*** is a unit matrix. This is equivalent to $x = b'$. The solution vector is therefore obtained directly from *b′*. The Gauss-Jordan method implements the same series of operations as implemented by Gauss elimination process. The main difference is that it applies these operations below as well as above the diagonal such that all off-diagonal elements of the matrix are reduced to

zero. Gauss-Jordan method also provides the inverse of the coefficient matrix A along with the solution vector $\{x\}$. The Gauss-Jordan method is highly used due to its stability and direct procedure. The Gauss-Jordan method requires more computational effort than Gauss elimination process.

Gauss-Jordan method is a modification of Gauss elimination method. The series of operations performed are quite similar to the Gauss elimination method. In the Gauss elimination method, an upper triangular matrix is derived while in the Gauss-Jordan method an identity matrix is derived. Hence, back substitutions are not required.

Exercise 2.5

Solve the following equations by Gauss-Jordan method.

$$\begin{aligned} x - 2y &= -4 \\ 5y + z &= -9 \\ 4x - 3z &= -10 \end{aligned}$$

Solution: (@ Lectures)

2.2.5 Cholesky's Triangularisation Method

In Cholesky's decomposition method there is no need for pivoting. If the decomposition fails, the matrix is not positive definite.

Consider a set of three simultaneous linear algebraic equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \qquad \textbf{(2.16)}$$

The above system can be written as

$$Ax = b \qquad \textbf{(2.17)}$$

where,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \; x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \; b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \qquad \textbf{(2.18)}$$

Let

$$A = LU \qquad \textbf{(2.19)}$$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \text{ and } U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} 2 & 1 & 4 \\ 8 & -3 & 2 \\ 4 & 11 & -1 \end{bmatrix}$$

$$A = \begin{bmatrix} 2 & 1 & 4 \\ 8 & -3 & 2 \\ 4 & 11 & -1 \end{bmatrix}, \; x = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \; b = \begin{bmatrix} 12 \\ 20 \\ 33 \end{bmatrix}$$

Equation (2.16) can be written as

$$LUx = b \qquad \textbf{(2.20)}$$

If we write $Ux = V$ **(2.21)**

Equation (2.20) becomes

$$LV = b \qquad \textbf{(2.22)}$$

Equation (2.22) is equivalent to the system

$$\begin{aligned} v_1 &= b_1 \\ l_{21}v_1 + v_2 &= b_2 \\ l_{31}v_1 + l_{32}v_2 + v_3 &= b_3 \end{aligned} \qquad \textbf{(2.23)}$$

The above system can be solved to find the values of v_1, v_2 and v_3 which give us the matrix V.

$Ux = V$

then becomes

$$\begin{aligned} u_{11}x_1 + u_{12}x_2 + u_{13}x_3 &= v_1 \\ u_{22}x_2 + u_{23}x_3 &= v_2 \\ u_{33}x_3 &= v_3 \end{aligned} \qquad \textbf{(2.24)}$$

which can be solved for x_3, x_2 and x_1 by the backward substitution process.

In order to compute the matrices L and U, we write Eq. (2.19) as

$$\begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Multiply the matrices on the left and equate the corresponding elements of both sides. Cholesky's triangularisation method is also known as Crout's triangularisation method or method of factorisation.

Exercise 2.6

Solve the following equations by Cholesky's triangularisation method.

$$2x - 6y + 8z = 24$$
$$5x + 4y + 3z = 2$$
$$3x + y + 2z = 16$$

Solution: (@ Lectures)

2.2.6 Eigenvalues and Eigenvectors

Matrices commonly appear in technological problems, for example those involving coupled oscillations and vibrations, and give rise to equations of the form

$$Ax = \lambda x \qquad \textbf{(2.25)}$$

where $A=\left(a_{ij}\right)$ is a square matrix, x is a column matrix $\left(x_i\right)$ and λ is a scalar quantity, i.e. a number.

For non – trivial solutions, i.e. for $x \neq 0$, the values of λ are called the *eigenvalues*, *characteristic values* or *latent roots* of the matrix A and the corresponding solutions of the given equations $Ax = \lambda x$ are called the *eigenvectors*, or *characteristic vectors* of A.
The set of equations

$$\begin{pmatrix} a_{11} & a_{12} & a_{1n} \\ a_{21} & a_{22} & a_{2n} \\ & & \\ a_{n1} & a_{n2} & a_{nn} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \\ x_n \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \\ \\ x_n \end{pmatrix} \qquad \textbf{(2.26)}$$

then simplifies

$$\begin{pmatrix} \left(a_{11}-\lambda\right) & a_{12} & a_{1n} \\ a_{21} & \left(a_{22}-\lambda\right) & a_{2n} \\ & & \\ a_{n1} & a_{n2} & \left(a_{nn}-\lambda\right) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \\ 0 \end{pmatrix} \qquad \textbf{(2.27)}$$

That is, $Ax = \lambda x$ becomes $Ax - \lambda x = 0$

i.e. $\left(A-\lambda I\right)x = 0$ **(2.28)**

the unit matrix I being introduced since we can subtract only a matrix from another matrix.

For this set of homogeneous linear equations (right – hand side constant terms all zero) to have non – trivial solutions

$\left|A-\lambda I\right|$ **must be zero.**

This is called the characteristic determinant of A and $\left|A-\lambda I\right| = 0$ is the characteristic equation, the solution of which gives the values of λ, i.e. the eigenvalues of A.

Exercise 2.7

Find the eigenvalues and corresponding eigenvectors of

$Ax = \lambda x$ where $A = \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix}$.

Solution: (@ Lectures)

2.2.7 Jacobi Iteration Method

This method is also known as *the method of simultaneous displacements*. Consider the system of linear equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \qquad \textbf{(2.29)}$$

Here, we assume that the coefficients a_{11}, a_{22} and a_{33} are the largest coefficients in the respective equations so that

$$\begin{aligned} |a_{11}| &> |a_{12}| + |a_{13}| \\ |a_{22}| &> |a_{21}| + |a_{23}| \\ |a_{33}| &> |a_{31}| + |a_{32}| \end{aligned} \qquad \textbf{(2.30)}$$

Jacobi's iteration method is applicable only if the conditions given in **Eq. (2.30)** are satisfied.

Now, we can write **Eq. (2.29)**

$$\begin{aligned} x_1 &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2 - a_{13}x_3\right) \\ x_2 &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_1 - a_{23}x_3\right) \\ x_3 &= \frac{1}{a_{33}}\left(b_3 - a_{31}x_1 - a_{32}x_2\right) \end{aligned} \qquad \textbf{(2.31)}$$

Let the initial approximations be x_1^0, x_2^0 and x_3^0 respectively. The following iterations are then carried out.

Iteration 1: The first improvements are found as

$$\begin{aligned} x_{11} &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2^0 - a_{13}x_3^0\right) \\ x_{21} &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_1^0 - a_{23}x_3^0\right) \\ x_{31} &= \frac{1}{a_{33}}\left(b_3 - a_{31}x_1^0 - a_{32}x_2^0\right) \end{aligned} \qquad \textbf{(2.32)}$$

Iteration 2: The second improvements are obtained as

$$\begin{aligned} x_{12} &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_{21} - a_{13}x_{31}\right) \\ x_{22} &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_{11} - a_{23}x_{31}\right) \\ x_{32} &= \frac{1}{a_{33}}\left(b_3 - a_{31}x_{11} - a_{32}x_{21}\right) \end{aligned} \qquad \textbf{(2.33)}$$

The above iteration process is continued until the values of x_1, x_2 and x_3 are found to a pre-assigned degree of accuracy. That is, the procedure is continued until the relative error between two consecutive vector norms is satisfactorily small. In Jacobi's method, it is a general practice to assume $x_1^0 = x_2^0 = x_3^0 = 0$. The method can be extended to a system of *n* linear simultaneous equations in *n* unknowns.

Exercise 2.8

Solve the following equations by Jacobi's method.

$$\begin{aligned} x + 2y + z &= 0 \\ 3x + y - z &= 0 \\ x - y + 4z &= 3 \end{aligned}$$

Solution:

Iterations results are shown in the table below:

r	x(r+1)	y(r+1)	z(r+1)
0	1	1	1
1	0	-1	0.75
2	0.583333	-0.375	0.5
3	0.291667	-0.54167	0.510417
4	0.350694	-0.40104	0.541667
5	0.314236	-0.44618	0.562066
6	0.336082	-0.43815	0.559896
7	0.332682	-0.44799	0.556442
8	0.33481	-0.44456	0.554832
9	0.333131	-0.44482	0.555157
10	0.333326	-0.44414	0.555512
11	0.333219	-0.44442	0.555632
12	0.33335	-0.44443	0.555591
13	0.333339	-0.44447	0.555556
14	0.333342	-0.44445	0.555548
15	0.333332	-0.44444	0.555553

so to four decimal places the approximate solutions are:

$x = 0.3333 \qquad y = -0.4444 \qquad z = 0.5555$

2.2.8 Gauss-Seidal Method

The Gauss-Seidal method is applicable to *predominantly diagonal systems*. A predominantly diagonal system has large diagonal elements. The absolute value of the diagonal element in each case is larger than the sum of the absolute values of the other elements in that row of the matrix **A**. For such predominantly diagonal systems, the Gauss-Seidal method always converges to the correct solution, irrespective of the choice of the initial estimates. Since the most recent approximations of the variables are used while proceeding to the next step, the convergence of the Gauss-Seidal method is twice as fast as in Jacobi's method. The Gauss-Seidal and Jacobi's methods converge for any choice of the initial approximations, if in each equation of the system, the absolute value

of the largest coefficient is greater than the sum of the absolute values of the remaining coefficients. In other words,

$$\sum_{\substack{i=1 \\ j\neq 1}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \leq 1 \qquad i = 1, 2, 3, \ldots, n$$

where the inequality holds in case of at least one equation. Convergence is assured in the Gauss-Seidal method if the matrix A is diagonally dominant and *positive definite*. If it is not in a diagonally dominant form, it should be connected to a diagonally dominant form by row exchanger, before starting the Gauss-Seidal iterative scheme.

Gauss-Seidal method is also an iterative solution procedure which is an improved version of Jacobi's method. The method is also known as the *method of successive approximations*.

Consider the system of linear equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \qquad \textbf{(2.34)}$$

If the absolute value of the largest coefficient in each equation is greater than the sum of the absolute values of all the remaining coefficients, then the Gauss-Seidal iteration method will converge. If this condition is not satisfied, then Gauss-Seidal method is not applicable. Here, in **Eq. (2.34)**, we assume the coefficient a_{11}, a_{22} and a_{33} are the largest coefficients.

We can rewrite **Eq. (2.34)** as

$$\begin{aligned} x_1 &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2 - a_{13}x_3\right) \\ x_2 &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_1 - a_{23}x_3\right) \\ x_3 &= \frac{1}{a_{33}}\left(b_3 - a_{31}x_1 - a_{32}x_2\right) \end{aligned} \qquad \textbf{(2.35)}$$

Let the initial approximations be x_1^0, x_2^0 and x_3^0 respectively. The following iterations are then carried out.

Iteration 1: The first improvements are found as

$$\begin{aligned} x_{11} &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2 - a_{13}x_3^0\right) \\ x_{21} &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_{11} - a_{23}x_3^0\right) \\ x_{31} &= \frac{1}{a_{33}}\left(b_3 - a_{31}x_{11} - a_{32}x_{21}\right) \end{aligned} \qquad \textbf{(2.36)}$$

Iteration 2: The second improvements are obtained as

$$\begin{aligned} x_{12} &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_{11} - a_{13}x_{31}\right) \\ x_{22} &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_{12} - a_{23}x_{31}\right) \\ x_{32} &= \frac{1}{a_{33}}\left(b_3 - a_{31}x_{12} - a_{32}x_{22}\right) \end{aligned} \qquad \textbf{(2.37)}$$

The above iteration process is continued until the values of x_1, x_2 and x_3 are obtained to a pre-assigned or desired degree of accuracy. That is, the procedure is continued until the relative error between two consecutive vector norms is satisfactorily small. In Gauss-Seidal method, it is a general practice to assume $x_1^0 = x_2^0 = x_3^0 = 0$. The convergence rate of Gauss-Seidal method is found to be twice to that of Jacobi's method. Like the Jacobi's method, Gauss-Seidal method can also be extended to *n* linear simultaneous algebraic equations in *n* unknowns.

Exercise 2.9

Using the Gauss-Seidal method solve the system of equations correct to three decimal places.

$$x + 2y + z = 0$$
$$3x + y - z = 0$$
$$x - y + 4z = 3$$

Solution:

Iterations results are shown in the table below:

r	x(r+1)	y(r+1)	z(r+1)
0	1	1	1
1	0	-0.5	0.625
2	0.375	-0.5	0.53125
3	0.34375	-0.4375	0.55469
4	0.33073	-0.4427	0.55664
5	0.33312	-0.4449	0.5555
6	0.33346	-0.4445	0.55551
7	0.33333	-0.4444	0.55556
8	0.33333	-0.4444	0.55556
9	0.33333	-0.4444	0.55556
10	0.33333	-0.4444	0.55556
11	0.33333	-0.4444	0.55556
12	0.33333	-0.4444	0.55556
13	0.33333	-0.4444	0.55556

Hence, the approximate solution is as follows:

$x = 0.3333$ $y = -0.4444$ $z = 0.5555$

2.2.9 Ill-Conditioning

An obvious question is: what happens when the coefficient matrix is almost singular; i.e., if |**A**| is very small? In order to determine whether the determinant of the coefficient matrix is "small" we need a reference against which the determinant can be measured. This reference is called the *norm* of the matrix, denoted by $\|A\|$. We can then say that the determinant is small if

$$|A| << \|A\|$$

Several norms of a matrix have been defined in existing literature, such as

$$\|A\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}^2} \qquad or \quad \|A\| = \max_{1\le i\le n}\sum_{j=1}^{n}|A_{ij}| \tag{2.38}$$

A formal measure of conditioning is the *matrix condition number*, defined as

$$cond(A) = \|A\|\|A^{-1}\| \tag{2.39}$$

If this number is close to unity, the matrix is well-conditioned. The condition number increases with the degree of ill-conditioning, reaching infinity for a singular matrix. Note that the condition number is not unique, but depends on the choice of the matrix norm. Unfortunately, the condition number is expensive to compute for large matrices. In most cases it is sufficient to gauge conditioning by comparing the determinant with the magnitudes of the elements in the matrix. If the equations are ill-conditioned, small changes in the coefficient matrix result in large changes in the solution.

As an illustration, consider the equations

$$2x + y = 3, \qquad 2x + 1.001y = 0$$

that have the solution $x = 1501.5,\ y = -3000$.

Since $|\boldsymbol{A}| = 2(1.001) - 2(1) = 0.002$ is much smaller than the coefficients, the equations are ill-conditioned. The effect of ill-conditioning can be verified by changing the second equation to $2x + 1.002y = 0$ and re-solving the equations. The result is $x = 751.5, y = -1500$. Note that a 0.1% change in the coefficient of y produced a 100% change in the solution.

Numerical solutions of ill-conditioned equations are not to be trusted. The reason is that the inevitable round off errors during the solution process is equivalent to introducing small changes into the coefficient matrix.
This in turn introduces large errors into the solution, the magnitude of which depends on the severity of ill-conditioning.

2.3 Summary

A matrix is a rectangular array of elements, in rows and columns. The elements of a matrix can be numbers, coefficients, terms or variables. This chapter provided the relevant and useful elements of matrix analysis for the solution of linear simultaneous algebraic equations. Topics covered include matrix definitions, matrix operations, determinants, matrix inversion, trace, transpose, and system of algebraic equations and solution.

The solution of n linear simultaneous algebraic equations in n unknowns is presented. There are two classes of methods of solving system of linear algebraic equations: direct and iterative methods. Direct methods transform the original equation into equivalent equations that can be solved more easily. Iterative or indirect methods start with a guess of the solution x, and then repeatedly refine the solution until a certain convergence criterion is reached. Four direct methods (matrix inversion method, Gauss elimination method, Gauss-Jordan method, Eigenvalues and Eigenvectors and Cholesky's triangularisation method) are presented. Two indirect or iterative methods (Jacobi's iteration method and Gauss-Seidal iteration method) are presented. The LU decomposition method is closely related to Gauss elimination method. LU decomposition is computationally very effective if the coefficient matrix remains the same but the right hand side vector changes. Cholesky's decomposition method can be used when the coefficient matrix A is symmetric and positive definite. Gauss-Jordan method is a very stable method for solving linear algebraic equations. Gauss-Seidal iterative substitution technique is very suitable for predominantly diagonal systems. It requires a guess of the solution.

EXERCISE 2

1. Classify the following matrices as **singular, ill-conditioned or well-conditioned**.

a) $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}$

b)

$B = \begin{bmatrix} 2.11 & -0.80 & 1.72 \\ -1.84 & 3.03 & 1.29 \\ -1.57 & 5.25 & 4.30 \end{bmatrix}$

c) $C = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$

d) $D = \begin{bmatrix} 4 & 3 & -1 \\ 7 & -2 & 3 \\ 5 & -18 & 13 \end{bmatrix}$

2. Use the method of **Gaussian elimination** to solve the following system of linear equations:

$$x_1 + x_2 + x_3 - x_4 = 2$$
$$4x_1 + 4x_2 + x_3 + x_4 = 11$$
$$x_1 - x_2 - x_3 + x_4 = 0$$
$$2x_1 + x_2 + 2x_3 - 2x_4 = 2$$

3. Using the **Gaussian elimination method**, solve the system of equations [**A**] $\{x\}$ = $\{b\}$ where:

$$[A] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & -1 & 3 & 0 \\ 0 & 2 & 0 & 3 \\ -1 & 0 & 2 & 1 \end{bmatrix} \text{ and } \{b\} = \begin{Bmatrix} 3 \\ 3 \\ 1 \\ 0 \end{Bmatrix}$$

4. Use Gauss elimination to solve the equations **AX** = **B**, where

$$A = \begin{bmatrix} 6 & -4 & 1 \\ -4 & 6 & -4 \\ 1 & -4 & 6 \end{bmatrix} \qquad B = \begin{bmatrix} -14 & 22 \\ 36 & -18 \\ 6 & 7 \end{bmatrix}$$

5. Solve the following set of simultaneous linear equations by the **matrix inverse method**.

a)

$$\begin{aligned} x_1 - x_2 + 3x_3 - x_4 &= 1 \\ x_2 - 3x_3 + 5x_4 &= 2 \\ x_1 + 2x_2 - x_4 &= 0 \\ x_1 + 2x_2 - x_4 &= -5 \end{aligned}$$

b)

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

6. Solve the following set of simultaneous linear equations by the **Cramer's Rule.**

a)

$$\begin{aligned} x_1 - x_2 + 3x_3 - x_4 &= 1 \\ x_2 - 3x_3 + 5x_4 &= 2 \\ x_1 + 2x_2 - x_4 &= 0 \\ x_1 + 2x_2 - x_4 &= -5 \end{aligned}$$

b)

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

7. Solve the following set of equations by **Gauss-Jordan method**.

a)

$$2x_1 + x_2 - 3x_3 = 11$$
$$4x_1 - 2x_2 + 3x_3 = 8$$
$$-2x_1 + 2x_2 - x_3 = -6$$

b)

$$4x_1 - 2x_2 - 3x_3 + 6x_4 = 12$$
$$-5x_1 + 7x_2 + 6.5x_3 - 6x_4 = -6.5$$
$$x_1 + 7.5x_2 + 6.25x_3 + 5.5x_4 = 16$$
$$-12x_1 + 22x_2 + 15.5x_3 - x_4 = 17$$

8. Solve the system of equations using **Cholesky's factorisations.**

a)

$$x_1 + x_2 + x_3 - x_4 = 2$$
$$x_1 - x_2 - x_3 + 2x_4 = 0$$
$$4x_1 + 4x_2 + x_3 + x_4 = 11$$
$$2x_1 + x_2 + 2x_3 - 2x_4 = 2$$

Answer: $x_4 = 0,\ x_3 = -1,\ x_2 = 2,\ x_1 = 1.$

b)

$$12x_1 - 6x_2 - 6x_3 + 1.5x_4 = 1$$
$$-6x_1 + 4x_2 + 3x_3 + 0.5x_4 = 2$$
$$-6x_1 + 3x_2 + 6x_3 + 1.5x_4 = 3$$
$$-1.5x_1 + 0.5x_2 + 1.5x_3 + x_4 = 4$$

9. Find the eigenvalues and corresponding eigenvectors of

$Ax = \lambda x$ where $A = \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix}$.

10. Find the eigenvalues and corresponding eigenvectors of

$Ax = \lambda x$ where $A = \begin{pmatrix} 1 & 0 & 4 \\ 0 & 2 & 0 \\ 3 & 1 & -3 \end{pmatrix}$.

11. Use **Jacobi iterative scheme** to obtain the solution of the system of equations correct to two decimal places.

$$\begin{bmatrix} 5 & -2 & 1 \\ 1 & 4 & -2 \\ 1 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \\ 17 \end{bmatrix}$$

Hence, the solution is given by $x = 1$, $y = 2$ and $z = 3$.

12. Solve the following equations by the **Gauss-Seidal method.**

a)

$$4x - y + z = 12$$
$$-x + 4y - 2z = -1$$
$$x + 4y - 2z = 5$$

We obtain the final values for x, y and z as $x = 3$, $y = 1$ and $z = 1$.

b)

$$4x_1 + 2x_2 = 4$$
$$2x_1 + 8x_2 + 2x_3 = 0$$
$$2x_2 + 8x_3 + 2x_4 = 0$$
$$2x_3 + 4x_4 = 14$$

13. Let

$$A = \begin{pmatrix} 2 & 3 & 1 \\ 0 & 1 & 2 \\ 1 & 1 & 4 \end{pmatrix}$$

Compute, directly from the definition, $cond_{\infty}(A)$.

14.

i. State the condition of convergence for **Gauss – Seidal** method.

ii. State the main difference between **Gauss elimination** and **Gauss – Jordan** method.

iii. Explain the method of **triangularisation**.

15. Determine the **rank** of the matrix **A**

$$A = \begin{bmatrix} 7 & 12 & 21 & 1 & 1 \\ 3 & 8 & 9 & 2 & 0 \\ 1 & 6 & 3 & 3 & 0 \\ 6 & 3 & 18 & 4 & 3 \\ 2 & 0 & 6 & 5 & 9 \end{bmatrix}$$

CHAPTER THREE

SOLUTION OF ALGEBRAIC AND TRANSCENDENTAL EQUATIONS

3.1 INTRODUCTION

One of the most common problems encountered in engineering analysis is that given a function $f(x)$, find the values of x for which $f(x)=0$. The solution (values of x) are known as the *roots* of the equation $f(x)=0$, or the *zeroes* of the function $f(x)$.

The roots of equations may be real or complex. In general, an equation may have any number of (real) roots or no roots at all. For example, $\sin(x)-x=0$ has a single root, namely, $x=0$, whereas $\tan(x)-x=0$ has infinite number of roots $(x=0,\ \pm 4.493,\ \pm 7.725,\ \ldots)$.

There are two types of methods available to find the roots of algebraic and transcendental equations of the form $f(x)=0$.

1. **Direct Methods:** Direct methods give the exact value of the roots in a finite number of steps. We assume here that there are no round - off errors. Direct methods determine all the roots at the same time.

2. **Indirect or Iterative Methods:** Indirect or iterative methods are based on the concept of successive approximations. The general procedure is to start with one or more initial approximation to the root and obtain a sequence of iterates (x_k) which in the limit converges to the actual or true solution to the root.

Indirect or iterative methods determine one or two roots at a time. The indirect or iterative methods are further divided into two categories: bracketing and open methods. The bracketing methods require the limits between which the root lies, whereas the open methods require the initial estimation of the solution. Bisection and False position methods are two known examples of the bracketing methods. Among the open methods, the Newton-Raphson and the method of successive approximation are most commonly used. The most popular method for solving a non-linear equation is the Newton-Raphson method and this method has a high rate of convergence to a solution.

In this chapter, we present the following indirect or iterative methods with illustrative examples:

1. Bisection Method
2. Method of False Position (Regular Falsi Method)
3. Newton-Raphson Method (Newton's method)
4. Successive Approximation Method.
5. Secant Method

3.2 BISECTION METHOD

After a root of $f(x) = 0$ has been bracketed in the interval (a,b). Bisection method can be used to close in on it. The Bisection method accomplishes this by successfully halving the interval until it becomes sufficiently small. Bisection method is also known as the *interval halving method*. Bisection method is not the fastest method available for

finding roots of a function, but it is the most reliable method. Once '*a*' has been bracketed, Bisection method will always close in on it.

We assume that $f(x)$ is a function that is real-valued and that x is a real variable. Suppose that $f(x)$ is continuous on an interval $a \leq x \leq b$ and that $f(a)\,f(b) < 0$. When this is the case, $f(x)$ will have opposite signs at the end points of the interval (a, b). As shown in Fig. 3.1 (*a*) and (*b*), if $f(x)$ is continuous and has a solution between the points $x = a$ and $x = b$, then either $f(a) > 0$ and $f(b) < 0$ or $f(a) < 0$ and $f(b) > 0$. In other words, if there is a solution between $x = a$ and $x = b$, then $f(a)\,f(b) < 0$.

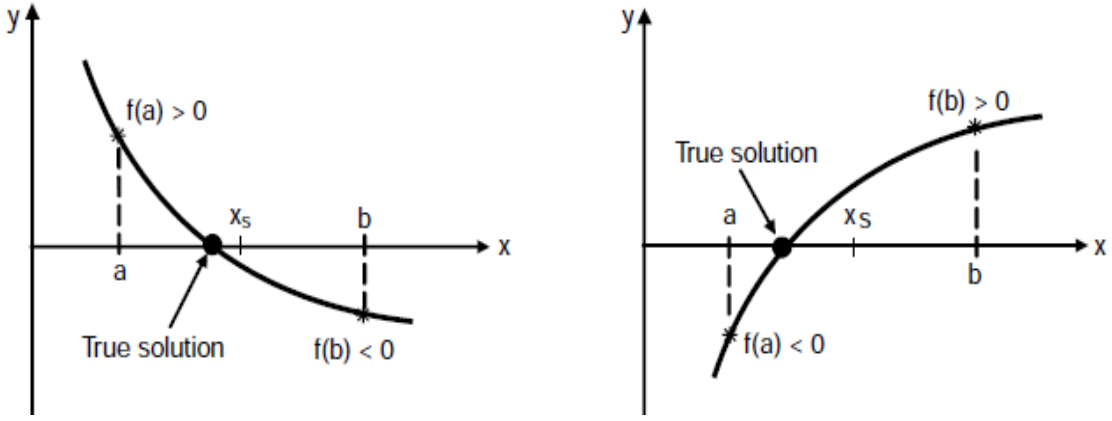

Fig. 3.1: Solution of $f(x) = 0$ between $x = a$ and $x = b$

The method of finding a solution with the Bisection method is illustrated in Fig. 3.2. It starts by finding points a and b that define an interval where a solution exists. The midpoint of the interval x_{s_1} is then taken as the first estimate for the numerical solution. The true solution is either in the portion between points a and x_{s_1}, or in the portion between points x_{s_1} and b. If the solution obtained is not accurate enough, a new interval that contains the true solution is defined. The new interval selected is the half of the original interval that contains the true solution, and its midpoint is taken as the new (second) estimate of the numerical solution. The procedure is repeated until the numerical solution is accurate enough according to a certain criterion that is selected.

The procedure or algorithm for finding a numerical solution with the Bisection method is given below:

Algorithm for the Bisection Method

1. Let $f(a)f(b)<0$
2. Compute the first estimate of the numerical solution x_{s_1} by

$$x_{s_1}=\frac{a+b}{2}$$

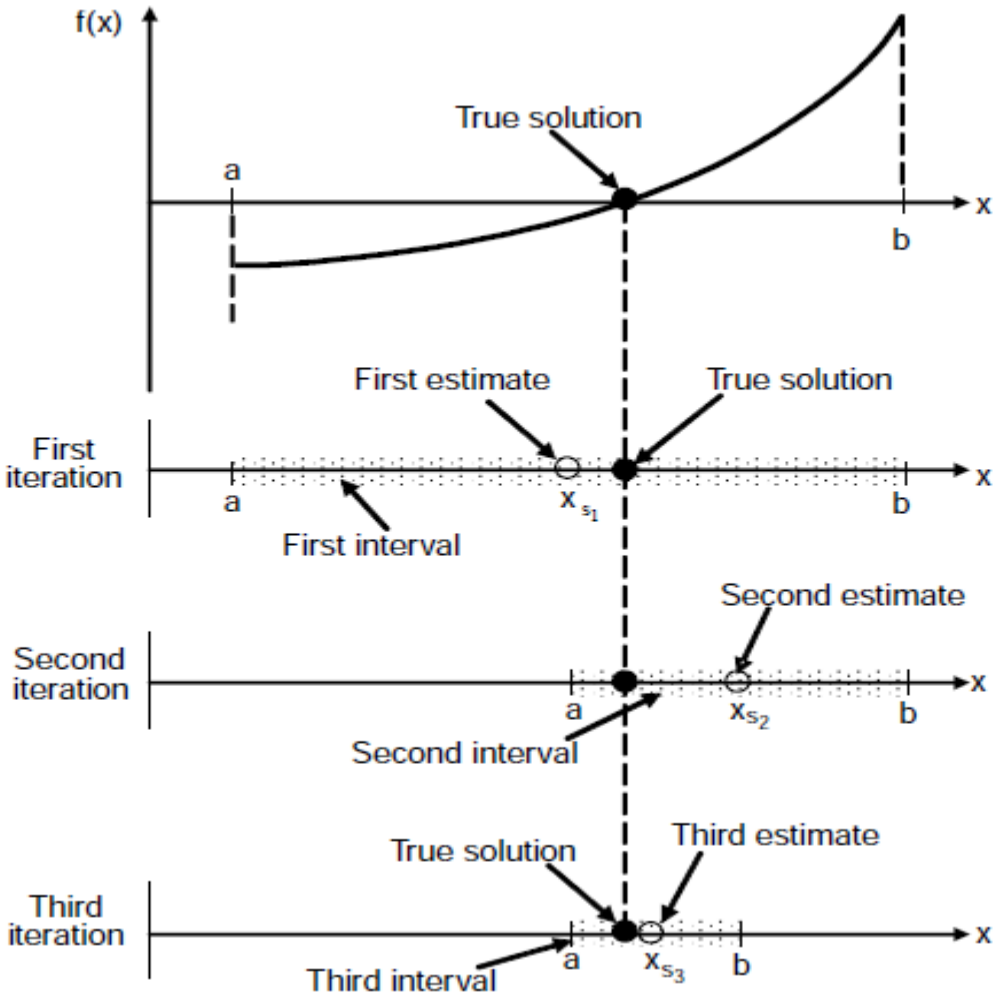

Fig. 3.2: Bisection Method

3. Determine whether the true solution is between a and x_{s_1} or between x_{s_1} and b by checking the sign of the product $f(a)f(x_{s_1})$:

   If $f(a)f(x_{s_1}) < 0$, the true solution is between a and x_{s_1}.

   If $f(a)f(x_{s_1}) \geq 0$, the true solution is between b and x_{s_1}.

   If $b - x_{s_1} \leq \in$, then accept x_{s_1} as the root and stop. $\in$ is the error tolerance, $\in > 0$.

4. Choose the subinterval that contains the true solution (a to x_{s_1} or x_{s_1} to b) as the new interval (a, b) and go back to step 2.

Steps 2 through 4 are repeated until a specified tolerance or error bound is attained.

3.2.1 Error Bounds

To find out how much iteration will be necessary, suppose we want to have

$$|b - x_{s_1}| \leq \in$$

This will be satisfied if

$$\frac{1}{2^n}(b-a) \leq \in \qquad \textbf{(3.1)}$$

Taking logarithms of both sides of Eq.(3.1), and simplifying the resulting expression, we obtain

$$n \geq \frac{\log\left(b-a/\in\right)}{\log 2} \qquad \textbf{(3.2)}$$

There are several advantages to the Bisection method. The method is guaranteed to converge. The method always converges to an answer, provided a root was bracketed in the interval (a, b) to start with. In addition, the error bound is guaranteed to decrease by one-half of each iteration. The method may fail when the function is tangent to the axis and does not cross the x-axis at $f(x) = 0$. The disadvantage of the Bisection method is that it generally converges more slowly than most other methods. For functions $f(x)$ that have a continuous derivative, other methods are usually faster. These methods may not always

converge. When these methods do converge, they are almost always much faster than the Bisection method.

Exercise 3.1

Use the Bisection method to find a root of the equation $x^3 - 4x - 8.95 = 0$ accurate to three decimal places.

Solution

The results of the algorithm for Bisection method are shown in Table 3.1.

n	a	b	Xs	f(a)	f(Xs)	f(a)*f(Xs)	b-Xs
1	2	3	2.5	-8.95	-3.325	29.7588	0.5
2	2.5	3	2.75	-3.325	0.84688	-2.8159	0.25
3	2.5	2.75	2.625	-3.325	-1.3621	4.52901	0.125
4	2.625	2.75	2.6875	-1.3621	-0.2891	0.3938	0.0625
5	2.6875	2.75	2.71875	-0.2891	0.27092	-0.0783	0.03125
6	2.6875	2.71875	2.70313	-0.2891	-0.0111	0.0032	0.015625
7	2.70313	2.71875	2.71094	-0.0111	0.12942	-0.0014	0.007813
8	2.70313	2.71094	2.70703	-0.0111	0.05905	-0.0007	0.003906
9	2.70313	2.70703	2.70508	-0.0111	0.02396	-0.0003	0.001953
10	2.70313	2.70508	**2.7041**	-0.0111	0.00643	-7E-05	0.000977
11	2.70313	2.7041	2.70361	-0.0111	-0.0023	2.6E-05	0.000488
12	2.70361	2.7041	2.70386	-0.0023	0.00205	-5E-06	0.000244
13	2.70361	2.70386	2.70374	-0.0023	-0.0001	3.2E-07	0.000122
14	2.70374	2.70386	2.7038	-0.0001	0.00096	-1E-07	6.1E-05
15	2.70374	2.7038	2.70377	-0.0001	0.00041	-6E-08	3.05E-05
16	2.70374	2.70377	2.70375	-0.0001	0.00014	-2E-08	1.53E-05
17	2.70374	2.70375	2.70374	-0.0001	6E-07	-8E-11	7.63E-06

Hence, the root is 2.704 accurate to three decimal places.

3.3 METHOD OF FALSE POSITION

The method of False Position (also called *the Regular Falsi method,* and *the linear interpolation method*) is another well-known bracketing method. It is very similar to Bisection method with the exception that it uses a different strategy to end up with its new root estimate. Rather than bisecting the interval (*a*, *b*), it locates the root by joining

$f(a_1)$ and $f(b_1)$ with a straight line. The intersection of this line with the *x*-axis represents an improved estimate of the root.

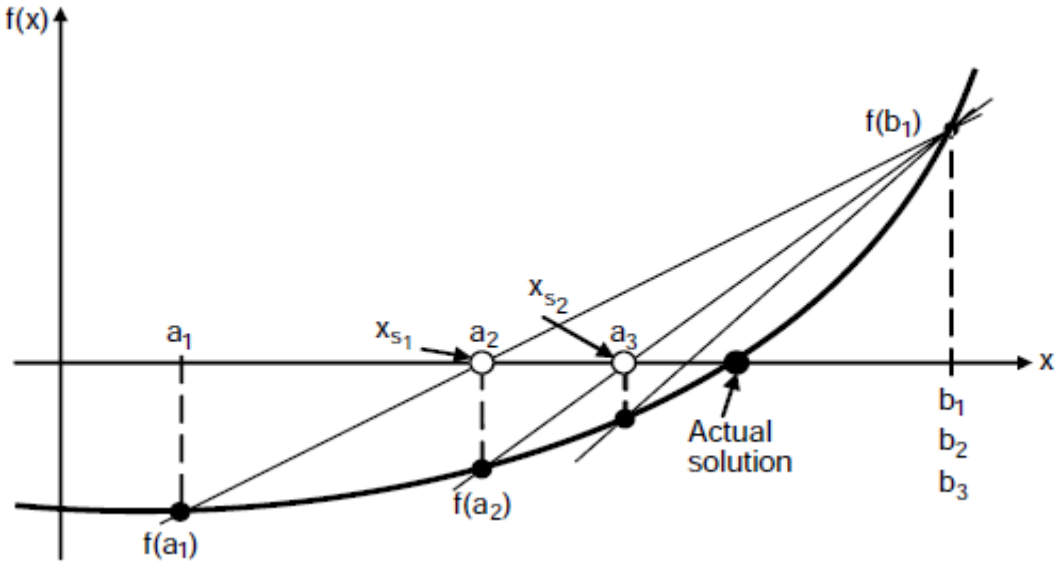

Fig. 3.3: Method of False Position

Here again, we assume that within a given interval (a, b), $f(x)$ is continuous and the equation has a solution. As shown in Fig. 3.3, the method starts by finding an initial interval (a_1, b_1) that brackets the solution.

$f(a_1)$ and $f(b_1)$ are the values of the function at the end points a_1 and b_1. These end points are connected by a straight line, and the first estimate of the numerical solution, x_{s_1}, is the point where the straight line crosses the axis. For the second iteration, a new interval (a_2, b_2) is defined. The new interval is either (a_1, x_{s_1}) where a_1 is assigned to a_2 and x_{s_1} to b_2 or (x_{s_1}, b_1) where x_{s_1} is assigned to a_2 and b_1 to b_2. The end points of the second interval are connected with a straight line, and the point where this new line crosses the *x*-axis is the second estimate of the solution, x_{s_1}. A new subinterval (a_3, b_3) is selected for the third iteration and the iterations will be continued until the numerical solution is accurate enough.

The equation of a straight line that connects points $\left(b, f\left(b\right)\right)$to point $\left(a, f\left(a\right)\right)$is given by

$$y = \frac{f\left(b\right) - f\left(a\right)}{b - a}\left(x - b\right) + f\left(b\right) \qquad \textbf{(3.3)}$$

The points x_s where the line intersects the *x*-axis is determined by substituting $y = 0$in Eq.(3.3) and solving the equation for *x*.

Hence,

$$x_s = \frac{af\left(b\right) - bf\left(a\right)}{f\left(b\right) - f\left(a\right)} \qquad \textbf{(3.4)}$$

The procedure (or algorithm) for finding a solution with the method of False Position is given below:

Algorithm for the method of False Position

1. Define the first interval (*a*, *b*) such that solution exists between them. Check $f\left(a\right)f\left(b\right) < 0$.
2. Compute the first estimate of the numerical solution x_s using Eq.(3.4).
3. Find out whether the actual solution is between a and x_{s_1} or between x_{s_1} and *b*. This is accomplished by checking the sign of the product $f\left(a\right)f\left(x_{s_1}\right)$.

   If $f\left(a\right)f\left(x_{s_1}\right) < 0$, the solution is between *a* and x_{s_1}.

   If $f\left(a\right)f\left(x_{s_1}\right) \geq 0$, the solution is between x_{s_1} and *b*.
4. Select the subinterval that contains the solution $\left(a\ to\ x_{s_1}, or\ x_{s_1}\ to\ b\right)$ is the new interval (*a*, *b*) and go back to step 2. Step 2 through 4 is repeated until a specified tolerance or error bound is attained.

The method of False Position always converges to an answer, provided a root is initially bracketed in the interval (a, b).

Exercise 3.2

Using the False Position method, find a root of the function $f(x) = e^x - 3x^2$ to an accuracy of 5 digits. The root is known to lie between 0.5 and 1.0.

Solution:

We apply the method of False Position with $a = 0.5$ and $b = 1.0$. Equation (3.4) is

$$x_s = \frac{af(b) - bf(a)}{f(b) - f(a)}$$

The calculations based on the method of False Position are shown in the Table 3.2.

n	a	b	f(a)	f(b)	Xs	f(Xs)	f(a)*f(Xs)	Rel. Error
1	0.5	1	0.89872	-0.2817	0.88067	0.085769907	0.07708324	1
2	0.88067	1	0.08577	-0.2817	0.90852	0.004414302	0.000378614	0.030654695
3	0.90852	1	0.00441	-0.2817	0.90993	0.000218671	9.65279E-07	0.00155095
4	0.90993	1	0.00022	-0.2817	0.91	1.08115E-05	2.36417E-09	7.67638E-05
5	0.91	1	1.1E-05	-0.2817	0.91001	5.34494E-07	5.7787E-12	3.7952E-06
6	0.91001	1	5.3E-07	-0.2817	**0.91001**	2.64238E-08	1.41234E-14	1.87624E-07
7	0.91001	1	2.6E-08	-0.2817	0.91001	1.30632E-09	3.4518E-17	9.2756E-09
8	0.91001	1	1.3E-09	-0.2817	0.91001	6.45808E-11	8.43631E-20	4.58559E-10
9	0.91001	1	6.5E-11	-0.2817	0.91001	3.19211E-12	2.06149E-22	2.267E-11
10	0.91001	1	3.2E-12	-0.2817	0.91001	1.58096E-13	5.0466E-25	1.12058E-12
11	0.91001	1	1.6E-13	-0.2817	0.91001	7.10543E-15	1.12334E-27	5.55107E-14
12	0.91001	1	7.1E-15	-0.2817	0.91001	0	0	2.44003E-15
13	0.91001	1	0	-0.2817	0.91001	0	0	1.22002E-16

The root is 0.91001 accurate to five digits.

3.4 NEWTON-RAPHSON METHOD

The Newton-Raphson method is the best-known method of finding roots of a function $f(x)$. The method is simple and fast. One drawback of this method is that it uses the derivative $f'(x)$ of the function as well as the function $f(x)$ itself. Hence, the Newton-Raphson method is usable only in problems where $f'(x)$ can be readily computed. Newton-Raphson method is also called *Newton's method*. Here, again we assume that $f(x)$ is continuous and differentiable and the equation is known to have a solution near a given point. Figure 3.4 illustrates the procedure used in Newton-Raphson method. The solution process starts by selecting point x_1 as the first estimate of the solution. The second estimate x_2 is found by drawing the tangent line to $f(x)$ at the point $(x_1, f(x_1))$ and determining the intersection point of the tangent line with the x-axis. The next estimate x_3 is the intersection of the tangent line to $f(x)$ at the point $(x_2, f(x_2))$ with the x-axis, and so on. The slope, $f'(x_1)$, of the tangent at point $(x_1, f(x_1))$ is written as

$$f'(x_1) = \frac{f(x_1) - 0}{x_1 - x_2} \qquad \textbf{(3.5)}$$

Rewriting Eq.(3.5) for x_2 gives

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \qquad \textbf{(3.6)}$$

Eq. (3.6) can be generalised for determining the next solution x_{i+1} from the current solution x_i as

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \qquad \textbf{(3.7)}$$

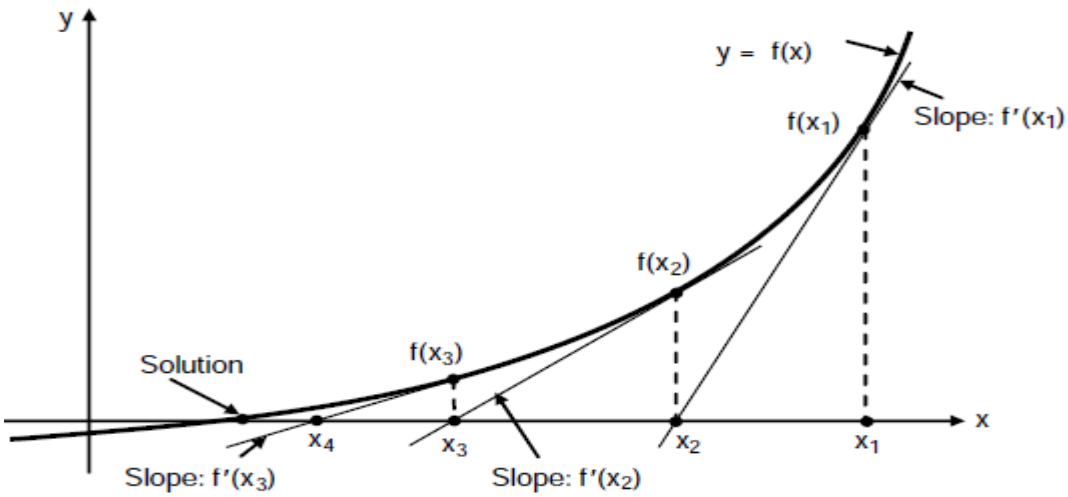

Fig. 3.4: Newton-Raphson method

The solution is obtained by repeated application of the iteration formula given by Eq.(3.7) for each successive value of 'i '.

Algorithm for Newton-Raphson Method:

1. Select a point x_1 as an initial guess of the solution.
2. For $i = 1, 2, \ldots$, until the error is smaller than a specified value, compute x_{i+1} by using Eq.(3.7).
3. Two error estimates that are generally used in Newton-Raphson method are given below:

   The iterations are stopped when the estimated relative error $\left|\frac{x_{i+1} - x_i}{x_i}\right|$ is smaller than a specified value $\in$.

   $$\left|\frac{x_{i+1} - x_i}{x_i}\right| \leq \in \qquad \textbf{(3.8)}$$

   The iterations are stopped when the absolute value of $f(x_i)$ is smaller than some number δ:

   $$f(x_i) \leq \delta \qquad \textbf{(3.9)}$$

4. The Newton-Raphson method, when successful, works well and converges fast. Convergence problems occur when the value of $f'(x)$ is close to zero in the vicinity of the solution, where $f(x) = 0$. Newton-Raphson method generally converges when $f(x), f'(x)$ and $f''(x)$ are all continuous, if $f'(x)$ is not zero at the solution and if the starting value x_1 is near the actual solution.

The Newton-Raphson's method converges if

$$\left| \frac{f(x)f''(x)}{[f'(x)]^2} \right| < 1 \qquad \textbf{(3.10)}$$

and its error ($\in_{i+1}$) formula is given by eq. (3.11)

$$\in_{i+1} \approx \frac{f''(\alpha)}{2f'(\alpha)} \qquad \textbf{(3.11)}$$

where α is the root of $f(x)$.

Equation (3.11) shows that the error at each stage is proportional to the sequence of the error in the previous stage. Hence, Newton-Raphson method has a quadratic convergence.

Exercise 3.3

Using Newton-Raphson method, find a root of the function $f(x) = e^x - 3x^2$ to an accuracy of 5 digits. The root is known to lie between 0.5 and 1.0.

Solution

The calculations based on the method of Newton-Raphson are shown in the Table 3.3.

i	Xi	f(Xi)	f'(Xi)	X(i+1)	ξ
0	1	-0.2817	-3.2817	0.91416	0.09391
1	0.91416	-0.0124	-2.9903	0.91002	0.00455
2	0.91002	-3E-05	-2.9757	0.91001	1.1E-05

3	0.91001	-2E-10	-2.9757	**0.91001**	6.6E-11
4	0.91001	0	-2.9757	0.91001	0
5	0.91001	0	-2.9757	0.91001	0
6	0.91001	0	-2.9757	0.91001	0
7	0.91001	0	-2.9757	0.91001	0
8	0.91001	0	-2.9757	0.91001	0

3.5 SUCCESSIVE APPROXIMATION METHOD

Suppose we are given an equation $f(x)=0$ whose roots are to be determined. The equation can be written as

$$x = f(x) \qquad \textbf{(3.12)}$$

Let $x = x_0$ be an initial approximation to the desired root α. Then, the first approximation x_1 is given by

$$x_1 = \psi(x_0)$$

The second approximation

$$x_2 = \psi(x_1)$$

The successive approximations are then given by

$$x_3 = \psi(x_2), x_4 = \psi(x_3), \quad , x_n = \psi(x_{n-1})$$

The sequence of approximations of $x_1, x_2, x_3, ..., x_n$ always converge to the root of $x = \psi(x)$ and it can be shown that if $|\psi'(x)| < 1$, when x is sufficiently close to the exact value c of the root and $x_n \to c$ as $n \to \infty$.

The convergence of $x_{i+1} = \psi(x_n)$, for $|\psi'(x)| < 1$

Exercise 3.4

Find a real root of $x^3 - 2x - 3 = 0$, correct to three decimal places using the Successive Approximation method.

Solution

The calculations are summarized in Table 3.4.

i	x_i	$\psi(x_i)$	ξ_i
0	1	1.70998	0.70998
1	1.70998	1.85856	0.08689
2	1.85856	1.88681	0.0152
3	1.88681	1.89208	0.0028
4	1.89208	1.89306	0.00052
5	1.89306	**1.89325**	9.6E-05
6	1.89325	1.89328	1.8E-05

3.6 SECANT METHOD

The secant method is very similar to the Newton-Raphson method. The main disadvantage of the Newton- Raphson method is that the method requires the determination of the derivatives of the function at several points. Often, the calculation of these derivatives takes too much time. In some cases, a closed-form expression for $f'(x)$ may be difficult to obtain or may not be available.

To remove this drawback of the Newton-Raphson method, the derivatives of the function being approximated by finite differences instead of being calculated analytically. In particular, the derivative $f'(x)$ is approximated by the backward difference

$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \qquad \textbf{(3.13)}$$

where x_i and x_{i-1} are two approximations to the root but does not require the condition

$f(x_i).\ f(x_{i-1}) < 0.$

Now, from the Newton-Raphson method, we have

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})} \qquad \textbf{(3.14)}$$

It should be noted here from Eq.(3.14) that this method requires two initial guess values x_0 and x_1 for the root.

The secant method evaluates the function only once in each iteration whereas the Newton-Raphson method evaluates two functions f and f' in each iteration. Therefore, the secant method is more efficient than the Newton-Raphson method.

Exercise 3.5

Find a root of the equation $x^3 - 8x - 5 = 0$ using the secant method.

Solution

The calculations are summarized in Table 3.5.

i	X(i-1)	f(X(i-1))	X(i)	f(X(i))	X(i+1)	f(X(i+1))
1	3	-2	3.5	9.875	3.08421	-0.3356
2	3.5	9.875	3.08421	-0.3356	3.09788	-0.0532
3	3.08421	-0.3356	3.09788	-0.0532	3.10045	0.00039
4	3.09788	-0.0532	3.10045	0.00039	3.10043	-4E-07
5	3.10045	0.00039	3.10043	-4E-07	**3.10043**	-4E-12
6	3.10043	-4E-07	3.10043	-4E-12	3.10043	0
7	3.10043	-4E-12	3.10043	0	3.10043	0

CHECK

Hence, the root is 3.1004 correct up to five significant figures.

3.7 SUMMARY

In this chapter, the techniques for the numerical solution of algebraic and transcendental equations have been presented. Numerical methods involving iterative solution of nonlinear equations are more powerful. These methods can be divided into two categories: Direct methods and Indirect (or iterative) methods. The indirect or iterative methods are further divided into two categories: bracketing and open method. The

bracketing methods require the limits between which the root lies, whereas the open methods require the initial estimation of the solution. Bisection and False Position methods are two known examples of the bracketing methods. Among the open methods, the Newton-Raphson and the method of Successive Approximation are most commonly used. The most popular method for solving a non-linear equation is the Newton-Raphson method and this method has a quadratic rate of convergence.

EXERCISE 3

1. Find one root of $e^x - 3x = 0$ correct to two decimal places using the method of **Bisection**.
2. Find the root of $\log x = \cos x$ correct to two decimal places using **Bisection method**
3. Find a real root of $\cos(x) - 3x + 5 = 0$. Correct to four decimal places using the method of **False Position** method.
4. Evaluate $\sqrt{29}$ to five decimal places by **Newton-Raphson** iterative method.
5. Find a real root of $2x - \log x - 9$ using the **Successive Approximation** method.
6. Determine a root of the equation $\sin(x) + 3\cos(x) - 2 = 0$ using the secant method.

   Answer: 1.2078
7. Find a root of the equation $x^6 - x - 1 = 0$ using the **secant method** approximations: $x_0 = 0$ and $x_1 = 1$.

CHAPTER FOUR

FINITE DIFFERENCES AND INTERPOLATION

4.1 INTRODUCTION

Interpolation is the technique of estimating the value of a function for any intermediate value of the independent variable. The process of computing or finding the value of a function for any value of the independent variable outside the given range is called *extrapolation*. Here, interpolation denotes the method of computing the value of the function $y = f(x)$ for any given value of the independent variable x when a set of values of $y = f(x)$ for certain values of x are known or given.

Hence, if (x_i, y_i), $i = 0, 1, 2, \ldots, n$ are the set of $(n + 1)$ given data points of the function $y = f(x)$, then the process of finding the value of y corresponding to any value of $x = x_i$ between x_0 and x_n, is called *interpolation.*

If the function $f(x)$ is known explicitly, then the value of y corresponding to any value of x can easily be obtained. On the other hand, if the function $f(x)$ is not known, then it is very hard to find the exact form of $f(x)$ with the tabulated values (x_i, y_i). In such cases, the function $f(x)$ can be replaced by a simpler, function, say, $\phi(x)$ which has the same values as $f(x)$ for $x_0, x_1, \ldots, x_n$. The function $\phi(x)$ is called the *interpolating* or *smoothing function* and any other value can be computed from $\phi(x)$.

If $\phi(x)$ is a polynomial, then $\phi(x)$ is called the *interpolating polynomial* and the process of computing the intermediate values of $y = f(x)$ is called the *polynomial interpolation*. In the study of interpolation, we make the following assumptions:

a) there are no sudden jumps in the values of the dependent variable for the period under consideration

b) the rate of change of figures from one period to another is uniform.

In this chapter, we present the study of interpolation based on the calculus of finite differences. The following important interpolation formulae obtained or derived based on forward, backward and central differences of a function are presented.

a) Newton's binomial expansion formula for equal intervals
b) Newton's forward interpolation formula for equal intervals
c) Newton's backward interpolation formula for equal intervals
d) Lagrange's formula for unequal intervals
e) Lagrange's formula for inverse interpolation
f) Gauss's forward interpolation formula
g) Gauss's backward interpolation formula

4.2 FINITE DIFFERENCE OPERATORS

Consider a function $y = f(x)$ defined on (a,b). x and y are the independent and dependent variables respectively.

If the points $x_0, x_1, ..., x_n$ are taken at equidistance i.e., $x_i = x_0 + ih, i = 0,1,2,...,n$, then the value of y, when $x = x_i$, is denoted as y_i, where $y_i = f(x_i)$. Here, the values of x are called arguments and the values of y are known as entries. The interval h is called the difference interval. The differences $y_1 - y_0, y_2 - y_1, ..., y_n - y_{n-1}$ are called the first differences of the function y. They are denoted by $\Delta y_0, \Delta y_1, ..., \Delta y_{n-1}$. That is

$$\begin{aligned} \Delta y_0 &= y_1 - y_0 \\ \Delta y_1 &= y_2 - y_1 \\ &\\ \Delta y_{n-1} &= y_n - y_{n-1} \end{aligned} \qquad \textbf{(4.1)}$$

The symbol Δ in Eq. (4.1) is called the difference operator.

4.2.1 Forward Differences

The *forward difference* or simply *difference* operator is denoted by Δ and may be defined as

$$\Delta f(x) = f(x+h) - f(x) \tag{4.2}$$

or writing in terms of y, at $x = x_i$, Eq.(4.2) becomes

$$\Delta f(x_i) = f(x_i + h) - f(x_i) \tag{4.3}$$

or $\Delta y_i = y_{i+1} - y_i,\ i = 0,1,2,...,n-1$

The differences of the first differences are called the *second differences* and they are denoted by $\Delta^2 y_0, \Delta^2 y_1, ..., \Delta^2 y_n$.

Generalising, we have

$$\begin{aligned} &\Delta^{n+1} f(x) = \Delta\left[\Delta^n f(x)\right] \ i.e.,\ \Delta^{n+1} y_i = \Delta\left[\Delta^n y_i\right],\ n = 0,1,2,... \\ &also,\ \Delta^{n+1} f(x) = \Delta^n\left[f(x+h) - f(x)\right] = \Delta^n f(x+h) - \Delta^n f(x) \end{aligned} \tag{4.4}$$

$$and, \qquad \Delta^{n+1} y_i = \Delta^n y_{i+1} - \Delta^n y_i,\ n = 0,1,2,... \tag{4.5}$$

The forward differences for the arguments $x_0, x_1, ..., x_5$ are shown in Table 4.1. Table 4.1 is called a *diagonal difference table* or *forward difference table*. The first term in Table 4.1 is y_0 and is called the *leading term*.

The differences $\Delta y_0, \Delta^2 y_0, \Delta^3 y_0, ...,$ are called the *leading differences*. Similarly, the differences with fixed subscript are called *forward differences*.

Table 4.1: Forward difference table

k	x_k	y_k	Δy_k	$\Delta^2 y_k$	$\Delta^3 y_k$	$\Delta^4 y_k$	$\Delta^5 y_k$
0	x_0	y_0					
			Δy_0				
1	x_1	y_1		$\Delta^2 y_0$			
			Δy_1		$\Delta^3 y_0$		
2	x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$	
			Δy_2		$\Delta^3 y_1$		$\Delta^5 y_0$
3	x_3	y_3		$\Delta^2 y_2$		$\Delta^4 y_1$	
			Δy_3		$\Delta^3 y_2$		
4	x_4	y_4		$\Delta^2 y_3$			
			Δy_4				
5	x_5	y_5					

4.2.2 Backward Differences

The *backward difference operator* is denoted by ∇ and it is defined as

$$\nabla f(x) = f(x) - f(x+h) \qquad \textbf{(4.6)}$$

Equation (4.6) can be written as

$$\nabla y_i = y_i - y_{i-1}, \quad i = n, n-1, ..., 1. \qquad \textbf{(4.7)}$$

The differences in Eq.(4.7) are called *first differences*. The *second differences* are denoted by $\nabla^2 y_0, \nabla^2 y_1, ..., \nabla^2 y_n$.

Generalising, we have

$$\nabla^k y_i = \nabla^{k-1} y_i - \nabla^{k-1} y_{i-1}, \quad i = n, n-1, \ldots, k.$$
$$\textit{where}, \nabla^0 y_i = y_i, \quad \nabla^1 y_i = \nabla y_i \tag{4.8}$$

The backward differences written in a tabular form is shown in Table 4.2. In Table 4.2, the differences $\nabla^n y$ with a fixed subscript 'i' lie along the diagonal upward sloping.

Table 4.2: Backward difference table

k	x_k	y_k	∇y_k	$\nabla^2 y_k$	$\nabla^3 y_k$	$\nabla^4 y_k$	$\nabla^5 y_k$
0	x_0	y_0					
			∇y_1				
1	x_1	y_1		$\nabla^2 y_2$			
			∇y_2		$\nabla^3 y_3$		
2	x_2	y_2		$\nabla^2 y_3$		$\nabla^4 y_4$	
			∇y_3		$\nabla^3 y_4$		$\nabla^5 y_5$
3	x_3	y_3		$\nabla^2 y_4$		$\nabla^4 y_5$	
			∇y_4		$\nabla^3 y_5$		
4	x_4	y_4		$\nabla^2 y_5$			
			∇y_5				
5	x_5	y_5					

4.2.3 Central Differences

The central difference operator is denoted by the symbol δ and is defined by

$$\delta f(x) = f\left(x + h/2\right) - f\left(x - h/2\right) \tag{4.9}$$

where h is the interval of differencing.

In terms of y, the first central difference is written as

$$\delta y_i = y_{i+\frac{1}{2}} - y_{i-\frac{1}{2}} \tag{4.10}$$

Generalising,

$$\delta^n y_i = \delta^{n-1} y_{i+\frac{1}{2}} - \delta^{n-1} y_{i-\frac{1}{2}} \tag{4.11}$$

The central difference table for the five arguments $x_0, x_1, ..., x_5$ is shown in Table 4.3.

Table 4.3: Central difference table

k	x_k	y_k	δy_k	$\delta^2 y_k$	$\delta^3 y_k$	$\delta^4 y_k$	$\delta^5 y_k$
0	x_0	y_0					
			$\delta y_{\frac{1}{2}}$				
1	x_1	y_1		$\delta^2 y_1$			
			$\delta y_{\frac{3}{2}}$		$\delta^3 y_{\frac{3}{2}}$		
2	x_2	y_2		$\delta^2 y_2$		$\delta^4 y_2$	
			$\delta y_{\frac{5}{2}}$		$\delta^3 y_{\frac{5}{2}}$		$\delta^5 y_{\frac{5}{2}}$
3	x_3	y_3		$\delta^2 y_3$		$\delta^4 y_3$	
			$\delta y_{\frac{7}{2}}$		$\delta^3 y_{\frac{7}{2}}$		
4	x_4	y_4		$\delta^2 y_4$			
			$\delta y_{\frac{9}{2}}$				
5	x_5	y_5					

It is noted in Table 4.3 that all odd differences have fraction suffices and all the even differences are with integral suffices.

Exercise 4.1

a) Construct the forward difference table and the horizontal table for the following data:

x	1	2	3	4	5
$y = f(x)$	4	6	9	12	17

b) Construct a forward difference table for the following data:

x	0	10	20	30
$y = f(x)$	0	0.174	0.347	0.518

Solution

a)

k	x_k	y_k	Δy_k	$\Delta^2 y_k$	$\Delta^3 y_k$	$\Delta^4 y_k$
0	1	4				
			2			
1	2	6		1		
			3		-1	
2	3	9		0		3
			3		2	
3	4	12		2		
			5			
4	5	17				

Same for the horizontal table (i.e. backward difference) but different headings.

4.2.4 Error Propagation in a Difference Table

Let $y_0, y_1, y_2, ..., y_n$ be the true values of a function and suppose the value y_5 to be affected with an error, so that its erroneous value is $y_5 + \in$. Then the successive differences of the y are as shown in Table 4.4.

Table 4.4: Error propagation in a difference table

k	x_k	y_k	Δy_k	$\Delta^2 y_k$	$\Delta^3 y_k$
0	x_0	y_0			
			Δy_0		
1	x_1	y_1		$\Delta^2 y_0$	
			Δy_1		$\Delta^3 y_0$
2	x_2	y_2		$\Delta^2 y_1$	
			Δy_2		$\Delta^3 y_1$
3	x_3	y_3		$\Delta^2 y_2$	
			Δy_3		$\Delta^3 y_2 + \in$
4	x_4	y_4		$\Delta^2 y_3 + \in$	
			$\Delta y_4 + \in$		$\Delta^3 y_3 - 3\in$
5	x_5	$y_5 + \in$		$\Delta^2 y_4 - 2\in$	
			$\Delta y_5 - \in$		$\Delta^3 y_4 + 3\in$

6	x_6	$y_6+\in$		$\Delta^2 y_5+\in$	
			Δy_6		
7	x_7	$y_7+\in$			

Table 4.4 shows that the effect of an error increases with the successive differences, that the coefficients of the $\in$'s are the binomial coefficients with alternating signs, and that the algebraic sum of the errors in any difference column is zero. The same effect is also true for the horizontal difference Table 4.2.

Exercise 4.2

Table below gives the values of a polynomial of degree five. It is given that $f(4)$ is in error. Correct the value of $f(4)$.

x	1	2	3	4	5	6	7
$y = f(x)$	0.975	-0.6083	-3.5250	-5.5250	-6.3583	4.2250	36.4750

Solution: (@ Lectures)

4.2.5 PROPERTIES OF THE OPERATOR Δ

1. If c is a constant then $\Delta c = 0$.

2. Δ is distributive, i.e.,

$$\Delta\left[f(x) \pm g(x)\right] = \Delta f(x) \pm \Delta g(x).$$

Similarly, we have

$$\Delta\left[f(x) - g(x)\right] = \Delta f(x) - \Delta g(x).$$

3. If c is a constant then

$$\Delta\left[cf(x)\right] = c\Delta f(x).$$

From properties 2 and 3 above, it is observed that Δ is a linear operator.

4. If m and n are positive integers then $\Delta^m \Delta^n f(x) = \Delta^{m+n} f(x)$.

In a similar manner, we can prove the following properties:

5. $\Delta\left[f_1(x) + f_2(x) + \quad f_n(x)\right] = \Delta f_1(x) + \Delta f_2(x) + \quad + \Delta f_n(x)$

6. $\Delta\left[f(x)g(x)\right] = f(x+h)\Delta g(x) + g(x)\Delta f(x)$.

7. $\Delta\left[\dfrac{f(x)}{g(x)}\right] = \dfrac{g(x)\Delta f(x) - f(x)\Delta g(x)}{g(x)g(x+h)}$.

Exercise 4.3

Evaluate $\Delta\left(e^{ax}\log bx\right)$.

Solution: (@ Lectures)

Exercise 4.4

Evaluate $\Delta\left[\dfrac{5x+12}{x^2+5x+6}\right]$, (Let $h = 1$)

Solution: (@ Lectures)

4.2.6 Difference Operator

a) Shift operator (E):

The shift operator is defined as

$$Ef(x) = f(x+h) \qquad \textbf{(4.12)}$$

or $Ey_i = y_{i+1}$ **(4.13)**

Hence, shift operator sifts the function value y_i to the next higher value y_{i+1}. The second shift operator gives

$$E^2 f(x) = E\left[Ef(x)\right] = E\left[f(x+h)\right] = f(x+2h) \qquad \textbf{(4.14)}$$

E is linear and obeys the law of indices.

Generalising,

$$E^n f(x) = f(x+nh) \text{ or } E^n y_i = y_{i+nh} \quad \textbf{(4.15)}$$

The inverse shift operator E^{-1} is defined as

$$E^{-1} f(x) = f(x-h) \quad \textbf{(4.16)}$$

In a similar manner, second and higher inverse operators are given by

$$E^{-2} f(x) = f(x-2h) \text{ and } E^{-n} f(x) = f(x-nh)$$

b) Average Operator (μ):

$$\mu f(x) = \frac{1}{2}\left[f\left(x + h/2\right) + f\left(x - h/2\right)\right]$$

i.e. $$\mu y_i = \frac{1}{2}\left[y_{i+1/2} + y_{i-1/2} \right] \quad \textbf{(4.17)}$$

Exercise 4.5

Evaluate $\left(\frac{\Delta^2}{E}\right)x^3$

Solution: (@ Lectures)

Exercise 4.6

Prove that $e^x = \frac{\Delta^2}{E} e^x . \frac{Ee^x}{\Delta^2 e^x}$, the interval of differencing being *h*.

Solution: (@ Lectures)

The relationships among the various operators are shown in Table 4.5.

Table 4.5.

	E	Δ	∇

E	E	$\Delta+1$	$(1-\nabla)^{-1}$
Δ	$E-1$	Δ	$(1-\nabla)^{-1}-1$
∇	$1-E^{-1}$	$1-(1+\Delta)^{-1}$	∇
μ	$\frac{1}{2}\left(E^{1/2}+E^{-1/2}\right)$	$\left(1+\frac{1}{2}\Delta\right)(1+\Delta)^{1/2}$	$\left(1-\frac{1}{2}\Delta\right)(1-\Delta)^{-1/2}$

4.3 INTERPOLATION WITH EQUAL INTERVAL

Here, we assume that for function $y=f(x)$, the set of $(n+1)$ functional values $y_0, y_1, \quad, y_n$ are given corresponding to the set of $(n+1)$ equally spaced values of the independent variable, $x_i = x_0 + ih, i = 0,1, \quad, n$, where h is the spacing**.**

4.3.1 Missing Values

Let a function $y=f(x)$ have equally spaced values $x_0, x_1, \quad, x_n$ of the argument and $y_0, y_1, \quad, y_n$ denote the corresponding values of the function. If one or more values of $y=f(x)$ is or are missing, we can determine the missing values by employing the relationship between the operators E and Δ.

4.3.2 Newton's Binomial Expansion Formula

Suppose $y_0, y_1, \quad, y_n$ denote the values of the function $y=f(x)$ corresponding to the values $x_0, x_0+h, x_0+2h, \quad, x_0+nh$ of x. Let one of the values of y be missing, since n values of the functions are known. Therefore, we have

$$\Delta^n y_0 = 0 \qquad \textbf{(4.18)}$$

$$\Rightarrow (E-1)^n y_0 = 0$$

Expanding Eq.(4.18), we have

$$\left[E^n - {}^nC_1E^{n-1} + {}^nC_2E^{n-2} + \quad + (-1)^n\right] y_0 = 0 \tag{4.19}$$

$$\Rightarrow E^n y_0 - nE^{n-1}y_0 + \frac{n(n-1)}{2!}E^{n-2}y_0 + \quad + (-1)^n y_0 = 0$$

$$\Rightarrow y_n - ny_{n-1} + \frac{n(n-1)}{2}y_{n-2} + \quad + (-1)^n y_0 = 0 \tag{4.20}$$

Equation (4.20) is quite useful in determining the missing values without actually constructing the difference table.

Exercise 4.7

Determine the missing entry in the following table.

x	0	1	2	3	4
$y = f(x)$	1	4	17	-	97

Solution: (@ Lectures)

4.3.3 Newton's Forward Interpolation Formula

Let $y = f(x)$, which takes the values $y_0, y_1, y_2, \quad, y_n$, that is the set of $(n+1)$ functional values $y_0, y_1, y_2, \quad, y_n$ are given corresponding to the set of $(n+1)$ equally spaced values of the independent variable,

$x_i = x_0 + ih,\ i = 0,1,2, \quad, n$ where h is the spacing. Let $\phi(x)$ be a polynomial of the n^{th} degree in x taking the same values as y corresponding to $x = x_0, x_1, \quad, x_n$. Then, $\phi(x)$ represents the continuous function $y = f(x)$ such that $f(x_i) = \phi(x_i)$ for $i = 0,1,2, \quad, n$ and at all other points $f(x) = \phi(x) + R(x)$ where $R(x)$ is called the *error term* (remainder term) of the interpolation formula.

Let

$$\phi(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)(x - x_1)(x - x_2) + \quad + a_n(x - x_0)(x - x_1)(x - x_2) \quad (x - x_{n-1})$$

(4.21)

and

$$\phi(x_i) = y_i;\ i = 0,1,2, \quad ,n \tag{4.22}$$

The constants $a_0, a_1, a_2, \quad , a_n$ can be determined as follows:

Substituting $x = x_0, x_1, x_2, \quad , x_n$ successively in Eq.(4.21), we get

$$a_0 = y_0 \tag{4.23}$$

$$y_1 = a_0 + a_1(x_1 - x_0) = y_0 + a_1(x_1 - x_0)$$

$$a_1 = \frac{y_1 - y_0}{x_1 - x_0} = \frac{\Delta y_0}{h} \tag{4.24}$$

$$y_2 = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)$$

$$y_2 - y_0 - a_1(x_2 - x_0) = a_2(x_2 - x_0)(x_2 - x_1)$$

$$(y_2 - y_0) - \frac{(y_1 - y_0)}{(x_1 - x_0)}(x_2 - x_0) = a_2(x_2 - x_0)(x_2 - x_1)$$

$$(y_2 - y_0) - \frac{(y_1 - y_0)}{h} 2h = a_2 2hh$$

$$a_2 = \frac{y_2 - 2y_1 + y_0}{2h^2} = \frac{\Delta^2 y_0}{2!h^2} \tag{4.25}$$

Similarly, we obtain

$$a_3 = \frac{\Delta^3 y_0}{3!h^3}, \quad , a_n = \frac{\Delta^n y_0}{n!h^n}$$

Hence, from Eq.(4.21), we have

$$\phi(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) + \quad + \frac{\Delta^n y_0}{n!h^n}(x - x_0)(x - x_1) \quad (x - x_{n-1})$$

(4.26)

Let $x = x_0 + uh$

or $x - x_0 = uh$

and

$$x - x_1 = (x - x_0) - (x_1 - x_0) = uh - h = (u-1)h$$

$$x - x_2 = (x - x_1) - (x_2 - x_1) = (u-1)h - h = (u-2)h, etc$$

(4.27)

Using the values from Eq.(4.27), Eq.(4.26) reduces to

$$\phi(x) = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + \quad + \frac{u(u-1)(u-2)\quad(u-(n-1))}{n!}\Delta^n y_0$$

(4.28)

This is sometimes written in operator form

$$\phi(x) = \left\{1 + u\Delta + \frac{u(u-1)}{2!}\Delta^2 + \frac{u(u-1)(u-2)}{3!}\Delta^3 + \quad + \frac{u(u-1)(u-2)\quad(u-(n-1))}{n!}\Delta^n\right\} y_0$$

which you no doubt recognize as the binomial expansion of

$$\phi(x) = (1 + \Delta)^u \, y_0$$

The formula given in Eq.(4.28) is called the *Newton's forward interpolation formula*. This formula is used to interpolate the values of y near the beginning of a set of equally spaced tabular values. This formula can also be used for extrapolating the values of y a little backward of y_0.

Exercise 4.8

Find $y = e^{3x}$ for $x = 0.05$ using the following table:

x	0	0.1	0.2	0.3	0.4
e^{3x}	1	1.3499	1.8221	2.4596	3.3201

Solution: (@ Lectures)

4.3.4 Newton's Backward Interpolation Formula

Newton's forward interpolation formula is not suitable for interpolation values of y near the end of a table of values.

Let $y = f(x)$, be a function which takes the values $y_0, y_1, y_2, \quad , y_n$, corresponding to the values $x_0, x_1, x_2, \quad , x_n$ of the independent variable x. Let the values of x be equally spaced with h as the interval of differencing.

That is, $x_i = x_0 + ih,\ i = 0,1,2, \quad , n$ where h is the spacing.

Let $\phi(x)$ be a polynomial of the n^{th} degree in x taking the same values as y corresponding to $x = x_0, x_1, \quad , x_n$. Then, $\phi(x)$ represents the continuous function $y = f(x)$ such that $f(x_i) = \phi(x_i)$ for $i = 0,1,2, \quad , n$ and at all other points $f(x) = \phi(x) + R(x)$ where $R(x)$ is called the *error term* (remainder term) of the interpolation formula.

Let

$$\phi(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + a_3(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \quad + a_n(x - x_n)(x - x_{n-1})(x - x_{n-2}) \quad (x - x_0)$$

(4.29)

and

$$\phi(x_i) = y_i;\ i = n, n-1, n-2, \quad , 0$$

(4.30)

$$x_{n-i} = x_{n-ih},\ i = 1,2, \quad , n$$

The constants $a_0, a_1, a_2, \quad , a_n$ can be determined as follows:

Substituting $x = x_n, x_{n-1}, x_{n-2}, \quad , x_0$ successively in Eq.(4.29), we get

$$a_0 = y_n \qquad \textbf{(4.31)}$$

$$y_{n-1} = a_0 + a_1(x_{n-1} - x_n) = y_n + a_1(x_{n-1} - x_n)$$

$$a_1 = \frac{y_{n-1} - y_n}{x_{n-1} - x_n} = \frac{\nabla y_n}{h}$$

(4.32)

Similarly, we obtain

$$a_2 = \frac{\nabla^2 y_n}{2!h^2}, \quad , a_n = \frac{\nabla^n y_n}{n!h^n}$$

Hence, from Eq.(4.29), we have

$$\phi(x) = y_n + \frac{\nabla y_n}{h}(x - x_n) + \frac{\nabla^2 y_0}{2!h^2}(x - x_n)(x - x_{n-1}) + \quad + \frac{\nabla^n y_n}{n!h^n}(x - x_n)(x - x_{n-1}) \quad (x - x_0)$$

(4.33)

Let

$$x = x_n + vh$$

or

$$x - x_n = vh$$

and

$$x - x_{n-1} = (v+1)h$$

$$x - x_0 = (v + n - 1)h$$

(4.34)

Using the values from Eq.(4.34), Eq.(4.33) reduces to

$$\phi(x) = y_n + v\nabla y_n + \frac{v(v+1)}{2!}\nabla^2 y_n + \frac{v(v+1)(v+2)}{3!}\nabla^3 y_n + \quad + v(v+1) \quad \frac{(v+(n-1))}{n!}\nabla^n y_n$$

(4.35)

where,

$$v = \frac{x - x_n}{h}$$

This is sometimes written in operator form

$$\phi(x) = \left\{1 + v\nabla + \frac{v(v+1)}{2!}\nabla^2 + \frac{v(v+1)(v+2)}{3!}\nabla^3 + \quad + v(v+1) \quad \frac{(v+(n-1))}{n!}\nabla^n \right\} y_n$$

which you no doubt recognize as the binomial expansion of

$$\phi(x) = (1 - \nabla)^{-v} y_n$$

The formula given in Eq.(4.35) is called the *Newton's backward interpolation formula*. This formula is used to interpolate the values of y near the ending of a set of equally spaced tabular values. This formula can also be used for extrapolating the values of y a little backward of y_n.

Exercise 4.9

Calculate the value of $f(84)$ for the data given in the table below:

x	40	50	60	70	80	90
$f(x)$	204	224	246	270	296	324

Solution: (@ Lectures)

4.3.5 Central Difference Interpolation Formulae

In this section, we derive some important interpolation formulae by means of central differences of a function, which are quite frequently employed in engineering and scientific computations.

In particular, we develop central difference formulae which are best suited for interpolation near the middle of a tabulated data set. The following central difference formulae are presented:

1. Gauss's forward interpolation formula
2. Gauss's backward interpolation formula

Let the function $y = y_x = f(x)$ be given for $(2n+1)$ equi-spaced values of argument $x_0, x_0 \pm h, x_0 \pm 2h, \quad , x_0, x_h$. The corresponding values of y be $y_i \left(i = 0, \pm 1, \pm 2, \quad , \pm n\right)$.

Also, let $y = y_0$ denote the central ordinate corresponding to $x = x_0$. We can then form the difference table as shown in Table 4.6. Table 4.7 shows the same Table 4.6 written using the Sheppard's operator δ, in which the relation $= \Delta E^{-1/2}$ was used. Tables 4.6 and 4.7 are known as central difference tables.

Tables 4.6: Central Difference Table

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$	$\Delta^6 y$
$x_0 - 3h$	y_{-3}						
		Δy_{-3}					
$x_0 - 2h$	y_{-2}		$\Delta^2 y_{-3}$				
		Δy_{-2}		$\Delta^3 y_{-3}$			
$x_0 - h$	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$		
		Δy_{-1}		$\Delta^3 y_{-2}$		$\Delta^5 y_{-3}$	
x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$		$\Delta^6 y_{-3}$

		Δy_0		$\Delta^3 y_{-1}$		$\Delta^5 y_{-2}$	
$x_0 + h$	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$		
		Δy_1		$\Delta^3 y_0$			
$x_0 + 2h$	y_2		$\Delta^2 y_1$				
		Δy_2					
$x_0 + 3h$	y_3						

Tables 4.7: Central Difference Table

x	y	δy	$\delta^2 y$	$\delta^3 y$	$\delta^4 y$	$\delta^5 y$	$\delta^6 y$
$x_0 - 3h$	y_{-3}						
		$\delta y_{-5/2}$					
$x_0 - 2h$	y_{-2}		$\delta^2 y_{-2}$				
		$\delta y_{-3/2}$		$\delta^3 y_{-3/2}$			
$x_0 - h$	y_{-1}		$\delta^2 y_{-1}$		$\delta^4 y_{-1}$		
		$\delta y_{-1/2}$		$\delta^3 y_{-1/2}$		$\delta^5 y_{-1/2}$	

BACKWARDS

x_0	y_0		$\delta^2 y_0$		$\delta^4 y_0$		$\delta^6 y_0$
		$\delta y_{1/2}$		$\delta^3 y_{1/2}$		$\delta^5 y_{1/2}$	
$x_0 + h$	y_1		$\delta^2 y_1$		$\delta^4 y_1$	FORWARDS	
		$\delta y_{3/2}$		$\delta^3 y_{3/2}$			
$x_0 + 2h$	y_2		$\delta^2 y_2$				
		$\delta y_{5/2}$					
$x_0 + 3h$	y_3						

4.3.5.1 Gauss's Forward Interpolation Formula

In deriving the Gauss's forward interpolation formula, we assume the differences lie on the bottom solid lines in Table 4.8 and they are of the form

$$y_u = y_0 + G_1\Delta y_0 + G_2\Delta^2 y_{-1} + G_3\Delta^3 y_{-1} + G_4\Delta^4 y_{-2} + \quad \textbf{(4.36)}$$

Table 4.8: Gauss's forward and backward interpolation formulae

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$	$\Delta^6 y$
$x_0 - 3h$	y_{-3}						
		Δy_{-3}					
$x_0 - 2h$	y_{-2}		$\Delta^2 y_{-3}$				
		Δy_{-2}		$\Delta^3 y_{-3}$			
$x_0 - h$	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$	BACKWARDS	

		Δy_{-1}		$\Delta^3 y_{-2}$		$\Delta^5 y_{-3}$	
x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$		$\Delta^6 y_{-3}$
		Δy_0		$\Delta^3 y_{-1}$		$\Delta^5 y_{-2}$	
$x_0 + h$	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$	**FORWARDS**	
		Δy_1		$\Delta^3 y_0$			
$x_0 + 2h$	y_2		$\Delta^2 y_1$				
		Δy_2					
$x_0 + 3h$	y_3						

Or, we assume the differences lie on the bottom solid lines in Table 4.7 and they are of the form

$$y_u = y_0 + G_1\delta y_{1/2} + G_2\delta^2 y_0 + G_3\delta^3 y_{1/2} + G_4\delta^4 y_0 + G_5\delta^5 y_{1/2} +$$

(4.37)

For $u \in [0,1]$, where $G_1, G_2, \quad , G_n$ are binomial coefficients to be determined by letting

$$G_{2n} = \binom{u+n-1}{2n}$$

$$G_{2n+1} = \binom{u+n}{2n+1}$$

$$for \qquad n \geq 0$$

These coefficients are derived from the Newton's forward interpolation formula.

Hence, the Gauss's forward interpolation formula can be written as

$$y_u = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_{-1} + \frac{(u+1)u(u-1)}{3!}\Delta^3 y_{-1} + \frac{(u+1)u(u-1)(u-2)}{4!}\Delta^4 y_{-2} +$$

Similarly,

$$y_u = y_0 + u\delta y_{1/2} + \frac{u(u-1)}{2!}\delta^2 y_0 + \frac{(u+1)u(u-1)}{3!}\delta^3 y_{1/2} + \frac{(u+1)u(u-1)(u-2)}{4!}\delta^4 y_0 +$$

(4.38)

4.3.5.2 Gauss's Backward Interpolation Formula

The Gauss's backward interpolation formula uses the differences which lie on the upper dashed line in Table 4.8 and can be assumed of the form

$$y_u = y_0 + G^*_1 \Delta y_{-1} + G^*_2 \Delta^2 y_{-1} + G^*_3 \Delta^3 y_{-2} + G^*_4 \Delta^4 y_{-2} +$$

(4.39)

Or, we assume the differences lie on the upper dashed lines in Table 4.7 and they are of the form

$$y_u = y_0 + G^*_1 \delta y_{-1/2} + G^*_2 \delta^2 y_0 + G^*_3 \delta^3 y_{-1/2} + G^*_4 \delta^4 y_0 + G^*_5 \delta^5 y_{-1/2} +$$

For $u \in [0,1]$, where $G^*_1, G^*_2, \quad , G^*_n$ are binomial coefficients to be determined by letting

$$G^*_{2n} = \binom{u+n}{2n}$$

$$G^*_{2n+1} = \binom{u+n}{2n+1}$$

$$for \qquad n \geq 0$$

These coefficients are derived from the Newton's backward interpolation formula.

Hence, the Gauss's backward interpolation formula can be written as

$$y_u = y_0 + u\Delta y_{-1} + \frac{u(u+1)}{2!}\Delta^2 y_{-1} + \frac{(u+1)u(u-1)}{3!}\Delta^3 y_{-2} + \frac{(u+2)(u+1)u(u-1)}{4!}\Delta^4 y_{-2} +$$

Similarly,

$$y_u = y_0 + u\delta y_{-1/2} + \frac{u(u+1)}{2!}\delta^2 y_0 + \frac{(u+1)u(u-1)}{3!}\delta^3 y_{-1/2} + \frac{(u+2)(u+1)u(u-1)}{4!}\delta^4 y_0 +$$

(4.40)

Exercise 4.10

Use Gauss's forward interpolation formula to find y for $x = 20$ given that

x	11	15	19	23	27
y	19.5673	18.8243	18.2173	17.1236	16.6162

Solution: (@ Lectures)

4.4 INTERPOLATION WITH UNEQUAL INTERVAL

The Newton's forward and backward interpolation formulae are applicable only when the values of *n* are given at equal intervals. In this section, we present Lagrange's formula for unequal intervals.

4.4.1 Lagrange's Formula for Unique Intervals

Suppose that we wish to interpolate arbitrary functions at a set of fixed nodes $x_0, x_1, \quad , x_n$. We can interpolate any function f by the Lagrange form of the interpolation polynomial as:

$$p_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i) \tag{4.41}$$

where, the function p_n is a linear combination of the polynomials f_i at nodes $x_0, x_1, \quad , x_n$ and the cardinal polynomial l_i, which is

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \left(\frac{x - x_j}{x_i - x_j} \right), \quad (0 \leq i \leq n) \tag{4.42}$$

equation (4.42) indicates that $l_i(x)$ is the product of n linear factors.

$$l_i(x) = \left(\frac{x - x_0}{x_i - x_0} \right) \left(\frac{x - x_1}{x_i - x_1} \right) \left(\frac{x - x_2}{x_i - x_2} \right) \quad \left(\frac{x - x_{i-1}}{x_i - x_{i-1}} \right) \left(\frac{x - x_{i+1}}{x_i - x_{i+1}} \right) \quad \left(\frac{x - x_n}{x_i - x_n} \right)$$

NB: ***The denominators are just numbers; the variable x occurs only in the numerators***

Exercise 4.11

Write out the cardinal polynomials appropriate to the problem of interpolating the following table and give the Lagrange form of the interpolating polynomial.

x	1/3	1/4	1
$f(x)$	2	-1	7

Solution: (@ Lectures)

4.5 SUMMARY

Interpolation is the method of computing the value of the function $y = f(x)$ for any given value of the independent variable x when a set of values of $y = f(x)$ for certain values of x are given. The study of interpolation is based on the assumption that there are no sudden jumps in the values of the dependent variable for the period under consideration. In this chapter, the study of interpolation was presented based on the calculus of finite differences. Some important interpolation formulae by means of forward, backward and central differences of a function, which are frequently used in scientific and engineering calculations, were also presented.

EXERCISE 4

1. The following is a table of values of a polynomial of degree 5. It is given that f (3) is in error. Correct the error.

x	0	1	2	3	4	5	6
$y = f(x)$	1	2	33	254	1054	3126	7777

2. Show that

   i. $\Delta \sin x = 2\cos\left(x + h/2\right)\sin h/2$

   ii. $\Delta \tan^{-1} x = \tan^{-1}\left[\dfrac{h}{1+hx+x^2}\right]$

   iii. $\Delta^3 = E^3 - 3E^2 + 3E - 1$

   iv. $\Delta \cosh\left(a+bx\right) = 2\sinh\dfrac{a+b\left(x+h\right)+a+bx}{2}\sinh\dfrac{a+b\left(x+h\right)-a-bx}{2}$

   v. $\Delta^r y_x = \nabla^r y_{x+r}$

3. Prove the following:

   a) $\nabla = \dfrac{E-1}{E}$

   b) $\left(1+\Delta\right)\left(1-\Delta\right) = 1$

   c) $\Delta\nabla = \Delta - \nabla$

4. Evaluate

   i. $\left(2\Delta^2 + \Delta - 1\right)\left(x^2 + 2x + 1\right)$

   ii. $\Delta^6\left(ax-1\right)\left(bx^2-1\right)\left(cx^3-1\right)$

   iii. $\Delta\left[\dfrac{x^2}{\cos 2x}\right]$

5. Determine the missing entry in the following table.

x	0	1	2	3	4	5
$y = f\left(x\right)$	1	3	11	-	189	491

6. The profits of a company (in thousands of GH¢) are given below:

year	1990	1993	1996	1999	2002
Profit y = f(x)	120	100	111	108	99

Calculate the total profit at 1991.

7. Given that $\sqrt{15500} = 124.4990$, $\sqrt{15510} = 124.5392$, $\sqrt{15520} = 124.5793$ and $\sqrt{15530} = 124.6194$, find the value of $\sqrt{15516}$.

8. A function $f(x)$ is defined by the following table

x	-4	-2	0	2	4	6	8
$f(x)$	277	51	1	-17	-147	-533	-1319

Find $f(4.4)$ using the Gregory – Newton backward difference formula.

9. Use Gauss's backward interpolation formula to find y for $x = 20$ from **Exercise 4.10.**

10. Use Gauss's backward interpolation formula to find the sales for the year 1986 from the following data:

Year	1951	1961	1971	1981	1991	2001
Sales (in thousands)	13	17	22	28	41	53

11. Using Lagrange's interpolation formula, find the value of y corresponding to $x = 10$ from the following data.

x	5	6	9	11
$y = f(x)$	380	-2	196	508

12. A slider in a machine moves along a fixed straight rod. Its distance $x(m)$ along the rod are given in the following table for various values of the time t (seconds).

$t(\text{sec})$	1	2	3	4	5	6
$x(m)$	0.0201	0.0844	0.3444	1.0100	2.3660	4.7719

Find the velocity and acceleration of the slider at time $t = 6$ sec.

13. Derive the first-order derivative formulas of a function *f*(*x*) and list the order of their error term used in
    (a) forward difference method
    (b) central difference method

14. A function $f(x)$ is defined by the following table

x	-4	-2	0	2	4	6	8
$f(x)$	277	51	1	-17	-147	-533	-1319

Find
    a) $f(-3)\ and\ f(1.6)$ using the **Gregory - Newton forward difference formula**
    b) $f(0.2) and\ f(3.1)$ using the **Gauss central difference formula**
    c) $f(4.4) and\ f(7)$ using the **Gregory – Newton backward difference formula**.

15. Given the table

x	$f(x)$

-2	-2.63906
0	-2.48491
5	-1.94591
6	-1.79176

Use **Lagrangian** interpolation to find the value of

a) $f(-0.8)$

b) $f(0.8)$

c) $f(5.5)$

CHAPTER FIVE

NUMERICAL INTEGRATION

5.1 INTRODUCTION

If $F(x)$ is a differentiable function whose derivative is $f(x)$, then we can evaluate the definite integral I as

$$I = \int_a^b f(x)dx = F(b) - F(a),\ F'(x) = f(x) \tag{5.1}$$

Equation (5.1) is known as the fundamental theorem of calculus. Most integrals can be evaluated by the formula given by Eq. (5.1) and there exists many techniques for making such evaluations. However, in many applications in science and engineering, most integrals cannot be evaluated because most integrals do not have anti-derivatives $F(x)$ expressible in terms of elementary functions.

In other circumferences, the integrands could be empirical functions given by certain measured values. In all these instances, we need to resort to numerical methods of integration. It should be noted here that, sometimes, it is difficult to evaluate the integral by analytical methods. Numerical integration (or numerical quadrature, as it is sometimes called) is an alternative approach to solve such problems. As in other numerical techniques, it often results in approximate solution. The integration can be performed on a continuous function or a set of data.

The integration given by Eq. (5.1) is shown in Fig. 5.1. The integration shown in Fig. 5.1 is called closed since the function values at the two points (a, b) where the limits of integration are located are used to find the integral. In open integration, information on the function at one or both limits of integration is not required.

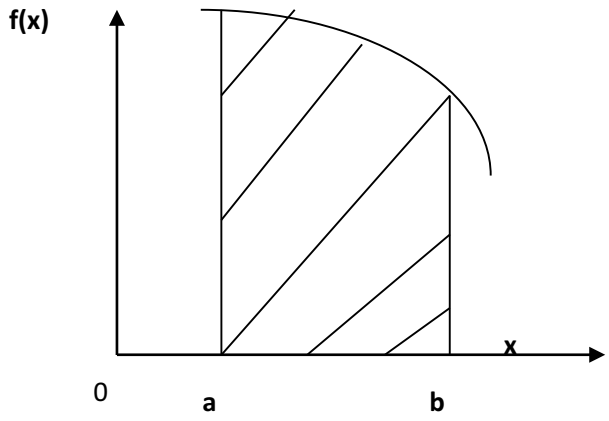

Fig: 5.1

The range of integration (b – a) is divided into a finite number of intervals in numerical integration. The integration techniques consisting of equal intervals are based on formulas known as Newton-Cotes closed quadrature formulas.

In this chapter, we present the following methods of integration with illustrative examples:

1. Trapezoidal rule.
2. Simpson's 1/3 rule.
3. Simpson's 3/8 rule.
4. Boole's and Weddle's rules.

5.2 RELATIVE ERROR

Suppose we are required to evaluate the definite integral

$$I = \int_a^b f(x)\,dx$$

In numerical integration, we approximate $f(x)$ by a polynomial $f(x)$ of suitable degree. Then, we integrate $f(x)$ within the limits (a,b). That is,

$$I = \int_a^b f(x)\,dx \cong \int_a^b \phi(x)\,dx$$

Here the exact value if

$$I = \int_a^b f(x)\,dx$$

Approximate value $= \int_a^b \phi(x)\,dx$

The difference $\left[\int_a^b f(x)\,dx - \int_a^b \phi(x)\,dx\right]$

is called the *error of approximation* and

$$\frac{\left[\int_a^b f(x)\,dx - \int_a^b \phi(x)\,dx\right]}{\int_a^b f(x)\,dx}$$

is called the *relative error of approximation*.

Hence, relative error of approximation $= \dfrac{exact\,value - approximate\,value}{exact\,value}$

5.3 NEWTON – COTES CLOSED QUADRATURE FORMULA

The general form of the problem of numerical integration may be stated as follows:

Given a set of data points (x_i, y_i), $i = 0,1,2,\quad ,n$ of a function $y = f(x)$, where $f(x)$ is not explicitly known. Here, we are required to evaluate the definite integral

$$I = \int_a^b y\,dx \qquad \textbf{(5.2)}$$

Here, we replace $y = f(x)$ by an interpolating polynomial $\phi(x)$ in order to obtain an approximate value of the definite integral of Eq.(5.2).

In what follows, we derive a general formula for numerical integration by using Newton's forward difference formula. Here, we assume the interval (a, b) is divided into n-equal subintervals such that

$$h = \frac{b-a}{n}$$

$$a = x_0 < x_2 < x_3 \quad < x_n = b$$

With $x_n = x_0 + nh$

Where

h = the internal size

n = the number of subintervals

a and b = the limits of integration with $b > a$.

Hence, the integral in Eq.(5.2) can be written as

$$I = \int_{x_0}^{x_n} y dx \tag{5.3}$$

Using Newton's forward interpolation formula, we have

$$I = \int_{x_0}^{x_n} \left[y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + \quad \right] dx \tag{5.4}$$

where $x = x_0 + uh$

$$I = h\int_0^n \left[y_0 + u\Delta y_0 + \frac{u^2-u}{2}\Delta^2 y_0 + \frac{u^3-3u^2+2u}{6}\Delta^3 y_0 + \quad \right] du \tag{5.5}$$

Hence, after simplification, we get

$$I = \int_{x_0}^{x_n} y dx = nh\left[y_0 + \frac{n}{2}\Delta y_0 + \frac{n(2n-3)}{12}\Delta^2 y_0 + \frac{n(n-2)^2}{24}\Delta^3 y_0 + \quad \right] \tag{5.6}$$

The formula given by Eq. (5.6) is known as *Newton-Cotes closed quadrature formula*. From the general formula Eq. (5.6), we can derive or deduce different integration formulae by substituting $n = 1, 2, 3,$, etc.

5.4 TRAPEZOIDAL RULE

In this method, the known function values are joined by straight lines. The area enclosed by these lines between the given end points is computed to approximate the integral as shown in Fig. 5.2.

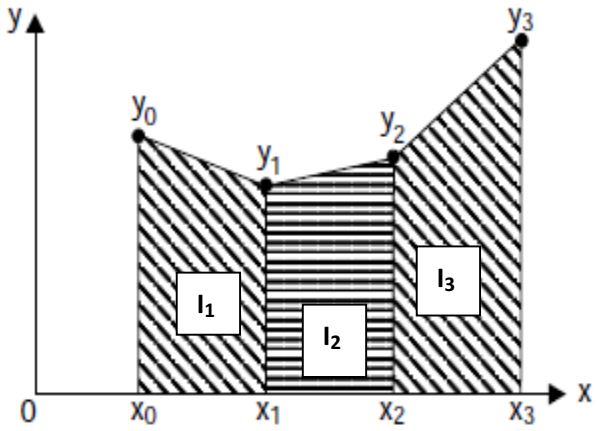

Fig: 5.2

Each subinterval with the line approximation for the function forms a trapezoid as shown in Fig. 5.2. The area of each trapezoid is computed by multiplying the interval size *h* by the average value of the function value in that subinterval. After the individual trapezoidal areas are obtained, they are all added to obtain the overall approximation to the integral.

Substituting $n = 1$ in Eq.(5.6) and considering the curve $y = f(x)$through the points (x_0, y_0) and (x_1, y_1)as a straight line (a polynomial of first degree so that the differences of order higher than first become zero), we get

$$I_1 = \int_{x_0}^{x_1} y dx = h\left[y_0 + \frac{1}{2}\Delta y_0\right] = \frac{h}{2}\left[y_0 + \frac{1}{2}(y_1 - y_0)\right] = \frac{h}{2}(y_0 + y_1) \qquad \textbf{(5.7)}$$

Similarly, we have

$$I_2 = \int_{x_1}^{x_2} y dx = \frac{h}{2}(y_1 + y_2)$$

$$I_3 = \int_{x_2}^{x_3} y dx = \frac{h}{2}\left(y_2 + y_3\right)$$

and so on

In general, we have

$$I_n = \int_{x_{n-1}}^{x_n} y dx = \frac{h}{2}\left(y_{n-1} + y_n\right) \qquad \textbf{(5.8)}$$

Adding all the integrals (Eq. (5.7), Eq. (5.8)) and using the interval additive property of the definite integrals, we obtain

$$I = \sum_{i=1}^{n} I_i = \int_{x_0}^{x_n} y dx = \frac{h}{2}\left(y_0 + 2\left(y_1 + y_2 + y_3 + \quad + y_{n-1}\right) + y_n\right) = \frac{h}{2}\left[X + 2I\right] \qquad \textbf{(5.9)}$$

where, X = sum of the end points

I = sum of the intermediate ordinates.

Equation (5.9) is known as the trapezoidal rule.

Summarising, the trapezoidal rule signifies that the curve $y = f\left(x\right)$ is replaced by $n-$ straight lines joining the points $\left(x_n, y_n\right), i = 0,1,2,3, \quad ,n$. The area bounded by the curve $y = f\left(x\right)$, the ordinates $x = x_0, x = x_n$ and the $x-$ axis is then approximately equivalent to the sum of the areas of the $n-$ trapezoids so obtained.

NB: The error estimate in trapezoidal rule is given by:

$$E = -\frac{\left(b-a\right)h^2}{12} f''\left(\xi\right), \quad a = x_0 < \xi < x_n = b$$

Therefore, the total error in the evaluation of the trapezoidal rule is of the order of h^2.

Exercise 5.1

Evaluate $\int_0^{12} \frac{1}{1+x^2} dx$ by using trapezoidal rule, taking $n = 6$, correct to five significant figures.

Solution: (@ Lectures)

5.5 SIMPSON'S 1/3 RULE

In Simpson's rule, the function is approximated by a second degree polynomial between successive points. Since a second degree polynomial contains three constants, it is necessary to know three consecutive function values forming two intervals as shown in Fig. 5.3.

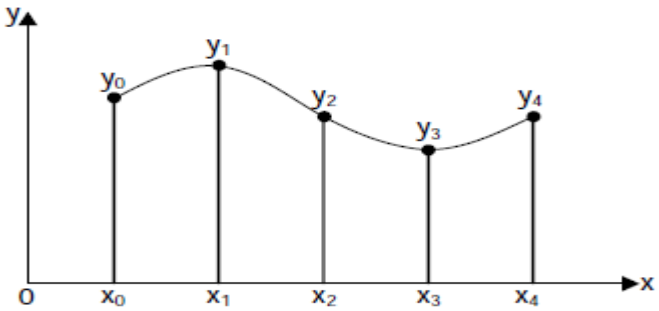

Fig. 5.3.

Consider three equally spaced points x_0, x_1 and x_2. Since the data are equally spaced,

let $h = x_{n+1} - x_n$ (see Fig. 5.4).

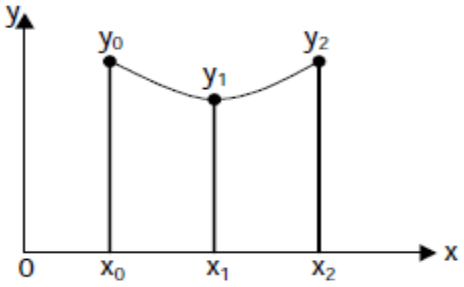

Fig. 5.4

Substituting $n = 2$ in Eq. (5.6) and taking the curve through the points $(x_0, y_0), (x_1, y_1)$ and (x_2, y_2)as a polynomial of second degree (parabola) so that the differences of order higher than two vanish, we obtain

$$I_1 = \int_{x_0}^{x_2} y dx = 2h\left[y_0 + \Delta y_0 + \frac{1}{6}\Delta^2 y_0\right] = \frac{h}{3}\left[y_0 + 4y_1 + y_2\right] \quad \textbf{(5.10)}$$

Similarly,

$$I_2 = \int_{x_2}^{x_4} y dx = \frac{h}{3}\left[y_2 + 4y_3 + y_4\right]$$

$$I_3 = \int_{x_4}^{x_6} y dx = \frac{h}{3}\left[y_4 + 4y_5 + y_6\right]$$

and so on.

In general, we can write

$$I_n = \int_{x_{2n-2}}^{x_n} y dx = \frac{h}{3}\left[y_{2n-2} + 4y_{2n-1} + y_{2n}\right]$$

(5.11)

Summing up all the above integrals, we obtain

$$I_n = \int_{x_0}^{x_n} y dx = \frac{h}{3}\left[y_0 + 4\left(y_1 + y_3 + y_5 + \quad + y_{2n-1}\right) + 2\left(y_2 + y_4 + y_6 + \quad + y_{2n-2}\right) + y_{2n}\right]$$

(5.12)

$$= \frac{h}{3}\left[X + 4O + 2E\right] \quad \textbf{(5.13)}$$

Where, X = sum of end ordinates

O = sum of odd ordinates

E = sum of even ordinates

Equation (5.13) is known as *Simpson's 1/3 rule*. Simpson's 1/3 rule requires the whole range (the given interval) to be divided into even number of equal subintervals.

NB: The error in the composite Simpson's 1/3 rule is

$$E = \frac{(b-a)h^4}{180} f^{(4)}(\xi)$$

where, $a = x_0 < \xi < x_n = b$

Exercise 5.2

Evaluate $\int_0^{12} \frac{1}{1+x^2} dx$ by using Simpson's 1/3 rule, taking $n = 6$, correct to five significant figures.

Solution: (@ Lectures)

5.6 SIMPSON'S 3/8 RULE

Putting n = 3 in Eq. (5.6) and taking the curve through (x_n, y_n), $n = 0,1,2,3$ as a polynomial of degree three such that the differences higher than the third order vanish, we obtain

$$I_1 = \int_{x_0}^{x_3} y dx = 3h\left[y_0 + \frac{3}{2}\Delta y_0 + \frac{1}{8}\Delta^3 y_0\right] = \frac{3}{8}h\left[y_0 + 3y_1 + 3y_2 + y_3\right] \qquad \textbf{(5.14)}$$

Similarly, we get

$$I_2 = \int_{x_3}^{x_6} y dx = \frac{3}{8}h\left[y_3 + 3y_4 + 3y_5 + y_6\right]$$

$$I_3 = \int_{x_6}^{x_9} y dx = \frac{3}{8}h\left[y_6 + 3y_7 + 3y_8 + y_9\right]$$

and so on.

Finally, we have

$$I_n = \int_{x_{3n-3}}^{x_{3n}} y dx = \frac{3}{8}h\left[y_0 + 3\left(y_1 + y_2 + y_4 + y_5 + y_7 + y_8 + \quad + y_{3n-2} + y_{3n-1}\right) + 2\left(y_3 + y_6 + y_9 + \quad + y_{3n-3}\right) + y_{3n}\right]$$

(5.15)

Equation (5.15) is called the Simpson's 3/8 rule. Here, the number of subintervals should be taken as multiples of 3. Simpson's 3/8 rule is not as accurate as Simpson's 1/3 rule. Simpson's 3/8 rule can be applied when the range (a, b) is divided into a number of subintervals, which must be a multiple of 3.

NB: Simpson's 1/3 rule requires the number of panels to be even. If this condition is not satisfied, we can integrate over the first (or last) three panels with Simpson's 3/8 rule.

Exercise 5.3

Evaluate $\int_0^{12} \frac{1}{1+x^2} dx$ by using Simpon's 3/8 rule and taking 7 ordinates, correct to five significant figures.

Solution: (@ Lectures)

5.7 BOOLE'S RULE

Substituting $n = 4$ in Eq.(5.6) and taking the curve through $(x_n, y_n), n = 0,1,2,3,4$ as a polynomial of degree 4, so that the difference of order higher than four vanish (or neglected), we obtain

$$\int_{x_0}^{x_4} y dx = 4h\left[y_0 + 2\Delta y_0 + \frac{5}{3}\Delta^2 y_0 + \frac{2}{3}\Delta^3 y_0 + \frac{7}{90}\Delta^4 y_0\right] \quad \textbf{(5.16)}$$

$$= \frac{2h}{45}\left[7y_0 + 32y_1 + 12y_2 + 32y_3 + 7y_4\right]$$

Likewise,

$$\int_{x_4}^{x_8} y dx = \frac{2h}{45}\left[7y_4 + 32\Delta y_5 + 12\Delta^2 y_6 + 32\Delta^3 y_7 + 7\Delta^4 y_8\right] \text{ and so on.}$$

Adding all the above integrals from x_0 to x_n, where n is a multiple of 4, we obtain

$$\int_{x_0}^{x_n} y dx = \frac{2h}{45}\left[7y_0 + 32\left(y_1 + y_3 + y_5 + y_7 + \quad\right) + 12\left(y_2 + y_6 + y_{10} + \quad\right) + 14\left(y_4 + y_8 + y_{12} + \quad\right) + 7y_n\right] \tag{5.17}$$

Equation (5.17) is known as *Boole's rule*. It should be noted here that the number of subintervals should be taken as a multiple of 4.

The leading term in the error of formula can be shown as

$$\frac{-8}{945} h^7 y^{(vi)}(\varsigma)$$

Exercise 5.4

Evaluate the integral $\int_0^{12} \frac{1}{1+x^2} dx$ by using Boole's rule using exactly five functional evaluations and correct to five significant figures.

Solution: (@ Lectures)

5.8 SUMMARY

In this chapter we have presented the various techniques on numerical integration. Integration methods such as the trapezoidal rule, Simpson's one-third rule, Simpson's three-eight's rule and Boole's rule.

EXERCISE 5

1. Evaluate $\int_2^6 \log_{10} x\, dx$ by using trapezoidal rule, taking n = 8, correct to five decimal places.
2. Evaluate the error bounds on $\int_0^{\pi} \sin(x) dx$ with the composite trapezoidal rule using
   i. eight panels and
   ii. sixteen panels.
3. Evaluate $\int_2^6 \log_{10} x\, dx$ by using trapezoidal rule, taking n = 6, correct to five decimal places.

4. Derive Simpson's 3/8 rule from Newton–Cotes formulae.
5. Evaluate the integral $\int_0^1 e^x dx$, by using Simpson's 3/8 rule and taking seven ordinates.
6. Evaluate the integral $\int_0^1 \left(1 + e^{-x} \sin 4x\right) dx$ using Boole's rule with h = 1/4.

CHAPTER SIX

NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATION

6.1 INTRODUCTION

Numerical methods are becoming more and more important in engineering applications, simply because of the difficulties encountered in finding exact analytical solutions but also, because of the ease with which numerical techniques can be used in conjunction with modern high-speed digital computers. Several numerical procedures for solving initial value problems involving first-order ordinary differential equations are discussed in this chapter.

In spite of the fact that the error analysis is an important part of any numerical procedure, the discussion in this chapter is limited primarily to the use of the procedure itself. The theory of errors and error analysis is sometimes fairly complex and goes beyond the intended scope of this chapter.

An ordinary differential equation is one in which an ordinary derivative of a dependent variable *y* with respect to an independent variable *x* is related in a prescribed manner to *x, y* and lower derivatives. The most general form of an ordinary differential equation of n^{th} order is given by

$$\frac{d^n y}{dx^n} = f\left(x, y, \frac{dy}{dx}, \frac{d^2 y}{dx^2}, \quad , \frac{d^{n-1} y}{dx^{n-1}}\right) \tag{6.1}$$

The Eq. (6.1) is termed as ordinary because there is only one independent variable. To solve an equation of the type (Eq.(6.1)), we also require a set of conditions. When all the conditions are given at one value x and the solution proceeds from that value of x, we have an *initial-value problem*. When the conditions are given at different values of x, we have a *boundary-value problem*.

For example,

$$y'' = -y, \quad y(0) = 1 \qquad y'(0) = 0$$

is an *initial value problem* since both auxiliary conditions imposed on the solution are given at $x = 0$. On the other hand,

$$y'' = -y, \quad y(0) = 1 \qquad y'(\pi) = 0$$

is a *boundary value problem* because the two conditions are specified at different values of x.

6.2 RUNGE – KUTTA METHODS

The aim of Runge–Kutta methods is to eliminate the need for repeated differentiation of the differential equations. Since no such differentiation is involved in the first-order Taylor series integration formula

$$y(x+h) = y(x) + y'(x)h = y(x) + F(x, y)h \tag{6.2}$$

it can be considered as the ***first-order Runge–Kutta method*** and it is also called ***Euler's method***. Due to excessive truncation error, this method is rarely used in practice.

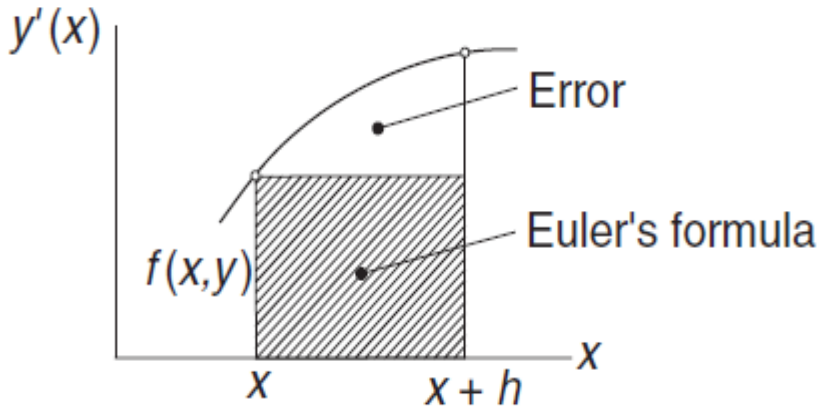

Fig 6.1: Graphical Representation of Euler's formula

Exercise 6.1

Use Euler's method to solve the following differential equation $\frac{dy}{dx} = \frac{1}{2}y,\ y(0) = 1$ and $0 \le x \le 1$. Use $h = 0.1$.

Solution: (@ Lectures)

6.2.1 Euler – Cauchy (Modified Euler) Method

Consider the differential equation

$$\frac{dy}{dx} = f(x, y) \qquad \textbf{(6.3)}$$

with the initial condition

$$y(x_0) = y_0$$

Integrating Eq.(6.3), we obtain,

$$y = y_0 + \int_{x_0}^{x} f(x, y)\,dx \qquad \textbf{(6.4)}$$

In modified Euler's method, instead of approximating (x, y) by $f(x_n, y_n)$ in Eq.(6.3), the integral in Eq.(6.4) is approximated using the trapezoidal rule.

Therefore, we have

$$y_1^{(n+1)} = y_0 + \frac{h}{2}\left[f(x_0, y_0) + f\left(x_1, y_1^{(n)}\right)\right],\ n = 0, 1, 2, 3, \qquad \textbf{(6.5)}$$

where $y_1^{(n)}$ **(i.e. Euler method)** is the n^{th} approximation to y_1.

Exercise 6.2

Use the modified Euler's method to obtain an approximate solution of $\frac{dy}{dt} = -2ty^2, y(0) = 1$, in the interval $0 \le t \le 1.0$ using $h = 0.1$. Compute the error and the percentage error. Given the exact solution is given by

$$y = \frac{1}{1+t^2}.$$

Solution: (@ Lectures)

6.2.2 Runge-Kutta Method of Order Two

In the Runge-Kutta method of order two, we consider up to the second derivative term in the Taylor series expansion and then substitute the derivative terms with the appropriate function values in the interval.

The integration formula can be conveniently evaluated by the following sequence of operations

$$K_1 = hf(x_n, y_n)$$

$$K_2 = hf\left(x_n + \frac{h}{2}, y_n + \frac{1}{2}K_1\right)$$

$$y(x+h) = y(x) + K_2 \qquad \textbf{(6.6)}$$

Runge-Kutta method of order two is also known as the *Midpoint method* because the derivative is replaced by functions evaluated at the midpoint $x_n + h/2$. The local truncation error in the Runge-Kutta method of order two is $O(h^3)$, and the global truncation error is $O(h^2)$.

NB:

The local truncation error is the error introduced by the approximation method at each step.

The global error is the absolute difference between the correct value and the approximate value.

Exercise 6.3

Use Runge-Kutta method of order two to integrate $\frac{dy}{dx} = \sin(y)$ with y(0) = 1 from $x = 0 \; to \; 0.5$ in steps of h = 0.1. Keep four decimal places in the calculations.

Solution: (@ Lectures)

6.2.3 Runge - Kutta Method of Order Four

The fourth-order Runge - Kutta method is obtained from the Taylor series along the same lines as the second-order method. Since the derivation is rather long and not very instructive, we skip it. The final form of the integration formula again depends on the choice of the parameters; that is, there is no unique Runge–Kutta fourth-order formula.

The most popular version, which is known simply as the Runge–Kutta method, entails the following sequence of operations:

$$K_1 = hf(x, y)$$

$$K_2 = hf\left(x + \frac{h}{2}, y + \frac{K_1}{2}\right)$$

$$K_3 = hf\left(x + \frac{h}{2}, y + \frac{K_2}{2}\right)$$

$$K_4 = hf(x + h, y + K_3)$$

$$y(x + h) = y(x) + \tfrac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4) \qquad \textbf{(6.7)}$$

The main drawback of this method is that it does not lend itself to an estimate of the truncation error. Therefore, we must guess the integration step size *h*, or determine it by trial and error. In contrast, the so-called *adaptive methods* can evaluate the truncation error in each integration step and adjust the value of *h* accordingly (but at a higher cost of computation).

Exercise 6.4

Use the Runge - Kutta method of order four with $h = 0.1$ to obtain an approximation to $y(1.5)$ for the solution of $\frac{dy}{dx} = 2xy, y(1) = 1$. The exact solution is given by $y = e^{x^2-1}$. Determine the percentage relative error.

Solution: (@ Lectures)

EXERCISE 6

1. Use Euler's method to solve the initial value problem $y' = 1 - t + 4y,\ y(0) = 1$, in the interval $0 \le t \le 0.5$. Use $h = 0.1$. The exact value is

   $$y = -\frac{9}{16} + \frac{1}{4}t + \frac{19}{16}e^{4t}$$

   Compute the error and the percentage error.

2. Use the modified Euler's method to find the approximate value of $y(1.5)$ for the solution of the initial value problem $\frac{dy}{dx} = 2xy, y(1) = 1$. Take $h = 0.1$. The exact solution is given by $y = e^{x^2-1}$. Determine the relative error and the percentage error.

3. Use Runge-Kutta method of order two to integrate $\frac{dy}{dx} = x\sin(y)$ with y(0) = 1 from $x = 0\ to\ 0.5$ in steps of h = 0.1. Keep four decimal places in the calculations.

4. Use the fourth-order Runge–Kutta method to integrate

$$y' = 3y - 4e^{-t}, \qquad y(0) = 1$$

from $x = 0$ *to* 1 in steps of h = 0.1. Compare the result with the analytical solution

$$y = e^{-t}.$$

CHAPTER SEVEN

SECOND ORDER PARTIAL DIFFERENTIAL EQUATION

7.1 INTRODUCTION

The most general form of a second-order partial differential equation is

$$a(x,y)\frac{\partial^2 f(x,y)}{\partial x^2} + b(x,y)\frac{\partial^2 f}{\partial x \partial y} + c(x,y)\frac{\partial^2 f}{\partial y^2} + d(x,y)\frac{\partial f}{\partial x} + e(x,y)\frac{\partial f}{\partial y} + g(x,y) = 0 \qquad \textbf{(7.1)}$$

Three types of equation are of a particular interest because they feature so prominently in engineering and science.

7.1.1 ELLIPTIC EQUATIONS

If $b^2 - 4ac < 0$ the partial differential equation is called an *elliptic* equation. Such equations arise out of steady-state problems as occur in potential or flow theory. Two examples are;

Poisson's equation

$$\frac{\partial^2 \phi(x,y)}{\partial x^2} + \frac{\partial^2 \phi(x,y)}{\partial y^2} = g(x,y)$$

Laplace's equation

$$\frac{\partial^2 \phi(x,y)}{\partial x^2} + \frac{\partial^2 \phi(x,y)}{\partial y^2} = 0$$

In both cases, $a = 1$, $b = 0$ and $c = 1$ and so $b^2 - 4ac = -4 < 0$.

7.1.2 HYPERBOLIC EQUATIONS

If $b^2 - 4ac > 0$ the partial differential equation is called an *hyperbolic* equation. Such equations arise out of vibrational and radiative problems as occur in wave mechanics. An example is

The wave equation

$$\frac{\partial^2 \phi(x,t)}{\partial x^2} = \frac{1}{k^2} \frac{\partial^2 \phi(x,t)}{\partial t^2}$$

Here a =1, b = 0 and c = $\frac{1}{k^2}$ and so $b^2 - 4ac > 0$

7.1.3 PARABOLIC EQUATIONS

If $b^2 - 4ac = 0$ the partial differential equation is called a *parabolic* equation. Such equations arise out of transient flow problems as occur in conduction or consolidation. An example is

The heat conduction or consolidation equation

$$\frac{\partial^2 \phi(x,t)}{\partial x^2} = \frac{1}{k}\frac{\partial^2 \phi(x,t)}{\partial t}$$

Here $a = 1$, $b = 0$ and $c = \frac{1}{\kappa}$ and so $b^2 - 4ac = 0$

In the equation above *a, b* and *c* are constant but in the general case they depend on *x* and *y* and so a given equation may change from one type to another within the same domain.

Exercise 7.1

Name the type of equation in each of the following

a) $2\frac{\partial f(x,y)}{\partial x} - 3y\frac{\partial f(x,y)}{\partial y} = 4xy$

b) $\frac{\partial f(x,y)}{\partial x} + \frac{\partial^2 f(x,y)}{\partial x \partial y} - \frac{\partial f(x,y)}{\partial y} = \frac{x}{y}$

c) $\frac{\partial^2 f(x,y)}{\partial x^2} - 2\frac{\partial^2 f(x,y)}{\partial x \partial y} + \frac{\partial^2 f(x,y)}{\partial y^2} = 0$

d) $\frac{\partial}{\partial x}\left[\frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y}\right] = \frac{x^2}{y^3}$

e) $3\frac{\partial^2 f(x,y)}{\partial x^2} - 2\frac{\partial^2 f(x,y)}{\partial x \partial y} + \frac{\partial^2 f(x,y)}{\partial y^2} = 3xy$

Solution: (@ Lectures)

RECOMMENDED TEXTBOOKS:

Atkinson, K.E., *An Introduction to Numerical Analysis, 2nd ed.,* Wiley, New York, NY, 1993.

Atkinson, K.E. and Han, W., *Elementary Numerical Analysis, 3rd ed.* Wiley, New York, 2004.

Atkinson, L.V. and Harley, P.J., *Introduction to Numerical Methods with PASCAL,* Addison Wesley, Reading, MA, 1984.

Atkinson, L.V., Harley, P.J. and Hudson, J.D., *Numerical Methods with FORTRAN 77,* Addison Wesley, Reading, MA, 1989.

Axelsson, K., *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1994.

Ayyub, B.M. and McCuen, R.H., *Numerical Methods for Engineers*, Prentice Hall, Upper Saddle River, NewJersey, NJ, 1996.

Chapra, S.C., *Applied Numerical Methods with MATLAB for Engineers and Scientists*, McGraw-Hill, NewYork, 2005.