NATIONAL OPEN UNIVERSITY OF NIGERIA

SCHOOL OF SCIENCE AND TECHNOLOGY

COURSE CODE: MTH 213

COURSE TITLE: NUMERICAL ANALYSIS 1

Course Code	MTH 213
Course Title	NUMERICAL ANALYSIS 1
Course Developer	Dr. Ajibola S. O. National Open University of Nigeria Lagos
Content Editor	Dr. ABIOLA. Bankole National Open University of Nigeria Lagos
Course Coordinator	Dr. Ajibola S. O. National Open University of Nigeria Lagos
Programme Leader	Dr. ABIOLA.Bankole National Open University of Nigeria Lagos

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
14/16 Ahmadu Bello Way
Victoria Island
Lagos

Abuja Office
5, Dar Es Salaam Street
Off Aminu Kano Crescent
Wuse II, Abuja
Nigeria.

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

First Printed 2008

ISBN:

Printed by: For
National Open University of Nigeria

CONTENT **PAGE**

COURSE GUIDE

MTH 213
NUMERICAL ANALYSIS 1

Course Developer	Dr. Ajibola S. O. National Open University of Nigeria Lagos
Content Editor	Dr. ABIOLA. Bankole National Open University of Nigeria Lagos
Course Coordinator	Dr. Ajibola S. O. National Open University of Nigeria Lagos
Programme Leader	Dr. ABIOLA. Bankole National Open University of Nigeria Lagos

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
14/16 Ahmadu Bello Way
Victoria Island
Lagos

Abuja Office
5, Dar Es Salaam Street
Off Aminu Kano Crescent
Wuse II, Abuja
Nigeria.

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

First Printed 2008

ISBN:

Printed by ……………..
For
National Open University of Nigeria

CONTENTS PAGE

Introduction

MTH 213: Discussion of Lagrange's form for; The technique of determining an approximate value of f(x) for a non-tabular value of x which lies in the internal [a, b] is called interpolation. The process of determining the value of f(x) for a value of x lying outside the interval [a, b] is called extrapolation.

The Lagrange's form of the interpolating polynomial derived above has same draw backs compared to Newton's form of interpolating polynomial. Before deriving Newton's general form of interpolating polynomial. We introduce the concept of divided difference and the tabular representation of divided differences.

Numerical solution of systems of linear algebraic equations play a prominent role in boundary value problems, for ordinary and partial differential equations, statistical influence, optimization theory, least square fittings of data etc.

Numerical methods for solving linear algebraic system may be divided into two types, direct and iterative. To understand the numerical methods for solving linear system of equations, it is necessary to have some knowledge of the properties of matrices. The prerequisite to the course shall be linear Algebra courses.

The Course

As a 3-credit unit course, 11 study units grouped into 3 modules of 3 units in module 1, 4 units in module 2 and 4 units in module 3.

This course guide gives a brief summary of the total contents contained in the course material. The fundamental theorem of algebra and its useful calories, inverse interpolation and errors. Newton's form of the interpolating polynomial features divided differences and interpolating polynomial error types. Likewise interpolating at equally spaced points, here we talked about differences.

For equally spaced nodes, we shall deal with three types of differences, namely forward, backward and central and discuss their representation in the form of a table. Also discussed her are some direct and iterative methods for finding the solution of system of linear algebraic equations.

Lastly, we discussed three fundamental theorems, namely; intermediate value theorem, Rolle's theorem and Lagrange's mean value theorem. All these theorems give properties of continues functions defined on a closed interval [a, b]. Although the theorems are not proved but their utility was illustrated with examples.

Course Aim & Objectives

On the completion of this course, you are expected to:

- find the Lagrange's form of interpolating polynomial
- complete the approximate value of f at a non-tabular point.
- Complete the error omitted in interpolation, if the function is known at a non-tabular point of interest.
- Find an upper bound in the magnitude of the error.
- Write forward, backward and central differences in terms of function values from a table of either difference and locate a difference of given order at given point.
- Obtain the interpolating polynomial of f(x) for a given data by applying any one of the interpolating formulae.
- Obtain the solution of systems of linear algebraic equations by using the direct methods such as Cramer's rule, Gauss elimination method Lu decomposition method.

Working through the Course

This course involves that you would be required to spend lot of time to read. The content of this material is very dense and require you spending great time to study it. This accounts for the great effort put into its development in the attempt to make it very readable and comprehensible. Nevertheless, the effort required of you is still tremendous. I would advice that you avail yourself the opportunity of attending the tutorial sessions where you would have the opportunity of comparing knowledge with your peers.

The Course Material

You will be provided with the following materials:

Course Guide
Study Units

In addition, the course comes with a list of recommended textbooks, which through are not compulsory for you to acquire or indeed read, are necessary as supplements to the course material.

Study Units

The following are the study units contained in this course. The units are arranged into 3 identifiable but readable modules.

Module 1

Unit 1 Interpolation (Lagrange's Form)

This unit takes one through the definition of interpolation, inverse interpolation and error.

Unit 2 Newton's Form of the Interpolating Polynomial

This unit is sub-divided into divided difference Newton's General Form of interpolating polynomial, and the error of the interpolating polynomial. Divided difference and derivative of the functions and further results on interpolations error.

Unit 3 Interpolation at Equally Spaced Points

This unit takes about the three types of differences i.e. forward, backward and central differences. Difference formulae which encompasses: Newton's Forward–Difference formula and Newton's Backward-Difference formula.

Module 2 Solution of Linear Algebraic Equations.

Unit 1 Direct Method

This unit entails the preliminaries, Cramer's rule, direct methods for special matrices. Gauss elimination methods and LU decomposition method.

Unit 2 Inverse of A Square Matrix

This unit is sub-divided into method of adjoints, the Gauss-Jordan reduction method and LU decomposition method.

Unit 3 Iterative Methods

This unit consists of the general iterative methods. The Jaccobi's iteration methods and the Gauss-Seidel iteration method.

Unit 4 Eigen-Values and Eigen-Vectors.

This unit focused on the Eigen value problem. The power method and the inverse power method.

Module 3

Unit 1 Review of Calculus

Here, the three fundamental theorems, Taylor's theorem, error (round off and truncation errors) are discussed.

Unit 2 Iteration Methods for Locating Root.

This unit discussed: The initial approximation to a root (tabulation and graphical methods). Bisection method and fixed point iteration method.

Unit 3 Chord Methods for Finding Root

This entails Repuler-Falsi method, Newton – Raphson method and convergence criterion.

Unit 4 Approximate Root of Polynomial Equation.

It can be sub-divided into some results on roots of polynomial equation. Birge-Vieta method and Graeffe’s Root squaring method.

Textbooks

More recent editions of these books are recommended for further reading.

Engineering Mathematics P. D. S. Verma.

Generalized functions in mathematical physics by V. S. Viadimirov.

Mathematical methods for science students by G. Stephenson.

Generalized functions by R. F. Hoskins.

Engineering mathematics by K. A. Strond.

Engineering Mathematics by Kreyszic.

Assessment

There are two components of assessment for this course. The Tutor Marked Assignment (TMAS) and the end of the course examination.

Tutor Marked Assignments (TMAs)

The (TMAS) is the continuous assessment component of your course. It accounts for 30% of the total score. You will be given 4 (TMAS) to answer. Three of these must be answered before you are allowed to sit for the end of course examination. The (TMAS) would be given to you by your facilitator and returned after you have done the assignment.

End Of Course Examination

This examination concludes the assessment for the course. It constitutes 70% of the whole course. You will be informed of the time for the examination. It may or may not coincide with the university semester examination.

Summary

In summary, we have seen how to desire the Lagrange's form of interpolating polynomial for a given data. It has been shown that the interpolating polynomial for a given data is unique. We have derived the general error formula and its use has been illustrated to judge the accuracy of our calculations.

For a system of 'n' equations $Ax = b$ in 'n' unknown, where A is a non-singular matrix, the methods of finding the solution vector x may be broadly classified into two types.

i) Direct methods and
ii) Iteration methods.

For larger systems, direct methods becomes more efficient if the coefficient matrix A is in one of the forms D (diagonal), L (lower triangular) or U (upper triangular).

We further discussed the following methods for finding approximate roots of polynomial questions: (Birge-Vieta and Graeffe's root squaring methods).

MODULE 1 INTERPOLATION

UNIT 1 INTERPOLATION (LAGRANGE'S FORM)

CONTENTS

1.0 INTRODUCTION

Let f be a real-valued function defined on the interval [a, b] and we denote $f(x_k)$ by f_k. suppose that the values of the function f(x) are given to be $f_0, f_1, f_2, \ldots, f_n$ when $x = x_0, x_1, x_2, \ldots, x_n$ respectively where $x_0 < x_1 < x_2 \ldots < x_n$ lying in the interval [a, b]. The function f(x) may not be known to us. The technique of determining an approximate value of f(x) for a non-tabular value of x which lies in the interval [a, b] is called interpolation. The process of determining the value of f(x) for a value of x lying outside the interval [a, b] is called extrapolation. In this unit, we derive a polynomial P(x) of degree n which agrees with the values of f(x) at the given (n + 1) distinct points, called nodes or abscissas. In other words, we can find a polynomial P(x) such that $P(x_j) = f_j$, j = 0, 1, 2, …, n. such a polynomial P(x) is called the interpolating polynomial of f(x).

In section 3.1 we prove the existence of an interpolating polynomial by actually constructing one such polynomial having the desired property. The uniqueness is proved by invoking the corollary of the fundamental theorem of Algebra. In section 3.2 we derive general expression for error in approximating the function by the interpolating polynomial at a point and this allows us to calculate a bound on the error over an interval. In proving this we make use of the general Rolle's theorem.

2.0 OBJECTIVES

After reading this unit, you should be able to:

- find the Lagrange's form of interpolating polynomial interpolating f(x) at n + 1 distinct nodal points
- compute the approximate value of f at a non-tabular point
- compute the value of $\bar{x}$ (approximately) given a number $\bar{y}$ such that $f(\bar{x}) = (\bar{y})$ (inverse interpolation)
- compute the error committed in interpolation, if the function is known, at a non-tabular point interest
- find an upper bound in the magnitude of the error.

3.0 MAIN CONTENT

3.1 Lagrange's Form

Let us recall the fundamental theorem of algebra and its useful corollaries.

Theorem 1

If P(x) is a polynomial of degree $n \geq 1$, that is $P(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$, $\ldots$, a_n real or complex numbers and $a_n \neq 0$, then P(x) has at least one zero, that is, there exists a real or complex number ξ such that $p(\xi) = 0$.

Lemma 1

If $z_1, z_2, \ldots, z_k$ are distinct zeros of the polynomial P(x), then

$$P(x) = (x - z_1)(x - z_2) \ldots (x - z_k)R(x)$$

for some polynomial R(x).

Corollary

If $P_k(x)$ and $Q_k(x)$ are the two polynomials of degree £ k which agree at the k + 1 distinct points $z_0, z_1, z_2, \ldots, z_k$ then $P_k(x) = Q_k(x)$ identically.

You have come across Rolle's Theorem in the perquisite course. But we need a generalized version of this theorem . (General Error Term). This is stated below.

Theorem 2

(Generalised Rolle's Theorem). Let f be a real-valued function defined on [a, b] which is n times differentiable on]a, b[. If f vanishes at the n + 1 distinct points $x_0, \ldots, x_n$ in [a, b], then a number c in]a, b[exists such that $f^{(n)}(c) = 0$.

We now show the existence of an interpolating polynomial and also show that it is unique. The form of the interpolating polynomial that we are going to discuss in this section is called the Lagrange's form of the interpolating polynomial. We start with a relevant theorem.

Theorem 3:

Let $x_0, x_1, \ldots x_n$ be n + 1 distinct points on the real line and let f(x) be a real-valued function defined on some interval I = [a, b] containing these points. Then, there exists exactly one polynomial $P_n(x)$ of degree n, which interpolates f(x) at $x_0, \ldots x_n$, that is, $P_n(x_j) = f(x_j)$, i = 0, 1, 2, …, n.

Proof:

First we discuss the uniqueness of the interpolating polynomial, and then exhibit one explicit construction of an interpolating polynomial (Lagrange's Form).

Let $P_n(x)$ and $Q_n(x)$ be two distinct interpolating polynomials of degree n, which interpolate f(x) at (n + 1) distinct points $x_0, x_1, \ldots x_n$. Let $h(x) = P_n(x) - Q_n(x)$. Note that h(x) is also a polynomial of degree $\leq$ n. Also

$$h(x_j) = P_n(x_j) - Q_n(x_j) = f(x_j) - f(x_j) = 0, \; i = 0, 1, 2, \ldots, n.$$

That is, h(x) has (n + 1) distinct zeros. But h(x) is of degree £ n and from the Corollary to Lemma 1, we have h(x) º 0. That is $P_n(x)$ $Q_n(x)$. This proves the uniqueness of the polynomial.

Since the data is given at the points $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n)$ let the required polynomial be written as

$$P_n(x_j) = L_0(x)f_0 + L_1(x)f_1 + \ldots + L_n(x)f_n = \sum_{i=0}^{n} L_i(x)f_i \qquad (1)$$

Setting $x = x_j$ in (1), we get

$$P_n(x_j) = \sum_{i=0}^{n} L_i(x_j)f_i \qquad (2)$$

Since this polynomial fits the data exactly, we must have

$$L_j(x_j) = 1$$

and $\quad L_j(x_j) = 0,\ i \neq j$

or $\quad L_j(x_j) = \partial_{ij} \quad (3)$

The polynomial $L_i(x)$ which are of degrees £ n are called the Lagrange fundamental polynomials. It is easily verified that these polynomial are given by

$$L_j(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(_i x - x_n)}$$

$$= \prod_{\substack{i=0 \\ i=j}}^{n} (x - x_j) \bigg/ \prod_{\substack{i=0 \\ i=j}}^{n} (x_i - x_j) \qquad (4)$$

Substituting of (4) in (1) gives the required Lagrange form of the interpolating polynomial.

Remark

The Lagrange form (Eqn. (1)) of interpolating polynomial makes it easy to show the existence of an interpolating polynomial. But its evaluation at a point x_i involves a lot computation.

A more serious drawback of the Lagrange form arises in practice due to the following: One calculates a linear polynomial $P_1(x)$, a quadratic polynomial $P_2(x)$ e.t.c., by increasing the number of interpolation points, until a satisfactory approximation to f(x) has been found. In such a situation Lagrange form does not take any advantage of the availability of $P_{k-1}(x)$ in calculating $P_k(x)$. Later on, we shall see how in this respect, Newton form, discussed in the next unit, is more useful.

Let us consider some example to construct this form of interpolation polynomials.

Example 1

If f(1) = -3, f(3) = 9, f(4) = 30 and f(6) = 132, find the Lagrange's interpolation polynomial of f(x).

Solution

We have $x_0 = 1$, $x_1 = 3$, $x_2 = 4$, $x_3 = 6$ and $f_0 = -3$, $f_1 = 9$, $f_2 = 30$, $f_3 = 132$.

The Lagrange's interpolating polynomial P(x) is given by

$$P(x) = L_0(x)f_0 + L_1(x)f_1 + L_2(x)f_2 + L_3(x)f_3 \qquad (5)$$

where

$$L_0(x) = \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)}$$

$$= \frac{(x - 3)(x - 4)(x - 6)}{(1 - 3)(1 - 4)(1 - 6)}$$

$$= \frac{1}{30}(x^3 - 13x^2 + 54x - 72)$$

$$L_1(x) = \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)}$$

$$= \frac{(x - 1)(x - 4)(x - 6)}{(3 - 1)(3 - 4)(3 - 6)}$$

$$= \frac{1}{6}(x^3 - 11x^2 + 34x - 24)$$

$$L_2(x) = \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)}$$

$$= \frac{(x - 1)(x - 4)(x - 6)}{(4 - 1)(4 - 3)(4 - 6)}$$

$$= \frac{1}{6}(x^3 - 10x^2 + 27x - 18)$$

$$L_3(x) = \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)}$$

$$= \frac{(x-1)(x-3)(x-4)}{(6-1)(6-3)(6-4)}$$

$$= \frac{1}{30}(x^3 - 8x^2 + 19x - 12)$$

Substituting $L_j(x)$ and $f_j j = 0, 1, 2, 3$ in Eqn. (5), we get

$$P(x) = -\frac{1}{30}[x^3 - 13x^2 + 54x - 72](-3) + \frac{1}{6}[x^3 - 11x^2 + 34x - 24] \quad (9)$$
$$-\frac{1}{6}[x^3 - 10x^2 + 27x - 18](30) + \frac{1}{30}[x^3 - 8x^2 + 19x - 12] \quad (132)$$
$$= \frac{1}{10}[x^3 - 13x^2 + 54x - 72] + \frac{2}{3}[x^3 - 11x^2 + 34x - 24]$$
$$-5[x^3 - 10x^2 + 27x - 18] + \frac{22}{5}[x^3 - 8x^2 = 19x - 12]$$

which gives on simplification

$$P(x) = x^3 - 3x^2 = 5x - 6$$

which is the Lagrange's interpolating polynomial of f(x).

Example 2

Using Lagrange's interpolation formula, find the value of f when x = 1.4 from the following table.

x	1.2	1.7	1.8	2.0
f	3.3201	5.4739	6.0496	7.3891

Solution

the Lagrange's interpolating formula with 4 points is

$$P(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} f_0 + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} f_1 +$$

$$\frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} f_2 + \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} f_3 \quad (6)$$

Substituting

$x_0 = 1.2$, $x_1 = 1.7$, $x_2 = 1.8$, $x_3 = 2.0$ and
$f_0 = 3.3201$, $f_1 = 5.4739$, $f_2 = 6.0496$, $f_3 = 7.3891$

in (6), we get

$$P(x) = \frac{(x - 1.7)(x - 1.8)(x - 2.0)}{(1.2 - 1.7)(1.2 - 1.8)(1.2 - 2.0)} * 3.3201 +$$

$$\frac{(x - 1.2)(x - 1.8)(x - 2.0)}{(1.7 - 1.2)(1.7 - 1.8)(1.7 - 2.0)} * 5.4739 +$$

$$\frac{(x - 1.2)(x - 1.7)(x - 2.0)}{(1.8 - 1.2)(1.8 - 1.7)(1.8 - 2.0)} * 6.0496 +$$

$$\frac{(x - 1.2)(x - 1.7)(x - 1.8)}{(2.0 - 1.2)(2.0 - 1.7)(2.0 - 1.8)} * 7.3891 \qquad (7)$$

Putting x = 1.4 on both sides of (7), we get

$$f(1.4) = P(1.4) = \frac{(1.4 - 1.7)(1.4 - 1.8)(1.4 - 2.0)}{(-0.5)(-0.6)(-0.8)} * 3.3201 +$$

$$\frac{(1.4 - 1.2)(1.4 - 1.8)(1.4 - 2.0)}{(0.5)(-0.1)(0.3)} * 5.4739 +$$

$$\frac{(1.4 - 1.2)(1.4 - 1.7)(1.4 - 2.0)}{(0.6)(0.1)(-0.2)} * 6.0496 +$$

$$\frac{(1.4 - 1.2)(1.4 - 1.7)(1.4 - 1.8)}{(0.8)(0.3)(0.2)} * 7.3891$$

$$= \frac{(-0.3)(-0.4)(-0.6)}{(-0.5)(-0.6)(-0.8)} * 3.3201 +$$

$$\frac{(0.2)(-0.4)(-0.6)}{(0.5)(-0.1)(-0.3)} * 5.4739 +$$

$$\frac{(0.2)(-0.3)(-0.6)}{(0.6)(0.1)(-0.2)} * 6.0496 +$$

$$\frac{(0.2)(-0.3)(-0.4)}{(0.8)(0.3)(0.2)} * 7.3891$$

$$= 0.99603 + 17.51648 – 18.1488 + 3.69455$$

$= 4.05826$

Therefore f(x) = 4.05826.

3.2 Inverse Interpolation

In inverse interpolation for a table of values of x and y = f(x), one is given a number $\bar{y}$ and wishes to find the point $\bar{x}$ so that $f(\bar{x}) = \bar{y}$, where f(x) is the tabulated function. This problem can always be solved if f(x) is (continuous/and) strictly increasing or decreasing (that is, the inverse of f exists). This is done by considering the table of values x_i, $f(x_i)$, i = 0, 1, …, n to be a table of values y_i $g(y_i)$, i = 0, 1, 2, …, n for the inverse function $g(y) = f^{n-1}(y) = x$ by taking $y_i = f(x_i)$, $g(y_i) = x_i$, i = 0, 1, 2, …, n. Then we can interpolate for the unknown value $g(\bar{y})$ in this table.

$$P_n(\bar{y}) = \sum_{i=0}^{n} x_i \prod_{\substack{i=0 \\ i=j}}^{n} \frac{(y - y_j)}{(y_i - y_j)}$$

and $\bar{x} = P_n(\bar{y})$. This process is called inverse interpolation.

Let us consider some examples.

Example 3

From the following table, find the Lagrange's interpolating polynomial which agrees with the values of x at the given values of y. Hence find the value of x when y = 2.

x	1	19	49	101
y	1	3	4	5

Solution

Let x = g(y). the Lagrange's interpolating polynomial P(y) of g(y) is given by

$$P(y) = \frac{(y-3)(y-4)(y-5)}{(1-3)(1-4)(1-5)} * 1 + \frac{(y-1)(y-4)(y-5)}{(3-1)(3-4)(3-5)} * 19$$

$$+ \frac{(y-1)(y-3)(y-5)}{(4-1)(4-3)(4-5)} * 49 + \frac{(y-1)(y-3)(y-4)}{(5-1)(5-3)(5-4)} * 101$$

$$= -\frac{1}{24}[y^3 - 12y^2 + 47y - 60] + \frac{19}{4}[y^3 - 10y^2 + 29y - 20]$$

$$-\frac{49}{3}[y^3 - 9y^2 + 23y - 15] + \frac{101}{8}[y^3 - 8y^2 + 19y - 12]$$

which, on simplification, gives

$P(y) = y^3 - y^2 + 1.$

The Lagrange's interpolating polynomial of x is given by P(y).

There fore, $x = P(y) = y^3 - y^2 + 1$

Therefore, when y = 2, x = P(2) = 5.

Example 4

Find the value of x when y = 3 from the following table of values.

x	4	7	10	12
y	-1	1	2	4

Solution

The Lagrange's interpolation polynomial of x is given by

$$P(y) = \frac{(y-1)(y-2)(y-4)}{(-2)(-3)(-5)}(4) + \frac{(y+1)(y-2)(y-4)}{2(1)(-3)}(7)$$

$$+ \frac{(y+1)(y-1)(y-4)}{(3)(1)(-2)}(10) + \frac{(y+1)(y-1)(y-2)}{(5)(3)(2)}(12)$$

$$\text{Therefore } P(3) = \frac{(2)(1)(-1)}{-(2)(3)(5)}(4) + \frac{(4)(1)(-1)}{(2)(3)}(7)$$

$$+ \frac{(4)(2)(-1)}{-(3)(2)}(10) + \frac{(4)(2)(1)}{(5)(3)(2)}(12)$$

$$= \frac{4}{15} - \frac{14}{3} + \frac{40}{3} + \frac{48}{15}$$

$$= \frac{182}{15} = 12.1333$$

Hence, x(3) = P(3) = 12. 1333.

Now we are going to find the error committed in approximating the value of the function by $P_n(x)$.

3.3 General Error Term

Let $E_n(x) = f(x) – P_n(x)$ be the error involved in approximating the function f(x) by an interpolating polynomial. We derive an expression for $E_n(x)$ in the following theorem. This result helps us in estimating a useful bound on the error as explained in an example.

Theorem 4

Let $x_0, x_1, \ldots, x_n$ be distinct numbers in the interval [a, b] and f has (continuous) derivatives upto order (n + 1) in the open interval]a, b[. if $P_n(x)$ is the interpolating polynomial of degree $\leq$ n, which interpolates f(x) at the points $x_0, \ldots, x_n$, then for each x $\in$ [a, b], a number $\xi(x)$ in]a, b[exists such that

$$E_n(x) = f(x) - P_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)(x-x_1)\ldots\ldots(x-x_n) \qquad (8)$$

Proof

If x $\neq$ x_k for any k = 0, 1, 2, …, n, define the function g for t in [a, b] by

$$g(t) = f(t) - P_n(t) - [f(x) - P_n(x)]\prod_{j=0}^{n}\frac{(t-x_j)}{(x-x_j)}.$$

since f(t) has continuous derivatives up to order (n + 1) and P(t) ha derivatives of all orders, g(t) has continuous derivatives up to (n + 1) order. Now, for k = 0, 1, 2, …, n, we have

$$g(x_k) = f(x_k) = P_n(x_k) - [f(x) - P_n(x)]\prod_{j=0}^{n}\frac{(x_k-x_j)}{(x-x_j)}.$$

$$= 0 - [f(x) - P_n(x)].0 = 0$$

Furthermore, $g(x) = f(x) - P_n(x) - [f(x) - P_n(x)]\prod_{j=0}^{n}\frac{(x-x_j)}{(x-x_j)}.$

$$= f(x) - P_n(x) - [f(x) - P_n(x)].\ 1 = 0$$

Thus g has continuous derivatives up to order (n + 1) and g vanishes at the (n + 2) distinct points x, x_0, …, x_n. By the generalized Rolle's Theorem (Theorem 2) there exists $\xi(x)$ in]a, b[for which $g^{(n+1)}\ \xi(x)$ = 0. Differentiating g(t), (n + 1) times (with respect to t) and evaluating at $\xi(x)$ i, we get

$$0 = g^{(n+1)}\ \xi(x) = f^{(n+1)}\ \xi(x) - (n+1)!\ \frac{[f(x) - P_n(x)]}{\prod_{i=0}^{n}(x - x_i)}$$

Simplifying we get (error at x = $\bar{x}$)

$$E_n(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = \prod_{i=0}^{n}(\bar{x} - x_i)\frac{f^{(n+1)}\varsigma(\bar{x})}{(n+1)!} \qquad (9)$$

The error formula (Eqn. (9)) derived above, is an important theoretical results because Lagrange interpolating polynomials are extensively used in deriving important formulae for numerical differentiation and numerical integration.

It is to be noted that $\xi = \xi(\bar{x})$ depends on the ;point $\bar{x}$ at which the error estimate is required. This dependence need not even be continuous. This error formula is of limited utility since $f^{(n+1)}(x)$ is not known (when we are given a set of data at specific nodes) and the point x is hardly known. But the formula can be used to obtain a bound on the error of interpolating polynomial. Let us see how, by an example.

Example 5

The following table gives the values of f(x) = e^x. If we fit an interpolating polynomial of degree four to the data, find the magnitude of the maximum possible error in the computed value of f(x) when x = 1.25.

x	1.2	1.3	1.4	1.5	1.6
y	3.3201	3.6692	4.0552	4.4817	4.9530

Solution

From Eqn. (9), the magnitude of the error associated with the 4th degree polynomial approximation is given by

$$|E_4(x)| = |(x - x_0)(x - x_1)(x - x_2)(x - x_3)(x - x_4)|\frac{f^{(5)}(\xi)}{5!}$$

$$= \left|(x-1.2)(x-1.3)(x-1.4)(x-1.5)(x-1.6)\right| \frac{f^{(5)}(\xi)}{5!} \qquad (10)$$

Since f(x) = c^x, $f^{(5)}(x) = e^x$.

When x lies in the interval [1.2, 1.6],

Max $|f^{(5)}(x)| = e^{1.6} = 4.9530$ (11)

Substituting (11) in (10), and putting x = 1.25, the upper bound on the magnitude of the error

$$= |(0.05\ (-0.05)\ (-0.15)\ (-0.25)\ (-0.35)| * \frac{4.9530}{120}$$

= 0.00000135.

4.0 CONCLUSION

Let us take a brief look at what you have studied in this unit as the concluding path of this unit to the summary.

5.0 SUMMARY

In this unit, we have seen how to derive the Lagrange's form of interpolating polynomial for a given data. It has been shown that he interpolating polynomial for a given data is unique. Moreover the Lagrange form of interpolating polynomial can be determined for equally spaced or unequally spaced nodes. We have also seen how the Lagrange's interpolation formula can be applied with y as the independent variable and x as the dependent variable so that the value of x corresponding to a given value of y can be calculated approximately when some conditions are satisfied. Finally, we have derived the general error formula and its use has been illustrated to judge the accuracy of our calculation. The mathematical formulae derived in this unit are listed below for your easy reference.

1) Lagrange's Form

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x)$$

where

$$L_i(x) = \left[\prod_{\substack{j=0 \\ j\neq i}}^{n} (x - x_j)\right] / \left[\prod_{\substack{j=0 \\ j\neq i}}^{n} (x_i - x_j)\right]$$

2) Inverse Interpolation

$$P_n(y) = \sum_{i=0}^{n} x_i \left[\prod_{\substack{j=0\\ j\neq i}}^{n} \frac{(y-y_j)}{(y_i-y_j)} \right]$$

3) Interpolation Error

$$E_n(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = \prod_{i=0}^{n} (\bar{x}-x_i) \frac{f^{(n+1)}\varsigma(\bar{x})}{(n+1)!}$$

6.0 TUTOR-MARKED ASSIGNMENT

1) Show that

i) $\sum_{i=0}^{n} L_i(x) = 1$

ii) $\sum_{i=0}^{n} L_i(x)\, x_i^k = x^k,\ k \le n$

where $L_i(x)$ are Lagrange fundamental polynomials

2) Let $w(x) = \prod_{k=0}^{n} (x - x_k)$. Show that the interpolating polynomial of degree $\le$ n with the nodes $x_0, x_1, \ldots, x_n$ can be written as

$$P_n(x) = w(x) \sum_{i=0}^{n} \frac{f(x_k)}{(x - x_k) w'(x_k)}$$

3) Find the Lagrange's interpolation polynomial of f(x) from the following data. Hence obtain f(2).

x	0	1	4	5
f(x)	8	11	68	123

4) Find the value of y when x = 6 from the following table:

x	1	2	7	8
y	4	5	5	4

5) Using the Lagrange's interpolation formula, find the value of y when x = 10.

x	5	6	9	11
y	12	13	14	16

6) For the data of Example 5 with last one omitted, i.e., considering only first four nodes, if we fit a polynomial of degree 3, find an estimate of the magnitude of the error in the computed value of f(x) when x = 1.25. Also find an upper bound in the magnitude of the error.

7) Find the value of x when y = 4 from the table given below:

x	8	16	20	72
y	-1	1	3	5

8) Using Lagrange's interpolation formula, find the value of f(4) from the following data:

x	8	16	20	72
y	-1	1	3	5

7.0 REFERENCES/FURTHER READINGS

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

UNIT 2 NEWTON FORM OF THE INTERPOLATING POLYNOMIAL

CONTENTS

1.0 INTRODUCTION

The Lagrange's form of the interpolating polynomial derived in Unit 1 has some drawbacks compared to Newton form of interpolating polynomial that we are going to consider now.

In practice, one is often not sure as to how many interpolation points to use. One often calculates $P_1(x)$, $P_2(x)$, … increasing the number of interpolation points, and hence the degrees of the interpolating polynomials till one gets a satisfactory approximation $P_k(x)$, no advantage is taken of the fact that one has already constructed $P_{k-1}(x)$, whereas in Newton form it is not so.

Before deriving Newton's general form of interpolating polynomial, we introduce the concept of divided difference and the tabular representation of divided differences. Also the error of the interpolating polynomial in this case is derived in terms of divided differences. Using the two different expressions for the error term we get a relationship between nth order divided difference and nth order derivative.

2.0 OBJECTIVES

After studying this unit, you should be able to:

- obtain a divided difference in terms of function values
- form a table of divided differences and find divided differences with a given set of arguments from the table

- show that divided difference is independent of the order of its arguments
- obtain the Newton's divided differences interpolating polynomial for a given data
- find an estimate of f(x) for a given non-tabular value of x from a table of values of x and y [f(x)]
- relate the k^{th} order derivative of f(x) with the k^{th} order divided difference from the expression for the error term.

3.0 MAIN CONTENTS

3.1 Divided Differences

Suppose that we have determined a polynomial $P_{k-1}(x)$ of degree £ k – 1 which interpolates f(x) at the points x_0, x_1, …x_{k-1}. In order to make use of $P_{k-1}(x)$ in calculating $P_k(x)$ we consider the following problem: What function g(x) should be added to $P_{k-1}(x)$ to get $P_k(x)$? Let g(x) = $P_k(x)$. - $P_{k-1}(x)$. Now, g(x0 is a polynomial of degree £ k and $g(x_i) = P_k(x_i) - P_{k-1}(x_i) = f(x_i) - f(x_i) = 0$ for i = 0, 1, …, k – 1.

Suppose that $P_n(x)$ is the Lagrange polynomial of degree at most n that agrees with the function f at the distinct numbers x_0, x_1, …x_n. $P_n(x)$ can have the following representation, called Newton form.

$$P_n(x) = A_0 + A_1 (x_1 – x_0) + A_1 (x_1 – x_0) (x – x_1) + \dots \\ + A_n (x – x_0)\dots(x – x_{n-1}) \qquad (1)$$

for appropriate constant A_0, A_1, …, A_n.

Evaluating $P_n(x)$ (Eqn. (1)) at x_0 we get $A_0 = P_n(x_0)$. Similarly when $P_n(x)$ is evaluated at x_1, we get $A_1 = \dfrac{f(x) - f(x_1)}{x_1 - x_0}$. Let us introduce the notation for divided differences and define it at this stage: The zeroeth divided difference of the function f, with respect to x_i, is denoted by $f[x_i]$ and is simply the evaluation of f at x_i, that is, $f[x_i] = f(x_i)$. the first divided difference of f with respect to x_i and x_{i+1} is denoted by $f[x_i, x_{i+1}]$ and defined as

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}$$

The remaining divided differences of higher orders are defined inductively as follows. The kth divided differences relative to x_i, x_{i+1}, …, x_{i+k} is defined as

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}\ldots, x_{i+k}] - f[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}.$$

where the (k – 1)st divided differences $f[x_i, \ldots.., x_{i+k}]$ have been determined. This shows that kth divided difference is the divided differences of (k – 1)st divided differences justifying the name. The divided difference $f[x_i, x_2, \ldots., x_k]$ is invariant under all permutations of the arguments x_i, x_2, …., x_k. To show this we proceed giving another expression for the divided difference.

For any integer k between 0 and n. let $Q_k(x)$ be the sum of the first k + 1 terms in form (1), i.e.

$Q_k(x) = A_0 + A_1 (x – x_0) + \ldots + A_k(x – x_0)\ldots(x – x_{k-1})$..

Since each of the remaining terms in Eqn. (1) has the factor $(x – x_0) (x – x_1)\ldots (x – x_k)$, Eqn. (1) can be rewritten as

$P_n(x) = Q_k(x) + (x – x_0)\ldots (x – x_k) R(x)$ for some polynomial R(x). as the term $(x – x_0) (x – x_1)\ldots (x – x_k) R(x)$ vanishes at each of the points x_0, … x_k, we have $f(x_i) = P_n(x_i) = Q_k(x_i)$, i = 0, 1, 2, …, k. Since $Q_k(x)$ is a polynomial of degree £ k, by uniqueness of interpolating polynomial $Q_k(x) = P_k(x)$.

This shows that $P_n(x)$ can be constructed step by step with the addition of the next term in Eqn. (1), as one construct the sequence $P_0(x)$, $P_1(x)$ … with $P_k(x)$ obtained from $P_{k-1}(x)$ in the form

$$P_k(x) = P_{k-1}(x) + A_k(x – x_0)\ldots (x – x_{k-1}) \quad (2)$$

That is, g(x) is a polynomial of degree £ k having (at least) the k distinct zeros x_0, …, x_{k-1}.

\ $P_k(x) - P_{k-1}(x) = g(x) = A_k(x – x_0)\ldots(x – x_{k-1})$, for some constant A_k. this constant A_k is called the kth divided difference of f(x) at the points x_0, …, x_k for reasons discussed below and is denoted by $f[x_0, x_1, \ldots, x_k]$. this coefficient depends only on the values of f(x) at the point x_0, …, x_k. thus Eqn. (2) can be written as

$P_k(x) =P_{k-1}(x) + f[x_0, \ldots, x_k] (x – x_0)\ldots (x – x_{k-1})$,

since $(x – x_0) (x – x_1)\ldots (x – x_{k-1}) = x^k$ + a polynomial of degree < k,
we can rewrite $P_k(x)$ s $P_k(x) = f[x_0, \ldots, x_k] x^k$ + a polynomial of degree < k (4)

(as $P_{k-1}(x)$ is a polynomial of degree < k).

But considering the Lagrange form of interpolating polynomial we have

$$P_k(x) = \sum_{i=0}^{k} f(x_i) \prod_{\substack{j=0 \\ j \neq i}}^{k} \frac{(x - x_j)}{(x_i - x_j)}$$

$$= \sum_{i=0}^{k} \left[\frac{f(x_i)^k}{\prod_{\substack{j=0 \\ i \neq j}}^{k} (x_i - x_j)} \right] x^k + \text{a polynomial of degree} < k.$$

Therefore, on comparison with Eqn. (4) we have

$$f[x_0, \ldots, x_k] = \sum_{i=0}^{k} \frac{f(x_i)}{(x_i - x_0)\ldots(x_1 - x_{i-1})\ldots(x_1 - x_{i+1})\ldots(x_i - x_k)} \quad (5)$$

This shows that

$f[y_0, \ldots, y_k] = f[x_0, \ldots, x_k]$

if $y_0, \ldots, y_k$ is a reordering of the sequence $x_0, \ldots, x_k$. We have defined the zeroeth divided difference of f(x) at x_0 by $f[x_0] = f(x_0)$ which is consistent with Eqn. (5).

For k = 1, we have from Eqn. (5)

$$f[x_0, x_k] = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0} + \frac{f(x_0) - f(x_1)}{x_0 - x_1} = \frac{f[x_1] - f[x_0]}{x_1 - x_0}$$

This shows that the first divided difference is really a divided difference of divided differences.

We show below in Theorem 1 that for k > 2

$$f[x_0, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0 \ldots, x_{k-1}]}{x_k - x_0} \quad (6)$$

This shows that the k^{th} divided difference is the divided difference of (k – 1)st divided differences justifying the name. If $M = (x_0, \ldots, x_n)$ and N denotes any n – 1 elements of M and the remaining two elements are denoted by a and b, then

$(f [x_0,.., x_n=$

$$\frac{[(n - 1\text{st divided difference on N and a} - (n - 1)\text{st divided difference on N and b}]}{a - b} \quad (7)$$

Theorem 1:

$$f[x_0,.., x_j] = \frac{f[x_1, ..., x_j] - f[x_0, x_1...., x_{j-1}]}{x_j - x_0} \quad (8)$$

Proof: Let $P_{i-1}(x)$ be the polynomial of degree £ i – 1 which interpolates f(x) at $x_0, \ldots, x_{i-1}$ and let $Q_{j-1}(x)$ be the polynomial of degree $\leq j - 1$ which interpolates f(x) at the points $x_1, \ldots, x_j$. Let us define P(x) as

$$P(x) = \frac{x - x_0}{x_j - x_0} Q_{j-1}(x) + \frac{x_j - x}{x_j - x_0} P_{j-1}(x).$$

This is a polynomial of degree £ j, and $P(x_i) = f(x_i)$ for i = 0, 1, …, j. By uniqueness of the interpolating polynomial we have $P(x) = P_j(x)$. Therefore

$$P_j(x) \frac{x - x_0}{x_j - x_0} Q_{j-1}(x) + \frac{x_j - x}{x_j - x_0} P_{j-1}(x).$$

Equating the coefficient of x^j from both sides of Eqn. (8), we obtain (leading) coefficient of

$$x^j \text{ in } P_j(x) = \frac{\text{leading coefficient of } Q_{j-1}(x)}{x_j - x_0} - \frac{\text{leading coefficient of } P_{j-1}(x)}{x_j - x_0}$$

That is $f[x_0, ..., x_j] = \dfrac{f[x_1, ..., x_j] - f[x_0, ..., x_{j-1}]}{x_j - x_0}$.

We now illustrate this theorem with the help of a few examples but before that we give the table of divided differences of various orders.

Table of divided differences

Suppose we denote, for convenience, a first order divided difference of f(x) with any two arguments by f[.,.], a second order divided difference with any three arguments by f[.,.,.] and so on. Then the table of divided difference can be written as follows

Table 1

x	f[.]	f[.,.]	f[.,.,.]	f[.,.,.,.]	f[.,.,.,.,.]
x_0	f_0				
		$f[x_0,x_1]$			
x_1	f_1		$f[x_0,x_1x_2]$		
		$f[x_1,x_2]$		$f[x_0,x_1x_2x_3]$	
x_2	f_2		$f[x_1,x_2x_3]$		$f[x_0,x_1x_2x_3x_4]$
		$f[x_2,x_3]$		$f[x_1x_2x_3x_4]$	
x_3	f_3		$f[x_2x_3x_4]$		
		$f[x_3,x_4]$			
x_4	f_4				

Example 1: If $f(x) = x^3$, find the value of f[a, b, c].

Solution: $f[a, b] = \frac{f(b) - f(a)}{b - a} = \frac{b^3 - a^3}{b - a}$

$$= b^2 + ba + a^2 = a^2 + ab + b^2$$

Similarly,

$$f[a, b] = c^2 + cb + b^2 = b^2 + bc + c^2$$

$$f[a, b, c] = \frac{f[b, c] - f[a, b]}{c - a}$$

$$= \frac{(b^2 + bc + c^2) - (a^2 + ab + b^2)}{c - a}$$

$$= \frac{(c^2 - a^2) + b(c - a)}{c - a}$$

$$= \frac{(c - a)(c + a + b)}{(c - a)}$$

$$= a + b + c$$

$$f[a, b, c] \quad = a + b + c.$$

Example 2: If $f(x) = \frac{1}{x}$, show that

$$f[a, b, c, d] = -\frac{1}{abcd}$$

Solution: $f[a, b] = \dfrac{\frac{1}{b} - \frac{1}{a}}{b - a} = \dfrac{a - b}{ab(b - a)} = -\dfrac{1}{ab}$

Similarly,

$f[b, c] = -\dfrac{1}{bc}$, $f[c, d] = -\dfrac{1}{cd}$

$$f[a, b, c] = \frac{\frac{1}{bc} + \frac{1}{ab}}{c - a} = \frac{\frac{1}{ab} - \frac{1}{bc}}{c - a}$$

$$= \left[\frac{\frac{c-a}{abc}}{c-a}\right] = \frac{1}{abc}$$

Similarly,

$$f[b, c, d] = \frac{1}{bcd}$$

however $f[a, b, c, d] = \left[\dfrac{\frac{c-a}{abc}}{c-a}\right] = \dfrac{1}{abc}$

$$= \left[\frac{\frac{a-d}{abcd}}{d-a}\right]$$

$$= -\frac{1}{abcd}$$

Consequently,

$$f[a, b, c, d] = -\frac{1}{abcd}$$

In next section we shall make use of the divided difference to derive Newton's general form of interpolating polynomial.

3.2 Newton's General Form of Interpolating Polynomial

In section 3.1 we have shown how $P_n(x)$ can be constructed step by step as one construct the sequence $P_0(x)$, $P_1(x)$, $P_2(x)$, ..., with $P_k(x)$ obtained from $P_{k-1}(x)$ with the addition of the next term in Eqn. (3), that is,

$P_k(x) = P_{k-1}(x) + (x - x_0)(x - x_1)...(x - x_{k-1}) f[x_0, ..., x_k]$
Using this Eqn. (1) ca be rewritten as

$$P_n(x) = f[x_0] + (x - x_0) f[x_0,x_1] + (x - x_0)(x - x_1) f[x_0,x_1,x_2] + ... + (x - x_0)(x - x_1)... (x - x_{n-1}) f[x_0,x_1,....,x_n]. \quad (9)$$

This can be written compactly as follows:

$$P_n(x) = \sum_{i=0}^{n} f[x_0,...,x_i]\prod_{j=0}^{j-1}(x - x_j) \quad (10)$$

This is the Newton's form of interpolating polynomial.

Example 3:

From the following table of values, find the Newton's form of interpolating polynomial approximating (x).

x	-1	0	3	6	7
f(x)	3	-6	39	822	1611

Solution:

We notice that the values of x are not equally spaced. We are required to find a polynomial which approximates f(x). We form the table of divided differences of f(x).

Table 2

x	f[.]	f[.,.]	f[.,.,.]	f[.,.,.,.]	f[.,.,.,.,.]
-1	3				
		9			
0	- 6		6		
		15		5	
3	39		41		1
		261		13	
6	822		132		
		789			
7	1611				

Since the divided difference up to order 4 are available, the Newton's interpolating polynomial $P_4(x)$ is given by

$$P_4(x) = f(x_0) + (x - x_0)\, f[x_0,x_1] + (x - x_0)(x - x_1)\, f[x_0,x_1,x_2] + (x - x_0)(x - x_1)(x - x_2)\, f[x_0,x_1,x_2,x_3] + (x - x_0)(x - x_1)(x - x_2)(x - x_3)\, f[x_0,x_1,x_2,x_3,x_4] \quad (11)$$

where $x_0 = -1$, $x_1 = 0$, $x_2 = 3$, $x_3 = 6$ and $x_4 = 7$.

The divided differences $f(x_0)$, $f[x_0,x_1]$, $f[x_0,x_1,x_2]$, $f[x_0,x_1,x_2,x_3]$ and $f[x_0,x_1,x_2,x_3,x_4]$ are those which lie along the diagonal at $f(x_0)$ as shown by the dotted line. Substituting the values of x_i and the values of the divided differences in Eqn. (11), we get

$$P_4(x) = 3 + (x + 1)(-9) + (x + 1)\, x\, (6) + (x + 1)\, x\, (x - 3)(5) + (x + 1)\, x\, (x - 3)(x - 6)(1)$$

which on simplification gives

$$P_4(x) = x^4 - 3x^3 + 5x^2 - 6$$

Therefore, $f(x) = P_4(x) = x^4 - 3x^3 + 5x^2 - 6$

We now consider an example to show how Newton's interpolating polynomial can be used to obtain the approximate value of the function f(x) at any non-tabular point.

Example 4:

Find the approximate values of f(x) at x = 2 and x = 5 in Example 3.

Solution: Since $f(x) = P_4(x)$, from Example 3, we get

$f(2) = P_4(2) = 16 - 24 + 20 - 6 = 6$

and

$f(5) = P(5) = 625 - 375 + 125 - 6 = 369$

Note 1: When the values of f(x) for given values of x are required to be found, it is not necessary to find the interpolating polynomial $P_4(x)$ in its simplified form given above. We can obtain the required values by substituting the values of x in Eqn. (11) itself. Thus,

$P_4(2) = 3 + (3)(-9) + (3)(2)(6) + (3)(2)(-1)(5) + (3)(2)(-1)(-4)\, 1$

Therefore, $P_4(2) = 3 - 27 + 36 - 30 + 24 = 6$.

Similarly,

$$P_4(5) = 3 + (6)(-9) + (6)(5)(6) + (6)(5)(2)(5) + (6)(5)(2)(-1)(1)$$
$$= 3 - 54 + 180 + 300 - 60 = 369.$$

Then $f(2) = P_4(2) = 6$

And

$f(5) = P(5) = 369$.

Example 5:

Obtain the divided differences interpolation polynomial and the Lagrange's interpolating polynomial of f(x) from the following data and show that they are same.

x	0	2	3	4
f(x)	-4	6	26	64

Solution:

(a) Divided differences interpolation polynomial:

Table 3

x	f[x]	f[.,.]	f[.,.,.]	f[.,.,.,.]
0	-4			
		5		
2	6		5	
		20		1
3	26		9	
		38		
4	64			

$$P(x) = -4 + x(5) + x(x-2)(5) + x(x-2)(x-3)(1)$$

$$= x^3 + x - 4$$

\ $P(x) = x^3 + x - 4$

b) Lagrange's interpolation polynomial:

$$P(x) = \frac{(x-2)(x-3)(x-4)}{(-2)(-3)(-4)}(-4) + \frac{x(x-3)(x-4)}{(2)(-1)(-2)}(6)$$

$$+ \frac{x(x - 2)(x - 4)}{(3)(1)(- 1)} (26) + \frac{x(x - 2)(x - 3)}{(4)(2)(1)} (64)$$

$$= \frac{1}{6} (x^3 – 9x^2 + 26x – 24) + \frac{3}{2} (x^3 – 7x^2 + 12x)$$

$$- \frac{26}{3} (x^3 – 6x^2 + 8x) + 8(x^3 – 5x^2 + 6x).$$

On simplifying, we get
$P(x) = x^3 + x – 4.$

Thus, we find that both polynomials are same.
In Unit 1 we have derived the general error term i.e. error committed in approximating f(x) by $P_n(x)$. In next section we derive another expression for the error term in term of divided difference.

3.3 The Error of the Interpolating Polynomial

Let $P_n(x)$ be the Newton form of interpolating polynomial of degree $\leq$ n which interpolates f(x) at x_0 ...., x_n.
The interpolating error $E_n(x)$ of $P_n(x)$ is given by

$$E_n(x) = f(x) – P_n(x) \tag{12}$$

Let $\bar{x}$ be any point different from $x_0, ..., x_n$. If $P_n(x)$ is the Newton form of interpolating polynomial which interpolates f(x) at x_0, ..., x_n and $\bar{x}$, then $P_{n+1}(\bar{x}) = f(\bar{x})$. Then by (10) we have

$$P_{n+1}(x) = P_n(x) + f[x_0, ..., x_n, \bar{x}] \prod_{j=0}^{n} (x - x_j)$$

Putting x = $\bar{x}$ in the above, we have

$$f(\bar{x}) = P_{n+1}(\bar{x}) = P_n(\bar{x}) + f[x_0, ..., x_n, \bar{x}] \prod_{j=0}^{n} (\bar{x} - x_j)$$

$$\text{i.e. } E_n(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = f[x_0, ..., x_n, \bar{x}] \prod_{j=0}^{n} (\bar{x} - x_j) \tag{13}$$

This shows that the error is like the next term in the Newton form.

3.4 Divided Difference and Derivative of the Function

Comparing Eqn. (13) with the error formula derived in Unit 1 Eqn. (9), we can establish a relationship between divided difference and the derivatives of the function

$$E_n(\bar{x}) = \frac{f^{(n+1)}[x(\bar{x})]}{(n+1)!} \prod_{j=0}^{n} \left(\bar{x} - x_j\right)$$

$$= f[x_0, x_1, ..., x_n, \bar{x}] \prod_{j=0}^{n} \left(\bar{x} - x_j\right)$$

Comparing, we have $f[x_0, x_1, ..., x_{n+1}] = \dfrac{f^{(n+1)}\varsigma}{(n+1)!}$

(considering $\bar{x} = x_{n+1}$)

Further it can be shown that x Î]min x_i, max x_i[.
We state these results in the following theorem.

Theorem 2:

Let f(x) be a real-valued function, defined on [a, b] and n times differentiable in]a, b[. If x_0, ..., x_n are n + 1 distinct points in [a, b], then there exists $\varsigma \in$]a, b[such that

$$f[x_0, ..., x_n] = \frac{f^{(n+1)}\varsigma}{n!}$$

Corollary 1:

If $f(x) = x^n$, then

$$f[x_0, ..., x_n] = \frac{n!}{n!} = 1.$$

Corollary 2:

If $f(x) = x^k$, $k < n$, then

$f[x_0, ..., x_k] = 0$

since nth derivative of x^k, $k < n$, is zero.

For example, consider the first divided difference

$$f[x_0,x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

by Mean Value Theorem $f(x_1) = f(x_0) + (x_1 - x_0) f'(\varsigma)$, $x_0 < \varsigma < x$,

substituting, we get

$f[x_0,x_1] = f'(\varsigma)$, $x_0 < \varsigma < x_1$.

Example 6:

If $f(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_0$, then find $f[x_0, x_1, ..., x_n] = a_n * \frac{n!}{n!} + 0 = a_n$.

Let us consider another example.

Example 7:

If $f(x) = 2x^3 + 3x^2 - x + 1$, find

f[1, -1, 2, 3], f[a, b, c, d], f[4, 6, 7, 8].

Solution:

Since f(x) is a cubic polynomial, the 3rd order divided differences of f(x) with any set of argument are constant and equal to 2, the coefficient of x^3 in f(x).

Thus, it follows that f[1, -1, 2, 3], f[a, b, c, d], and f[4, 6, 7, 8] are each equal to 2.

In the next section, we are going to discuss about bounds on the interpolation error.

3.5 Further Results on Interpolation Error

We have derived error formula

$$E_n(x) = f(x) - P_n(x) = \prod_{i=0}^{n} (\bar{x} - x_i) \frac{f^{(n+1)}\varsigma(\bar{x})}{(n+1)!},$$

We assume that f(x) is (n + 1) times continuously differentiable in the interval of interest [a, b] = I that contains x_0, ..., x_n and x. since $\varsigma(x)$ is known we may replace $f^{(n+1)}(\varsigma(x))$ by $\max_{x \in I}\left|f^{(n+1)}(x)\right|$. If we denote (x - x_0) (x – x_1)...(x – x_n) by $\psi_n(x)$ then we have

$$\left|E_n(x)\right| = |f(x) - P_n(x)| \le \frac{\max_{x \in I}\left|f^{(n+1)}(t)\right|}{(n+1)!} \max_{x \in I}\left|\psi_n(t)\right| \qquad (14)$$

Consider now the case when the nodes are equally spaced, that is (m $x_j = x_0 + jh$), j = 0,.....,N, and h is the spacing between consecutive nodes. For the case n = 1 we have linear interpolation. If $x \in [x_{i-1}, x_i]$, then we approximate f(x) by $P_1(x)$ which interpolates at

x_{i-1}, and x_i. From Eqn. (14) we have $|E_n(x)| \le \frac{1}{2} \max_{t \in I}|f''(t)| \max_{t \in I}|\psi_1(t)|$

where $\psi_1(x) = (x - x_{i-1})(x - - x_i)$.

Now,

$$\frac{d\psi_1}{dx} = x - x - - = 0$$

gives $x = (x_{i-1} - x_i)/2$.

Hence, the maximum value of $(x - x_{i-1})(x - - x_i)|$ occurs at
$x = x^* = (x_{i-1} - x_i)/2$.

The maximum value is given by

$$|\psi_1(x^*)| = \frac{(x_i - x_{i-1})^2}{4} = \frac{h^2}{4}.$$

Thus, we have for linear interpolation, for nay $x \in I$

$$|E_1(x)| = |f(x) - P_1(x)| \le \frac{(x_i - x_{i-1})^2}{4} \frac{1}{2} \max_{x \in I}|f''(x)|$$

$$= \frac{h^2}{8} M. \qquad (15)$$

where $|f''(x)| \le M$ on I.

For the case n = 2, it can be shown that for any $x \in [x_{i-1}, x_{i+1}]$.

$$|E_2(x)| \le \frac{h^2 M}{9\sqrt{3}} \text{ where } |f'''(x)| \le M \text{ on I.} \tag{16}$$

Example 8:

Determine the spacing h in table of equally spaced values of the function of $f(x) = \sqrt{x}$ between 1 and 2, so that interpolation with a first degree polynomial in this table will yield seven place accuracy.

Solution: Here

$$f''(x) = -\frac{1}{4}x^{-3/2}$$

$$\max_{1 \le x \le 2} |f''(x)| = \frac{1}{4}.$$

and $|E_1(x)| \le \frac{h^2}{32}$.

For seven place accuracy, h is to be chosen such that

$$\frac{h^2}{32} < 5.10^{-8}.$$

or $h^2 < (160)10^{-8}$ that is $h < .0013$.

4.0 CONCLUSION

This unit shall be concluded by giving a summary of what we have covered in it.

5.0 SUMMARY

In this unit we have derived a form of interpolating polynomial called Newton's general form, which has some advantage over the Lagrange's form discussed in Unit 1. This form is useful in deriving some other interpolating formulas. We have introduced the concept of divided differences and discussed some of its important properties before deriving Newton's general form. The error term has also been derived and utilizing the error term we have established a relationship between the divided difference and the derivative of the function f(x) for which the interpolating polynomial has been obtained. The main formula derived are listed below:

1) $$f[x_0,....,x_j] = \frac{f[x_1,....,x_j] - f[x_0,....,x_{j-1}]}{x_j - x_0}$$

2) $P_n(x) = \sum_{i=0}^{n} f[x_0,...,x_i]\prod_{j=0}^{j-1}(x - x_j)$

3) $E_n(x) = f[x_0,...,x_n,x]\prod_{j=0}^{n}(x - x_j)$

4) $f[x_0,....,x_n] = \dfrac{f^{(n)}(\xi)}{n!}$, $\xi \in$]min x_i, $\max_i$[

6.0 TUTOR-MARKED ASSIGNMENT

1) Find the Lagrange's interpolating polynomial of f(x) from the table of values given below and show that it is the same as the Newton's divided differences interpolating polynomial.

x	0	1	4	5
f(x)	8	11	68	123

2) Form the table of values given below, obtain the value of y when x = 1.5 using

a) divided differences interpolation formula.

b) Lagrange's interpolation formula.

x	0	1	2	4	5
f(x)	5	14	41	98	122

3) Using Newton's divided difference interpolation formula, find the values of f(8) and f(15) from the following table.

x	4	5	7	10	11	13
f(x)	48	100	294	900	1210	2028

4) If $f(x) = 2x^3 – 3x^2 + 7x + 1$, what is the value of f[1, 2, 3, 4]?

5) If $f(x) = 3x^2 – 2x + 5$, find f[1, 2], f[2, 3] and f[1, 2, 3].

6) If f(x) takes the values -21, 15, 12 and 3 respectively when x assumes the values -1, 1, 2 and 3, find the polynomial which approximates f(x).

7) Find the polynomial which approximate f(x), tabulated below

x	-4	-1	0	2	5
f(x)	1245	33	5	9	1335

Also find an approximate value of f(x) at x = 1 and x = -2.

8) From the following table, find the value of y when x = 102

x	93.0	96.2	100.0	104.2	108.7
y	11.38	12.80	14.70	17.07	19.91

7.0 REFERENCES/FURTHER READINGS.

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

UNIT 3 INTERPOLATION AT EQUALLY SPACED POINTS

CONTENTS

1.0 INTRODUCTION

Suppose that y is a function of x. The exact functional relation y = f(x) between x and y may or may not be known. But, the values of y at (n + 1) equally spaced of x are supposed to be known, i.e., (x_i, y_i); i = 0, ..., n are known where $x_i – x_{i-1} = h$ (fixed), i = 1, 2, ..., n. Suppose that we are required to determine an approximate value of f(x) or its derivative f'(x) for some values of x in the interval of interest. The methods for solving such problems are based on the concept of finite differences. We have introduced the concept of forward, backward and central differences and discussed their interrelationship in the previous unit

We have already introduced two important forms of the interpolating polynomial in Units 1 and 2. These forms simply when the nodes are equidistant. For the case of equidistant nodes, we have derived the Newton's forward, backward difference forms and Stirling's central difference form of interpolating, each suitable for use under a specific situation. We have derived these methods in the previous unit and also given the corresponding error term.

2.0 OBJECTIVES

After reading this unit, you should be able to:

- write a forward difference in terms of function values from a table of forward differences and locate a difference of given order at a given point
- write a backward difference in terms of function values from a table of backward differences and identify differences of various orders at any given point from the table
- expand a central difference in terms of function values and form a table of central differences
- establish relations between V, ∇, d and divided difference
- obtain the interpolating polynomial of f(x) for a given data by applying any one of the interpolating formulas
- compute f(x) approximately when x lies near the beginning of the table and estimate the error
- compute f(x) approximately when x lies near the end of the table and estimate the error
- estimate the value of f(x) when x lies near the middle of the table and estimate the error.

3.0 MAIN CONTENTS

3.1 Differences

Suppose that we are given a table of values (x_i, y_i), i = 0, 1, 2, ..., N where $y_i = f(x_i) = f_j$.
Let the nodal points be equidistant. That is

$$x_i = a + ih,\ i = 0, ..., N, \text{ with } N = (b - a)/h \qquad (1)$$

For simplicity we introduce a linear change of variables

$$s = s(x) = \frac{x - x_0}{h}, \text{ so that } x = x(s) = x_0 + sh \qquad (2)$$

and introduce the notation

$$f(x) = f(x_0 + sh) = f_s \qquad (3)$$

The linear change of variables in Eqn. (2) transforms polynomials of degree n in x into polynomials of degree n is s. we have already introduced the divided-difference table to calculate a polynomial of degree £ n which interpolates f(x) at $x_0, x_1, ..., x_n$. For equally spaced nodes, we shall deal with three types of differences, namely, forward, backward and central and discuss their representation in the form of a table. We shall also derive the relationship of these differences with divided differences and their interrelationship.

3.1.1 Forward Differences

We denote the forward differences of f(x) if ith order at $x = x_0 + sh$ by $\Delta^i f_s$ and define it as follows:

$$\Delta^i f_s = \begin{cases} f_s & i = 0 \\ V(V^{i-1} f_s) = V^{i-1} f_{s+1} - V^{i-1} f_s, & i > 0. \end{cases}$$

Where V denotes forward difference operator.

When s = k, that is, $x = x_k$, we have

for i = 1 $\quad \Delta f_k = f_{k+1} - f_k$

for i = 2 $\quad \Delta^2 f_k = f_{k+1} - f_k$

$$= f_{k+2} - f_{k+1} - [f_{k+1} - f_k]$$

$$= f_{k+2} - f_{k+1} + f_k$$

Similarly $\quad \Delta^3 f_k = f_{k+3} - 3f_{k+2} + 3f_{k+1} - f_k$

We recall the binomial theorem

$$(a + b)^s = \sum_{j=0}^{n} \begin{bmatrix} s \\ j \end{bmatrix} a^j b^{r-j} \qquad (4)$$

where s is a real non-negative integer.

We give below in Lemma 1 the relationship between the forward and divided differences. This relation will be utilized to derive the Newton's forward difference formula which interpolates f(x) at $x_k + ih$, i = 0, 1, ..., n.

Lemma 1: For all $i \geq 0$

$$f[x_k, ..., k_{k+1}] = \frac{1}{i!h^i} \Delta^i f_k \qquad (5)$$

Proof:

We prove the result by induction.
For i = 0, both sides of relation (5) are same by convention, that is,

$f[x_k] = f(x_k) = f_k = \Delta^0 f_k$.

Assuming that relation (5) holds for $i = n \geq 0$, we have for i = n + 1

$$f[x_k, x_{k+1}, ..., k_{k+n+1}] = \frac{f[x_{k+1}, ..., x_{k+n+1}] - f[x_k, ..., x_{k+n}]}{x_{k+n+1} - x_k}$$

$$= \frac{[\Delta^n f_{k+1} / n! h^n] - [\Delta^n f_k / n! h^n]}{x_0 + (k+n+1)h - x_0 - kh}$$

$$= \frac{\Delta^n f_{k+1} - \Delta^n f_k}{(n+1)! h^{n+1}} = \frac{\Delta^{n+1} f_k}{(n+1)! h^{n+1}}$$

This shows that relation (5) holds for i = n + 1 also. Hence (5) is proved. We now give a result which immediately follows from this theorem in the following corollary.

Corollary:

If $P_n(x)$ is a polynomial of degree n with leading coefficient a_n, and x_0 is an arbitrary point, then

$\Delta^n P_n(x_0) = a_n n!\, h^n$

and $\Delta^{n+1} P_n(x_0) = 0$, i.e., all higher differences are zero.

Proof: Taking k = 0 in relation (5) we have

$$f[x_0, ..., x_i] = \frac{1}{1!\,h^i} \Delta^i f_0. \qquad (6)$$

Let us recall that

$$f[x_0, ..., x_i] = \frac{f^{(1)}(\xi)}{1!} \qquad (7)$$

where f(x) is a real-valued function defined on [a, b] and i times differentiable in]a, b[and $\xi \in$]a, b[.

Taking i = n and $f(x) = P_n(x)$ in Eqns. (6) and (7), we get

$$\Delta^{n} P_n(x_0) = n!h^{n} P_n[x_0, ..., x_n] = n!h^n \frac{P_n^{(n)}(x)}{n!}$$
$$= h^n n! a_n.$$

Since $\Delta^{n+1} P_n(x_0) = \Delta^{n} P_n(x_1) - \Delta^{n} P_n(x_0)$

$$= h^n n! a_n - h^n n! a_n = 0.$$

This completes the proof

The shift operator E is defined as

$$Ef_i = f_{i+1} \tag{8}$$

In general $Ef(x) = f(x + h)$.

We have $E^s f_i = f_{i+s}$

For example,

$E^3 f_i = f_{i+3}$, $E^{1/2} f_i = f_{i+1/2}$ and $E^{-1/2} f_i = f_{i-1/2}$

Now,

$\Delta f_i = f_{i+1} - Ef_i - f_i = (E - 1)f_i$

Hence the shift and forward difference operations are related by

$$\Delta = E - 1$$

or $\quad E = 1 + \Delta$

Operating s times, we get

$$\Delta^s = (e - 1)^s = \sum_{j=0}^{n} \binom{s}{j} E^j (-1)^{r-1} \tag{9}$$

Making use of relation (8) in Eqn. (9), we get

$$\Delta^s f_i = \sum_{j=0}^{n} (-1)^{r-1} \binom{s}{j} f_{j+1}$$

We now give in Table 1, the forward differences of various orders using 5 values.

Table 1: Forward Difference Table

x	f(x)	$\Delta^1 f$	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
x_0	f_0				
		Δf_0			
x_1	f_1		$\Delta^2 f_0$		
		Δf_1		$\Delta^3 f_0$	
x_2	f_2		$\Delta^2 f_1$		$\Delta^4 f_0$
		Δf_2		$\Delta^3 f_1$	
x_3	f_3		$\Delta^2 f_2$		
x_4	f_4	Δf_3			

Note that the forward difference $\Delta^k f_0$ lie on a straight line sloping downward to the right.

3.1.2 Backward Differences

Let f be a real-valued function of x. let the values of f(x) at n + 1 equally spaced points $x_0, x_1, \ldots, x_n$ be $f_0, f_1, \ldots, f_n$ respectively.

The backward differences of f(x) of ith order at $x_k = x_0 + kh$ are denoted by $\nabla^i f_k$. They are defined as follows:

$$\nabla^i f_k = \begin{cases} f_k, & 1 = 0 \\ \nabla^{i-1}(\nabla f_k) = \nabla^{i-1}[f_k - f_{k-1}], & i \geq 1 \end{cases} \qquad (10)$$

where ∇ denotes backward difference operator.

Using (10), we have for

$i = 1;\ \nabla f_k = f_k - f_{k-1}$

$$\begin{aligned} i = 2;\ \nabla^2 f_k &= \nabla(f_k - f_{k-1}) \\ &= \nabla f_k - f_{k-1} \\ &= f_k - 2f_{k-1} + f_{k-2} \end{aligned}$$

$$\begin{aligned} i = 3;\ \nabla^3 f_k &= \nabla^2[f_k - f_{k-1}] = \nabla^2 f_k - \nabla^2 f_{k-1} = \nabla[f_k] - \nabla[f_{k-1}] \\ &= \nabla[f_k - f_{k-1}] - \nabla[f_{k-1} - f_{k-2}] \\ &= \nabla f_k - \nabla f_{k-1} - \nabla f_{k-1} + f_{k-2} \\ &= f_k - f_{k-1} - 2[f_{k-1} + f_{k-2}] + f_{k-2} - f_{k-3} \\ &= f_k - 3f_{k-2} + 3f_{k-2} - f_{k-3} \end{aligned}$$

By induction we can prove the following lemma which connects the divided difference with the backward difference.

Lemma 2: The following relation holds

$$f[x_{n-k}, ..., x_n] = \frac{1}{k!h^k} \nabla^k f(x_n) \qquad (11)$$

The relation between the backward difference operator ∇ and the shift operator E is given by

$$\nabla = 1 E^{-1} \text{ or } E = (i - \nabla)^{-1}$$

Since $\nabla f_k = f_k - f_{k-1} = f_k - E^{-1} f_k = [1 - E] f_k$.

Operating s times, we get

$$\nabla^s f_k = [1 - E]^s f_k = \left[\sum_{j=0}^{n} \binom{s}{m} E^{-m} (-1)^m\right] f_k$$

$$= \sum_{j=0}^{n} \binom{s}{m} (-1)^m f_{k-m} \qquad (12)$$

We can extend the binomial coefficient notation to include negative numbers, by letting

$$\begin{bmatrix} s \\ i \end{bmatrix} = \frac{-s(-s-1)(-s-2)....(-s-i+1)}{i!} = (-1)^i \frac{s(s+1)....(s+i-1)}{i!}$$

The backward differences of various orders with 5 nodes are given in Table 2.

Table 2: Backward Difference Table

x	f(x)	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$
x_0	f_0				
		∇f_1			
x_1	f_1		$\nabla^2 f_2$		
		∇f_2		$\nabla^3 f_3$	
x_2	f_2		$\nabla^2 f_3$		$\nabla^4 f_4$
		∇f_3		$\nabla^3 f_4$	
x_3	f_3		$\nabla^2 f_4$		
		∇f_4			
x_4	f_4				

Let us consider the following example:

Example 1: Evaluate the differences

(a) $\nabla^3[a_2x^2 + a_1x + a_0]$

(b) $\nabla^3[a_3x^3 + a_2x^2 + a_3x + a_0]$.

Solution:

(a) $\nabla^3[a_2x^2 + a_1x + a_0] = 0$

(b) $\nabla^3[a_3x^3 + a_2x^2 + a_3x + a_0]$.

$= a_3\nabla^3(x^3) + \nabla^3[a_2x^2 + a_1x + a_0]$

$= a_3.3 ! h^2$

Note that the backward differences $\nabla^k f_4$ lie on a straight line sloping upward to the right.

Also note that $V f_k = \nabla f_{k+1} = f_{k+1} – f_k$.

Try to show that $V^4 f_0 = \nabla^4 f_4$.

Let us now discuss about the central differences.

3.1.3 Central Differences

The first order central difference of f at x_k, denoted by df_k, is defined as

$df = f(x + h/2) – f(x – h/2) = f_{k+1/2} – f_{k-1/2}$.

Operating with d, we obtain the higher order central differences as

$d^s f_k = f_k$ when $s = 0$.

The second order central difference is given by

$$d^2 f_k = d[f_{k+1/2} – f_{k-1/2}] = d[f_{k+1/2}] - d[f_{k-1/2}]$$
$$= f_{k+1} – f_k – f_k + f_{k-1}$$
$$= f_{k+1} – 2f_k + f_{k-1}$$

Similarly,

$d^3 f_k = f_{k+3/2} - 3f_{k+1/2} + 3f_{k-1/2} - f_{k-3/2}$

and $d^4 f_k = f_{k+2} - 4f_{k+1} + 6f_k - 4f_{k-1} + f_{k-2}$.

Notice that the even order differences at a tabular value x_k are expressed in terms of tabular values of f and odd order differences at a tabular value x_k are expressed in terms of non-tabular value of f. also note that the coefficients of $d^s f_k$ are the same as those of the binomial expansion of $(1 – x)^s$, s = 1, 2, 3, ...

Since
$d f_k \; f_{k+1/2} – f_{k-1/2} = (E^{1/2} – E^{-1/2})f_k$

We have the operation relation

$$d = E^{1/2} – E^{-1/2} \tag{14}$$

The central differences at a non-tabular point $x_{k+1/2}$ can be calculated in a similar way.

For example,

$d f_{k+1/2} = f_{k+1} - f_k$

$$\begin{aligned} d^2 f_{k+1/2} &= f_{k+3/2} - 2f_{k+1/2} + f_{k-1/2} \\ d^3 f_{k+1/2} &= f_{k+2} - 3f_{k+1} + 3f_k - f_{k-1} \\ d^4 f_{k+1/2} &= f_{k+3/2} - 4f_{k+3/2} + 6f_{k+1/2} - 4f_{k-1/2} + f_{k-3/2} \end{aligned} \tag{15}$$

Relation (15) can be obtained easily by using the relation (14)

We have

$d^s f_k = [E^{1/2} – E^{-1/2}]^s f_k$

$$= \left[\sum_{i=0}^{n} \binom{s}{i} E^{-i/2} E^{(n-i)/2} (-1)^i \right] f_k$$

$$= \left[\sum_{i=0}^{n} \binom{s}{i} (-1)^i \right] f_{k+(n/2)-1} \tag{16}$$

The following formulas can also be established:

$$f[x_0, ..., x_{2m}] = \frac{1}{(2m)!h^{2m}} d^{2m} f_m \tag{17}$$

$$f[x_0, ..., x_{2m+1}] = \frac{1}{(2m+1)!h^{2m+1}} d^{2m+1} f_{m+1/2} \tag{18}$$

$$f[x_{-m}, ...,x_0, ..., x_m] = \frac{1}{(2m)!h^{2m}} d^{2m} f_0 \tag{19}$$

$$f[x_{-m}, ..., x_0, ..., x_{m+1}] = \frac{1}{(2m+1)!h^{2m+1}} d^{2m+1} f_{1/2} \tag{20}$$

$$f[x_{-(m+1)}, ..., x_0, ..., x_m] = \frac{1}{(2m+1)!h^{2m+1}} d^{2m+1} f_{-1/2} \tag{21}$$

We now give below the central difference table with 5 nodes.

Table 3: Central Difference Table

x	f	df	d^2f	d^3f	d^4f
x_{-2}	f_{-2}				
		$df_{-3/2}$			
x_{-1}	f_{-1}		d^2f_{-1}		
		$df_{-1/2}$		$d^3f_{-1/2}$	
x_0	f_0		d^2f_0		d^4f_0
		$df_{1/2}$		$d^3f_{1/2}$	
x_1	f_1		d^2f_1		
		$df_{3/2}$			
x_2	f_2				

Note that the difference $d^{2m}f_0$lie on a horizontal line shown by the dotted lines.

Table 4: Central Difference Table

x	f	df	d^2f	d^3f	d^4f
x_0	f_0				
		$df_{1/2}$			
x_1	f_1		d^2f_1		
		$df_{3/2}$		$d^3f_{3/2}$	
x_2	f_2		d^2f_2		d^4f_2
		$df_{5/2}$		$d^3f_{5/2}$	
x_3	f_3		d^2f_3		
		$df_{7/2}$			
x_4	f_4				

Note that the difference $d^{2m}f_2$ lie on a horizontal line.

We now define the mean operator $m$as follows

$$mf_k = \frac{1}{2}[f_{k+1/2} + f_{k-1/2}]$$

$$= \frac{1}{2}[E^{1/2} + E^{-1/2}]f_k.$$

Hence

$$m = \frac{1}{2}[E^{1/2} + E^{-1/2}]$$

Relation Between the Operators V, ∇, d and m

We have expressed V, ∇, d and m in terms of the operator E as follows

$$V = E - 1$$

$$\nabla = 1 - E^{-1}$$

$$d = E^{1/2} - E^{-1/2}$$

$$m = \frac{1}{2}[E^{1/2} + E^{-1/2}]$$

$$V = E(1 = E^{-1}) = E\nabla$$

$$= E^{1/2}(E^{1/2} - E^{-1/2})\ E^{1/2}d$$

Also $E^{1/2} = m + \frac{d}{2}$

$E^{-1/2} = m - \frac{d}{2}$

Example 2:

(a) Express $V^3 f_1$ as a backward difference.

(b) Express $V^3 f_1$ as a central difference.

(c) Express $d^2 f_2$ as a forward difference.

Solution:

(a) $\Delta^3 f_1 = (E\nabla)^3 f_1 = E^3\nabla^3 f_1 = \nabla^3 E^3 f_1 = \nabla^3 f_4$ $\quad (\Delta = E\nabla)$

(b) $\Delta^3 f_1 = [E^{1/2}\partial]^3 f_1 = E^{3/2}\partial^3 f_1 = \partial^3 E^{3/2} f_1 = \partial^3 f_{5/2}$ $\quad (\nabla = E^{1/2}\partial)$

(c) $\partial^2 f_2 = [E^{-1/2}\partial]^2 f_2 = E^{-1}\Delta^2 f_2 = \Delta^2 E^{-1} f_2 = \Delta^2 f_1$ $\quad (\partial = E^{-1/2}\Delta)$

Example 3: Prove that

(a) $m^2 = 1 + \frac{\partial^2}{4}$

(b) $md = \frac{1}{2}(\Delta + \nabla)$

(c) $\sqrt{1 + m^2d^2} = 1 + \frac{\partial^2}{2}$

Solution:

(a) We have $m = \frac{1}{2}[E^{1/2} + E^{-1/2}]$

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

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

$$= 1 + \frac{\partial^2}{4}$$

(b) L.H.S.

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

R.H.S.

$$\frac{1}{2}(\Delta + \nabla) = \frac{1}{2}[(E-1) + (1-E^{-1})] = \frac{1}{2}(E - E^{-1}).$$

Hence, the result.

(c) We have

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

$$\backslash \quad 1 + m^2d^2 = 1 + \frac{(E - E^{-1})^2}{4} = \frac{(E - E^{-1})^2 + 4}{4} = \frac{(E + E^{-1})^2}{4}$$

$$\backslash \quad \sqrt{1 + m^2d^2} = \frac{E + E^{-1}}{2} = \frac{(E^{1/2} - E^{-1/2})^2 + 2}{2}$$

$$= \frac{d^2 + 2}{2} = 1 + \frac{\partial^2}{4}$$

3.2 Difference Formulas

We shall now derive different difference formulas using the results obtained in the preceding section (Section 3.2).

3.2.1 Newton's Forward-Difference Formula

In Unit 2, we have derived Newton's form of interpolating polynomial (using divided differences). We have also established in Section 3.2 1, the following relationship between divided differences and forward differences

$$f[x_k, ..., x_{k+n}] = \frac{1}{n!h^n} V^n f_k \qquad (21)$$

Substituting the divided differences in terms of the forward differences in the Newton's form, and simplifying we get Newton's forward-difference form. The Newton's form of interpolating polynomial interpolating at x_k, x_{k+1}, ..., x_{k+n} is

$$P_n(x) = \sum_{i=0}^{n} (x - x_k)(x - x_{k+1})..........(x - x_{k+i-1}) f[x_k, ...x_{k+1}]$$

Substituting (22), we obtain

$$P_n(x) = \sum_{i=0}^{n} (x - x_k)(x - x_{k+1})..........(x - x_{k+i-1}) \frac{1}{1!h^1} \Delta^i f_k \qquad (23)$$

Setting k = 0, we have the form

$$P_n(x) = \sum_{i=0}^{n} \frac{1}{1!h^1} (x - x_0)(x - x_1)..........(x - x_{i-1}) \Delta^i f_0$$

$$= f_0 + \frac{(x - x_0)}{1!} \frac{\Delta f_0}{h} + \frac{(x - x_0)(x - x_1)}{h^2} \frac{\Delta^2 f_0}{h^2} + ...$$
$$+ \frac{(x - x_0)....(x - x_{n-1})}{n!} \frac{\Delta^n f_0}{h^n} \qquad (24)$$

Using the transformation (2), we have

$$x - x_{k+j} = x_0 + sh - [x_0 + (k + j)h] = (s - k - i + 1) \; V^i f_k$$

$$= \sum_{i=0}^{n} \Delta^1 f_k \begin{bmatrix} s-k \\ i \end{bmatrix}$$

$$= f_k + (s - k) \; \Delta f_k + \frac{(s - k)(s - k - 1)}{2!} \Delta^2 f_k + ...$$

$$+ \frac{(s-k)(s-n-1)}{n!} \Delta^n f_k \qquad (25)$$

of degree £ n.

Setting k = 0 in (25) we get the formula

$$P_n(x_0 + sh) = \sum_{i=0}^{n} \Delta^i f_0 \begin{bmatrix} s \\ i \end{bmatrix} \qquad (26)$$

The form (23), (24), (25) or (26) is called the Newton's forward-difference formula.

The error term is now given by

$$E_n(x) = \begin{bmatrix} s \\ n+1 \end{bmatrix} h^{n+1} f^{n+1}(x)$$

-

Example 4:

Find the Find the Newton's forward-difference interpolating polynomial which agrees with the table of values given below. Hence obtain the value of f(x) at x = 1.5.

x	1	2	3	4	5	6
f(x)	10	19	40	79	142	235

Solution: We form a table of forward differences of f(x).

Table 5: Forward differences

x	f(x)	Δf	$\Delta^2 f$	$\Delta^3 f$
1	10			
		9		
2	19		12	
		21		6
3	40		18	
		39		6
4	79		24	
		63		6
5	142		30	
		93		
6	235			

Since the third order differences are constant, the higher order differences vanish and we can infer that f(x) is a polynomial of degree 3 and the Newton's forward-differences interpolation polynomial exactly represents f(x) and is not an approximation to f(x). The step length in the data id h = 1. Taking x_0 = 1 and the subsequent values of x as x_1, x_2, ..., x_5 the Newton's forward-differences interpolation polynomial.

$$f(x) = f_0 + (x - 1)\ V f_0 + \frac{(x-1)(x-2)}{2!} V^2 f_0 + \frac{(x-1)(x-2)(x-3)}{3!} V^3 f_0$$

becomes

$$f(x) = 10 + (x - 1)(9) + \frac{(x-1)(x-2)}{2}(12) + \frac{(x-1)(x-2)(x-3)}{6}(6)$$

l= 10 + (x – 1) + 6(x – 1) (x – 2) + (x – 1) (x – 2) (x – 3)

which on simplification gives

$f(x) = x^3 + 2x + 7$
\ $f(1.5) = (1.5)^3 + 2(1.5) + 7$
= 3.375 + 3 + 7 = 13.375

Note:
If we want only the value of (1.5) and the interpolation polynomial is not needed, we can use the formula (26). In this case,

$$s = \frac{x - x_0}{h} = \frac{1.5 - 1}{1} = 0.5$$

and

$$f(1.5) = 10 + (0.5)(9) + \frac{(0.5)(-0.5)}{2}(12) + \frac{(0.5)(-0.5)(-1.5)}{6}(6)$$

= 10 + 4.5 – 1.5 + 0.375

= 13.375.

Example 5:

From the following table, find the number of students who obtained less than 45 marks.

Marks	30 - 40	40 - 50	50 – 60	60 - 70	70 - 80
No. of students	31	42	51	35	31

Solution:

We form a table of the number of students' f(x) whose marks are less than x. In other words, we form a cumulative frequency table.

Table 6: Frequency Table

x	f(x)	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$
40	31				
		42			
50	73		9		
		51		-25	
60	124		-16		37
		35		12	
70	159		-4		
		31			
80	190				

We have $x_0 = 40$, $x = 45$ and $h = 10$

$\therefore$ $s = 0.5$

$$\therefore f(45) \approx 31 + (0.5)(42) + \frac{(0.5)(-0.5)}{2}(9) + \frac{(0.5)(-0.5)(-1.5)}{6}(-25) + \frac{(0.5)(-0.5)(-1.5)(-2.5)}{24}(37)$$

$= 31 + 21 - 1.125 - 1.5625 - 1.4453$
$= 47.8672 \approx 48$

$\therefore$ The number of students who obtained less than 45 marks is approximately 48.

3.2.2 Newton's Backward-Difference Formula

Reordering the interpolating nodes as $x_n, x_{n-1}, ..., x_0$ and applying the Newton's divided difference form, we get

$$P_n(x) = f[x_n] + (x - x_n) f[x_{n-1}, x_n] + (x - x_{n-1}) f[x_{n-2}, x_{n-1}, x_n] + ... + (x - x) ... (x - x_n) f[x_0, ..., x_n] \quad (27)$$

We may also write

$$P_n(x) = P_n\left[x_n + \frac{x - x_n}{h} h\right]$$

$$= P_n[x_n + sh] = \sum_{i=0}^{n} (x - x_n)(x - x_{n-1})\ldots(x - x_{n-i+1})f[x_n, \ldots, x_{n-1}]$$

$$= \sum_{i=0}^{n} \frac{1}{i!h^i}(x - x_n)(x - x_{n-1})\ldots(x - x_{n-i+1})\nabla^i f_n \qquad (28)$$

Set $x = x_n + sh$, then

$x - x_i = x_n + sh - [x_n - (n - i)h] = (s + n - i)h$

$x - x_{n-j} = (s + n - n + j)h = (s + j)h$

and

$(x - x_n)(x - x_{n-1}) \ldots (x - x_{n-i+1}) = s(s + 1) \ldots s(s + i - 1)h^i$

Equation (28) becomes

$$P_n(x) = \sum_{i=0}^{n} \frac{1}{i!!} s(s+1)\ldots\ldots(s+i-1) f_n$$

$$= f_n + s\nabla f_n + \frac{s(s+1)}{2!}\nabla^2 f_n + \frac{s(s+1)\ldots(s+n-1)}{n!}\nabla^n f_n \qquad (29)$$

We have seen already that

$$\begin{bmatrix} s \\ k \end{bmatrix} = (-1)^k \frac{s(s+1)\ldots(s+k-1)}{k!}$$

Hence, equation (29) ca be written as

$$P_n(x) = f(x_n) + (-1)\begin{bmatrix} s \\ 1 \end{bmatrix}\nabla f(x_n)\ (-1) + (-1)^2\begin{bmatrix} s \\ 2 \end{bmatrix}\nabla^2 f(x_n)\ (-1)^2$$

$$+ \ldots + (-1)^k\begin{bmatrix} s \\ k \end{bmatrix}\nabla^k f(x_n)$$

or

$$P_n(x) = \sum_{k=0}^{n} (-1)^k \begin{bmatrix} s \\ k \end{bmatrix}\nabla^k f(x_n) \qquad (30)$$

Equation (27), (28) or (29) is called the Newton’s backward-difference form.

In this case error is given by

$$E_n(x) = (-1)^{n+1} \frac{s(s+1)\ldots(s+n)}{(n+1)!} h^{n+1} f^{n+1}(x). \quad (31)$$

The backward-difference form is suitable for approximating the value of the function at x that lies towards the end of the table.

Example 6:

Find the Newton's backward differences interpolating polynomial for the data of Example 4.

Solution:

We form the table of backward differences of f(x).

Table 7: Backward Difference Table

x	f(x)	∇f	$\nabla^2 f$	$\nabla^3 f$
1	10			
		9		
2	19		12	
		21		6
3	40		18	
		39		6
4	79		24	
		63		6
5	142		30	
		93		
6	235			

Tables 5 and 7 are the same except that we consider the differences of Table 7 as backward differences. If we name the abscissas as $x_0, x_1, \ldots, x_5$, then $x_n = x_5 = 6$, $f_n = f_5 = 235$. with h = 1, the Newton's backward differences polynomial for the given data is given by

$$P(x) = f_5 + (x - x_5)\nabla f_5 + \frac{(x-x_5)(x-x_4)}{2!}\nabla^2 f_5 + \frac{(x-x_5)(x-x_4)(x-x_3)}{3!}\nabla^3 f_5$$

$$= 235 + (x-6)(93) + \frac{(x-6)(x-5)}{2}(30) + \frac{(x-6)(x-5)(x-4)}{6}(6)$$

$$= 235 + 93(x-6) + 15(x-6) + (x-4)(x-5)(x-6)$$

which on simplification gives

$P(x) = x^3 + 2x + 7,$

which is the same as the Newton's forward differences interpolation polynomial in Example 4.

Example 7:

Estimate the value of (1.45) from the data given below:

x	1.1	1.2	1.3	1.4	1.5
f(x)	1.3357	1.5095	1.6984	1.9043	2.1293

Solution:

We form the backward differences table for the data given.

Table 8: Backward Differences Table

x	f(x)	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$
1.1	1.3357				
		0.1738			
1.2	1.5095		0.0151		
		0.1889		0.0019	
1.3	1.6984		0.0170		0.0002
		0.2059		0.0021	
1.4	1.9043		0.0191		
		0.2250			
1.5	2.1293				

Here $x_n = 1.5$, $x = 1.45$, $h = 0.1$

Hence, $s = \dfrac{x - x_n}{h} = \dfrac{1.45 - 1.5}{0.5} = -0.5$

The Newton's backward differences interpolation formula gives

$$f(x) = f_n + s\nabla f_n + \frac{s(x+1)}{2!}\nabla^2 f_n + \frac{s(s+1)(s+2)}{3!}\nabla^3 f_n + \frac{s(s+1)(s+2)(s+3)}{4!}\nabla^4 f_n$$

$$= 2.1293 + (-0.5)(0.2250) + \frac{(-0.5)(0.5)}{2}(0.0191) + \frac{(-0.5)(0.5)(1.5)}{6}(0.0021) + \frac{(-0.5)(0.5)(2.5)}{24}(0.0002)$$

= 2.1293 – 0.1125 – 0.00239 – 0.00013 – 0.0000078

= 2.01427 » 2.0143

3.3.3 Stirling's Central Difference Form

A number of central difference formulas are available which can be used according to a situation to maximum advantage. But we shall consider only one such method known as Stirling's method. This formula is used whenever interpolation is required of x near the middle of the table of values.

For the central difference formulas, the origin x_0, is chosen near the point being approximated and points below x_0 are labelled as x_1, x_2, ... and those directly above as x_{-1}, x_{-2}, ... (as in Table 3). Using this convention, Stirling's formula for interpolation is given by

$$P_n(x) = f(x_0) + \frac{s}{2}[df_{1/2} + df_{-1/2}] + \frac{s^2}{2!}d^2f_0$$

$$+ \frac{s(s^2 - 1^2)}{3!}\,\frac{1}{2}[d^3f_{1/2} + d^3f_{-1/2}] + ...$$

$$+ \frac{s(s^2 - 1^2)s(s^2 - 2^2)\ldots[s^2 - (p - 1)^2]}{(2p - 1)!}\,\frac{1}{2}[d^{2p-1}f_{1/2} + d^{2p-1}f_{-1/2}]$$

$$+ \frac{s(s^2 - 1^2)\ldots[s^2 - (p - 1)^2]}{(2p)!}d^{2p}f_0$$

$$+ \frac{s(s^2 - 1^2)\ldots s(s^2 - p^2)}{(2p + 1)!}\,\frac{1}{2}[d^{2p+1}f_{1/2} + d^{2p+1}f_{-1/2}] \qquad (32)$$

where $s = (x – x_0)/h$ and if $n = 2p + 1$ is odd.

If $n = 2p$ is even, then the same formula is used deleting the last term.

The Stirling's interpolation is used for calculation when x lies between $x_0 - \frac{1}{4}h$ and $x_0 + \frac{1}{4}h$.

It may be noted from the Table 3, that the odd order differences at $x_{-1/2}$ are those which lie along the horizontal line between x_0 and x_{-1}. Similarly, the odd order differences at $x_{1/2}$ are those which lie along the horizontal line between x_0 and x_1. even order differences at x_0 are those which lie along the horizontal line through x_0.

Example 8: Using Stirling's formula, find the value of (1.32) from the following table of values.

x	1.1	1.2	1.3	1.4	1.5
f(x)	1.3357	1.5095	1.6984	1.9043	2.1293

Solution:

Table 9: Central Difference

x	f(x)	df	d^2f	d^3f	d^4f
1.1	1.3357				
		0.1738			
1.2	1.5095		0.0151		
		0.1889		0.0019	
1.3	1.6984		0.0170		0.0002
		0.2059		0.0021	
1.4	1.9043		0.0191		
		0.2250			
1.5	2.1293				

Choose $x_0 = 1.3$

Therefore $\quad s = \dfrac{(x - x_0)}{h} = \dfrac{1.32 - 1.3}{0.1} = 0.2.$

From Eqn. (32), we have

$$f(x) \text{ » } f_0 + \frac{s}{2}[df_{-1/2} + df_{1/2}] + \frac{s^2}{2!} d^2f_0 + \frac{s(s^2 - 1^2)}{3!}\frac{1}{2}[d^3f_{-1/2} + d^3f_{1/2}] + \frac{s^2(s^2 - 1^2)}{4!} d^4f_0.$$

Now,

$$\frac{1}{2}[df_{-1/2} + df_{1/2}] = \frac{1}{2}(0.1889 + 0.2059) = 0.1974$$

$$\frac{1}{2}[d^3f_{-1/2} + d^3f_{1/2}] = \frac{1}{2}(0.0019 + 0.0021) = 0.0020$$

Also $d^2f_0 = 0.0170$, $d^4f_0 = 0.0002$.

Substituting in the above equation, we get

$$f(x) = 1.6984 + (0.2)(0.1974) + \frac{0.04}{2}(0.0170) + \frac{(0.2)(-0.96)}{6}(0.0020)$$

$$+\ \frac{(0.04)(-\ 0.96)}{24}\ (0.0002)$$

$= 1.6984 + 0.03948 + 0.00034 – 0.00006 – 0$

$= 1.73816$; 1.7382.

4.0 CONCLUSION

As in the summary.

5.0 SUMMARY.

In this unit, we have derived interpolation formulas for data with equally spaced values of the argument. We have seen how to find the value of f(x) for a given value of x by applying an appropriate interpolation formula derived in this section. The application of the formulas derived in this section is easier when compared to the application of the formulas derived in Units1 and 2. However, the formulas derived in this unit can only be applied to data with equally spaced arguments whereas the formulas derived in Units 1 and 2 can be applied for data with equally spaced or unequally spaced arguments. Thus, the formulas derived in Units 1 and 2 are of a more general nature than those of Unit 3. The interpolation polynomial which fits a given data can be determined by using any of the formulas derived in this section which will be unique whatever be the interpolation formula that is used.

The interpolation formulas derived in this unit are listed below:

1) Newton's forward difference formula:

$$P_n(x) = P_n(x_0 + sh) = \sum_{i=0}^{n} \begin{bmatrix} s \\ k \end{bmatrix} \nabla^i f_0$$

$$f_0 + s\,V\,f_0 + \frac{s(s-\ 1)}{2\,!}\,V^2 f_0 + \ldots + \frac{s(s-\ 1)\ldots s(s-\ n\ +\ 1)}{n\,!}\,V^n f_0$$

where $s = (x – x_0)/h$.

2) Newton's backward difference formula:

$$P_n(x) = P_n(x_n + sh) = \sum_{k=0}^{n} (-1)^k \begin{bmatrix} s \\ k \end{bmatrix} \nabla^k f_n \quad \text{where } s = (x – x_0)/h$$

3) Stirling's central difference formula:

$$P_n(x) = P_n(x_0 + sh) = f_0 + \frac{s}{2}[df_{1/2} + df_{-1/2}] + \frac{s^2}{2!}d^2f_0 + \frac{s(s^2 - 1^2)}{3!}\frac{1}{2}[d^3f_{1/2} + d^3f_{-1/2}] + \ldots + \frac{s^2(s^2 - 1^2)\ldots(s^2 - (p-1)^2)s^2f_0}{(2p)!}$$
$$+ \frac{s^2(s^2 - 1^2)\ldots(s^2 - p^2)}{(2p+1)!}[d^{2p+1}f_{1/2} + d^{2p+1}f_{-1/2}]$$

if $n = 2p + 1$ is odd. If $n = 2p$ is even, the same formula is used deleting the last term.

6.0 TUTOR-MARKED ASSIGNMENT.

1) Express $\nabla^4 f_5$ in terms of function values.

2) Show that $(E + 1)\ d = 2(E - 1)\ m$.

3) The population of a town in the decimal census was given below. Estimate population for the year 1915.

Year x	1911	1921	1931	1941	1951
Population: y (in thousands)	46	66	81	93	101

4) from the following table, find the value of y (0.23):

x	0 .20	0.22	0.24	0.26	0.28	0.30
y	1.6596	1.6698	1.6804	1.6912	1.7024	1.7139

5) Find the number of men getting wages between Rs. 10 and Rs. 15 from the following table.

Wages in Rs. x	0 - 10	10 - 20	20 - 30	30 -40
No. of men y	9	30	35	42

6) The area A of a circle of diameter d is given in the following table. Find the area of the circle when the diameter is 82 units.

d	80	85	90	95	100
A	5026	5674	6362	7088	7854

7) From the table of values of 3a, find the value of y when x = 0.29.

8) Using the backward differences interpolation, find the polynomial which agree with the values of $y(x)$ where

$y(0) = 1$, $y(1) = 0$, $y(2) = 1$ and $y(3) = 10$.

9) In 3c, find the number of candidates whose marks are less than or equal to (i) 70, (ii) 89.

10) Find f(1.725) from the following table.

x	1.5	1.6	1.7	1.8	1.9
f(x)	4.4817	4.9530	5.4739	6.0496	6.6859

11) Evaluate f(4.325) from the following.

x	4.1	4.2	4.3	4.4	4.5
f(x)	30.1784	33.3507	36.8567	40.7316	45.0141

12) Find the approximate value of y(2.15) from the table

x	0	1	2	3	4
f(x)	6.9897	7.4036	7.7815	8.1281	8.4510

7.0 REFERENCES/FURTHER READINGS.

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

MODULE 2 SOLUTION OF LINEAR ALGEBRAIC EQUATIONS

UNIT 1 DIRECT METHOD

CONTENTS

Notations and Symbols

$A = [a_{ik}]$	Matrix with the elements a_{ik}
det A = |A|	Determinant of a square matrix A
¥	infinity
r	Rho
u	Nu
m	Mu
l	Lambda
||A||	Norm of a matrix A
i	Imaginary unit, $i^2 = -1$.

Also see the list given in Block 1.

1.0 INTRODUCTION

One of the commonly occurring problems in applied mathematics is finding one or more roots of an equation f(x) = 0. In most cases explicit solutions are not available and we are satisfied with being able to find one or more roots to a specified degree of accuracy. In Block 1, we have discussed various numerical methods for finding the roots of an equation f(x) = 0. There we have also discussed the convergence of these methods. Another important problem of applied mathematics is to find the solution of systems of linear equations arise in a large number of areas, both directly in modelling physical situations and indirectly in the numerical solution of other mathematical models. These applications occur in all areas of the physical, biological and engineering sciences. For instance, in physics, the problem of steady state temperature in a plate is reduced to solving linear equations.

Engineering problems such as determining the potential in certain electrical networks, stresses in a building frame, flow rates in a hydraulic system etc. are all reduced to solving a set of algebraic equations simultaneously. Linear algebraic systems are also involved in the optimization theory, least squares fitting of data, numerical solution of boundary value problems for ordinary and partial differential equations, statistical inference etc. Hence, the numerical solution of systems linear algebraic equations plays a very important role.

Numerical methods for solving linear algebraic systems may be divided into two types, direct and iterative. Direct methods are those which, in the absence of round-off or other errors, yield the exact solution in a finite number of elementary arithmetic operations. Iterative methods start with an initial approximation.

To understand the numerical methods for solving linear system of equations it is necessary to have some knowledge of the properties of matrices. You might have already studied matrices, determinants and their properties in your linear algebra courses. However, we begin with a quick recall of few definitions here. In this unit, we have also discussed some direct methods for finding the solution of system of linear algebraic equations.

2.0 OBJECTIVES

After studying this unit, you should be able to:

- state the difference between the direct and iterative methods of solving the system of linear algebraic equations
- obtain the solution of system of linear algebraic equations by using the direct method
- use the pivoting technique while transforming the coefficient matrix to upper or lower triangular matrix.

3.0 MAIN CONTENTS

3.1 Preliminaries

As we have mentioned earlier, you might be already familiar with vectors, matrices, determinants and their properties (Ref. Linear algebra MTE-02). A rectangular array of (real or complex) numbers of the from

$$\begin{bmatrix} a_{11} & a_{12} \ldots\ldots & a_{2n} \\ a_{21} & a_{22} \ldots\ldots & a_{2n} \\ & & \\ a_{n1} & a_{n2} \ldots\ldots & a_{nn} \end{bmatrix}$$

is called a matrix. The numbers a_{11}, a_{12}, ..., a_{nn} are the elements of the matrix. The horizontal lines are called rows and the vertical lines called columns of the matrix. A matrix with m rows and n columns is called an $m \times n$ matrix (read as m by n matrix). We usually denote matrices by capital letters A, b etc., or by (a_{jk}), (b_{ik}) etc.

If the matrix has the same number of rows and columns, we call it a square matrix and the number of rows or columns is called its order. If a matrix has only one column it is a column matrix or column vector and if it has only one row it is a row matrix or row vector.

The matrices $A = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} = [a_{11}, a_{21}, \ldots a_{n1}]^T$ and

$B = [a_{11}, a_{12}, \ldots, a_{1n}]$ are respectively the column and row matrices. We give below some special square matrices $A = (a_{ij})$ of order n.

1) A matrix $A = (a_{ij})$ in which $a_{ij} = 0$ (i, j = 1, 2 ....., n) is called a null matrix and is denoted by 0.
   e.g.,
   $A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ is a 2×2 null matrix.

2) A matrix A in which all the non-diagonal elements vanish i.e., $a_{ij} = 0$ for $i \neq j$ is called a diagonal matrix.

E.g., A = $\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$

is a 3×3 diagonal matrix.

3) The identity matrix I is a diagonal matrix in which all the diagonal elements are equal to one. The identity matrix of order 4 is

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

4) A square matrix is lower triangular if all the elements above the main diagonal vanish i.e., $a_{ij} = 0$ for $j > i$. A lower triangular matrix of order 3 has the form

$$A = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Similarly upper triangular matrices are matrices in which,
$a_{ij} = 0$ for $i > j$.

e.g., $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$

Two matrices A = (a_{ij}) and B = (b_{ij}) are equal iff they have the same number of rows and columns and their corresponding elements are equal, that is $a_{ij} = b_{ij}$ for all i, j.

You must also be familiar with the addition and multiplication of matrices.

Addition of matrices is defined only for matrices of same order. The sum C = A + B of two matrices A and B, is obtained by adding the corresponding elements of A and B, i.e., $c_{ij} = a_{ij} + b_{ij}$.

For example, if A = $\begin{bmatrix} 4 & 6 & 3 \\ 0 & 1 & 2 \end{bmatrix}$ and B = $\begin{bmatrix} 5 & -1 & 0 \\ 3 & 1 & 0 \end{bmatrix}$ then

$$A + B = \begin{bmatrix} 1 & 5 & 3 \\ 3 & 2 & 2 \end{bmatrix}$$

Product of an m ´ n matrix A = (a_{ij}) and an n ´ p matrix B = (b_{ij}) is an m ´ p matrix C. C = AB, whose (i, k)th entry is

$$c_{ij} = \sum_{j=1}^{0} a_{ij} b_{ij} = a_{ij}\, b_{ij} + a_{i2}b_{i2} + ... + a_{in}\, b_{nk}$$

That is, to obtain the (i, k)th element of AB, take the ith row of A and kth column of B, multiply their corresponding elements and add up all these products. For example, if

$$A = \begin{bmatrix} 2 & 3 & -1 \\ 1 & 0 & 2 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \text{ then (1, 2)the element}$$

of AB is

$$[2 \;\; 3 \;\; -1] \begin{bmatrix} 1 \\ 4 \\ 2 \\ \; \end{bmatrix} = 2 * 1 + 3 * 4 + (-1) *2 = 12$$

Note that two matrices A and B can be multiplied only if the number of columns of A equals the number of rows of B. In the above example the product BA is not defined.

The matrix obtained by interchanging the rows and columns of A is called the transpose of A and is denoted by A^T

$$\text{If } A = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix} \text{ then } A^T = \begin{bmatrix} 2 & -1 \\ 3 & 1 \end{bmatrix}$$

Determinant is a number associated with square matrices.

$$\text{For a } 2 ´ 2 \text{ matrix } A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\det(A) = \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

For a 3 ´ 3 matrix A = $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$

det(A) = a_{11} det $\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}$ - a_{12} det $\begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix}$ + a_{13} det $\begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$

A determinant can be expanded about any row or column. The determinant of an n ´ n matrix A = (a_{ij}) is given by det(A) = $(-1)^{i+1}a_{ij}$ det(A_{ij}) + $(-1)^{i+2}a_{i2}$ det(A_{i2}) + ... + $(-1)^{i+n}a_{in}$ det(A_{in}), where the determinant is expanded about the ith row and A_{ij} is the (n – 1) ´ (n – 1) matrix obtained from A by deleting the ith row and jth column and i £ i £ n. Obviously, computation is simple if det(A) is expanded along a row or column that has maximum number of zeros. This reduces the number of terms to be computed.

The following example will help you to get used to calculating determinants.

Example 1:

If A = $\begin{bmatrix} 1 & 2 & 6 \\ 5 & 4 & 1 \\ 7 & 3 & 2 \end{bmatrix}$ calculated det (A).

Solution:

Let us expand by the first row. We have

$|A_{11}|$ = $\begin{bmatrix} 4 & 1 \\ 3 & 2 \end{bmatrix}$ = 4 * 2 – 1 * 3 = 5, $|A_{12}|$ = $\begin{bmatrix} 5 & 1 \\ 7 & 2 \end{bmatrix}$ = 5 * 2 = 7 * 1 = 3
,
$|A_{13}|$ = $\begin{bmatrix} 5 & 4 \\ 7 & 3 \end{bmatrix}$ = 5 * 3 – 4 * 7 = -13.

Thus,

$|A| = (-1)^{1+1}*1*|A_{11}|+(-1)^{1+2}*2*|A_{12}|+(-1)^{1+3}*6*|A_{13}|$=5–6–78 = -79

If the determinant of a square matrix A has the value zero, then the matrix A is called a singular matrix, otherwise, A is called a nonsingular matrix.
We shall now give some more definitions.

Definition:

The inverse of an n ´ n nonsingular matrix A is an n ´ n matrix B having the property

A B = B A = i
where I is an identity matrix of order n ´ n.

the inverse matrix B if it exists, is denoted by A^{-1} and is unique.

Definition:

For a matrix A = (a_{ij}), the cofactor A_{ij} of the element a_{ij} is given by

$A_{ij} = (-1)^{i+j} M_{ij}$

where M_{ij} (minor) is the determinant of the matrix of order (n – 1) ´ (n – 1) obtained from A after deleting its ith row and the jth column.

Definition:

The matrix of cofactors associated with the n ´ n matrix A is an n ´ n matrix A^c obtained from A by replacing each element of A by its cofactor.

Definition:

The transpose of the cofactor matrix A^c of A is called the adjoint of A and is written as adj(A). Thus

$\text{adj}(A) = (A^c)^T$

Let us now consider a system of n linear algebraic equations in n unknowns

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

where the coefficients a_{ij} and the constant b_i (i = 1, ..., n) are real and known. This system of equations in matrix from may be written as

$$A\,x = b \qquad (2)$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ & & & \\ & & & \\ a_{n1} & a_{n2} & & a_{nn} \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \\ \\ x_n \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \\ \\ b_n \end{bmatrix}$$

A is called the coefficient matrix and has real elements.

Our problem is to find the values x_i, i = 1, 2 ...., n if they exist, satisfying Eqn. (2). Before we discuss some methods of solving the system (2), we give the following definitions.

Definition:

A system of linear Eqns. (2) is said to be consistent if it has at least one solution. If no solution exists, then the system is said to be inconsistent.

Definition:

The system of Eqns. (2) is said to be homogeneous if b = 0, that is, all the elements b_i, b_2, ..., b_n are zero, otherwise the system is called non-homogeneous.

In this unit, we shall consider only non-homogeneous systems.

You also know from you linear algebra that the non-homogeneous system of Eqns. (2) has a unique solution, if the matrix A is nonsingular. You may recall the following basic theorem on the solvability of linear systems (Ref. Theorem 4, Sec. 5.0, Unit 1, Block 3, Module 1).

Theorem 1:

A non-homogeneous system of n linear equations in n known has a unique solution if and only if the coefficient matrix A is nonsingular.

If A is nonsingular, thenA^{-1} exists, and the solution of system (2) can be expressed as

$x = A^{-1}b$.

In case the matrix A is singular, then the system (2) has no solution if $b \neq 0$ or has an infinite number of solutions if b = 0. here we assume that A is a nonsingular matrix.

As we have already mentioned in the introduction, the methods of solution of the system (2) may be classified into two types:

i) Direct Methods: which in the absence of round-off errors give the exact solution in a finite number of steps.

ii) Iterative Methods: Starting with an approximate solution vector $x^{(0)}$, these methods generates a sequence of approximate solution vectors $\{x^{(k)}\}$ which converge to the exact solution vector x as the number of iterations $k \to \infty$. Thus iterative methods are infinite processes. Since we perform only a finite number of iterations, these methods can only find some approximation to the solution vector x. We shall discuss iterative methods later in Units 4 and 5.

In this unit we shall discuss only the direct methods. You are familiar with one such method due to the mathematician Cramer and known as Cramer's Rule. Let us briefly review it.

3.2 Cramer's Rule

In the system (2), let $d = \det(A) \neq 0$ and $b \neq 0$. Then the solution of the system is obtained as

$$x_i = d_i/d, \ i = 1, 2, ..., n \qquad (3)$$

where d_i is the determinant of the matrix obtained from A by replacing the ith column of A by the column vector b. let us illustrate the method through an example.

Example 2:

Solve the system of equations.

$3x_1 + x_2 + 2x_3 = 3$
$2x_1 - 3x_2 - x_3 = -3$
$x_1 - 2x_2 - x_3 = 4$
using Cramer's rule.

Solution: We have,

$$d = |A| = \begin{vmatrix} 3 & 1 & 2 \\ 1 & -3 & -1 \\ 1 & 2 & 1 \end{vmatrix} = 8$$

$$d_1 = \begin{vmatrix} 3 & 1 & 2 \\ -3 & -3 & -1 \\ 4 & 2 & 1 \end{vmatrix}$$

= 8 (first column in A is replaced by the column vector b)

$$d_2 = \begin{vmatrix} 3 & 3 & 2 \\ 2 & -3 & -1 \\ 1 & 4 & 1 \end{vmatrix}$$

= 16 (second column in A is replaced by the column vector b

$$d_3 = \begin{vmatrix} 3 & 1 & 3 \\ 2 & -3 & -3 \\ 1 & 2 & 4 \end{vmatrix}$$

= -8 (third column in A is replaced by the column vector b)

Using (3), we get the solution

$x_1 = d_1/d = 1$; $x_2 = d_2/d = 2$; $x_3 = d_3/d = -1$

While going through the example and attempting the self assessment exercises you must have observed that in Cramer's methods we need to evaluate n + 1 determinants each of order n, where n is the number of equations. If the number of operations required to evaluate a determinant is measured in terms of multiplications only, then to evaluate a determinant of second order, i.e.,

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

we need two multiplications or (2 – 1) 2! multiplications. To evaluate a determinant of third order

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = (a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31})$$

we need 12 multiplication or (3 – 1)3! multiplications. In general, to evaluate a determinant of nth order we need (n – 1)n! multiplications.

Also for a system of n equations, Cramer's rule requires n + 1determinants each of order n and performs n divisions to obtain x_i, i = 1, 2, ..., n. Thus the total number of multiplications and divisions needed to solve a system of n equations, using Cramer's rule becomes

M = total number of multiplications + total number of divisions
= (n + 1) (n - 1)n! + n

In Table 1, we have given the values of M for different values of n.

Table 1

Number of equations N	Number of operations n
2	8
3	51
4	364
5	2885
6	25206
7	241927
8	2540168
9	29030409
10	359251210

From the table, you will observe that as n increases, the number of operations required for Cramer's rule increases very rapidly. For this reason, Cramer's rule is not generally used for n > 4. hence for solving large systems, we need more efficient methods. In the next section we describe some direct methods which depend on the form of the coefficient matrix.

3.3 Direct Methods for Special Matrices

We now discuss three special forms of matrix A in Eqn. (2) for which the solution vector x can be obtained directly.

Case 1:

A = D, where D is diagonal matrix. In this case the systems of Eqns. (2) are of the form

$$\begin{array}{llllll} a_{11}x_1 & \ldots\ldots\ldots & & & = & b_1 \\ . & a_{22}x_2 & & . & = & b_2 \\ . & . & & . & = & . \\ . & & . & . & & . \\ . & & . & . & & . \\ . & & & a_{nn}x_n & = & b_n \end{array}$$

and det (A) – $a_{11}\ a_{22}\ ....\ a_{nn}$

Since the matrix A is nonsingular, a_{11} ¹ 0 for 1, 2, ..., n and we obtain the solution as

$x_i = b_i/a_{ii}$, i = 1, 2, ..., n.

Note that in this case we need only n divisions to obtain the solution vector.

Case 2 :

A = L, where L is a lower triangular matrix ($a_{ij} = 0$, j > i). The system of Eqns. (2) is now of the form

$$
\begin{array}{ll}
a_{11}x_1 & = b_1 \\
a_{21}x_1 + a_{22}x_2 & = b_2 \\
a_{31}x_1 + a_{32}x_2 + a_{33}x_3 & = b_3 \\
. & \\
. & \\
. & \\
a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + ... + a_{nn}x_n & = b_n
\end{array}
\qquad (4)
$$

and det (A) = $a_{11}a_{22}...a_{nn}$.

You may notice here that the first equation of the system (4) contains only x_1, the second equation contains only x_1 and x_2 and so on. Hence, we find x_1 from the first equation, x_2 from the second equation and proceed in that order till we get x_n from the last equation.

Since the coefficient matrix A is nonsingular, a_{11} ¹ 0, i = 1, 2, ..., n. we thus obtain

$x_1 = b_1/a_{11}$

$x_2 = (b_2 - a_{21}x_1)/a_{22}$

$x_3 = (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33}$

.

.

.

$$x_n = (b_n - \sum_{j=1}^{n-1} a_{ij}\ x_j)/a_{nn}$$

In general, we have for any i

$$x_i = (b_i - \sum_{j=1}^{n-1} \left(a_{ij}x_j\right)/a_{ii} \quad i = 1, 2, ..., n. \qquad (5)$$

For example, consider the system of equations

$$\begin{aligned} 5x_1 \quad &= 5 \\ -x_1 - 2x_2 \quad &= -7 \\ -x_1 + 3x_2 + 2x_3 &= 5 \end{aligned}$$

From the first equation we have,

$$x_1 = 1$$

From the second equation we get,

$$x_2 = \frac{-7 + x_1}{-2} = 3$$

and from the third equation we have,

$$x_3 = \frac{5 + x_1 - 3x_2}{2} = -\frac{3}{2}.$$

Since the unknowns in this methods are obtained in the order $x_1, x_2, ..., x_n$, this method is called the forward substitution method.

The total number of multiplications and divisions needed to obtain the complete solution vector x, using this method is

$$M = 1 + 2 + ... + n = n(n + 1)/2.$$

Case 3:

A = U, where U is an upper triangular matrix ($a_{ij} = 0$, $j < 1$). The system (2) is now of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n &= b_1 \\ a_{22}x_2 + a_{23}x_3 + ... + a_{2n}x_n &= b_2 \\ a_{33}x_3 + ... + a_{3n}x_n &= b_3 \qquad (6) \end{aligned}$$

$$\begin{aligned} a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n &= b_{n-1} \\ a_{nn} \quad x_n &= b_n \end{aligned}$$

and det (A) = $a_{11}a_{22}...a_{nn}$.

You may notice here that the nth (last) equation contains only x_n, the (n – 1)th equation contains x_n and x_{n-1} and so on. We can obtain x_n from the nth equation, x_{n-1} from the (n – 1)th equation and proceed in that order till we get x_1 from the first equation. Since the coefficient matrix A is nonsingular, $a_{ii} \neq 0$, i = 1, 2, ..., n and we obtain

$x_n = b_n/a_{nn}$
$x_{n-1} = (b_{n-1} – a_{n-1,n}x_n)/a_{n-1,n-1}$

$$x_1 = (b_i - \sum_{j=2}^{n} a_{ij}\, x_j)/a_{11}$$

or in general

$$x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}\, x_j)/a_{ii} \quad i = 1, 2, ..., n \tag{7}$$

Since the unknowns in this method are determined in the order x_n, x_{n-1}, ..., x_1, this method is called the back substitution method. The total number pf multiplications and divisions needed to obtain the complete solution vector x using this method is again n(n + 1)/2.

Let us consider the following example.

Example 3:

Solve the linear system of equations

$$\begin{aligned} 2x_1 + 3x_2 – x_3 &= 5 \\ -2x_2 – x_3 &= -7 \\ -5x_3 &= -15 \end{aligned}$$

Solution:

From the last equation, we have

$x_3 = 3$.

From the second equation, we have

$$x_2 = \frac{b_2 - a_{23}x_3}{a_{22}} = \frac{(-7 + 3)}{(-2)} = 2.$$

Hence from the first equation, we get

$$x_1 = \frac{b_1 - a_{12}x_2 - a_{13}x_3}{a_{11}} = \frac{(5 - 3.2 + 3)}{2} = 1$$

In the above discussion you have observed that the system of Eqns. (2) can be easily solved if the coefficient matrix A in Eqns. (2) has one of the three forms D, L or U or if it can be transformed to one of these forms. Now, you would like to know how to reduce the given matrix A into one of these three forms? One such method which transforms the matrix A to the form U is the Gauss elimination method which we shall describe in the next section.

3.4 Gauss Elimination Method

Gauss elimination is one of the oldest and most frequently used methods for solving systems of algebraic equations. It is attributed to the famous German mathematician, Carl Fredrick Gauss (1777 – 1855). This method is the generalization of the familiar method of eliminating one unknown between a pair of simultaneous linear equations. You must have learnt this method in your linear algebra course (MTH 122). In this method the matrix A is reduced to the form U by using the elementary row operations which include:

i) interchanging any two rows

ii) multiplying (or dividing) any row by a non-zero constant

iii) adding (or subtracting) a constant multiple of one row to another row.

The operation $R_i + mR_j$ is an elementary row operation, that means, add to the elements of the ith row m times the corresponding elements of the jth row. The elements in the jth row remain unchanged.

If any matrix A is transformed into another matrix B by a series of elementary row operations, we say that A and B are equivalent matrices. Consequently, we have the following definition.

To understand the Gauss elimination method let us consider a system of three equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \qquad (8)$$

Let $a_{11} \neq 0$. In the first stage of elimination we multiply the first equation in Eqns. (8) by $m_{21} = (-a_{21}/a_{11})$ and add to the second equation. Then multiply the first equation by $m_{31} = (-a_{31}/a_{11})$ and add to the third equation. This eliminates x_1 from the second and third equations. The new system called the first derived system then becomes

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1$$

$$a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 = b_2^{(1)} \qquad (9)$$

$$a_{32}^{(1)} x_2 + a_{33}^{(1)} x_3 = b_3^{(1)}$$

where,

$$a_{22}^{(1)} = a_{22} - \frac{a_{21}}{a_{11}} a_{12}$$
$$a_{23}^{(1)} = a_{23} - \frac{a_{21}}{a_{11}} a_{13}$$
$$b_2^{(1)} = b_2 - \frac{a_{21}}{a_{11}} b_1$$
$$a_{32}^{(1)} = a_{32} - \frac{a_{31}}{a_{11}} a_{12}$$
$$a_{33}^{(1)} = a_{33} - \frac{a_{31}}{a_{11}} a_{13}$$
$$b_3^{(1)} = b_3 - \frac{a_{31}}{a_{11}} b_1$$

In the second stage of elimination we multiply the second equation in (9) by $m_{32} = (-a_{32}^{(1)}/a_{22}^{(1)})$, $a_{22}^{(1)} \neq 0$ and add to the third equation. This eliminates x_2 from the third equation. The new system called the second derived system becomes

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 &= b_2^{(1)} \\ a_{33}^{(2)} x_3 &= b_3^{(2)} \end{aligned} \qquad (11)$$

where

$$a_{33}^{(2)} = a_{33}^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}} a_{23}^{(1)}$$
$$b_3^{(2)} = b_3^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}} b_2^{(1)} \qquad (12)$$

You may note here that the system of Eqns. (11) is an upper triangular system of the form (6) and can be solved using the back substitution provided method $a_{33}^{(2)} \neq 0$.

Let us illustrate the method through an example.

Example 4:

Solve the following linear system

$$\begin{aligned} 2x_1 + 3x_2 - x_3 &= 5 \\ 4x_1 + 4x_2 - 3x_3 &= 3 \\ -2x_1 + 3x_2 - x_3 &= 1 \end{aligned} \qquad (13)$$

using Gauss elimination method.

Solution:

To eliminate x_1 from the second and third equations of the system (13) add $\frac{-4}{2}$ = -2 times the first equation to the second equation and add -(-2)/2 = 1 times the first equation to the third equation. We obtain the new system as

$$\begin{aligned} 2x_1 + 3x_2 - x_3 &= 5 \\ -2x_2 - x_3 &= -7 \\ 6x_2 - 2x_3 &= 6 \end{aligned} \qquad (14)$$

In the second stage, we eliminate x_2 from the third equation of system (14). Adding -6/(-2) = 3 times the second equation to the third equation, we get

$$\begin{aligned} 2x_1 + 3x_2 - x_3 &= 5 \\ -2x_2 - x_3 &= -7 \\ -5x_3 &= -15 \end{aligned} \qquad (15)$$

System (15) is in upper triangular form and its solution is

$x_3 = 3$, $x_2 = 2$, $x_1 = 1$.

You may observe that we can write the above procedure more conveniently in matrix form. Since the arithmetic operations we have performed here affect only the elements of the matrix A and the vector b, we consider the augmented matrix i., [A|b] (matrix A augmented by the vector b) and perform the elementary now operations on the augmented matrix.

$$[A|b] = \left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{array}\right] R_2 - \frac{a_{21}}{a_{11}} R_1,\ R_3 - \frac{a_{31}}{a_{11}} R_1$$

$$\sim \left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ & a_{22}^{(1)} & a_{32}^{(1)} & b_2^{(1)} \\ & a_{32}^{(1)} & a_{33} & b_3 \end{array}\right] \quad R_3 - \frac{a_{32}^{(1)}}{a_{22}^{(1)}} R_2$$

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

which is in the desired from where, $a_{22}^{(1)}$, $a_{23}^{(1)}$, $a_{32}^{(1)}$, $a_{33}^{(1)}$, $b_2^{(1)}$, $b_3^{(1)}$, $a_{33}^{(2)}$, $a_3^{(2)}$ are given by Eqns. (10) and (12).

Definition: The diagonal elements a_{11}, $a_{22}^{(1)}$ and $a_{33}^{(2)}$ which are used as divisors are called pivots.

You might have observed here that for a linear system of order 3, the elimination was performed in 3 – 1 = 2 stages. In general for a system of n equations given by Eqns. (2) the elimination is performed in (n – 1) stages. At the ith stage of elimination, we eliminate x_i, starting from (i + 1)th row up to the nth row. Sometimes, it may happen that the elimination process stops in less than (n – 1) stages. But this is possible only when no equations containing the unknowns are left or when the coefficients of all the unknowns in remaining equations become zero. Thus if the process stops at the rth stage of elimination then we get a derived system of the form

$$\begin{aligned} & a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ & a_{22}^{(1)} x_2 + \ldots + a_{2n}^{(1)} x_n = b_2^{(1)} \\ & . \\ & . \\ & . \\ & a_{rr}^{(r-1)} x_r + \ldots + a_{rn}^{(r-1)} x_n = b_r^{(r-1)} \\ & \qquad\qquad 0 = b_{r+1}^{(r-1)} \\ & \qquad\qquad . \quad\; . \\ & \qquad\qquad . \quad\; . \\ & \qquad\qquad . \quad\; . \\ & \qquad\qquad 0 = b_n^{(r-1)} \end{aligned} \tag{16}$$

Where $r \leq n$ and $a_{11} \neq 0$, $a_{22}^{(1)} \neq 0$, ..., $a_{rr}^{(r-1)} \neq 0$.

In the solution of system of linear equations we can thus expect two different situations

1) $r = n$
2) $r < n$.

Let us now illustrate these situations through examples.

Example 5:

Solve the system of equations

$$4x_1 + x_2 + x_3 = 4$$
$$x_1 + 4x_2 – 2x_3 = 4$$
$$-x_1 + 2x_2 – 4x_3 = 2$$

using Gauss elimination method

Solution:

Here we have

$$[A|b] = \left[\begin{array}{ccc|c} 4 & 1 & 1 & 4 \\ 1 & 4 & -2 & 4 \\ 1 & 2 & -4 & 2 \end{array}\right] R_2 - \frac{1}{4} R_1, R_3 + \frac{1}{4} R_1$$

$$= \left[\begin{array}{ccc|c} 4 & 1 & 1 & 4 \\ 0 & 15/4 & 9/4 & 3 \\ 0 & 9/4 & 15/4 & 3 \end{array}\right] R_3 - \frac{3}{5} R_2$$

$$= \left[\begin{array}{ccc|c} 4 & 1 & 1 & 4 \\ 0 & 15/4 & -9/4 & 3 \\ 0 & 0 & -12/5 & 6/5 \end{array}\right]$$

using back substitution method, we get
$x_3 = -1/2$; $x_2 = 1/2$; $x_1 = 1$

Also, $\det (A) = 4 * \frac{15}{4} * \frac{(-\ 12)}{5} = -36$

Thus in this case we observe that $r = n = 3$ and the given system of equations has a unique solution. Also the coefficient matrix A in this case is nonsingular. Let us look at another example.

Example 6:

Solve the system of equations

$$3x_1 + 2x_2 + x_3 = 3$$
$$2x_1 + x_2 + x_3 = 0$$
$$6x_1 + 2x_2 + 4x_3 = 6$$

using Gauss elimination method. Does the solution exist?

Solution: We have

$$[A|b] = \left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 2 & 1 & 1 & 0 \\ 6 & 2 & 4 & 6 \end{array}\right] \; R_2 - \frac{2}{3}R_1,\; R_3 - 2R_1$$

$$= \left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -1/3 & 1/3 & -2 \\ 0 & -2 & 2 & 0 \end{array}\right] \; R_3 - 6R_2$$

$$= \left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -1/3 & 1/3 & -2 \\ 0 & 0 & 0 & 12 \end{array}\right]$$

In this case you can see that $r < n$ and elements b_1, $b_2^{(1)}$ and $b_3^{(2)}$ are all non-zero.

Since we cannot determine x_3 from the last equation, the system has no solution. In such a situation we say that the equations are inconsistent. Also note that det (A) = 0 i.e., the coefficient matrix is singular.

We now consider a situation in which not all b's are non-zero.

Example 7: Solve the system of equations

$$16x_1 + 22x_2 + 4x_3 = -2$$
$$4x_1 - 3x_2 + 2x_3 = 9$$
$$12x_1 + 25x_2 + 2x_3 = -11$$

using gauss elimination method.

Solution:

In this case we have

$$[A|b] = \left[\begin{array}{ccc|c} 6 & 22 & 4 & -2 \\ 4 & -3 & 2 & 9 \\ 12 & 25 & 2 & -11 \end{array}\right] R_2 - \frac{1}{4} R_1,\ R_3 - \frac{3}{4} R_1$$

$$= \left[\begin{array}{ccc|c} 6 & 22 & 4 & -2 \\ 0 & -17/2 & 1 & 19/2 \\ 0 & 17/2 & -1 & -19/2 \end{array}\right] R_3 + R_2$$

$$= \left[\begin{array}{ccc|c} 6 & 22 & 4 & -2 \\ 0 & -17/2 & 1 & 19/2 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Now in this case r < n and elements b_1, $b_2^{(1)}$ are non-zero, but $b_3^{(2)}$ is zero. Also the last equation is satisfied for any value of x_3. Thus, we get

x_3 = any value

$$x_2 = -\frac{2}{17}\left(\frac{19}{2} - x_3\right)$$

$$x_1 = \frac{1}{16}(-2 - 22x_2 - 4x_3)$$

Hence the system of equations has infinitely many solutions.

Note that in this case also det(A) = 0.

The conclusions derived from Examples 4, 5 and 6 are true for any system of linear equations. We now summarize these conclusions as follows:

i) If r = n, then the system of Eqns. (2) has a unique solution which can be obtained using the back substitution method. Moreover, the coefficient matrix A in this case is nonsingular.

ii) If r < n and all the elements $b_{r+1}^{(r-1)}$, $b_{r+2}^{(r-1)}$, ..., $b_n^{(r-1)}$ are zero then the system has no solution. In this case we say that the system of equations inconsistent.

iii) If r < n and all the elements $b_{r+1}^{(r-1)}$, $b_{r+2}^{(r-1)}$, ..., $b_n^{(r-1)}$, if present, are zero, then the system has infinite number of solutions. In this case the system has only r linearly independent rows.

In both the cases (ii) and (iii), the matrix A is singular.

Now we estimate the number of operations (multiplication and division) in the Gauss elimination method for a system of n linear equations in n unknowns as follows:

No. of divisions
1st step of elimination $(n - 1)$ divisions
2nd step of elimination $(n - 2)$ divisions

$(n - 1)$th step of elimination 1 divisions
$\therefore$ Total number of divisions $= (n - 1) + (n - 2) + \ldots + 1$

$$= \sum (n - 1) = \frac{n(n - 1)}{2}$$

No. of multiplications
1st step of elimination $n(n - 1)$ multiplications
2nd step of elimination$(n - 1)(n - 2)$ multiplications
$(n - 1)$th step of elimination 2.1 multiplications
$\therefore$ Total number of multiplications $= n(n - 1) + (n - 1)(n - 1) + \ldots + 2.1$

$$\begin{aligned} &= \sum n(n - 1) \\ &= \sum n^2 - \sum n \\ &= \frac{n(n + 1)(2n + 1)}{6} - \frac{n(n + 1)}{2} \\ &= \frac{1}{3} n(n + 1)(n - 1) \end{aligned}$$

Also the back substitution adds n divisions (one division at each step) and the numbers of multiplications added are
$(n - 1)$th equation 1 multiplication
$(n - 2)$th equation 2 multiplication

1st equation $(n - 1)$ multiplication
$\therefore$ Total multiplications $= \sum (n - 1) = \frac{n(n - 1)}{2}$

Total operation added by back substitution $= \frac{n(n - 1)}{2} + n = \frac{n(n + 1)}{2}$

You can verify these results for $n = 3$ from Eqns. (9) and (11).

Thus to find the solution vector x using the Gauss elimination method, we need

$$M = \frac{n(n-1)}{2} + \frac{1}{3}n(n^2 - 1) + \frac{n}{2}(n+1)$$
$$= \frac{n}{6}[2n^2 + 6n - 2]$$
$$= \frac{n^3}{6} + n^2 - \frac{n}{3}$$

operations. For large n, we may say the total number of operations needed is $\frac{1}{3}n^3$ (approximately). Thus, we find that Gauss elimination method needs much lesser number of operations compared to the Cramer's rule.

It is clear from above that you can apply Gauss elimination method to a system of equations of any order. However, what happens if one of the diagonal elements i.e., the pivots in the triangularization process vanishes? Then the method will fail. In such situations we modify the Gauss elimination method and this procedure is called pivoting.

Pivoting

In the elimination procedure the pivots a_{11}, $a_{22}^{(1)}$, ..., $a_{nn}^{(n-1)}$ are used as divisors. If at any stage of the elimination one of these pivots say $a_{ii}^{(i-1)}$, ($a_{11}^{(0)} = a_{11}$), vanishes then the elimination procedure cannot be continued further (see Example 8). Also, it may happen that the pivot $a_{ii}^{(i-1)}$, though not zero, may be very small in magnitude compared to the remaining elements in the ith column. Using a small number as a divisor may lead to the growth of the round-off error. In such cases the multipliers (e.g. $\frac{-a_{i-1,i}^{(i-2)}}{a_{ii}^{(i-1)}}$, $\frac{-a_{i-2,i}^{(i-3)}}{a_{ii}^{(i-1)}}$) will be larger than one in magnitude. The use of large multiplier will lead to magnification of error both during the elimination phase and during the back substitution phase of the solution. To avoid this we rearrange the remaining rows (ith row upto nth row) so as to obtain a non-vanishing pivot or to make it the largest element in magnitude in that column. The strategy is called pivoting (see Example 9). The pivoting is of the two types; partial pivoting and complete pivoting.

Partial Pivoting

In the first stage of elimination, the first column is searched for the largest element in magnitude and this largest element is then brought at the position of the pivot by interchanging the first row with the row having the largest element in magnitude in the first column. In the second stage of elimination, the second column is searched for the largest element in magnitude among the (n – 1) elements leaving the first element and then this largest element in magnitude is brought at the position of the second pivot by interchanging the second row with the row having the largest element in the second column. This searching and interchanging of rows is repeated in all the n – 1 stages of the elimination. Thus we have the following algorithm to find the pivot.

For i = 1, 2, ..., n, find j such that

$$\left|a_{ji}^{(i-1)}\right| = \max_{k} \left|a_{ki}^{(i-1)}\right|, \; i \leq k \leq n,$$

and interchange rows i and j.

Complete Pivoting

In the first stage of elimination, we search the entire matrix A for the largest element in magnitude and bring it at the position of the pivot. In the second stage of elimination we search the square matrix of order n – 1 (leaving the first row and the first column) for the largest element in magnitude and bring it to the position of second pivot and so on. This requires at every stage of elimination not only the interchanging of rows but also interchanging of columns. Complete pivoting is much more complicated and is not often used.

In this unit, by pivoting we shall mean only partial pivoting.

Let us now understand the pivoting procedure through examples.

Example 8:

Solve the system of equations

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

Using Gauss elimination method with partial pivoting.

Solution:

Let us first attempt to solve the system without pivoting. We have

$$[A|b] = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 3 & 3 & 4 & 20 \\ 2 & 1 & 3 & 13 \end{array}\right] \quad R_2 - 3R_1,\; R_3 - 2R_1$$

$$= \left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & 0 & 1 & 2 \\ 0 & -1 & 1 & 1 \end{array}\right]$$

Note that in the above matrix the second pivot has the value zero and the elimination procedure cannot be continued further unless, pivoting is used.

Let us now use the partial pivoting. In the first column 3 is the largest element. Interchanging the rows 1 and 2, we have

$$[A|b] = \left[\begin{array}{ccc|c} 3 & 3 & 4 & 20 \\ 1 & 1 & 1 & 6 \\ 2 & 1 & 3 & 13 \end{array}\right] R_2 - \frac{1}{3} R_1,\ R_3 - \frac{2}{3} R_1$$

$$= \left[\begin{array}{ccc|c} 3 & 3 & 4 & 20 \\ 0 & 0 & -1/3 & -2/3 \\ 0 & -1 & 1/3 & -1/3 \end{array}\right]$$

In the second column, 1 is the largest element in magnitude leaving the first element. Interchanging the second and third rows we have

$$[A|b] = \left[\begin{array}{ccc|c} 3 & 3 & 4 & 20 \\ 0 & -1 & 1/3 & -1/3 \\ 0 & 0 & -1/3 & -2/3 \end{array}\right]$$

You may observe here that the resultant matrix is in triangular form and no further elimination is required. Using back substitution method, we obtain the solution

$x_3 = 2$, $x_2 = 1$, $x_1 = 3$.

Let us consider another example.

Example 9:

Solve the system of equations

$$0.0003\, x_1 + 1.566\, x_2 = 1.569$$
$$0.3454\, x_1 - 0.436\, x_2 = 3.018 \tag{17}$$

using Gauss elimination method with and pivoting. Assume that the numbers in arithmetic calculations are rounded to four significant digits. The solution of the system (17) is $x_1 = 10$, $x_2 = 1$.

Solution:

Without Pivoting

$$m_{21} = -\frac{a_{21}}{a_{11}} = -\frac{0.3454}{0.0003} = -1151.0 \text{ (rounded to four places)}$$

$$\begin{aligned} a_{22}^{(1)} &= -0.436 - 1.566 \times 1151 \\ &= -0.436 - 1802.0 - 1802.436 \\ &= -1802.0 \end{aligned}$$

$$\begin{aligned} b_{2}^{(1)} &= 3.018 - 1.569 \times 1151.0 \\ &= 3.018 - 1806.0 \\ &= -1803.0 \end{aligned}$$

Thus, we get the system of equations

$$0.0003\, x_1 + 1.566\, x_2 = 1.569$$
$$-1802.0\, x_2 = -1803.0$$

which gives

$$x_2 = \frac{1803.0}{1802.0} = 1.001$$

$$x_1 = \frac{1.569 - 1.566 \times 1.001}{0.0003} = \frac{1.569 - 1.568}{0.0003}$$
$$= 3.333$$

which is highly inaccurate compared to the exact solution.

We interchange the first and second equations in (17) and get

$0.3454\, x_1 - 0.436\, x_2 = 3.018$

$0.0003\, x_1 + 1.566\, x_2 = 1.569$

we obtain

$$m_{21} = -\frac{a_{21}}{a_{11}} = -0.0009$$

$$\begin{aligned} a_{22}^{(1)} &= 1.566 - 0.0009 \times (0.436) \\ & 1.566 - 0.0004 \\ &= 1.566 \end{aligned}$$

$$\begin{aligned} b_{2}^{(1)} &= 1.569 - 3.018 \times (0.0009) \\ &= 1.569 - -.0027 \\ &= 1.566 \end{aligned}$$

Thus, we get the system of equations

$$0.3454\, x_1 - 0.436\, x_2 = 3.018$$
$$1.566\, x_2 = 1.566$$

which gives

$x_2 = i$

$$x_1 = \frac{3.018 + 0.436}{0.3454} = \frac{3.454}{0.3454} = 10$$

which is the exact solution.

We now make the following two remarks about pivoting.

Remark: If the matrix A is diagonally dominant i.e.,

$|a_{ii}| \geq \sum_{\substack{i=1 \\ j=1}}^{n} |a_{ii}|$, then no pivoting is needed. See Example 5 in which A is diagonally dominant.

Remark:

If exact arithmetic is used throughout the computation, pivoting is not necessary unless the pivot vanishes. However, if computation is carried up to a fixed number of digits, we get accurate results if pivoting is used.

There is another convenient way of carrying out the pivoting procedure. Instead of physically interchanging the equations all the time, the n original equations and the various changes made in them can be recorded in a systematic way. Here we use an $n \times (n + 1)$ working array or matrix which we call W and is same as our augmented matrix [A|b]. Whenever some unknown is eliminated from an equation, the changed coefficients and right side for this equation are calculated and stored in the working array W in place of the previous coefficients and right side. Also, we use an n-vector which we call $p = (p_i)$ to keep track of which equations have already been used as pivotal equation (and therefore should not be changed any further) and which equations are still to be modified. Initially, the ith entry p_i pf p contains the integer i, i = 1, ..., n and working array W is of the form

$$W = (w_{ij}) = \left[\begin{array}{cccc|c} a_{11} & a_{12} & & a_{1n} & b_1 \\ a_{21} & a_{22} & & a_{2n} & b_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & & \\ a_{n1} & a_{n2} & & a_{nn} & b_n \end{array}\right]$$

Further, one has to be careful in the selection of the pivotal equation for each step. For each step the pivotal equation must be selected on the basis of the current state of the system under consideration i.e. without foreknowledge of the effect of the i = 1, ..., n, where d_i is the number
$d_i = \max |a_{ij}|$
$1 \leq j \leq n$

At the beginning of say kth step of elimination, e pick as pivotal equation that one from the available n – k, which has the absolutely largest coefficient of x_k relative to the size of the equation. This means that the integer j is selected between k and n for which

$$\frac{|w_{p_{jk}}|}{d_{p_j}} \; 3 \; \frac{|w_{ik}|}{d_i}, \text{ " } i = p_k, \ldots\ldots, p_n$$

We can also store the multipliers in the working array W instead of storing zeros. That is, if p_i is the first pivotal equation and we use the multipliers $m_{pi,1}$, i = 2, ..., n to eliminate x_1 from the remaining (n – 1) positions of the first column then in the first column we can store the multipliers $m_{pi,1}$, i = 2, ..., n, instead of storing zeros.

Let us now solve the following system of linear equations by scaled partial pivoting by storing the multipliers and maintaining pivotal vector.

Example 10:

Solve the following system of linear equations with pivoting

$x_1 – x_2 + 3x_3 = 3$
$2x_1 + x_2 + 4x_3 = 7$
$3x_1 + 5x_2 – 2x_3 = 6$

Solution:

Here the working matrix is

$$W = \begin{bmatrix} 1 & -1 & 3 & 3 \\ 2 & 1 & 4 & 7 \\ 3 & 5 & -2 & 6 \end{bmatrix} \quad p = [p_1, p_2, p_3]^T = [1, 2, 3]^T$$

and $d_1 = 3$, $d_2 = 4$ and $d_3 = 5$.

Note that d's will not change in the successive steps.

Step 1: Now $\frac{|w_{p1,1}|}{d_1} = \frac{1}{3}\ \frac{|w_{p2,1}|}{d_2} = \frac{2}{4} = \frac{1}{2},\ \frac{|w_{p3,1}|}{d_3} = \frac{3}{5}$.

Since $\frac{3}{5} > \frac{1}{2}, \frac{1}{3}$,

Hence, $p_1 = 3$, $p_2 = 2$ and $p_3 = 1$.

We use the third equation to eliminate x_1 from first and second equations and store corresponding multipliers instead of storing zeros in the working matrix.

The multipliers are $m_{pi,1} = \frac{w_{p_{i,1}}}{w_{p_{1,1}}}$, i = 2, 3

Therefore, $m_{2,1} = \frac{w_{p_{2,1}}}{w_{p_{1,1}}} = \frac{w_{2,1}}{w_{3,1}} = \frac{2}{3}$

and $m_{1,1} = \frac{w_{p_{3,1}}}{w_{p_{1,1}}} = \frac{w_{1,1}}{w_{3,1}} = \frac{1}{3}$

After the first step the working matrix is transformed to

$$W^{(1)} = \begin{bmatrix} (1/3) & -8/3 & 11/3 & 1 \\ (2/3) & -7/3 & 16/3 & 3 \\ \lceil 3 \rceil & 5 & -2 & 6 \end{bmatrix} \quad p = (p_1, p_2, p_3)^T = (3, 2, 1)^T$$

Step 2: $\frac{|w_{p_{2,2}}|}{dp_2} = \frac{|w_{2,2}|}{d_2} = \frac{7/3}{4} = \frac{7}{12}$

$\frac{|w_{p_{3,2}}|}{dp_3} = \frac{|w_{1,2}|}{d_1} = \frac{8/3}{3} = \frac{8}{9}$

Now $\frac{8}{9} > \frac{7}{12}$ so that we have $p = (p_1, p_2, p_3)^T = (3, 2, 1)^T$.

Multiplier is $m_{p_{i,2}} = \frac{w_{p_{i,2}}}{w_{p_{2,2}}}$, i = 3

$$\Rightarrow \quad m_{p_{3,2}} = \frac{w_{p_{i,2}}}{w_{p_{2,2}}} = \frac{-7/3}{-8/3} = \frac{7}{8}.$$

That is, we use the first equation as pivotal equation to eliminate x_2 from second equation and also we store the multiplier. After the second step, we have the following working matrix.

$$W^{(2)} = \begin{bmatrix} \frac{1}{3} & -\frac{8}{3} & \frac{11}{3} & 1 \\ \frac{2}{3} & \frac{7}{8} & \frac{51}{24} & \frac{17}{8} \\ 3 & 5 & -2 & 6 \end{bmatrix} \quad p = [3, 1, 2]^T$$

In the working matrix the circled numbers denote multipliers and squared ones denote pivotal elements. Rearranging the equations (i.e., 3rd equation becomes the first equation, 1st becomes the 2nd and 2nd becomes the third) we get the reduced upper triangular system which can be solved by back substitution.

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

$$-\frac{8}{3}x_2 + \frac{11}{3}x_3 = 1$$

$$\frac{51}{24}x_3 = \frac{17}{8}$$

By back substitution, we get $x_1 = 1$, $x_2 = 1$ and $x_3 = 1$.

We now make the following two remarks.

Remark:

We do not interchange rows in Step 1 and 2, instead we maintain a pivotal vector and use it at the end to get upper triangular system.

Remark:

We store multipliers in the working matrix so that we can easily solve Ax = c, once we have solved Ax = b. This will be explained to you in detail in Unit 2 when we discuss the method of obtaining inverse of a matrix A.

We shall now describe the triangularization method which is also a direct method for the solution of system of equations.

In this method the matrix of coefficients of the linear system being solved is factored into the product of two triangular matrices. This method is frequently used to solve a large system of equations. We shall discuss the method in the next section.

3.5 LU Decomposition Method

Let us consider the system of Eqns. (2), where A is a non-singular matrix. We first write the matrix A as the product of a lower triangular matrix L and an upper triangular matrix U in the form

A = LU

or in matrix form we write (18)

$$\begin{bmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ & & & \\ a_{n1} & a_{n2} & & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & & 0 \\ l_{21} & l_{22} & & 0 \\ & & & \\ l_{n1} & l_{n2} & & l_{nn} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & & u_{1n} \\ 0 & u_{22} & & u_{2n} \\ & & & \\ 0 & 0 & & u_{nn} \end{bmatrix} \tag{19}$$

The left side matrix A has n^2 elements, whereas L and U have 1 + 2 + ... + n = n(n + 1)/2 elements each. Thus, we have n^2 + n unknowns in L and U which are to be determined. On comparing the corresponding elements on two sides in Eqn. (19), we get n^2 equations in n^2 + n unknowns and hence n unknowns are determined. Thus, we get a solution in terms of these n unknowns i.e., we get a n parameter family of solutions. In order to obtain a unique solution we either take all the diagonal elements of L as 1, or all the diagonal elements of U as 1.

For u_{ij} = 1, i = 1, 2, ..., n, the method is called the Crout LU decomposition method. For 1_{ii} = 1, i = 1, 2, ..., n we have Doolittle LU decomposition method. Usually Crout's LU decomposition method is used unless it is specifically mentioned. We shall now explain the method for n = 3 with u_{ii} = 1, i = 1, 2, 3. We have

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

or

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

On comparing the elements of the first column, we obtain
$1_{11} = a_{11}$, $1_{21} = a_{21}$, $1_{31} = a_{31}$ (20)
i.e., the first column of L is determined.

On comparing the remaining elements of the first row, we get
$1_{11}u_{12} = a_{12}$; $1_{11}u_{13} = a_{13}$
which gives
$u_{12} = a_{12}/1_{11}$; $u_{13} = a_{13}/1_{11}$ (21)

Hence the first row of U is determined

On comparing the elements of the second column, we get
$1_{21}u_{12} + 1_{22} = a_{22}$
$1_{31}u_{12} + 1_{32} = a_{32}$
which gives

$$\begin{bmatrix} l_{22} & = a_{22} - l_{21}u_{12} \\ l_{32} & = a_{32} - l_{31}u_{12} \end{bmatrix} \qquad (22)$$

Now the second column of L is determined.

On comparing the elements of the second row, we get
$1_{21}u_{13} + 1_{22}u_{23} = a_{23}$
which gives $u_{23} = (a_{23} - 1_{21}\, u_{13})/1_{22}$ (23)
and the second row of U is determined.

On comparing the elements of the third column, we get
$1_{31}u_{13} + 1_{32}u_{23} + 1_{33} = a_{33}$
which gives $1_{33} = a_{33} - 1_{31}u_{13} - 1_{32}u_{23}$ (24)
You must have observed that in this method, we alternate between getting a column of L and a row of U in that order. If instead of u_{ii} = 1 1, 2, ..., n, we

take 1_{ii} = 1, i = 1, 2, ..., n, then we alternative between getting a row of U and a column of L in that order.

Thus, it is clear from Eqns. (20) – (24) that we can determine all the elements of L and U provided the nonsingular matrix A is such that

$$a_{11} \ {}^{1} \ 0, \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} {}^{1} \ 0.$$

Similarly, for the general system of Eqns. (2), we obtain the elements of L and U using the relations

$$1_{ij} = a_{ij} - \sum_{i=1}^{j=1} 1_{ik}u_{kj},\ i \ {}^{3} \ j$$

$$u_{ij} = (a_{ij} - \sum_{i=1}^{j=1} 1_{ik}u_{kj})/1_{ii},\ i \ {}^{3} \ j$$

$u_{ii} = 1$

Also, det (A) = $1_{11}1_{22}$ ....., 1_{nn}.

Thus w can say that every nonsingular matrix A can be written as the product of a lower triangular matrix and an upper triangular matrix if all principal minors of A are nonsingular, i.e., if

$$a_{11} \ {}^{1} \ 0, \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} {}^{1} \ 0, \begin{matrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{matrix} {}^{1} \ 0, \ldots\ldots |A| \ {}^{1} \ 0.$$

Once we have obtained the elements of the matrices L and U, we write the system of equations

A x = b (25)

in the form

L U x = b (26)

The system (26) may be further written as the following two systems

U x = y (27)

L y = b (28)

Now, we first solve the system (28), i.e.,

L y = b,

using the forward substitution method to obtain the solution vector y. Then using this y, we solve the system (27), i.e.,

U x = y,

using the backward substitution method to obtain the solution vector x.

The number of operations for this method remains the same as that in the Gauss-elimination method.

We now illustrate this method through an example.

Example 11:

Use the LU decomposition method to solve the system of equations

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

Solution: Using $1_{ii} = 1$, i = 1, 2, 3, we have

$$\begin{bmatrix} 1 & 1 & 1 \\ 4 & 3 & -1 \\ 3 & 5 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{31} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

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

On comparing the elements of row and column alternatively, on both sides, we obtain

first row : $u_{11} = 1$, $u_{12} = 1$, $u_{13} = 1$
first column : $1_{21} = 4$, $1_{31} = 3$
second row : u_{22} - -1, $u_{23} = -5$
second column : $1_{32} = -2$
third row : $u_{33} = -10$

Thus, we have

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 3 & -2 & 1 \end{bmatrix} \quad U = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -5 \\ 0 & 0 & -10 \end{bmatrix}$$

Now from the system
L y = b

or

$$\begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 3 & -2 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 6 \\ 4 \end{bmatrix}$$

we get
$y_1 = 1$, $y_2 = 2$, $y_3 = 5$
and from the system
$U x = y$
or

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -5 \\ 0 & 0 & -10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

we get
$x_3 = -1/2$, $x_2 = 1/2$, $x_1 = 1$.

4.0 CONCLUSION

Same as in the summary.

5.0 SUMMARY

In this unit we have covered the following:

1) For a system of n equations
$Ax = b$ (see Eqn. (2))

in n unknowns, where A is n ´ n non-singular matrix, the methods of finding the solution vector x may be broadly classified into two types: (1) direct methods and (ii) iterative methods

2) Direct methods produce the exact solution in a finite number of steps provided there are no round-off errors. Cramer's rule is one such method. This method gives the solution vector as

$$x_i = \frac{d_i}{d} \quad i = 1, 2, ..., n$$

where $d = |A|$ and d_i is the determinant pf the matrix obtained from A by replacing the ith column of A by the column vector b. Total number of operations required for Cramer's rule in solving a system of n equations are

$$M = (n + 1)(n - 1)n! + n$$

Since the number M increases very rapidly, Cramer's rule is not used for n > 4.

3) For larger systems, direct methods becomes more efficient if the coefficient matrix A is in one of the forms D (diagonal), L (lower triangular) or U (upper triangular).

4) Gauss elimination method is another direct method for solving large systems (n > 4). In this method the coefficient matrix A is reduced to the form U by using the elementary row operations. The solution vector x is then obtained by using the back substitution method. For large n, the total numbers of operations required in Gauss elimination method are $\frac{1}{3}n^3$ (approximately).

5) In Gauss elimination method if at any stage of the elimination any of the pivots vanishes or become small in magnitude, elimination procedure cannot be continued further. In such cases pivoting is used to obtain the solution vector x.

6) Every non-singular matrix A can be written as the product of a lower triangular matrix and an upper triangular matrix, by the LU decomposition method, if all the principal minors of A are non-singular. Thus, LU decomposition method, which is a modification of the Gauss elimination method can be used to obtain the solution vector x.

6.0 TUTOR-MARKED ASSIGNMENT (TMA)

1) If $A = \begin{bmatrix} 3 & -2 & 0 & 2 \\ 2 & 1 & 0 & -1 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & -3 & 1 \end{bmatrix}$ calculate det (A).

2) Solve the system of equations

$3x_1 + 5x_2 = 8$
$-x_1 + 2x_2 - x_3 = 0$
$3x_1 - 6x_2 + 4x_3 = 1$

using Cramer's rule.

3) Solve the system of equations

$$\begin{aligned} x_1 + 2x_2 - 3x_3 + x_4 &= -5 \\ x_2 + 3x_3 + x_4 &= 6 \\ 2x_1 + 3x_2 + x_3 + x_4 &= 4 \\ x_1 \qquad + x_3 + x_4 &= 1 \end{aligned}$$

using Cramer's rule.

4) Solve the system of equations

$$\begin{aligned} x_1 \qquad &= 1 \\ 2x_1 = x_2 \qquad &= 1 \\ 3x_1 - x_2 - 2x_3 \qquad &= 0 \\ 4x_1 + x_2 - 3x_3 + x_4 \qquad &= 3 \\ 5x_1 - 2x_2 - x_3 - 2x_4 + x_5 \qquad &= 1 \end{aligned}$$

using forward substitution method.

5) Solve the system of equations

$$\begin{aligned} x_1 - 2x_2 + 3x_3 - 4x_4 + 5x_5 &= 3 \\ x_2 - 2x_3 + 3x_4 - 4x_5 &= -2 \\ x_3 - 2x_4 + 3x_5 &= 2 \\ x_4 - 2x_5 &= -1 \\ x_5 &= 1 \end{aligned}$$

using backward substitution method.

6) Use Gauss elimination method to solve the system of equations

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 3 \\ 3x_1 - 2x_2 - 4x_3 &= -2 \\ 2x_1 + 3x_2 - x_3 &= -6 \end{aligned}$$

7) Solve the system of equations

$$\begin{bmatrix} 1 & 2 & -3 & 1 \\ 0 & 1 & 3 & 1 \\ 2 & 3 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \\ 4 \\ 1 \end{bmatrix}$$

8) Use Gauss elimination method to solve the system of equations

$$\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ 1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

9) Solve the system of equations

0.729x + 0.81y + 0.9z = 0.6867
x + y + z = 0.8338
1.331x + 1.21y + 1.1z = 1.000

using gauss eliminating method with and without pivoting. Round off the numbers in arithmetic calculations to four significant digits. The exact solution of the system rounded to four significant digit is

x = 0.2245, y = 0.2814 z = 0.3279

10) Use the LU decomposition method with u_{ii} = 1, i = 1, 2, 3 to solve the system of equations given in Example 11.

11) Use the LU decomposition method with 1_{ii} = 1, i = 1, 2, 3 to solve the system of equations given in TMA Question 4 no. 1.

12) Use L U decomposition method to solve the system of equations given in TMA Question 4 no. 3.

7.0 REFERENCES/FURTHER READINGS.

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

UNIT 2 DIRECT METHOD

CONTENTS

1.0 INTRODUCTION

In the previous unit, you have studied the Gauss elimination and LU decomposition methods for solving systems of algebraic equations A x = , when A is a n ´ n nonsingular matrix. Matrix inversion is another problem associated with the problem of finding solutions of a linear system. If the inverse matrix A^{-1} of the coefficient matrix A is known then the solution vector x can be obtained from $x = A^{-1}b$. In genral, inversion of matrices for solving system of equations should be avoided whenever possible. This is because, it involves greater amount of work and also it is difficult to obtain the inverse accurately in many problems. However, there are two cases in which the explicit computation of the inverse is desirable. Firstly, when several systems equations, having the same coefficient matrix A but different right hand side b, have to b e solved. Then computations are reduced if we first find the inverse matrix and then find the solution. Secondly, when the elements of A^{-1} themselves have some special physical significance. For instance, in the statistical treatment of the fitting of a function to observational data by the method of least squares, the elements of A^{-1} give information about the kind and magnitude of errors in the data.

In this unit, we shall study a few important methods for finding the inverse of a nonsingular square matrix.

2.0 OBJECTIVES

After studying this unit, you should be able to:

- obtain the inverse by adjoint method for n < 4
- obtain the inverse by the Gauss-Jordan and LU decomposition methods

- obtain the solution of a system of linear equations using the inverse method.

3.0 MAIN CONTENTS

3.1 The Method of Adjoints

You already know that the transpose of the matrix of the cofactors of elements of A is called the adjoint matrix and is denoted by adj(A).

Formally, we have the following definition.

Definition:

The transpose of the cofactor matrix A^c of A is called the adjoint of A and is written a adj(A).

$$\text{adj}(A) = (A^c)^T$$

The inverse of a matrix can be calculated using the adjoint of a matrix.

E obtain the inverse matrix A^{-1} of A from

$$A^{-1} = \frac{1}{\det(A)}\,\text{adj}(A) \tag{1}$$

This method of finding the inverse of a matrix is called the method of adjoints.

Note that det(A) in Eqn. (1) must not be zero and therefore the matrix A must be nonsingular.

We shall not be going into the details of the method here. We shall only illustrate it through examples.

Example 1: Find A^{-1} for the matrix

$$A = \begin{bmatrix} 5 & 8 & 1 \\ 0 & 2 & 1 \\ 4 & 3 & -1 \end{bmatrix}$$

and solve the system of equations

$$A\,x = b \tag{2}$$

for

i) $b = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$ ii) $b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ iii) $b = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

Solution:

Since det(A) = -1 ¹ 0, the inverse of A exists. We obtain the cofactor matrix A^c from A by replacing each element of A by its cofactor as follows:

$$A^c = \begin{bmatrix} 5 & 4 & -8 \\ 11 & -9 & 17 \\ 6 & -5 & 10 \end{bmatrix}$$

\ $\text{adj}(A) = (A^c)^T = \begin{bmatrix} 5 & 11 & 6 \\ 4 & -9 & -5 \\ 8 & 17 & 10 \end{bmatrix}$

Now $A^{-1} = \dfrac{1}{\det(A)}\,\text{adj}(A)$

\ $A^{-1} = - = \begin{bmatrix} 5 & 11 & 6 \\ 4 & -9 & -5 \\ 8 & 17 & 10 \end{bmatrix} = \begin{bmatrix} 5 & -11 & -6 \\ 4 & 9 & 5 \\ 8 & -17 & -10 \end{bmatrix}$

Also the solution of the given system of equations are

i) $x = A^{-1}b = \begin{bmatrix} 5 & -11 & -6 \\ 4 & 9 & 5 \\ 8 & -17 & -10 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix}$

ii) $x = A^{-1}b = \begin{bmatrix} 5 & -11 & -6 \\ 4 & 9 & 5 \\ 8 & -17 & -10 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 4 \\ 8 \end{bmatrix}$

iii) $x = A^{-1}b = \begin{bmatrix} 5 & -11 & -6 \\ 4 & 9 & 5 \\ 8 & -17 & -10 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 12 \end{bmatrix}$

We now take up an example in which the given matrix A is lower triangular and we shall show that its inverse is also a lower triangular matrix.

Example 2: Find A^{-1} for the matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{bmatrix}$$

Solution:

We have

$\det(A) = 18 \neq 0$. Thus A^{-1} exists.

Now

$$A^c = \begin{bmatrix} 18 & -12 & -2 \\ 0 & 6 & -5 \\ 0 & 0 & 3 \end{bmatrix}$$

$$\therefore \quad A^{-1} = \frac{(A^c)^T}{\operatorname{adj}(A)} = \frac{1}{18} \begin{bmatrix} 18 & 0 & 0 \\ 12 & 6 & 0 \\ -2 & -5 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2/3 & 1/3 & 0 \\ 1/9 & -5/18 & 1/6 \end{bmatrix}$$

Thus, A^{-1} is again a lower triangular matrix. Similarly, we can illustrate that the inverse of an upper triangular matrix is again upper triangular.

Example 3:

Find A^{-1} for the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$$

Solution:

Since, $\det(A) = 24 \neq 0$, A^{-1} exists.

We obtain

$$A^c = \begin{bmatrix} 24 & 0 & 0 \\ 12 & 6 & 0 \\ -2 & -5 & 4 \end{bmatrix}$$

$$\therefore \quad A^{-1} = \frac{1}{24}\begin{bmatrix} 24 & -12 & -2 \\ 0 & 6 & -5 \\ 0 & 0 & 4 \end{bmatrix} = \begin{bmatrix} 1 & -1/2 & -1/12 \\ 0 & 1/4 & -5/24 \\ 0 & 0 & 1/6 \end{bmatrix}$$

which is again an upper triangular matrix.

The method of adjoints provides a systematic procedure to obtain the inverse of a given matrix and for solving systems of linear equations. To obtain the inverse of an $n \times n$ matrix, using this method, we need to evaluate one determinant of order n, n determinants each of order n – 1 and perform n^2 divisions. In addition, if this method is used for solving a linear system we also need matrix multiplication. The number of operations (multiplications and divisions) needed, for using this method, increases very rapidly as n increases. For this reason, this method is not used when $n > 4$.

For large n, there are methods which are efficient and are frequently used for finding the inverse of a matrix and solving linear systems. We shall now discuss these methods.

3.2 The Gauss-Jordan Reduction Method

This method is a variation of the Gauss elimination method. In the Gauss elimination method, using elementary row operations, we transform the matrix A to an upper triangular matrix U and obtain the solution by using back substitution method. In Gauss-Jordan reduction not only the elements below the diagonal but also the elements above the diagonal of A are made zero at the same time. In other words, we transform the matrix A to a diagonal matrix D. This diagonal matrix may then be reduced to an identity matrix by dividing each row by its pivot element.

Alternately, the diagonal elements can also be made unity at the same time when the reduction is performed. This transforms the coefficient matrix into an identity matrix. Thus, on completion of the Gauss-Jordan method, we have

$$[A|b] \Longrightarrow [I|d] \tag{3}$$

The solution is then given by

$$x_i = d_i,\ i = 1, 2, \ldots\ldots, n \tag{4}$$

In this method also, we use elementary row operations that are used in the Gauss elimination method. We apply these operations both below and above the diagonal in order to reduce all the off-diagonal elements of the matrix to zero. Pivoting can be used to make the pivot non-zero or make it the largest element in magnitude in that column as discussed. We illustrate the method through an example.

Example 4: Solve the system of equations

$x_1 + x_2 + x_3 = 1$
$4x_1 + 3x_2 – x_3 = 6$
$3x_1 + 5x_2 + 3x_3 = 4$
using Gauss-Jordan method with pivoting.

Solution: We have

$$[A|b] = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 4 & 3 & -1 & 6 \\ 3 & 5 & 3 & 4 \end{array}\right] \quad \text{(interchanging first and second row)}$$

$$» \left[\begin{array}{ccc|c} 4 & 3 & -1 & 6 \\ 1 & 1 & 1 & 1 \\ 3 & 5 & 3 & 4 \end{array}\right] R_2 - \frac{1}{4} R_1,\ R_3 - \frac{3}{4} R_1$$

$$» \left[\begin{array}{ccc|c} 4 & 3 & -1 & 6 \\ 0 & 1/4 & 5/4 & -1/2 \\ 0 & 11/4 & 15/4 & -1/2 \end{array}\right] \quad \text{(interchanging second and third row)}$$

$$» \left[\begin{array}{ccc|c} 4 & 3 & -1 & 6 \\ 0 & 11/4 & 15/4 & -1/2 \\ 0 & 1/4 & 5/4 & -1/2 \end{array}\right] R_3 – 1/11\ R_2,\ R_1 - \frac{12}{11} R_2$$

$$\approx \left[\begin{array}{ccc|c} 4 & 0 & -56/11 & 72/11 \\ 0 & 11/4 & 15/4 & -1/2 \\ 0 & 0 & 10/11 & -5/11 \end{array}\right] \quad R_1 + \frac{56}{10}R_3,\ R_2 - \frac{33}{8}R_3$$

$$\approx \left[\begin{array}{ccc|c} 4 & 0 & 0 & 4 \\ 0 & 11/4 & 0 & 11/8 \\ 0 & 0 & 10/11 & 5/11 \end{array}\right]$$

$R1/4$ (divide first row by 4),

$\frac{4}{11}R_2$ (divide second row by 11/4),

$\frac{11}{10}R_3$ (divide third row by 10/11).

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

which is the desired form.

Thus, we obtain

$x_1 = 1,\ x_2 = \frac{1}{2},\ x_3 = -\frac{1}{2}$.

The method can be easily extended to a general system of n equations. Just as we calculated the number of operations needed for Gauss elimination method in the same way you can verify that the total number of operations needed for this method is $M = \frac{1}{2}n^3 + \frac{n^2}{2} + n$.

Clearly this method requires more number of operations compared to the Gauss elimination method. We therefore, do not use this method generally for solving system of equations but is very commonly used for finding the inverse matrix. This is don by augmenting the matrix A by the identity matrix I of the order same as that of A. Using elementary row operations on the augmented matrix [A|I] we reduce the matrix A to the form I and in the process the matrix I is transformed to A^{-1}

That is

$$[A|I] \Longrightarrow [I|A^{-1}] \qquad (5)$$

We now illustrate the method through examples.

Example 5: Find the inverse of the matrix

$$A = \begin{bmatrix} 3 & 1 & 2 \\ 2 & -3 & -1 \\ 1 & -2 & 1 \end{bmatrix}$$

using the Gauss-Jordan method.

Solution: We have

$$[A|I] = \left[\begin{array}{ccc|ccc} 3 & 1 & 2 & 1 & 0 & 0 \\ 2 & -3 & -1 & 0 & 1 & 0 \\ 1 & -2 & 1 & 0 & 0 & 1 \end{array}\right]_{R/3}$$

$$» \left[\begin{array}{ccc|ccc} 1 & 1/3 & 2/3 & 1/3 & 0 & 0 \\ 2 & -3 & -1 & 0 & 1 & 0 \\ 1 & -2 & 1 & 0 & 0 & 1 \end{array}\right]_{R - 2R,\ R - R}$$

$$» \left[\begin{array}{ccc|ccc} 1 & 1/3 & 2/3 & 1/3 & 0 & 0 \\ 0 & -11/3 & -7/3 & -2/3 & 1 & 0 \\ 0 & -7.3 & 1/3 & -1/3 & 0 & 1 \end{array}\right] \quad 3R/11$$

$$» \left[\begin{array}{ccc|ccc} 1 & 1/3 & 2/3 & 1/3 & 0 & 0 \\ 0 & 1 & 7/11 & 2/11 & -3/11 & 0 \\ 0 & -7/3 & 1/3 & -1/3 & 0 & 1 \end{array}\right] \quad R_1 - \frac{1}{3}R_2,\ R_3 + \frac{7}{3}R_2$$

$$» \left[\begin{array}{ccc|ccc} 1 & 0 & 5/11 & 3/11 & 1/11 & 0 \\ 0 & 1 & 7/11 & 2/11 & -3/11 & 0 \\ 0 & 0 & 20/11 & 1/11 & -7/11 & 1 \end{array}\right] \quad \frac{11}{20}R_3$$

$$» \left[\begin{array}{ccc|ccc} 1 & 0 & 5/11 & 3/11 & 1/11 & 0 \\ 0 & 1 & 7/11 & 2/11 & -3/11 & 0 \\ 0 & 0 & 0 & 1/20 & -7/20 & 11/20 \end{array}\right] \quad R_1 - \frac{5}{11}R_3,\ R_2 - \frac{7}{11}R_3$$

$$\gg \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/4 & 1/4 & -1/4 \\ 0 & 1 & 0 & 3/20 & -1/20 & -7/20 \\ 0 & 0 & 1 & 1/20 & -7/20 & 11/20 \end{array}\right]$$

Thus, we obtain

$$A^{-1} = \begin{bmatrix} 1/4 & 1/4 & -1/4 \\ 3/20 & -1/20 & -7/20 \\ 1/20 & -7/20 & 11/20 \end{bmatrix}$$

Example 6: Find the inverse of the matrix

$$A = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 1 & 1/2 & 0 & 0 \\ 1 & 0 & -3 & 0 \\ 1 & -7/2 & -17 & 55/3 \end{bmatrix}$$

using the Gauss-Jordan method

Solution:

Here we have

$$[A|I] = \left[\begin{array}{cccc|cccc} 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1/2 & 0 & 0 & 0 & 1 & 0 & 0 \\ 2 & 0 & -3 & 0 & 0 & 0 & 1 & 0 \\ 1 & -7/2 & -17 & 55/3 & 0 & 0 & 0 & 1 \end{array}\right] \quad \frac{1}{2}R_1$$

$$\gg \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 1 & 1/2 & 0 & 0 & 0 & 1 & 0 & 0 \\ 2 & 0 & -3 & 0 & 0 & 0 & 1 & 0 \\ 1 & -7/2 & -17 & 55/3 & 0 & 0 & 0 & 1 \end{array}\right]$$

$R_2 – R_1, R_3 – 2R_1, R_4 – R_1$

$$\gg \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & -1/2 & 1 & 0 & 0 \\ 0 & 0 & -3 & 0 & -1 & 0 & 1 & 0 \\ 0 & -7/2 & -17 & 55/3 & -1/2 & 0 & 0 & 1 \end{array}\right] \quad 2R_2$$

$$\approx \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 2 & 0 & 0 \\ 0 & 0 & -3 & 0 & -1 & 0 & 1 & 0 \\ 0 & -7/2 & -17 & 55/3 & -1/2 & 0 & 0 & 1 \end{array}\right] \quad R_4 + \frac{7}{2}R_2$$

$$\approx \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 2 & 0 & 0 \\ 0 & 0 & -3 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -17 & 55/3 & -4 & 7 & 0 & 1 \end{array}\right] \quad \left(-\frac{1}{3}R_3\right)$$

$$\approx \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1/3 & 0 & -1/3 & 0 \\ 0 & 0 & -17 & 55/3 & -4 & 7 & 0 & 1 \end{array}\right] \quad \left(-\frac{1}{17}R_4\right)$$

$$\approx \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1/3 & 0 & -1/3 & 0 \\ 0 & 0 & -17 & 55/3 & 4/17 & -7/17 & 0 & -1/17 \end{array}\right] \quad R_4 - R_3$$

$$\approx \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1/3 & 0 & -1/3 & 0 \\ 0 & 0 & 0 & -55/51 & -5/51 & -7/17 & 1/3 & -1/17 \end{array}\right] \quad \left(-\frac{51}{55}R_4\right)$$

$$\approx \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1/3 & 0 & -1/3 & 0 \\ 0 & 0 & 0 & 1 & 1/11 & 21/55 & -17/55 & 3/55 \end{array}\right]$$

Hence

$$A^{-1} = \begin{bmatrix} 1/2 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 1/3 & 0 & -1/3 & 0 \\ 1/11 & 21/55 & -17/55 & 3/55 \end{bmatrix}$$

is the inverse of the given lower triangular matrix.

Let us now consider the problem of finding the inverse of an upper triangular matrix.

Example 7:

Find the inverse of the matrix

$$A = \begin{bmatrix} 1 & 3/2 & 2 & 1/2 \\ 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & 2/3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Using the Gauss-Jordan method.

$$[A|I] = \left[\begin{array}{cccc|cccc} 1 & 3/2 & 2 & 1/2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -4 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2/3 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right] \quad R_1 - \frac{3}{2}R_2$$

$$\text{»} \left[\begin{array}{cccc|cccc} 1 & 0 & 8 & -1 & 1 & -3/2 & 0 & 0 \\ 0 & 1 & -4 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2/3 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right] \quad R_1 - 8R_3,\ R_2 + 4R_3$$

$$\text{»} \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & -19/3 & 1 & -3/2 & -8 & 0 \\ 0 & 1 & 0 & 11/3 & 0 & 1 & 4 & 0 \\ 0 & 0 & 1 & 2/3 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right]$$

$$R_1 + \frac{19}{3}R_4,\ R_2 - \frac{11}{3}R_4,\ R_3 - \frac{2}{3}R_4$$

$$\text{»} \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & -3/2 & -8 & 19/3 \\ 0 & 1 & 0 & 0 & 0 & 1 & 4 & -11/3 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & -2/3 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right]$$

Hence

$$A^{-1} = \begin{bmatrix} 1 & -3/2 & -8 & 19/3 \\ 0 & 1 & 4 & -11/3 \\ 0 & 0 & 1 & -2/3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

which is the inverse of the given upper triangular matrix.

Note that in Example 2, 3, 6 and 7, the inverse of a lower/upper triangular matrix is again a lower/upper triangular matrix. There is another method of finding the inverse of a matrix A which uses the pivoting strategy. Recall that in Sec. 3.4 of Unit 1, for the solution of system of linear algebraic equation Ax = b, we showed you how the multipliers $m_{p,i,k}$'s can be stored in working array W during the process of elimination. The main advantage of storing these multipliers is that if we have already solved the linear system of equations Ax = b or order n, by the elimination method and we want to solve the system Ax = c with the same coefficient matrix A, only the right side being different, then we do not have to go through the entire elimination process again. Since we have saved in the working matrix W all the multipliers used and also have saved the p vector, we have only to repeat the operations on the right hand side to obtain β, such that $Ux = \beta$ is equivalent to Ax = c.

In order to understand the calculations necessary to derive β , from c consider the changes made in the right side b during the elimination process. Let k be an integer between 1 and n, and assume that the ith equation was used as pivotal equation during step k of the elimination process. Then $i = p_k$. initially, the right side of equation i is just b_i.

If k > 1, then after Step 1, the right side is

$b_i^{(1)} = b_i - m_{i1}\, b_{p_1}$

If k > 2, then after Step 2, the right side is

$b_i^{(2)} = b_i^{(1)} = m_{i2}\, b_{p_2}^{(1)}$

$\quad = b_i - m_{i1}\, b_{p1} - m_{i2}\, b_{p_2}^{(1)}$

In the same manner, we have the right side of equation $i = p_k$ as

$$b_i^{(k-1)} = b_i - m_{i1}\, b_{p1} - m_{i2}\, b_{p_2}^{(1)} - \ldots.. - m_{i,k-1}\, b_{p_{k-1}}^{(k-2)} \qquad (6)$$

Replacing i by p_k in Eqn. (6), we get

$$b_{p_k}^{(k-1)} = b_{p_k} - m_{p_k{}^{'}1}\, b_{p1} - m_{p_k{}^{'}2}\, b_{p_2}^{(1)} - \ldots. - m_{p_k{}^{'}k-1}\, b_{p_{k-1}}^{(k-2)} \qquad (7)$$

$$k = 1, 2, \ldots\ldots, n.$$

Also, since $\tilde{b}_j = b_{p_j}^{(j-1)}$, j = 1, 2, ..., n, we can rewrite Eqn. (7) as

$$\tilde{b}_k = \tilde{b}_{p_k} - m_{p_{k,1}}\tilde{b}_1 - m_{p_{k,2}}\tilde{b}_2 - \ldots\ldots - m_{p_{k,k-1}}\tilde{b}_{k-1} \qquad (8)$$

$$k = 1, \ldots., n.$$

Eqn. (8) can then be used to calculate the entries of $\tilde{b}$. But since the multipliers m_{ij}'s are stored in entries w_{ij}'s of the working matrix W, we can also write Eqn. (8) in the form

$$\tilde{b}_k = \tilde{b}_{p_k} - \sum_{j=1}^{k-1} W_{pkj}\tilde{b}_j,\ k = 1, \ldots., n \qquad (9)$$

Hence, if we just know the final content of the first n columns of W and the pivoting strategy p then we can calculate the solution x of Ax = b by using the back substitution method and writing

$$x_k = \frac{\tilde{b}_k - \sum_{j=k+1}^{n} W_{p_k j}x_j}{W_{p_k k}},\ k = n, n-1, \ldots\ldots, 1 \qquad (10)$$

The vector $x = [x_1\ x_2 \ldots\ldots x_n]^T$ will then be the solution of Ax = b.

For finding the inverse of an n $\times$ n matrix A, we use the above algorithm. We first calculate the final contents of the n columns of the working matrix W and the pivoting vector p and then solve each of the n systems

$$Ax = e_j,\ j = 1, \ldots\ldots, n \qquad (11)$$

where $e_1 = [1\quad 0 \ldots\ldots 0]^T$, $e_2 = [0\quad 1\quad 0 \ldots\ldots 0]^T$, ..., $e_n = [0\quad 0 \ldots\ldots 1]^T$, with the help of Eqn. (9) and (10). Then for each j = 1, ..., n the solution of system of system (11) will be the corresponding column of the inverse matrix A^{-1}. The following example will help you to understand the above procedure.

Example 8:

Find the inverse of the matrix

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

using partial pivoting.

Solution:

Initially $p = [p_1, p_2, p_3]^T = [1, 2, 3]^T$ and the working matrix is

$$W^{(0)} = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ 1 & 1 & 2 \end{bmatrix}$$

Now $d_1 = 2$, $d_2 = 2$, $d_3 = 2$.

Step 1: $\dfrac{|W_{p_{1,1}}|}{d_1} = \dfrac{1}{2}, \dfrac{|W_{p_{2,1}}|}{d_2} = \dfrac{2}{2} = 1, \dfrac{|W_{p_{3,1}}|}{d_3} = \dfrac{1}{2}$

$1 > \dfrac{1}{2}, \dfrac{1}{2} \quad \therefore \quad p_1 = 2, p_2 = 1, p_3 = 3$

We use the second equation to eliminate x_1 from first and third equations and store corresponding multipliers instead of storing zeros in the working matrix. The multipliers are

$m_{p_{i,1}} = \dfrac{w_{p_{i,1}}}{w_{p_{1,1}}}$, i = 2, 3

$\therefore \quad m_{p_{2,1}} = m_{11} = \dfrac{w_{p_{2,1}}}{w_{p_{1,1}}} = \dfrac{1}{2}$

$m_{p_{3,1}} = m_{31} = \dfrac{w_{p_{3,1}}}{w_{p_{1,1}}} = -\dfrac{1}{2}$

we get the following working matrix

$$W^{(1)} = \begin{bmatrix} 1/2 & 3/2 & -1 \\ \boxed{2} & 1 & 0 \\ 1/2 & 3/2 & 2 \end{bmatrix}, \ p = (2, 1, 3)^T$$

Step 2: $\dfrac{|w_{p_{2,2}}|}{dp_2} = \dfrac{|w_{p_{1,2}}|}{d_1} = \dfrac{3/2}{2} = \dfrac{3}{4}$

$\dfrac{|w_{p_{3,2}}|}{dp_3} = \dfrac{|w_{p_{3,2}}|}{d_3} = \dfrac{3/2}{2} = \dfrac{3}{4}$

Since $\dfrac{3}{4} = \dfrac{3}{4}$ so we take $p = (2, 1, 3)^T$

Now $m_{p_{i,2}} = \dfrac{w_{p_{i,2}}}{w_{p_{2,2}}}$, i = 3

\ $m_{p_{3,2}} = m_{32} = \dfrac{w_{p_{3,2}}}{w_{p_{1,2}}} = \dfrac{3/2}{3/2} = 1$

We use the first equation as pivotal equation to eliminate x_2 from the third equation and also store the multipliers. After the second step we have the following working matrix

$$W^{(2)} = \begin{bmatrix} 1/2 & \boxed{3/2} & -1 \\ \boxed{2} & 1 & 0 \\ 1/2 & 1 & 3 \end{bmatrix}, \; p = (2, 1, 3)^T$$

Now in this case, $w^{(2)}$ is our final working matrix with pivoting strategy $p = (2, 1, 3)^T$

Note that circled ones denote multipliers and squared ones denote pivot elements in the working matrices.

To find the inverse of the given matrix A, we have to solve
$Ax = e_1 = [b_1\ b_2\ b_3]^T$
$Ax = e_2 = [b_1\ b_2\ b_3]^T$
$Ax = e_3 = [b_1\ b_2\ b_3]^T$
where $e_1 = [1\ 0\ 0]^T$, $e_2 = [0\ 1\ 0]^T$, $e_3 = [0\ 0\ 1]^T$

First we solve the system $Ax = e_1$ and consider

$$\begin{bmatrix} 1/2 & \boxed{3/2} & -1 \\ 2 & 1 & 0 \\ 1/2 & 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ x \\ x \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \; p = (2, 1, 3)^T \qquad (12)$$

Using Eqn. (9), we get

with $p_1 = 2$, $\tilde{b}_1 = b_2 = 0$
with $p_2 = 1$, $\tilde{b}_2 = b_1 - w_{11}\tilde{b}_1$

$$= 1 - \left[\frac{1}{2}\right]0$$
$$= 1$$

with $p_3 = 3$, $\tilde{b}_3 = b_3 - w_{31}\tilde{b}_1 - w_{32}\tilde{b}_2$

$$= 0 - \left[\frac{1}{2}\right].0 - 1.1 = -1$$

Using Eqn. (10), we then get the following system of equations

$3x_3 = -1$

$\frac{3}{2}x_2 - x_3 = 1$

$2x_1 + x_2 = 0$

which gives $x_3 = -\frac{1}{3}$, $x_2 = \frac{4}{9}$ and $x_1 = -\frac{2}{9}$

i.e., vector x = $\left[\frac{1}{2}.\frac{4}{9}-\frac{1}{3}\right]^T$ is the solution of system (12).

Remember that the solution of system (12) constitutes the first column of the inverse matrix A^{-1}.

In the same way we solve the system of equations $Ax = e_2$ and $Ax = e_3$, or

$$\begin{bmatrix} 1/2 & \boxed{3/2} & -1 \\ \boxed{2} & 1 & 0 \\ 1/2 & 1 & 3 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \ p = (2, 1, 3)^T \qquad (13)$$

and

$$\begin{bmatrix} 1/2 & 3/2 & -1 \\ 2 & 1 & 0 \\ 1/2 & 1 & 3 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \ p = (2, 1, 3)^T \qquad (14)$$

Using Eqns (9) and (10), we obtain the solution of system (13) as

x = $\left[\frac{5}{9}.\frac{1}{9}-\frac{1}{3}\right]^T$ which is the second column of A^{-1} and the solution of system (14), i.e., x = $\left[\frac{1}{9}.\frac{2}{9}-\frac{1}{3}\right]^T$ as the third column of A^{-1}

$$\text{Hence } A^{-1} = \begin{bmatrix} 2/9 & 5/9 & -1/9 \\ 4/9 & -1/9 & 2/9 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

You may recall that in Sec. 3.5 of Unit 1 we discussed the LU decomposition method. Using this method we can factorise any non-singular square matrix A into the product of a lower triangular matrix L and upper triangular matrix U. That is, we can write

$$A = L\,U. \qquad \ldots \ (15)$$

In the next section we shall discuss how form (15) can be used to find the inverse of non-singular square matrices.

3.3 L U Decomposition Method

Let us consider Eqn. (15) and take the inverse on both the sides. If we use the fact that the inverse of the product of matrices is the product of their inverses takes in reverse order, then we obtain

$$A^{-1} = (L\,U)^{-1} = U^{-1}\,L^{-1} \qquad (16)$$

We can now find the inverse of U and L separately and obtain the inverse matrix A^{-1} from Eqn. (16).

Remark: It may appear to you that finding an inverse of a matrix by this method is a lengthy process. But, in practice, this method is very useful because of the fact that here we deal with triangular matrices and triangular matrices are easily invertible. It involves only forward and backward substitutions.

Let us now consider an example to understand how the method works.

Example 9:

Find the inverse of the matrix

$$A = \begin{bmatrix} 3 & 1 & 2 \\ 2 & -3 & -1 \\ 1 & -2 & 1 \end{bmatrix}$$

Using LU decomposition method.

Solution:

We write,

$$A = \begin{bmatrix} 3 & 1 & 2 \\ 2 & -3 & -1 \\ 1 & -2 & 1 \end{bmatrix} = LU = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & u & u \\ 0 & 1 & u \\ 0 & 0 & 1 \end{bmatrix} \qquad (17)$$

Comparing the coefficients on both sides of Eqn. (17), we obtain

$1_{11} = 3, 1_{21} = 2, 1_{31} = 1$ (multiplying the rows of L by the first column of U)

$1_{11}u_{12} = 1, u_{12} = \frac{1}{3}$ (multiplying the rows of L by the

$1_{11}u_{13} = 2, u_{13} = 2/3$ second and third column of U)

The second column of L is obtained from

$1_{21}u_{12} + 1_{22} = a_{22}, 1_{22} = -3 - \frac{2}{3} = -\frac{11}{3}$

$1_{31}u_{12} + 1_{32} = a_{32}, 1_{32} = -2 - \frac{1}{3} = -\frac{7}{3}$

u_{23} is obtained from

$1_{21}u_{13} + 1_{22}u_{23} = a_{23}, u_{23} = \frac{-1 - 2(2/3)}{-11/3} = \frac{7}{11}$

1_{33} is obtained from

$1_{31}u_{13} + 1_{32}u_{23} + 1_{33} = 1, 1_{33} = \frac{20}{11}$

Thus we have

$$L = \begin{bmatrix} 3 & 0 & 0 \\ 2 & -11/3 & 0 \\ 1 & -7/3 & 20/11 \end{bmatrix} \text{ and } U = \begin{bmatrix} 1 & 1/3 & 2/3 \\ 0 & 1 & 7/11 \\ 0 & 0 & 1 \end{bmatrix}$$

Now since L is a lower triangular matrix L^{-1} is also a lower triangular matrix. Let us assume that

$$L^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Using the identity LL^{-1}, we have

$$LL^{-1} = \begin{bmatrix} 3 & 0 & 0 \\ 2 & -11/3 & 0 \\ 1 & -7/3 & 20/11 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 2 & -11/3 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Comparing the coefficients, we get

$$1'_{11} = \frac{1}{3}, \ 1'_{22} = -\frac{3}{11}, \ 1'_{33} = \frac{11}{20}$$

Also,

$$2\, l'_{11} - \frac{11}{3}\, l'_{21} = 0,\; l'_{21} = \frac{6}{33} = \frac{2}{11}$$

$$l'_{11} - \frac{7}{3} l'_{21} + \frac{20}{11}\, l'_{31} = \frac{1}{20}$$

$$-\frac{7}{3}\, l'_{22} + \frac{20}{11} l'_{32} = 0,\; l'_{32} = -\frac{7}{20}$$

\ $$L^{-1} = \begin{bmatrix} 1/3 & 0 & 0 \\ 2/11 & -3/11 & 0 \\ 1/20 & -7/20 & 11/20 \end{bmatrix}$$

Similarly, since U is an upper triangular matrix, U^{-1} is also upper triangular matrix. Using $UU^{-1} = I$, we obtain by backward substitution.

$$U = \begin{bmatrix} 1 & 1/3 & 2/3 \\ 0 & 1 & 7/11 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } U^{-1} = \begin{bmatrix} 1 & -1/3 & -5/11 \\ 0 & 1 & -7/11 \\ 0 & 0 & 1 \end{bmatrix}$$

Therefore, we have from Eqn. (16)

$$A^{-1} = U^{-1} L^{-1} = \begin{bmatrix} 1 & -1/3 & -5/11 \\ 0 & 1 & -7/11 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 0 & 0 \\ 2/11 & -3/11 & 0 \\ 1/20 & -7/20 & 11/20 \end{bmatrix}$$

$$= \begin{bmatrix} 1/4 & 1/4 & -1/4 \\ 3/20 & -1/20 & -7/20 \\ 1/20 & -7/20 & 11/20 \end{bmatrix}$$

4.0 CONCLUSION

We now end this unit by giving a summary of what we have covered in it.

5.0 SUMMARY

In this unit we have covered the following:

1) Using the method of adjoints, the inverse of a given non-singular matrix A can be obtained from

$$A^{-1} = \frac{1}{\det(A)}\,\mathrm{adj}(A) \qquad \text{(see Eqn. (1))}$$

Since the number of operations in the adjoint method to find the inverse of $n \times n$ non-singular matrix A increases rapidly as n increases, the method is not generally used for $n > 4$.

2) For large n, the Gauss-Jordan reduction method, which is an extension of the Gauss elimination method can be used for finding the inverse matrix and solve the linear systems.

$$Ax = b \qquad \text{(see Eqn. (2))}$$

using the Gauss-Jordan method.

a) the solution of system of Eqns (2) can be obtained by using elementary row operations

$$[A|b] \xrightarrow{\text{reduced to}} [I|d]$$

b) the inverse matrix A^{-1} can be obtained by using elementary

row operations $[A|I] \xrightarrow{\text{reduced to}} [I|A^{-1}]$

3) For large n, another useful method of finding the inverse matrix A^{-1} is LU decomposition method. Using this method any non-singular matrix A is first decomposed into the product of a lower triangular matrix L and an upper triangular matrix U. That is

$A = LU$

U^{-1} and L^{-1} can be obtained by backward and forward substitutions. Then the inverse can be found from

$A^{-1} = U^{-1} L^{-1}$

6.0 TUTOR-MARKED ASSIGNMENT

1) Solve the system of equations

$3x_1 + x_2 + 2x_3 = 3$

$2x_1 – x_2 – x_3 = 1$

$x_1 – 2x_2 + x_3 = -4$

using the method of adjoints.

2) Solve the system of equations

$$\begin{bmatrix} 2 & 3 & 4 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 3 & 1 & -1 \\ 1 & -2 & -1 & 4 \end{bmatrix} \begin{bmatrix} X1 \\ X2 \\ X3 \\ X4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 1 \\ 5 \end{bmatrix}$$

using the method of adjoints.

3) Verify that the total number of operations needed for Gauss-Jordan reduction methods is $\frac{1}{2}$ n^3 + $\frac{n^2}{2}$ + n.

4) In example 6 and 7 verify that
$A\,A^{-1} = A^{-1}\,A = I$.

5) Solve the system of equation
$x_1 + 2x_2 + x_3 = 0$
$2x_1 + 2x_2 + 3x_3 = 3$
$-x_1 - 3x_2 = 2$
using the Gauss-Jordan method with pivoting.

6) Find the inverse of the matrix

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

using the Gauss-Jordan method.

7) Find the inverse of the matrix

$$A = \begin{bmatrix} 5 & 8 & 1 \\ 0 & 2 & 1 \\ 4 & 3 & -1 \end{bmatrix}$$

using the LU decomposition method.

8) Find the inverse of the matrix

$$A = \begin{bmatrix} 3 & 1 & 2 \\ 2 & -1 & -1 \\ 1 & -2 & 1 \end{bmatrix}$$

Using the LU decomposition method.

7.0 REFERENCES/FURTHER READINGS

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

UNIT 3 ITERATIVE METHODS

CONTENTS

1.0 INTRODUCTION

In the previous two units, you have studied direct methods for solving linear system of equations $Ax = b$, A being $n \times n$ non-singular matrix. Direct methods provide the exact solution in a finite number of steps provided exact arithmetic is used and there is no round-off error. Also, direct methods are generally used when the matrix A is dense or filled, that is, there are few zero elements, and the order of the matrix is not very large say $n < 50$.

Iterative methods, on the other hand, start with an initial approximation and by applying a suitably chosen algorithm, lead to successively better approximations. Even if the process converges, it would give only an approximate solution. These methods are generally used when the matrix A is sparse and the order of the matrix A is very large say $n > 50$. Sparse matrices have very few non-zero elements. In most cases these non-zero elements lie on or near the main diagonal giving rise to tri-diagonal, five diagonal or band matrix systems. It may be noted that there are no fixed rules to decide when to use direct methods and when to use iterative methods. However, when the coefficient matrix is sparse or large, the use of iterative methods is ideally suited to find the solution which take advantage of the sparse nature of the matrix involved.

In this we shall discuss two iterative methods, namely, Jacobi iteration and Gauss-Seidel iteration methods which are frequently used for solving linear system of equations.

2.0 OBJECTIVES

After studying this unit, you should be able to:

- obtain the solution of system of linear equations, $Ax = b$, when the matrix A is large or sparse, by using the iterative method viz; Jacobi method or the Gauss-Seidel method
- tell whether these iterative methods converges or not
- obtain the rate of convergence and the approximate number of iterations needed for the required accuracy of these iterative methods.

3.0 MAIN CONTENT

3.1 The General Iteration Method

In iteration methods as we have already mentioned, we start with some initial approximate solution vector $x^{(0)}$ an generate a sequence of approximation $\{x^{(k)}\}$ which converge to the exact solution vector x as $k \to \infty$. If the method is convergent, each iteration produces a better approximation to the exact solution. We repeat the iterations till the required accuracy is obtained. Therefore, in an iterative method the amount of computation depends on the desired accuracy whereas in direct methods the amount of computation is fixed. The number of iterations needed to obtain the desired accuracy also depends on the initial approximation, closer the initial approximation to the exact solution, faster will be the convergence.

Consider the system of equations

$$Ax = b \qquad \text{... (1)}$$

where A is an $n \times n$ non-singular matrix.

Writing the system in expanded form, we get

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

We assume that the diagonal coefficients $a_{ii} \neq 0$, $(i = 1, \ldots.., n)$. If some of $a_{ii} = 0$, then we arrange the equations so that this condition holds. We then rewrite system (2) as

$$x_1 = -\frac{1}{a_{11}}(a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n) + \frac{b_1}{a_{11}}$$

$$x_2 = -\frac{1}{a_{22}}(a_{21}x_1 + a_{23}x_3 + ... + a_{2n}x_n) + \frac{b_2}{a_{22}} \qquad (3)$$

$$x_n = -\frac{1}{a_{nn}}(a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn-1}x_{n-1}) + \frac{b_n}{a_{nn}}$$

In matrix form, system (3) can be written as

$x = Hx + c$

where

$$H = \begin{bmatrix} 0 & \frac{-a12}{a11} & \frac{-a13}{a11} & \frac{-a1n}{a11} \\ \frac{a21}{a22} & 0 & \frac{-a23}{a22} & \frac{-a2n}{a22} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{an1}{ann} & \frac{-an2}{ann} & \frac{-an,n-1}{ann} & 0 \end{bmatrix} \qquad (4)$$

and the elements of c are $c_i = \frac{b_i}{a_{ii}}$ (i = 1, 2, ..., n)

To solve system (3) we make an initial guess $x^{(0)}$ of the solution vector and substitute into the r.h.s. of Eqn. (3). The solution of Eqn. (3) will then yield a vector $x^{(1)}$, which hopefully is a better approximation to the solution than $x^{(0)}$. We then substitute $x^{(1)}$ into the r.h.s. of Eqn. (3) and get another approximation, $x^{(2)}$. We continue in this manner until the successive iterations $x^{(k)}$ have converged to the required number of significant figures.

In general we can write the iteration method for solving the linear system of Eqns. (1) in the form

$$x^{(k+1)} = Hx^{(k)} + c, \ k = 0, 1...... \qquad (5)$$

where $x^{(k)}$ and $x^{(k+1)}$ are the approximations to the solution vector x at the kth and the (k + 1)th iterations respectively. H is called the iteration matrix and depends on A. c is a column vector and depends on both A and b. The matrix H is generally a constant matrix.

When the method (5) is convergent, then

$$\lim_{k \to \infty} x^{(k)} = \lim_{k \to \infty} x^{(k+1)} = x$$

and we obtain from Eqn. (5)

$$x = Hx + c \qquad (6)$$

If we define the error vector at the kth iteration as

$$\varepsilon^{(k)} = x^{(k)} - x \tag{7}$$

then subtracting Eqn. (6) from Eqn. (5), we obtain

$$\varepsilon^{(k+1)} = H\,\varepsilon^{(k)} \tag{8}$$

Thus, we get from Eqn. (8)

$$\varepsilon^{(k)} = H\,\varepsilon^{(k-1)} = H^2\,\varepsilon^{(k-2)} = \ldots = H^k\,\varepsilon^{(0)} \tag{9}$$

Where $\varepsilon^{(0)}$ is the error in the initial approximate vector. Thus, for the convergence of the iterative method, we must have

$$\lim_{k \to \infty} \varepsilon^{(k)} = 0$$

independent of $\varepsilon^{(0)}$.

Before we discuss the above convergence criteria, let us recall the following definitions from linear algebra.

Definition:

For a square matrix A of order n, and a number λ the value of λ for which the vector equation $Ax = \lambda x$ has non-trivial solution $x \neq 0$, is called an eigenvalue or characteristic value of the matrix A.

Definition:

The largest eigenvalue in magnitude of A is called the spectral radius of A ad is denoted by p(A).

The eigenvalues of the matrix A are obtained from the characteristic equation

$\det(A - \lambda I) = 0$

which is an nth degree polynomial in λ. The roots of this polynomial $\lambda_1, \lambda_2, \ldots\ldots, \lambda_n$ are the eigenvalues of A. Therefore, we have

$$\rho(A) = \max_i |\lambda_i| \tag{10}$$

We now state a theorem on the convergence of the iterative methods.

Theorem 1:

An iteration method of the form (5) is convergent for arbitrary initial approximate vector $x^{(0)}$ if and only if $\rho(H) < 1$.

We shall not be proving this theorem here as its proof makes use of advanced concepts from linear algebra and is beyond the scope of this course.

We define the rate of convergence as follows:

Definition:

The number $\nu = -\log_{10} \rho(H)$ is called the rate of convergence of an iteration method.

Obviously, smaller the value of $\rho(H)$, larger is the value of ν.

Definition:

The method is said to have converged to m significant digits if $\max_i |\varepsilon_i^{(k)}| \leq 10^{-m}$, that is, largest element in magnitude, of the error vector $\varepsilon^{(k)} \leq 10^{-m}$. Also the number of iterations k that will be needed to make $\max_i |\varepsilon_i^{(k)}| \leq 10^{-m}$ is given by

$$k = \frac{m}{\nu} \tag{11}$$

Therefore, the number of iterations that are required to achieved the desired accuracy depends on ν. For a method having higher rate of convergence, lesser number of iterations will be needed for a fixed accuracy and fixed initial approximation.

There is another convergence criterion for iterative methods which is based on the norm of a matrix.

The norm of a square matrix A of order n can be defined in the same way as we define the norm of an n-vector by comparing the size of Ax with the size of x (an n-vector) as follows:

i) $\|A\|_2 = \max \dfrac{\|Ax\|_2}{\|x\|_2}$

based on the Euclidean vector norm, $\|x\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2}$ and

ii) $||A||_\infty = \max \frac{||Ax||_\infty}{||x||_\infty}$, based on the maximum vector norm, $||x||_\infty = \max\limits_{1 \le i \le n} |x_i|$.

In (i) and (ii) above the maximum is taken over all (non zero) n-vector. The most commonly used norms is the maximum norm $||A||_\infty$, as it is easier to calculate. It can be calculated in any of the following two ways:

$$||A||_\infty = \max_x \sum_i |a_{ik}| \text{ (maximum absolute column-sum)}$$

Or

$$||A||_\infty = \max_i \sum_k |a_{ik}| \text{ (maximum absolute row sum)}$$

The norm of a matrix is a non-negative number which in addition to the property $||AB|| \le ||A||\,||B||$
satisfies all the properties of a vector norm, viz.,

a) $||A|| \ge 0$ and $||A|| = 0$ if $A = 0$

b) $||\alpha A|| = |\alpha|\,||A||$, for all numbers α.

c) $||A + B|| \le ||A|| + ||B||$
where A and B are square matrices of order n.

We no state a theorem which gives the convergence criterion for iterative methods in terms of the norm of a matrix.

Theorem 2:

The iteration method of the form (5) for the solution of system (1) converges to the exact solution for any initial vector, if $||H||<1$.

Also note that
$||H|| \ge \rho(H)$.

This ca be easily proved by considering the eignevalue problem $Ax = \lambda x$.
Then $||A|| = ||\lambda x|| = |\lambda|\,||x||$
or $|\lambda|\,||x|| = ||Ax|| \le ||A||\,||x||$
i.e., $|\lambda| \le ||A||$ since $||x|| \ne 0$

Since this results is true for all eignevalue, we have

$\rho(A) \le \|A\|$.

The criterion given in Theorem 2 is only a sufficient condition, it is not necessary. Therefore, for a system of equations for which the matrix H is such that either $\max\limits_{i} \sum\limits_{k=1}^{n} |h_{ik}| < 1$, the iteration always converges, but if the condition is violated it is not necessary that the iteration diverges.

There is another sufficient condition for convergence as follows:

Theorem 3:

If the matrix A is strictly diagonally dominant that is,

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}|, \; i = 1, 2, ..., n.$$

Then the iteration method (5) converges for nay initial approximation x_{10}. If no better initial approximation is known, we generally take $x^{(0)} = 0$.

We shall mostly use the criterion given in Theorem 1, which is both necessary and sufficient.

For using the iteration method (5), we need the matrix H and the vector c which depend on the matrix A and the vector b. the well-known iteration methods are based on the splitting of the matrix A in the form

$$A = D + L + U \tag{12}$$

where D is the diagonal matrix, L and U are respectively the lower and upper triangular matrices with zero diagonal elements. Based on the splitting (12), we now discuss two iteration methods of the form (5).

3.2 The Jacobi's Iteration Method

We write the system of Eqn. (1) in the form (2), viz.,

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

We assume that a_{11}, a_{22}, ... a_{nn} are pivot elements and $a_{ii} \ne 0$, i = 1, 2, ..., n. if any of the pivots is zero, we can interchange the equations to obtain non-zero pivots (partial pivoting).

Note that, A being a non-singular matrix, it is possible for us to make all the pivots non-zero. It is only when the matrix A is singular that even complete pivoting may not lead to all the non-zero pivots.

We rewrite system (2) in the form (3) and define the Jacobi iteration method as

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

$$x_2^{(k+1)} = -\frac{1}{a_{22}}(a_{21}x_2^{(k)} + a_{23}x_3^{(k)} + ... + a_{2n}x_n^{(k)} - b_2)$$

.

.

.

$$x_n^{(k+1)} = -\frac{1}{a_{ii}}(a_{n1}x_i^{(k)} + a_{n2}x_2^{(k)} + ... + a_{n,n-1}x_{n-1}^{(k)} - b_n)$$

$$\text{or } x_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{j=1}^{n} a_{ij}x_j^{(k)} - b_i, \text{ i = 1, 2, ... n, k = 0, 1, ...} \qquad (13)$$

The method (13) can be put in the matrix form as

$$\begin{bmatrix} x_1^{(k+1)} \\ x_1^{(k+1)} \\ \\ x_1^{(k+1)} \end{bmatrix} = -\begin{bmatrix} \frac{1}{a11} & & & \\ & \frac{1}{a22} & & \\ & & \cdots & \\ & & & \frac{1}{ann} \end{bmatrix} \begin{bmatrix} 0 & a_{12} & .... & a_{1n} \\ a_{21} & 0 & .... & a_{2n} \\ & & \cdots & \\ a_{n1} & a_{n2} & .... & 0 \end{bmatrix} \begin{bmatrix} x_1^{-(k)} \\ x_2^{-(k)} \\ \\ x_n^{-(k)} \end{bmatrix} - \begin{bmatrix} b_1 \\ b_2 \\ \\ b_n \end{bmatrix} \text{ r}$$

$$x^{(k+1)} = -D^{-1}(L + U)x^{(k)} + D^{-1}b, \text{ k = 0, 1, ...} \qquad (14)$$

where

$$D = \begin{bmatrix} a_{11} & 0 & \cdots\cdots & 0 \\ 0 & a_{22} & \cdots\cdots & 0 \\ 0 & \cdots\cdots & & a_{nn} \end{bmatrix}, \quad L = \begin{bmatrix} 0 & 0 & \cdots\cdots & 0 \\ a_{21} & 0 & \cdots\cdots & 0 \\ a_{31} & a_{32} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n,\,n-1} & 0 \end{bmatrix}$$

$$\text{and } U = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots\cdots & a_{1n} \\ 0 & 0 & a_{23} & & a_{2n} \\ & \cdots\cdots & & & \\ & & & & a_{n-1,n} \\ 0 & 0 & 0 & \cdots\cdots & 0 \end{bmatrix}$$

The method (14) is for the form (5), where

$H = -D^{-1}(L + U)$ and $c = D^{-1}b$

For computation purpose, we obtain the solution vector $x^{(k+1)}$ at the (k + 1)th iteration, element by element using Eqn. (13). For large n, we rarely use the method in its matrix form as given by Eqn. (14).

In this method in the (k + 1)th iteration we use the values, obtained at the kth iteration viz., $x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}$ on the right hand side of Eqn. (13) and obtain the solution vector $x^{(k+1)}$. We then replace the entire vector $x^{(k)}$ on the right side of Eqn. (13) by $x^{(k+1)}$ to obtain the solution at the next iteration. In other words each of the equations is simultaneously changed by using the most recent set of x-values. It is for this reason this method is also known as the method of simultaneous displacements.

Let us now solve a few examples for better understanding of the method and its convergence.

Example 1:

Perform four iterations of the Jacobi method for solving the system of equations

$$\begin{bmatrix} 8 & 1 & 1 \\ 1 & -5 & 1 \\ 1 & 1 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 16 \\ 7 \end{bmatrix} \qquad (15)$$

with $x^{(0)} = 0$, the exact solution is $x = [-1\ -4\ -3]^T$.

Solution:

The Jacobi method when applied to the system of Eqns. (15) becomes

$$
\begin{aligned}
x_1^{(k+1)} &= \frac{1}{8}[x_2^{(k)} + x_3^{(k)} - 1] \\
x_2^{(k+1)} &= \frac{1}{5}[x_1^{(k)} + x_3^{(k)} - 16] \qquad (16) \\
x_3^{(k+1)} &= \frac{1}{4}[x_1^{(k)} + x_2^{(k)} - 7], \; k = 0, 1, ...
\end{aligned}
$$

Starting with $x^{(0)} = [0\ 0\ 0]^T$, we obtain form Eqns. (16), the following results:

k = 0

$$
\begin{aligned}
x_1^{(1)} &= \frac{1}{8}[0 + 0 - 1] = -0.125 \\
x_2^{(1)} &= \frac{1}{5}[0 + 0 - 16] = -3.2 \\
x_3^{(1)} &= \frac{1}{4}[0 + 0 - 7] = -1.75
\end{aligned}
$$

k = 1

$$
\begin{aligned}
x_1^{(2)} &= \frac{1}{8}[-3.2 - 1.75 - 1] = -0.7438 \\
x_2^{(2)} &= \frac{1}{5}[-0.125 - 1.75 - 16] = 3.5750 \\
x_3^{(2)} &= \frac{1}{4}[-0.125 - 3.2 - 7] = -2.5813
\end{aligned}
$$

k = 2

$$
\begin{aligned}
x_1^{(3)} &= \frac{1}{8}[-3.5750 - 2.5813 - 1] = -0.8945 \\
x_2^{(3)} &= \frac{1}{5}[-0.7438 - 2.5813 - 16] = -3.8650 \\
x_3^{(3)} &= \frac{1}{4}[-0.7438 - 3.5750 - 7] = 2.8297
\end{aligned}
$$

k = 3

$$
\begin{aligned}
x_1^{(4)} &= \frac{1}{8}[-3.8650 - 2.8297 - 1] = 0.9618 \\
x_2^{(4)} &= \frac{1}{5}[-0.8945 - 2.8297 - 16] = -3.9448 \qquad (17)
\end{aligned}
$$

$$x_3^{(4)} = \frac{1}{4}[-0.8945 - 3.8650 - 7] = -2.9399$$

Thus, after four iterations we get the solution as given in Eqns (17). We find that after iteration, we get better approximation to the exact solution.

Example 2:

Jacobi method is used to solve the system of equations

$$\begin{bmatrix} 4 & -1 & 1 \\ 4 & -8 & 1 \\ -2 & 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 21 \\ 15 \end{bmatrix} \qquad (18)$$

Determine the rate of convergence of the method and the number of iterations needed to make $\max_i |\hat{\imath}_i^{(k)}|$, 10^{-2}

Perform these number of iteration starting with initial approximation $x^{(0)}$ = $[1\ 2\ 2]^T$ and compare the result with the exact solution $[2, 4\ 3]^T$

Solution:

The Jacobi method when applied to the system of Eqns. (18), gives the iteration matrix

$$H = -\begin{bmatrix} \begin{matrix} \frac{1}{a_{11}} & 0 & 0 \\ 0 & \frac{1}{a_{22}} & 0 \\ 0 & 0 & \frac{1}{a_{33}} \end{matrix} \left| \begin{matrix} 1 & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & a_{32} & 0 \end{matrix} \right. \end{bmatrix}$$

$$= -\begin{bmatrix} \begin{matrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{8} & 0 \\ 0 & 0 & \frac{1}{5} \end{matrix} \left| \begin{matrix} 0 & -1 & 1 \\ 4 & 0 & 1 \\ 2 & 1 & 0 \end{matrix} \right. \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1/4 & -1/4 \\ 1/2 & 0 & 1/8 \\ 2/5 & -1/5 & 0 \end{bmatrix}$$

The eignevalues of the matrix H are the roots of the characteristic equation.

$\det (H - \lambda I) = 0$

Now

$$\det (H - \lambda I) = \begin{bmatrix} 1 & 1/4 & -1/4 \\ 1/2 & -1 & 1/8 \\ 2/5 & -1/5 & -1 \end{bmatrix} = \lambda^3 - \frac{3}{80} = 0$$

All the three eigenvalues of the matrix H are equal and they are equal to

$\lambda = 0.3347$

The spectral radius is

$$\rho (H) = 0.3347 \qquad (19)$$

We obtain the rate of convergence as

$\nu = -\log_{10}(0.3347) = 0.4753$

The number of iterations needed for the required accuracy is given by

$$k = \frac{2}{\nu} \approx 5 \qquad (20)$$

The Jacobi method when applied to the system of Eqns. (18) becomes

$$x^{(k+1)} = \begin{bmatrix} 0 & 1/4 & -1/4 \\ 1/2 & 0 & 1/8 \\ 2/5 & -1/5 & 0 \end{bmatrix} x^{(k)} + \begin{bmatrix} 7/4 \\ 21/8 \\ 3 \end{bmatrix}, \ k = 0, 1, \ldots \qquad (21)$$

starting with the initial approximation $x^{(0)} = [1 \ 2 \ 2]^T$, we get from Eqn. (21)

$x^{(1)} = [1.75 \quad 3.375 \quad 3.0]^T$
$x^{(2)} = [1.8437 \quad 3.875 \quad 3.025]^T$
$x^{(3)} = [1.9625 \quad 3.925 \quad 2.9625]^T$
$x^{(4)} = [1.9906 \quad 3.9766 \quad 3.0000]^T$
$x^{(5)} = [1.9941 \quad 3.9953 \quad 3.0009]^T$

which is the result after five iterations. Thus, you can see that result obtained after five iterations is quite close to the exact solution $[2 \ 4 \ 3]^T$

Example 3:

Perform four iterations of the Jacobi method for solving the system of equations

$$\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad (22)$$

With $x^{(0)} = [0.5 \quad 0.5 \quad 0.5 \quad 0.5]^T$. What can you say about the solution obtained if the exact solution is $x = [1 \ 1 \ 1 \ 1]^T$?

Solution:

The Jacobi method when applied to the system of Eqns. (22) becomes

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{2}[1 + x_2^{(k)}] \\ x_2^{(k+1)} &= \frac{1}{2}[x_1^{(k)} + x_3^{(k)}] \\ x_3^{(k+1)} &= \frac{1}{2}[x_2^{(k)} + x_4^{(k)}] \qquad (23) \\ x_4^{(k+1)} &= \frac{1}{2}[1 + x_3^{(k)}], \ k = 0, 1, ... \end{aligned}$$

Using $x^{(0)} = [0.5 \ 0.5 \ 0.5 \ 0.5]^T$, we obtain

$x^{(1)} = [0.75 \ 0.5 \ 0.5 \ 0.75]^T$
$x^{(2)} = [0.75 \ 0.625 \ 0.625 \ 0.75]^T$
$x^{(3)} = [0.8125 \ 0.6875 \ 0.6875 \ 0.8125]^T$
$x^{(4)} = [0.8438 \ 0.75 \ 0.75 \ 0.8438]^T$

You may notice here that the solution is improving after each iteration. Also the solution obtained after four iterations is not a good approximation to the exact solution $x = [1 \ 1 \ 1 \ 1]^T$. this shows that we require a few more iterations to get a good approximation.

Example 4:

Find the spectral radius of the iteration matrix when the Jacobi method, is applied to the system of equations

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

Verify that the iterations do not converge to the exact solution x = $[1\ \ 3\ \ -1]^T$.

Solution:

The iteration matrix H in this case becomes

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

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

and c = $[-1\ 5\ -3]^T$

The eigenvalue of H are roots of the characteristic equation
det (H - λ I) = 0. This gives us

$-\lambda(\lambda^2 - 4) = 0$
i.e., $\lambda = 0, \pm 2$
$\therefore\ \rho(H) = 2 > 1$.

Thus, the condition in Theorem 1 is violated. The iteration method does not converge.

We now perform few iteration and see what happens actually. Taking $x^{(0)} = 0$ and using the Jacobi method

$$x^{(k+1)} = \begin{bmatrix} 0 & 0 & -2 \\ 0 & 0 & 2 \\ 1 & 1 & 0 \end{bmatrix} x^{(k)} + \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}$$

we obtain

$x^{(1)} = (-1\ 5\ -3)^T$
$x^{(2)} = (5\ -1\ 3)^T$
$x^{(3)} = (-7\ 11\ -9)^T$

$x^{(4)} = (17\ -13\ 15)^T$
$x^{(5)} = (-31\ 35\ -33)^T$

and so on, which shows that the iterations are diverging fast. You may also try to obtain the solution with other initial approximations.

Let us now consider an example to show that the convergence criterion given in Theorem 3 is only a sufficient condition. That is, there are systems of equation which are not diagonally dominant but, the Jacobi iteration method converges.

Example 5:

Perform iterations of the Jacobi method for solving the system of equations

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

With $x^{(0)} = [0\ \ 1\ \ 1]^T$. What can you say about the solution obtained if the exact solution is $x = [0\ 1\ 2]^T$?

Solution:

The Jacobi method when applied to the given system of equations becomes

$$x_1^{(k+1)} = [3 - x_2^{(k)} - x_3^{(k)}]$$
$$x_2^{(k+1)} = 1$$
$$x_3^{(k+1)} = [-1 + 3x_2^{(k)}],\ k = 0, 1, ...$$

Using $x^{(0)} = [0\ 1\ 1]^T$, we obtain
$x^{(1)} = [1\ 1\ 2]^T$
$x^{(2)} = [0\ 1\ 2]^T$
$x^{(3)} = [0\ 1\ 2]^T$

You may notice here that the coefficient matrix is not diagonally dominant but the iterations converge to the exact solution after only two iterations.

We have already mentioned that iterative methods are usually applied to large linear system with a sparse coefficient matrix. For sparse matrices, the number of non-zero entries is small, and hence the number of arithmetic operations to be performed per step is small. However, iterative methods may not always converge, and even when they converge, they may require a large number of iterations.

We shall now discuss the Gauss-Seidel method which is a simple modification of the method of simultaneous displacements and has improved rate of convergence.

3.3 The Gauss-Seidel Iteration Method

Consider the system of Eqns. (2) written in form (3). For this system of equations, we define the Gauss-Seidel method as:

$$
\begin{aligned}
x_1^{(k+1)} &= -\frac{1}{a_{11}}(a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \dots + a_{1n}x_n^{(k)} - b_1)\\
x_2^{(k+1)} &= -\frac{1}{a_{22}}(a_{21}x_1^{(k+1)} + a_{23}x_3^{(k)} + \dots + a_{2n}x_n^{(k)} - b_2)\\
&\;.\\
&\;.\\
&\;.\\
x_n^{(k+1)} &= -\frac{1}{a_{nn}}(a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + \dots + a_{n,n-1}x_{n-1}^{(k+1)} - b_n)
\end{aligned}
\tag{24}
$$

$$\text{or } x_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{j=1}^{n} a_{ij}x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} - b_i \text{ , i = 1, 2, ... n}$$

You may notice here that in the first equation of system (24), we substitute the initial approximation ($x_2^{(0)}$, $x_3^{(0)}$, ..., $x_n^{(0)}$) on the right hand side. In the second equation w substitute ($x_1^{(1)}$, $x_3^{(0)}$, ..., $x_n^{(0)}$) on the right hand side. In the third equation, we substitute ($x_1^{(1)}$, $x_2^{(1)}$, $x_4^{(0)}$, ..., $x_n^{(0)}$) on the right hand side. We continue in this manner until all the components have been improved. At the end of this first iteration, we will have an improved vector ($x_1^{(1)}$, $x_2^{(1)}$, ..., $x_n^{(1)}$). The entire process is then repeated. In other words, the method uses an improved component as soon as it becomes available. It is for this reason the method is also called the method of successive displacements.

We can also write the system of Eqns. (24) as follows:

$$a_{11}x_1^{(k+1)} = -a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \dots a_{1n}x_n^{(k)} + b_1$$

$$a_{21}x_2^{(k+1)} + a_{21}x_2^{(k+1)} = -a_{23}x_3^{(k)} - \dots - a_{2n}x_n^{(k)} + b_2$$

.

.

.

$a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + ...+ a_{nn}x_n^{(k+1)} \quad b_n$

In matrix form, this system can be written as

$$(D + L)\, x^{(k+1)} = -U\, x^{(k)} + b \qquad (25)$$

where D is the diagonal matrix

$$D = \begin{bmatrix} a_{11} & & & 0 \\ 0 & a_{22} & & .. \\ & & a_{33} & .. \\ & & & .. \\ & & & \\ 0 & & & a_{nn} \end{bmatrix}$$

and L and U are respectively the lower and upper triangular matrices with the zeros along the diagonal and are of the form

$$L = \begin{bmatrix} 0 & 0 & 0 & ...... & 0 & 0 \\ a_{21} & 0 & 0 & ...... & & 0 \\ a_{31} & a_{32} & 0 & 0 & .... & 0 \\ ...... & & & & & .. \\ .... & & & & & .. \\ a_{n1} & a_{n2} & & & ...... & a_{nn} \end{bmatrix} \quad U = \begin{bmatrix} 0 & a_{12} & a_{13} & ..... & a_{1n} \\ 0 & 0 & a_{23} & ..... & a_{2n} \\ 0 & 0 & 0 & ..... & a_{3n} - \\ & & & & ..... \\ & & & & a_{n-1,n} \\ 0 & & & & 0 \end{bmatrix}$$

From Eqn. (25), we obtain

$$x^{(k+1)} = -\, (D + L)^{-1}\, Ux^{(k)} + (D + L)^{-1}b \qquad (26)$$

which is of the form (5) with

$H = -(D + L)^{-1}\, U$ and $c = (D + L)^{-1}b$.

It may again be noted here, that if A is diagonally dominant then the iteration always converges.

Gauss-Seidel method will generally converge if the Jacobi method converges, and will converge at a faster rate. For symmetric A, it can be shown that

r (Gauss-Seidel iteration method) = [r (Jacobi iteration method)]^2

Hence the rate of convergence of the Gauss-Seidel method is twice the rate of convergence of the Jacobi method. This result is usually true even when A is not symmetric.

We shall illustrate this fact through examples.

Example 6:

Perform four iterations (rounded to four decimal places) using the Gauss-Seidel method for solving the system of equations

$$\begin{bmatrix} 8 & 1 & 1 \\ 1 & -5 & 1 \\ 1 & 1 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 16 \\ 7 \end{bmatrix} \qquad (27)$$

with $x^{(0)} = 0$. The exact solution is $x = (-1\ -4\ -3)^T$.

Solution: The Gauss-Seidel method, for the system (25) is

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{8}[x_2^{(k)} + x_3^{(k)} - 1] \\ x_2^{(k+1)} &= \frac{1}{5}[x_1^{(k+1)} + x_3^{(k+1)} - 16] \\ x_3^{(k+1)} &= \frac{1}{4}[x_1^{(k+1)} + x_2^{(k+1)} - 7], \ k = 0, 1, ... \end{aligned} \qquad (28)$$

Taking $x^{(0)} = 0$, we obtain the following iterations.

k = 0

$$\begin{aligned} x_1^{(1)} &= \frac{1}{8}[0 + 0 - 1] = -0.125 \\ x_2^{(1)} &= \frac{1}{5}[-0.125 + 0 - 16] = -3.225 \\ x_3^{(1)} &= \frac{1}{4}[-0.125 - 3.225 - 7] = -2.5875 \end{aligned}$$

k = 1

$$\begin{aligned} x_1^{(2)} &= \frac{1}{8}[-3.225 - 2.5875 - 1] = -0.8516 \\ x_2^{(2)} &= \frac{1}{5}[-0.8516 - 2.5875 - 16] = 3.8878 \\ x_3^{(2)} &= \frac{1}{4}[-0.8516 - 3.8878 - 7] = -2.9349 \end{aligned}$$

k = 2

$$x_1^{(3)} = \frac{1}{8}[-3.8878 - 2.9349 - 1] = -0.9778$$

$$x_2^{(3)} = \frac{1}{5}[-0.9778 - 2.9349 - 16] = -3.9825$$
$$x_3^{(3)} = \frac{1}{4}[-0.9778 - 3.9825 - 7] = 2.9901$$

k = 3

$$x_1^{(4)} = \frac{1}{8}[-3.9825 - 2.9901 - 1] = 0.9966$$
$$x_2^{(4)} = \frac{1}{5}[-0.9966 - 2.9901 - 16] = -3.9973$$
$$x_3^{(4)} = \frac{1}{4}[-0.996 - 3.9973 - 7] = -2.9985$$

which is a good approximation to the exact solution x = $(-1\ -4\ -3)^T$ with maximum absolute error 0.0034. Comparing with the results obtained in Example 1, we find that the values of x_i, i = 1, 2, 3 obtained here are better approximation to the exact solution than the one obtained in Example 1.

Example 7:

Gauss-Seidel method is used to solved the system of equations

$$\begin{bmatrix} 4 & -1 & 1 \\ 4 & -8 & 1 \\ 2 & 1 & 5 \end{bmatrix} \begin{bmatrix} x \\ x \\ x \end{bmatrix} = \begin{bmatrix} 7 \\ 21 \\ 15 \end{bmatrix} \qquad (29)$$

Determine the rate of convergence of the method and the number of iterations needed to make $\max_i |\hat{I}_i^{(k)}|$, 10^{-2}. Perform these number of iterations with $x^{(0)} = [1\ 2\ 2]^T$ and compare the results with the exact solution x = $[2\ 4\ 3]^T$.

Solution: The Gauss-Seidel method (26) when applied to the system of Eqns. (29) gives the iteration matrix.

$$H = -\begin{bmatrix} 4 & 0 & 0 \\ 4 & -8 & 0 \\ 2 & 1 & 5 \end{bmatrix} \begin{bmatrix} 0 & -1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Since the inverse of a lower triangular matrix let

$$L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 4 & -8 & 0 \\ 2 & 1 & 5 \end{bmatrix}$$

Then

$$\begin{bmatrix} 4 & 0 & 0 \\ 4 & -8 & 0 \\ 2 & 1 & 5 \end{bmatrix} \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

\ $41_{11} = 1,\ 1_{11} = \frac{1}{4}$

$41_{11} - 81_{21} = 0,\ 1_{21}\frac{1}{8}$

$-81_{22} = 1,\ 1_{22} = -\frac{1}{8}$

$-21_{11} + 1_{21} + 51_{31} = 0,\ 1_{31} = \frac{3}{40}$

$-1_{22} + 51_{32} = 0,\ 1_{32} = \frac{1}{40}$

$51_{33} = 1,\ 1_{33} = \frac{1}{5}$

\ $L = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ \frac{1}{8} & \frac{-1}{8} & 0 \\ \frac{3}{40} & \frac{1}{40} & \frac{-1}{5} \end{bmatrix}$

Hence

$$H = \begin{bmatrix} \frac{-1}{4} & 0 & 0 \\ \frac{-1}{8} & \frac{1}{8} & 0 \\ \frac{3}{40} & \frac{-1}{40} & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 0 & -1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{8} & 0 \\ 0 & \frac{3}{40} & \frac{-1}{10} \end{bmatrix}$$

The eigenvalues of the matrix H are the roots of the characteristic equation

$$\det(H - \lambda I) = \begin{vmatrix} -\lambda & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{8} - \lambda & 0 \\ 0 & \frac{3}{40} & -\left(\frac{1}{10} + \lambda\right) \end{vmatrix} = 0$$

We have

$\lambda(80\lambda^2 - 2\lambda - 1) = 0$
which gives

$\lambda = 0, 0.125, -0.1$

Therefore, we have
$\rho(H) = 0.125$

The rate of convergence of the method is given by
$\nu = -\log_{10}(0.125) = 0.9031$

The number of iterations needed for obtaining the desired accuracy is given by

$$k = \frac{2}{\nu} = \frac{2}{0.9031} \approx 3$$

The Gauss-Seidel method when applied to the system of Eqns. (29) becomes

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{4}[7 - x_3^{(k)} + x_2^{(k)}] \\ x_2^{(k+1)} &= \frac{1}{8}[-21 - 4x_1^{(k+1)} - x_3^{(k)}] \\ x_3^{(k+1)} &= \frac{1}{5}[15 + 2x_1^{(k+1)} - x_2^{(k+1)}] \end{aligned} \qquad (30)$$

The successive iterations are obtained as

$x^{(1)} = [1.75 \quad 3.75 \quad 2.95]^T$
$x^{(2)} = [1.95 \quad 3.9688 \quad 2.95]^T$
$x^{(3)} = [1.9956 \quad 3.9961 \quad 2.9990]^T$

which is an approximation to the exact solution after three iterations. Comparing the results obtained in Example 2, we conclude that the Gauss-Seidel method converges faster than the Jacobi method.

Example 8:

Use the Gauss-Seidel method for solving the following system of equations.

$$\begin{bmatrix} 2 & -1 & 0 & 1 \\ 1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{31}$$

with $x^{(0)} = [0.5 \;\; 0.5 \;\; 0.5 \;\; 0.5]^T$. Compare the results with those obtained in Example 3 after four iterations. The exact solution is $x = [1 \;\; 1 \;\; 1 \;\; 1]^T$.

Solution:

Use the Gauss-Seidel method, when applied to the system of Eqns. (31) becomes

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{2}\,[1 + x_2^{(k)}] \\ x_2^{(k+1)} &= \frac{1}{2}\,[x_1^{(k+1)} + x_3^{(k)}] \\ x_3^{(k+1)} &= \frac{1}{2}\,[x_2^{(k+1)} + x_4^{(k)}] \\ x_4^{(k+1)} &= \frac{1}{2}[1 + x_3^{(k+1)}],\ k = 0, 1, \ldots \end{aligned} \tag{32}$$

Starting with the initial approximation $x^{(0)} = [0.5 \;\; 0.5 \;\; 0.5 \;\; 0.5]^T$, we obtain the following iterates

$x^{(1)} = [0.75 \quad 0.625 \quad 0.5625 \quad 0.7813]^T$
$x^{(2)} = [0.8125 \quad 0.6875 \quad 0.7344 \quad 0.8672]^T$
$x^{(3)} = [0.8438 \quad 0.7891 \quad 0.8282 \quad 0.9141]^T$
$x^{(4)} = [0.8946 \quad 0.8614 \quad 0.8878 \quad 0.9439]^T$

In Example 3, the result obtained after four iterations by the Jacobi method was

$x^{(4)} = [0.8438 \quad 0.75 \quad 0.75 \quad 0.8438]^T$

Remark:

The matrix formulations of the Jacobi and Gauss-Seidel methods are used whenever we want to check whether the iterations converge or to find the rate of convergence. If we wish to iterate and find solutions of the systems, we shall use the equation form of the methods.

4.0 CONCLUSION

We now end this unit by giving a summary of what we have covered in it.

5.0 SUMMARY

In this unit, we have covered the following:

1) Iterative methods for solving linear system of equations
   $Ax = b$ (see Eqn. (1))
   where A is an $n \times n$, non-singular matrix. Iterative methods are generally used when the system is large and the matrix A is sparse. The process is started using an initial approximation and lead to successively better approximations.

2) General iterative method for solving the linear system of Eqn. (1) can be written in the form
   $x^{(k+1)} = Hx^{(k)} + c$, $k = 0, 1, \ldots$ (see Eqn. (5))
   where $x^{(k)}$ and $x^{(k+1)}$ are the approximation to the solution vector x at the kth and the (k + 1)th iterations respectively. H is the iteration matrix which depends on A and is generally a constant matrix. c is a column vector and depends on both A and b.

3) Iterative method of the form given in 2) above converges for any initial vector, if $||H|| < 1$, which is a sufficient condition for convergence. The necessary and sufficient condition for convergence is $\rho(H) <$, where $\rho(H)$ is the spectral radius of H.

4) In the Jacobi iteration method or the method of simultaneous displacements.
   $H = -D^{-1}(L + U)$; $c = D^{-1}b$
   where D is a diagonal matrix, L and U are respectively the lower and upper triangular matrices with zero diagonal elements.

5) In the Gauss-Seidel iteration method or the method of successive displacements
   $H = -(D + L)^{-1}U$ and $c = (D + L)^{-1}b$.

6) If the matrix A in Eqn. (1) is strictly diagonally dominant then the Jacobi and Gauss-Seidel methods converge Gauss-Seidel method converges faster than the Jacobi method.

6.0 TUTOR-MARKED ASSIGNMENT (TMA)

1) Perform five iteration of the Jacobi method for solving the system of equations given in Example 4 with $x^{(0)} = [1\ \ 1\ \ 1]^T$.
2) Perform four iterations of the Jacobi method for solving the system of equations

$$\begin{bmatrix} 5 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 6 \\ 4 \end{bmatrix}$$

with $x^{(0)} = 0$. Exact solution is $x = (1\ -1\ -1)^T$

3) Perform four iterations of the Jacobi method for solving the system of equations

$$\begin{bmatrix} 5 & -1 & -1 & -1 \\ 1 & 10 & -1 & -1 \\ 1 & -1 & 5 & -1 \\ 1 & -1 & -1 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 12 \\ 8 \\ 34 \end{bmatrix}$$

with $x^{(0)} = 0$. The exact solution is $x = [1\ 2\ 3\ 4]^T$

4) Set up the Jacobi method in matrix form for solving the system of equations

$$\begin{bmatrix} 1 & 0 & -1/4 & -1/4 \\ 0 & 1 & -1/4 & -1/4 \\ 1/4 & -1/4 & 1 & 0 \\ 1/4 & -1/4 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

and perform four iterations. Exact solution is $x = (\ 1\ 1\ 1\ 1)^T$. Take $x^{(0)} = 0$.

5) Perform four iterations of the Gauss-Seidel method for solving the system of equations given in no. 3.

6) Perform four iterations of the Gauss-Seidel method for solving the system of equations given in no. 4.

7) Gauss-Seidel method is used to solve the system of equations given in no. 4. Determine the rate of convergence and the number of iterations needed to make $\max\limits_{1}|\hat{I}_{i}^{(k)}|$, 10^{-2}. Perform four iterations and compare the results with the exact solution.

7.0 REFERENCES/FURTHER READINGS

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

UNIT 4 EIGENVALUES AND EIGENVECTORS

CONTENTS

1.0 INTRODUCTION

In Unit 7, you have seen that eigenvalues of the iteration matrix play a major role in the study of convergence of iterative methods for solving linear system of equations. Eigenvalues are also of great importance in many physical problems. The stability of an aircraft is determined by the location of the eigenvalues of a certain matrix in the complex plane. The natural frequencies of the vibrations of a beam are actually eigenvalues of a matrix. Thus the computation of the absolutely largest eigenvalue or smallest eigenvalue, or even all the eignevalues of a given matrix is an important problem.

For a given system of equation of the form

$Ax = l\ x$

(1)

Or

$(A - l\ I)x = 0$ (2)

the values of the parameter l , for which the system of Eqn. (2) has a nonzero solution, are called the eigenvalues of A. Corresponding to these eigenvalues, the nonzero solutions of Eqn. (2) i.e. the vectors x, are called the eigenvectors of A. The problem of finding the eigenvalues and the eigenvectors of a square matrix A is known as the eigenvalue problem. In this unit, we shall discuss the eigenvalue problem. To begin with, we shall give you some definitions and properties related to eigenvalues.

2.0 OBJECTIVES

After studying this unit, you should be able to:

- solve simple eigenvalue problems
- obtain the largest eigenvalue in magnitude and the corresponding eigenvector of a given matrix by using the power method
- obtain the smallest eigenvalue in magnitude and an eigenvalue closest to any chosen number along with the corresponding eigenvector of a given matrix by using the inverse power method.

3.0 MAIN CONTENT

3.1 The Eigenvalue Problem

In the previous three units, we were concerned with the non-homogeneous system of linear equations, $Ax = b$. We know that this system has a unique solution if the matrix A is nonsingular. But, if the vector $b = 0$, then the system reduces to the homogeneous system

$$Ax = 0 \qquad (3)$$

If the coefficient matrix A, in Eqn. (3) is nonsingular, then system has only the zero solution, $x = 0$. for the homogeneous system (3) to have a nonzero solution is not unique.

The homogeneous system of Eqn. (2) will have a nonzero solution only when the coefficient matrix $(A - \lambda I)$ is singular, that is,

$$\det (A - \lambda I) = 0 \qquad (4)$$

If the matrix A is an $n \times n$ matrix then Eqn. (4) gives a polynomial of degree n in λ. This polynomial is called the characteristic equation of A. The n roots $\lambda_1, \lambda_2, ..., \lambda_n$ of this polynomial are the eigenvalues of A. for each eigenvalue λ_i, there exists a vector x_i (the eigenvector) which is the nonzero solution of the system of equations

$$(A - \lambda_i)x_i = 0 \qquad (5)$$

The eigenvalues have a number of interesting properties. We shall now state and prove a few of these properties which we shall be using frequently.

P1: A matrix A is singular if and only if it has a zero eigenvalue.

Proof: If A has a zero eigenvalue then

$\det(A - 0I) = 0$

$\Rightarrow \det(A) = 0$

$\Rightarrow$ A is singular.

Conversely, if A is singular then

$\det(A) = 0$

$\Rightarrow \det(A - 0I) = 0$

$\Rightarrow$ 0 is an eigenvalue of the matrix A.

P2: A and A^T have the same eigenvalues.

Proof: If λ is an eigenvalue of A then

$\det(A - \lambda I) = 0$

$\Rightarrow \det(A - \lambda I)^T = 0$

$\Rightarrow \det(A^T - \lambda I^T) = 0$

$\Rightarrow \det(A^T - \lambda I) = 0$

$\Rightarrow \lambda$ is an eigenvalue of A^T

Hence the result.

However, the eigenvectors and A and A^T are not the same.

P3: If the eigenvalue of a matrix A are $\lambda_1, \lambda_2, ..., \lambda_n$ then the eigenvalues of A^m, m any positive integer, are $\lambda_1^m, \lambda_2^m, ..., \lambda_n^m$. Also both the matrices A and A^m have the same set of eigenvectors.

Proof: Since λ_i (i = 1, 2, ..., n) are the eigenvalues of A, we have

$$Ax = \lambda_i x, \; i = 1, 2, ..., n \tag{6}$$

Pre-multiplying Eqn. (6) by A on both sides, we get

$$A^2x = A\lambda_i x = \lambda_i(Ax) = \lambda_i^2 x \tag{7}$$

which implies that $\lambda_1^2, \lambda_2^2, ..., \lambda_n^2$ are the eigenvalues of A^2. further, A and A^2 have the same eigenvectors. Pre-multiplying Eqn. (7) (m – 1) times by A on both sides the general result follows.

P4: If $\lambda_1, \lambda_2, ..., \lambda_n$ are the eigenvalues of A, then $1/\lambda_1, 1/\lambda_2, ..., 1/\lambda_n$ are the eigenvalues pf A^{-1}. Also both the matrices A and A^{-1} have the same set of eigenvectors.

Proof: Since l_i (i = 1, 2, ..., n), are the eigenvalues of A, we have

$$Ax = l_i x, \; i = 1, 2, ..., n \tag{8}$$

Pre-multiplying Eqn. (8) on both sides by A^{-1}, we get

$$A^{-1}A\,x = l_i\, A^{-1}x$$

which gives

$$x = l_i\, A^{-1}x$$

or $A^{-1}x = \frac{1}{l_i} x$

and hence the result.

P5: If $l_1, l_2, ..., l_n$ are the eigenvalues of A, then $l_i - q$, i = 1, 2, ..., n are the eigenvalues of $A - qI$ for any real number q. Both the matrices A and $A - qI$ have the same set of eigenvectors.

Proof: Since l_i is an eigenvalues of A, we have

$$Ax = l_i x, \; i = 1, 2, ..., n \tag{9}$$

Subtracting q x from both sides of Eqn. (9), we get

$$Ax - qx = l_i x - qx$$

which gives

$$(A - qI)x = (l_i - q)x$$

and the results follows.

P6: If l_i, i = 1, 2, ..., n are the eigenvalues of A then $\frac{1}{l_i - q}$, i = 1, 2, ..., n are the eigenvalues of $(A - qI)^{-1}$ for any real number q. Both the matrices A and $(A - qI)^{-1}$ have the same set of eigenvectors.

P6 can be proved by combining P4 and P5. we leave the proof to you.

We now give you a direct method of calculating the eigenvalues and eigenvectors of a matrix.

Example 1:

Find the eigenvalues of the matrix

a) $A =; \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

b) $A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{bmatrix}$

c) $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$

Solution:

a) Using Eqns. (4), we obtain the characteristic equations as

$$\det(A - \lambda I) = \begin{bmatrix} 1-\lambda & 0 & 0 \\ 0 & 2-\lambda & 0 \\ 0 & 0 & 3-\lambda \end{bmatrix} = 0$$

which gives $(1 - \lambda)(2 - \lambda)(3 - \lambda) = 0$.

and hence the eigenvalues of A are $\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 3$.

b) $$\det(A - \lambda I) = \begin{bmatrix} 1-\lambda & 0 & 0 \\ 2 & 3-\lambda & 0 \\ 4 & 5 & 6-\lambda \end{bmatrix} = 0$$

which gives $(1 - \lambda)(3 - \lambda)(6 - \lambda) = 0$.

and hence the eigenvalues of A are $\lambda_1 = 1, \lambda_2 = 3, \lambda_3 = 6$.

c) $$\det(A - \lambda I) = \begin{bmatrix} 1-\lambda & 2 & 3 \\ 0 & 4-\lambda & 5 \\ 0 & 0 & 6-\lambda \end{bmatrix} = 0$$

Therefore, $(1 - \lambda)(4 - \lambda)(6 - \lambda) = 0$.

Eigenvalues of A are $\lambda_1 = 1$, $\lambda_2 = 4$, $\lambda_3 = 6$.

Remark: Observe that in Example 1 (a), the matrix A is diagonal and in parts (b) and (c), it is lower and upper triangular respectively. In these cases the eigenvalues of A are the diagonal elements. This is true for any diagonal, lower triangular or upper triangular matrix. Formally, we give the result in the following theorem.

Theorem 1:

The eigenvalues of a diagonal, lower triangular or an upper triangular matrix are the diagonal elements themselves. Let us consider another example.

Example 2:

Find the eigenvalues and the corresponding eigenvectors of the matrices.

a) $\begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}$;

b) A = $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$

and

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

Solution:

a) Using Eqns. (4), we obtain the characteristic equation as

$$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} = 0,$$

which gives the polynomial

$$\lambda^2 - 5\lambda + 4 = 0$$

i.e., $(\lambda - 1)(\lambda - 4) = 0$

The matrix A has two distinct real eigenvalues $l_1 = 1$, $l_2 = 4$. To obtain the corresponding eigenvectors we solve the system of Eqn. (5) for each value of l .

For l = 1, we obtain the system of equations

$$x_1 + 2x_2 = 0$$
$$x_1 + 2x_2 = 0$$

which redices to a single equation

$$x_1 + 2x_2 = 0$$

Taking $x_2 = k$, we get $x_1 = -2k$, k being arbitrary nonzero constant. Thus, the eigenvector is of the form

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{bmatrix} k \\ k \end{bmatrix} = k \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

For l = 4, we obtain the system of equations

$$-2x_1 + 2x_2 = 0$$
$$x_1 - x_2 = 0$$

which reduces to a single equation

$$x_1 - x_2 = 0$$

Taking $x_2 = k$, we get $x_1 = k$ and the corresponding eigenvector is

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Note: In practice we usually omit k and say that $[-2\ 1]^T$ and $[1\ 1]^T$ are the eigenvectors of A corresponding to the eigenvalues l = 1 and l = 4 respectively. Moreover, the eigenvectors in this case are linearly independent.

b) The characteristic equation in this case becomes

$$(l - 1)^2 = 0$$

Therefore, the matrix A has a repeated real eigenvalue. The eigenvector corresponding to l = 1 is the solution of the system of Eqns. (5), which reduces to a single equation

$x_2 = 0$

Taking $x_1 = k$, we obtain the eigenvector as

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = k \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Note: that, in this case of repeated eigenvalues, we got linearly dependent eigenvectors.

c) The characteristic equation in this case becomes

$$\lambda^2 - 2\lambda + 5 = 0$$

which gives two complex eigenvalues $\lambda - 1 \pm 2i$.

The eigenvector corresponding to $\lambda = 1 + 2i$ is the solution of the system of Eqns. (5). In this case we obtain the following equations

$$ix_1 + x_2 = 0$$
$$x_1 - ix_2 = 0$$

which reduces to the single equation

$$x_1 - ix_2 = 0$$

Taking $x_2 = k$, we get the eigenvector

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Similarly, for $\lambda = 1 - 2i$, we obtain the eigenvector

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

In the above problem you may note that corresponding to complex eigenvalues, we got complex eigenvectors. Let us now consider an example of 3×3 matrix.

Example 3:

Determine the eigenvalues and the corresponding eigenvectors for the matrices

a) $A = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$;

$A = \begin{bmatrix} 6 & -2 & 2 \\ 2 & 3 & -1 \\ 2 & -1 & 3 \end{bmatrix}$

Solution:

a) The characteristic equation in this case becomes

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

which gives the polynomial
$(2 - l)(l^2 - 4l + 2) = 0$

Therefore, the eigenvalues of A are 2, $2 + \sqrt{2}$ and $2 - \sqrt{2}$.

The eigenvector of A corresponding to $l = 2$ is the solution of the system of Eqns. (5), which reduces to

$x_2 = 0$
$x_1 + x_3 = 0$

Taking $x_3 = k$, we obtain the eigenvector

$$\begin{bmatrix} X1 \\ X2 \\ X3 \end{bmatrix} = k \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

The eigenvector of A corresponding to $l = 2 + \sqrt{2}$ is the solution of the system of equations

$$\begin{bmatrix} \sqrt{2} & -1 & 0 \\ -1 & \sqrt{2} & -1 \\ 0 & -1 & \sqrt{2} \end{bmatrix} \begin{bmatrix} X1 \\ X2 \\ X3 \end{bmatrix} = k \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad (10)$$

To find the solution of system of Eqns. (10), we use Gauss elimination method.

Performing $R_2 - \frac{1}{\sqrt{2}} R_1$, we get

$$\begin{bmatrix} \sqrt{2} & -1 & 0 \\ 0 & -1/\sqrt{2} & -1 \\ 0 & -1 & -\sqrt{2} \end{bmatrix} \begin{bmatrix} X1 \\ X2 \\ X3 \end{bmatrix} = k \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Again performing $R_3 - \sqrt{2} R_2$, we get

$$\begin{bmatrix} \sqrt{2} & -1 & 0 \\ 0 & -1/\sqrt{2} & -1 \\ 0 & 0 & 0 \end{bmatrix}_0 \begin{bmatrix} X1 \\ X2 \\ X3 \end{bmatrix} = k \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Which give the equations

$-\sqrt{2}\ x_1 - x_2 = 0$
$-x_2 - \sqrt{2}\ x_3 = 0$

Taking $x_3 = k$, we obtain the eigenvector

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = k \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \end{bmatrix}$$

Similarly, corresponding to the eigenvalue l $= 2 - \sqrt{2}$, the eigenvector is the solution of system of equations

$$\begin{bmatrix} \sqrt{2} & -1 & 0 \\ -1 & \sqrt{2} & -1 \\ 0 & -1 & \sqrt{2} \end{bmatrix} \begin{bmatrix} X1 \\ X2 \\ X3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Using the Gauss elimination method, the system reduces to the equations
$\sqrt{2}\ x_1 - x_2 = 0$
$x_2 - \sqrt{2}\ x_3 = 0$

Taking $x_3 = k$, we obtain the eigenvector

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = k \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \end{bmatrix}$$

b) The characteristic equation in this case becomes
$(l - 8)(l - 2)^2 = 0$

Therefore the matrix A has the real eigenvalues 8, 2 and 2. The eigenvalue 2 is repeated two times.

The eigenvector corresponding to $l = 8$ is solution of system of Eqns. (5), which reduces to

$$\begin{aligned} x_1 + x_2 - x_3 &= 0 \\ 2x_1 + 5x_2 + x_3 &= 0 \\ 2x_1 - x_2 - 5x_3 &= 0 \end{aligned} \tag{11}$$

Subtracting the last equation of system (11) from the second equation we obtain the system of equations

$$\begin{aligned} x_1 + x_2 - x_3 &= 0 \\ x_2 + x_3 &= 0 \end{aligned}$$

Taking $x_3 = k$, the eigenvector is

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = k \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

The eigenvector corresponding to $l = 2$ is the solution of system of Eqns. (5), which reduces to a single equation.

$$2x_1 - x_2 + x_3 = 0 \tag{12}$$

We can take any values for x_1 and x_2 which need not be related to each other. The two linearly independent solutions can be written as:

$$k \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \text{ or } k \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

Note that in Eqn. (12), it is not necessary that we always assign values to x_1 and x_2. we can assign values to any of the two variables and obtain the corresponding value of the third variable.

On the basis of Example 2 and 3, we can make in general, the following observations:

For a given $n \times n$ matrix A, the characteristic Eqn. (4) is a polynomial of degree n in λ. The n roots of this polynomial λ_1, ..., λ_n, called the eigenvalues of A may be real or complex, distinct or repeated. Then,

i) For distinct, real eigenvalues we, obtain linearly independent eigenvectors. (Examples 2(a) and 3(a))

ii) For a repeated eigenvalue, there may or may not be linearly independent eigenvectors. (Examples 2(b) and 3(b))

iii) For a complex eigenvalue, we obtain a complex eigenvector.

iv) An eigenvector is not unique. Any non-zero multiple of it is again an eigenvector.

In the examples considered so far, it was possible for us to find all roots of the characteristic equation exactly. But this may not always be possible. This is particularly true for $n > 3$. In such cases some iterative method like Newton-Raphson method may have to be used to find a particular eigenvalue or all the eigenvalues from the characteristic equation. However, in many practical problems, we do not require all the eigenvalues but need only a selected eigenvalue. For example, when we use iterative methods for solving a non-homogeneous system of linear equations $Ax = b$, we need to know only the largest eigenvalue in magnitude of the iteration matrix H, to find out whether the method converges or not. One iterative method, which is frequently used to determine the largest eigenvalue in magnitude (also called the dominant eigenvalue) and the corresponding eigenvector for a given square matrix A is the power method. In this method we do not find the characteristic equation. This method is applicable only when all the eigenvalues are real and distinct. If the magnitude of two or more eigenvalues is the same then the method converges slowly.

3.2 The Power Method

Let us consider the eigenvalue problem

$$Ax = \lambda x.$$

Let $\lambda_1, \lambda_2, \ldots\ldots, \lambda_n$ be the n real and distinct eigenvalues of A such that

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$$

Therefore, λ_1 is the dominant eigenvalue of A.

In this method, we start with an arbitrary nonzero vector $y^{(0)}$ (not an eigenvector), and form a sequence of vectors $(y^{(k)})$

$$y^{(k+1)} = Ay^{(k)}, \; k = 0, 1, \ldots. \qquad (13)$$

In the limit as $k \to \infty$, $y^{(k)}$ converges to the eigenvector corresponding to the dominant eigenvalue of the matrix A. we can stop the iteration when the largest element in magnitude in $y^{(k+1)} - y^{(k)}$ is less than the predefined error tolerance. For simplicity, we usually take the initial vector $y^{(0)}$ with all its elements equal to one.

Note that in the process of multiplying the matrix A with the vector $y^{(k)}$, the elements of the vector $y^{(k+1)}$ may become very large. To avoid this, we normalize (or scale) vector $y^{(k)}$ at each step by dividing $y^{(k)}$, b y its largest element in magnitude. This will make the largest element in magnitude in the vector $y^{(k+1)}$ as one and the remaining elements less than one.

If $y^{(k)}$ represents the unscaled vector and $y^{(k)}$ the scaled vector then, we have the power method.

$$y^{(k+1)} = Av^{(k)} \qquad (14)$$

$$v^{(k+1)} = \frac{1}{m_{m+1}} y^{(k+1)}, \; k = 0, 1, \ldots \qquad (15)$$

with, $v^{(0)} = y^{(0)}$ and m_{k+1} being the largest element in magnitude of $y^{(k+1)}$. We then obtain the dominant eigenvalue by taking the limit

$$\lambda_1 = \lim_{k \to \infty} \frac{(y^{(k+1)})r}{(v^{(k)})r} \qquad (16)$$

where r represents the rth component of that vector. Obviously, there are n ratios of numbers. As $k \to \infty$ all these ratios tend to the same value, which is the largest eigenvalue in magnitude i.e., λ_1. The iteration is stopped when the magnitude of the difference of any two ratios is less than the prescribed tolerance.

The corresponding eigenvector is then $v^{(k+1)}$ obtained at the end of the last iteration performed.

We now illustrate the method through an example.

Example 4:

Find the dominant eigenvalue and the corresponding eigenvector correct to two decimal places of the matrix

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

Using the power method.

Solution:

We take
$y^{(0)} = v^{(0)} = (1\ \ 1\ \ 1)^T$

Using Eqn. (14), we obtain

$$y^{(1)} = Av^{(0)} = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

Now $m_1 = 1$ and $v^{(1)} = \dfrac{1}{m_1}\ y^{(1)} = (1\ \ 0\ \ 1)^T$.
Again,

$$y^{(2)} = Av^{(1)} = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}$$

$m_2 = 2$ and $v^{(2)} = \dfrac{1}{m_2}\ y^{(2)} = (1\ \ -1\ \ 1)^T$.

Proceeding in this manner, we have

$y^{(3)} = Av^{(2)} = [3\ \ -4\ \ 3]^T$
$m_3 = 4$
$v^{(3)} = \dfrac{1}{4}y^{(3)} = [0.75\ \ -1\ \ 0.75]^T$

$y^{(4)} = Av^{(3)} = [2.5 \;\; -3.5 \;\; 2.5]^T$

$m_4 = 3.5$

$v^{(4)} = \frac{1}{3.5} y^{(4)} = [0.7143 \;\; -1 \;\; 0.7143]^T$

$y^{(5)} = Av^{(4)} = [2.4286 \;\; -3.4286 \;\; 2.4286]^T$

$m_5 = 3.4286$

$v^{(5)} = \frac{1}{3.4286} y^{(5)} = [0.7083 \;\; -1 \;\; 0.7083]^T$

$y^{(6)} = Av^{(6)} = [2.4166 \;\; -3.4166 \;\; 2.4166]^T$

$m_6 = 3.4166$

$v^{(6)} = \frac{1}{3.4166} y^{(6)} = [0.7073 \;\; -1 \;\; 0.7073]^T$

$y^{(7)} = Av^{(6)} = [2.4146 \;\; -3.4146 \;\; 2.4146]^T$

$m_7 = 3.4146$

$v^{(7)} = \frac{1}{3.4146} y^{(7)} = [0.7071 \;\; -1 \;\; 0.7071]^T$

After 7 iterations, the ratios $\frac{(y^{(7)})r}{(v^{(6)})r}$ are given as 3.4138, 3.4146 and 3.4138. The maximum error in these ratios is 0.0008. Hence the dominant eigenvalue can be taken as 3.414 and the corresponding eigenvector is $[0.7071 \;\; -1 \;\; 0.7071]^T$

Note that the exact dominant eigenvalue of A as obtained in Example 3 was $2 + \sqrt{2} = 3.4142$ and the corresponding eigenvector was $[1 \;\; -\sqrt{2} \;\; 1]^T$ which can also be written as $[\frac{1}{\sqrt{2}} \;\; -1 \;\; \frac{1}{\sqrt{2}}]^T = [0.7071 \;\; -1 \;\; 0.7071]^T$

You must have realized that an advantage of the power method is that the eigenvector corresponding to the dominant eigenvalue is also generated at the same time. Usually, for most of the methods of determining eigenvalues, we need to do separate computations to obtain the eigenvector.

In some problems, the most important eigenvalue is the least magnitude. We shall discuss now the inverse power method which gives the least eigenvalue in magnitude.

We first note that if λ is the smallest eigenvalue in magnitude of A, then $\frac{1}{\lambda}$ is the largest eigenvalue in magnitude of A^{-1}. The corresponding eigenvectors are same. If we apply the power method to A^{-1}, we obtain its largest eigenvalue and the corresponding eigenvector. This eigenvalue is then the smallest eigenvalue in magnitude of A and the eigenvector is same. Since power method is applied to A^{-1}, it is called the inverse power method.

Consider the method

$$y^{(k+1)} = A^{-1}v^{(k)}, \; k = 0, 1, 2, \ldots\ldots\ldots. \tag{17}$$

$$v^{(k+1)} = \frac{1}{m_{k+1}} y^{(k+1)} \text{ with } v^{(0)} = y^{(0)}$$

where $y^{(0)}$ is an arbitrary nonzero vector different from the eigenvector of A.

However, algorithm (17) is not in suitable form, as one has to find A^{-1}. Alternately, we write Eqn. (17) as

$$Ay^{(k+1)} = v^{(k)}$$

$$v^{(k+1)} = \frac{1}{m_{k+1}} y^{(k+1)}, \; k = 0, 1, 2, \ldots\ldots\ldots. \tag{18}$$

We now need to solve a system of equations for $y^{(k+1)}$, which can be obtained using any of the method discussed in the previous units. The largest eigenvalue of A^{-1} is again given by

$$\mu = \lim_{k \to \infty} \frac{(y^{(k+1)})_r}{(v^{(k)})_r}$$

The corresponding eigenvector is $v^{(k+1)}$.
We now illustrate the method through an example.

Example 5:

Find the smallest eigenvalue in magnitude and the corresponding eigenvector of the matrix.

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

using four iterations of the inverse power method.

Solution:

Taking $v^{(0)} = [1 \;\; 1 \;\; 1]^T$, we write

First iteration

$Ay^{(1)} = v^{(0)}$

or

$$\begin{bmatrix} 2 & -1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad (19)$$

For solving the system of Eqns. (19), we use the LU decomposition method. We write

$$A = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} = LU = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix} \qquad (20)$$

comparing the coefficient on both sides of Eqns. (20), we obtain

$$A = LU = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 3/2 & 0 \\ 0 & -1 & 4/3 \end{bmatrix} \begin{bmatrix} 1 & -1/2 & 0 \\ 0 & 1 & -2/3 \\ 0 & -1 & 4/3 \end{bmatrix}$$

Solving $Lz = v^{(0)}$

and then $Uy^{(1)} = z$

we obtain

$y^{(1)} = [3/2 \quad 2 \quad 3/2] = [1.5 \quad 2.0 \quad 1.5]^T$

$m_1 = 2.0$

\ $v^{(1)} = \dfrac{1}{m_1} y^{(1)} = [0.75 \quad 1.0 \quad 0.75]^T$

Second iteration

$Ay^{(2)} = v^{(1)}$

Solving $Lz = v^{(1)}$

and $Uy^{(2)} = z$

we obtain

$y^{(2)} = [1.25 \quad 1.75 \quad 1.25]^T$

$m_2 = 1.75$

$v^{(2)} = \frac{1}{m_2} y^{(2)} = [0.7143 \quad 1 \quad 0.7143]^T$

Third iteration

$Ay^{(3)} = v^{(2)}$

$y^{(3)} = [1.2143 \quad 1.7143 \quad 1.2143]^T$

$m_3 = 1.7143$

$v^{(3)} = \frac{1}{m_3} y^{(3)} = [0.7083 \quad 1 \quad 0.7083]^T$

Fourth iteration

$Ay^{(4)} = v^{(3)}$

$y^{(4)} = [1.2083 \quad 1.7083 \quad 1.2083]^T$

$m_4 = 1.7083$

$v^{(4)} = \frac{1}{m_4} y^{(4)} = [0.7073 \quad 1 \quad 0.7073]^T$

after 4 iterations, the ratios $\frac{(y^{(4)})r}{(v^{(3)})r}$ are given as 1.7059, 1.7083, 1.7059. The maximum error in these ratios is 0.0024. hence the dominant eigenvalue of A^{-1} can be taken as 1.70. Therefore, $\frac{1}{1.70}$ = 0.5882 is the smallest eigenvalue of A in magnitude and the corresponding eigenvector is given by $[0.7073 \quad 1 \quad 0.7073]^T$.

Note that the smallest eigenvalue in magnitude of A as calculated in Example 3 was $2 - \sqrt{2} = 0.5858$ and the corresponding eigenvector was $[1 \ \sqrt{2} \ 1]^T$ or $[0.7071 \ 1 \ 0.7071]^T$.

The inverse power method can be further generalized to find some other selected eigenvalues of A. For instance, one may be interested to find the eigenvalue of A which is nearest to some chosen number q. You know from P6 of Sec. 3.1 that the matrices A and A - qI have the same set of eigenvectors. Further, for each eigenvalue λ_i of A, $\lambda_i - q$ is the eigenvalue of A – qI.

We can therefore use the iteration

$$y^{(k+1)} = (A - qI)^{-1}v^{(k)} \tag{21}$$

with scaling as described in Eqns. (14) – (16). We determine the dominant eigenvalue μ of $(A - qI)^{-1}$ using the procedure given in eqns. (18), i.e.

$$(A - qI)\, y^{(k+1)} = v^{(k)}$$

$$v^{(k+1)} = \frac{1}{\mu_{k+1}}\, y^{(k+1)} \tag{22}$$

Using P6, we have the relation

$$\mu = \frac{1}{\lambda - q}, \text{ where } \lambda \text{ is an eigen value of A.}$$

$$\text{i.e., } \lambda = \frac{1}{\mu} + q \tag{23}$$

Now since μ is the largest eigenvalue in magnitude of $(A - qI)^{-1}$, $\frac{1}{\mu}$ must be the smallest eigenvalue in magnitude of A – qI. Hence, the eigenvalue $\frac{1}{\mu} + q$ of A is closest to q.

Example 6:

Find the eigenvalue of the matrix A, nearest to 3 and also the corresponding eigenvector using four iterations of the inverse power method where,

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

Solution:

In this case q = 3. Thus we have

$$A - 3I = \begin{bmatrix} 1 & -1 & 0 \\ 1 & -1 & -1 \\ 0 & -1 & -1 \end{bmatrix}$$

To find $y^{(k+1)}$, we need to solve the system

$$\begin{bmatrix} 1 & -1 & 0 \\ 1 & -1 & -1 \\ 0 & -1 & -1 \end{bmatrix} y^{(k+1)} = v^{(k)} \tag{24}$$

and normalize $y^{(k+1)}$ as given in Eqn. (22).

First iteration

Starting with $v^{(0)} = [1 \;\; 1 \;\; 1]^T$ and using the Gauss elimination method to solve the system (24), we obtain

$y^{(1)} = [0 \;\; -1 \;\; 0]^T$

$m_1 = 1$

$v^{(1)} = \dfrac{1}{m_1} y^{(1)} = [0 \;\; -1 \;\; 0]^T$

Second iteration

$Ay^{(2)} = v^{(1)}$

$y^{(2)} = [1 \;\; -1 \;\; 1]^T$

$m_2 = 1$

$v^{(2)} = \dfrac{1}{m_2} y^{(2)} = [1 \;\; -1 \;\; 1]^T$

Third iteration

$Ay^{(3)} = v^{(2)}$

$y^{(3)} = [2 \;\; -3 \;\; 2]^T$

$m_3 = 3$

$$v^{(3)} = \frac{1}{m_3} y^{(3)} = [\frac{2}{3} \;\; -1 \;\; \frac{2}{3}]^T$$

Fourth iteration

$Ay^{(4)} = v^{(3)}$

$$y^{(4)} = [\frac{5}{3} \;\; -\frac{7}{3} \;\; \frac{5}{3}]^T$$

$$m_4 = \frac{7}{3} = 2.333$$

$$v^{(4)} = \frac{1}{m_4} y^{(4)} = [\frac{5}{7} \;\; -1 \;\; \frac{5}{7}]^T$$

After four iterations, the ratios $\frac{(y^{(4)})r}{(v^{(3)})r}$ are given as 2.5, 2.333, 2.5. The maximum error in these ratios is 0.1667. Hence the dominant eigenvalue of $(A – 31)^{-1}$ can be taken as 2. Thus the eigenvalue l of A closest to 3 as given by Eqn. (23) is

$$l = \frac{1}{m} + 3$$

$$= \frac{1}{2} + 3 = \frac{7}{2} = 3.5$$

and the corresponding eigenvector is $v^{(4)} = [5/7 \;\; -1 \;\; 5/7]$ = $[0.7143 \;\; -1 \;\; 0.7143]^T$. Note that the eigenvalue of A closest to 3 as obtained in Example 3 was $2 + \sqrt{2} = 3.4142$. The eigenvector corresponding to this eigenvalue was $[0.7071 \;\; -1 \;\; 0.7071]^T$

The eigenvalues of a given matrix can also be estimated. That is, for a given matrix A, we can find the region in which all its eigenvalues lie. This can be done as follows:

Let λ_i be an eigenvalue of A and x_i be the corresponding eigenvector, i.e.,

$$Ax_i = \lambda_i x_i \tag{25}$$

or

$$\begin{aligned}
a_{11}x_{i,1} + a_{12}x_{i,2} + \ldots\ldots + a_{1n}x_{i,n} &= \lambda_i x_{i,1} \\
a_{21}x_{i,1} + a_{22}x_{i,2} + \ldots\ldots + a_{2n}x_{i,n} &= \lambda_i x_{i,2} \\
\vdots \qquad \vdots \qquad\qquad \vdots &\qquad \vdots \\
a_{k1}x_{i,1} + a_{k2}x_{i,2} + \ldots\ldots + a_{kn}x_{i,n} &= \lambda_i x_{i,k} \\
\vdots \qquad \vdots \qquad\qquad \vdots &\qquad \vdots \\
a_{n1}x_{i,1} + a_{n2}x_{i,2} + \ldots\ldots + a_{nn}x_{i,n} &= \lambda_i x_{i,n}
\end{aligned} \tag{26}$$

Let $|x_{i,k}|$ be the largest element in magnitude of the vector $[x_{i,1}, x_{i,2}, \ldots\ldots, x_{i,n}]^T$. Consider the kth equation of the system (26) and divide it by $x_{i,k}$. We then have

$$a_{k1}\left(\frac{x_{i,1}}{x_{i,k}}\right) + a_{k2}\left(\frac{x_{i,2}}{x_{i,k}}\right) + \ldots. + a_{kk} + \ldots. + a_{kn}\left(\frac{x_{i,n}}{x_{i,k}}\right) = \lambda_i \tag{27}$$

Taking the magnitudes on both sides of Eqn. (27), we get

$$\begin{aligned}
|\lambda_i| &\le |a_{k1}|\left|\frac{x_{i,1}}{x_{i,k}}\right| + |a_{k2}|\left|\frac{x_{i,2}}{x_{i,k}}\right| + \ldots.. + |a_{kk}| + \ldots. + |a_{kn}| \\
&\le |a_{k1}| + a_{k2}| + \ldots.. + |a_{kk}| + \ldots. + |a_{kn}|
\end{aligned} \tag{28}$$

since $\left|\frac{x_{i,j}}{x_{i,k}}\right| \le 1$ for j = 1, 2, ... n.

Since eigenvalues of A and A^T are same Ref. P2), Eqn. (28) can also be written as

$$|\lambda_i| \le |a_{1k}| + |a_{2k}| + \ldots.. + |a_{kk}| + \ldots.. + |a_{nk}| \tag{29}$$

Since $|x_{i,k}|$, the largest element in magnitude, is unknown, we approximate Eqns. (28) and (29) by

$$|1 \;|\,, \quad \max_{i} \sum_{\substack{i=1 \\ j=i}}^{n} \left|a_{ij}\right| \text{ (maximum absolute row sum)} \qquad (30)$$

and

$$|l\ | \; f \quad \max_{j} \sum_{\substack{i=1 \\ j=i}}^{n} |a_{ij}| \text{ (maximum absolute column sum)} \qquad (31)$$

We can also rewrite Eqn. (27) in the form

$$|l_i - a_{kk}| = a_{k1}\left(\frac{x_{i,1}}{x_{i,k}}\right) + a_{k2}\left(\frac{x_{i,2}}{x_{i,k}}\right) + ... + a_{kn}\left(\frac{x_{i,n}}{x_{i,k}}\right)$$

and taking magnitude on both sides, we get

$$|l_i - a_{kk}| \;,\; \sum_{\substack{i=1 \\ j=i}}^{n} |a_{ij}| \qquad (32)$$

Again, since A and A^T have the same eigenvalues Eqn. (32) can be written as

$$|l_i - a_{kk}| \;,\; \sum_{\substack{i=1 \\ j=i}}^{n} |a_{ij}| \qquad (33)$$

Note that since the eigenvalues can be complex, the bounds (30), (31), (32) and (33) represents circles in the complex plane. If the eigenvalues are real, then they represent intervals. For example, when A is symmetric then the eigenvalues of A are real.

Again in Eqn. (32), since k is not known, we replace the circle by the union of the n circle

$$|l_i - a_{ii}| \;,\; \sum_{\substack{i=1 \\ j=i}}^{n} |a_{ij}|, \; i = 1, 2, ..., n. \qquad (34)$$

Similarly from Eqn. (33), we have that eigenvalues of A lie in the union of circles

$$|l_i - a_{ii}| \sum_{\substack{i=1 \\ j=i}}^{n} |a_{ij}| \;,\; , \; i = 1, 2, ..., n. \qquad (35)$$

The bounds derived in Eqns. (30), (31), (34) and (35) for eigenvalues are all independent bounds. Hence the eigenvalues must lie in the intersection of these bounds. The circles derived above are called the Gerschgorin circles and the bounds are called the Gerschgorin bounds.

Let us now consider the following examples:

Example 7:

Estimate the eigenvalues of the matrix

$$A = \begin{bmatrix} 1 & -1 & 2 \\ 2 & 1 & 3 \\ 1 & 3 & 2 \end{bmatrix}$$

using the Gerschgorin bounds.

Solution:

The eigenvalues of A lie in following regions:

i) absolute row sums are 4, 6 and 6. Hence
|l | , max [4, 6, 6] = 6 (36)

ii) absolute column sums are 4, 5 and 7. Hence
|l | , 7 (37)

iii) union of the circles [using (35)]
|l - 1| , 3
|l - 1| , 4
|l - 2| , 5
union of circles in (iii) is |l - 1| , 5 (38)
union of circles in (iv) is |l - 2| , 5 (39)

The eigenvalues lie in all circles (36), (37), (38) and (39) i.e., in the intersection of these circles as shown by shaded region in Fig. 1.

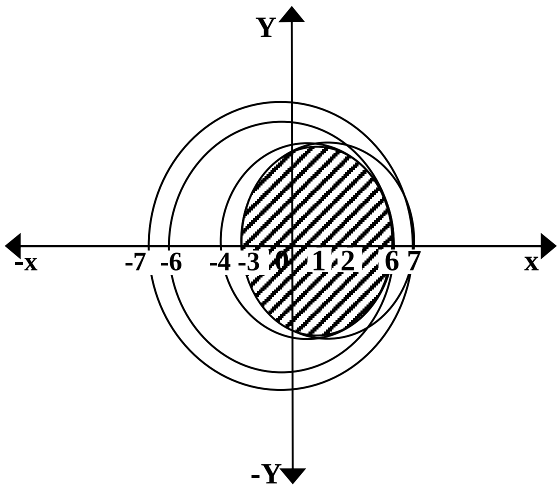

Fig. 1

Example 8:

Estimate the eigenvalues of the symmetric matrix

$$A = \begin{bmatrix} 1 & -1 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & -2 \end{bmatrix}$$

by the Gerschgorin bounds.

Solution:

The eigenvalues lie in the following regions:

i) | l | , max [4, 4, 6] = 6

ii) union of the circles
  a) | l - 1| , 3
  b) | l - 1| , 3
  c) | l + 1| , 4

Since A is symmetric, it has real eigenvalues. Therefore, the eigenvalues lie in the intervals

i) -6 , l , 6

ii) union of
  a) -3 , l -1 , 3, i.e. -2 , l , 4
  b) -4 , l +2 , 4, i.e. -6 , l , 2
  union of (a) and (c) is -6 , l , 4.

Intersection of (i) and (ii) is -6 , l , 4. Hence the eigenvalues of A lie i the interval -6 , l , 4.

Note that in Example 8, since the matrix A is symmetric, the bounds (30) and (31) are same and also the bounds (34) and (35) are same.

You may now try the following self assessment exercise.

4.0 CONCLUSION

We can now conclude as in summary.

5.0 SUMMARY

In this unit, we have covered the following:

1) For a given system of equations of the form

   $Ax = \lambda x$ (see Eqn. (1)).

   the values of λ for which Eqn. (1) has a nonzero solution are called the eigenvalues and the corresponding nonzero solutions (which are not unique) are called the eigenvectors of the matrix A.

2) The following are the steps involved in solving an eigenvalue problem

   i) Find the nth degree polynomial (called the characteristic equation) in λ from det $(A - \lambda I) = 0$.
   ii) Find the n roots λ_i, $i = 1, 2, ..., n$ of the characteristic equation.
   iii) Find the eigenvectors corresponding to each λ_i.

3) For $n \geq 3$, it may not be possible to find the roots of the characteristic equation exactly. In such cases, we use some iterative method like Newton Raphson method to find these roots. However,

   i) when only the largest eigenvalue in magnitude is to be obtained, we use the power method. In this method we obtain a sequence of vectors $\{y^{(k)}\}$, using the iiteative scheme

      $y^{(k+1)} = A y^{(k)}, k = 0, 1, ...$ (see Eqn. (13))

      which in the limit as $k \to \infty$, converges to the eigenvector corresponding to the dominant eigenvalue of the matrix A. The vector $y^{(0)}$ is an arbitrary non-zero vector (different from with the eigenvector of A).

   ii) we use the inverse power method with the iteration scheme

$$y^{(k+1)} = (A - qI)^{-1} v^{(k)},$$

i.e., $(A - qI)^{(k+1)} = v^{(k)}$, k = 0, 1, 2, ...
where $y^{(0)} = v^{(0)}$ is an arbitrary non-zero vector (not an eigenvector)

a) with q = 0, if only the least eigenvalue of A in magnitude and the corresponding eigenvector are to be obtained and

b) with any q, if the eigenvalue of A, nearest to some chosen number q and the corresponding eigenvector are to be obtained.

6.0 TUTOR-MARKED ASSIGNMENT (TMA)

1) Determine the Eigenvalues and the corresponding eigenvectors of the following

$$A = \begin{bmatrix} 1 & \sqrt{2} & 2 \\ \sqrt{2} & 3 & \sqrt{2} \\ 2 & \sqrt{2} & 1 \end{bmatrix}$$

2) $$A = \begin{bmatrix} 15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix}$$

3) $$A = \begin{bmatrix} 2 & 2 & -3 \\ 2 & 1 & -6 \\ 1 & -2 & 0 \end{bmatrix}$$

4) $$A = \begin{bmatrix} 2 & -1 & -1 \\ 3 & -2 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

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

6) $A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ 1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$

7) Find the smallest eigenvalue in magnitude and the corresponding eigenvector of the matrix

$$A = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}$$

with $v^{(0)}$ – [-1 1]^T, using four iterations of the power method.

8) Find the eigenvalue which is nearest to -1 and the corresponding eigenvector for the matrix

$$A = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}$$

with $v^{(0)}$ = [-1 1]^T, using four iterations of the inverse power method.

8) Using four iterations of the inverse power method, find the eigenvalue which is nearest to 5 and the corresponding eigenvector for the matrix

$A = \begin{bmatrix} 3 & 2 \\ 3 & 4 \end{bmatrix}$ (exact eigenvalues are = 1 and 6)

with $v^{(0)}$ = [1 1]^T

10) Estimate the eigenvalues of the matrix A given in Example 3(a) and 3(b), using the Gerschgorin bounds.

7.0 REFERENCES/FURTHER READINGS

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

MODULE 3

UNIT 1 REVIEW OF CALCULUS

CONTENTS

NUMERICAL ANALYSIS

Mathematical modelling of physical/biological problems generally gives rise to ordinary or partial differential equations or an integral equation or in terms of a set of such equation. A number of these problems can be solved exactly by mathematical analysis but most of them cannot be solved exactly. Thus, a need arises to devise numerical methods to solve these problems. These methods for solution of mathematical methods may give rise to a system of algebraic equations or a non-linear equation or system of non-linear equations. The numerical solution of these systems of equations is quantitative in nature but when interpreted give qualitative results and are very useful. Numerical analysis deals with the development and analysis of the numerical methods. We are offering this course of numerical analysis to students entering the Bachelor's Degree Programme as an elective subject.

It was in the year 1624 that the English mathematician, Henry Briggs used a numerical procedure to construct his celebrated table of logarithms. The interpolation problem was first taken up by Briggs but was solved by the 17th century mathematicians and physicists, Sir Isaac Newton and James Gregory. Later on, other problems were considered and solved by more and more efficient methods. In recent years the invention and development of electronic calculators/computers have strongly influenced the development of numerical analysis.

This course assumes the knowledge of the course MTH 112, MTH 122. They are prerequisite for this course. Number of results from linear algebra are also used in this course. These results have been stated wherever required. For details of these results our linear algebra course MTH 121 may be referred. This course is divided into 4 blocks. The first block, deals with the problem of finding approximate roots of a non-linear equation in one unknown. We have started the block with a recall of four important theorems from calculus which are referred to throughout the course. After introducing the concept of 'error' that arise due to approximations, we have discussed two basic approximation methods namely, bisection and fixed point iteration methods and two commonly used methods, namely. secant and Newton-Raphson methods. In Block 2, we have considered the problem of finding the solution of system of linear equations. We have discussed both direct and iterative methods of solving system of linear equations.

Block 3 deals with the theory of interpolation. Here, we are concerned only with polynomial interpolation. The existence and uniqueness of interpolating polynomials are discussed. Several form of interpolating polynomials like Lagrange's and Newton's divided difference forms with error terms are discussed. This block concludes with a discussion on Newton's forward and backward difference form.

In Block 4, using interpolating polynomials we have obtained numerical differentiation and integration formulae together with their error terms. After a brief introduction to difference equations the numerical solution of the first order ordinary differential equation is dealt with. More precisely, Taylor series, Euler's and second order Runge Kutta methods are derived with error terms for the solution of differential equations.

Each block consists 4 units. All the concepts given in the units are followed by a number of examples well as exercises. These will help you get a better grasp of the techniques discussed in this course. We have used a scientific calculator for doing computations throughout the course. While attempting the exercises given in the units, you would also need a calculator which is available at your study centre. The solutions/answers to the exercises in a unit are given at the end of the unit. We suggest that you look at them only after attempting the exercises. A list of symbols and notations are also given in for your reference.

You ma like to look up some more books on the subject and try to solve some exercises given in them. This will help you get a better grasp of the techniques discussed in this course. We are giving you a list of titles which will be available in your study centre for reference purposes.

Some useful books

Numerical Methods for Scientific and Engineering Computation by
M. K. Jain, S.R.K. Iyengar, R.K. Jain.

Elementary Numerical Analysis by
Samuel D. Conte and Carl de Boor.

NOTATION AND SYMBOLS

Î	belong to
É	contains
< (£)	less than (less than or equal to)
>(³)	greater than (greater than or equal to)
R	set of real numbers
C	set of complex numbers
n!	n(n-1) ... 3. 2. 1 (n factorial)
[ ]	closed interval
] [	open interval
|x|	absolute value of a number x
i.e.	that is
$\overset{n}{\underset{j=1}{\text{å}}} a_i$	$a_1 + a_2 + ... + a_n$
x® a	x tends to a
$\lim_{x\text{®} a} f(x)$	limit of f(x) as x tends to a
$P_n(x)$	nth degree polynomial
f'(x)	derivative of f(x) with respect to x
»	approximately equal to
a	alpha
b	beta
g	gamma
e	epsilon
p	pi
å	capital sigma
z	zeta

BLOCK INTRODUCTION

This is the first of the four blocks which you will be studying in the Numerical Analysis course. In this block we shall be dealing with the problem of finding approximate roots of a non-linear equation in one unknown. In the Elementary Algebra course you have studied some methods for solving polynomial equations of degree up to and including four. In this block we shall introduce you to some numerical methods for finding solutions of equation. These methods are applicable to polynomial and transcendental equations.

This block consists of four units. In Unit 1, we begin with a recall of our important theorems from calculus which are referred to throughout the course. We then introduce you to the concept of ‘error’ that arise due to approximation. In Unit 2, we shall discuss two types of errors that are common in numerical approximation methods, namely, bisection method and fixed point iteration method. Each of these methods involve a process that is repeated until an answer or required accuracy is achieved. These methods are known as iteration methods. We shall also discuss two accurate methods, namely, secant and Newton-Raphson methods in Unit 3. Unit 4, which is the last unit of this block, deals with the solutions of the most well-known class of equations, the polynomial equations. For finding the roots of polynomial equations we shall discuss Birge-Vieta and Graeffe’s root squaring methods.

As already mentioned in the course introduction, we shall be using a scientific calculator for doing computations throughout the block. While attempting the exercises given in this block, you would also need a calculator which is available at your centre. We therefore suggest you to go through the instructions manual, supplied with the calculator, before using it.

Lastly we remind you to through the solved examples carefully, and to attempt all exercises in each unit. This will help you to gain some practice over various methods discussed in this block.

1.0 INTRODUCTION

The study of numerical analysis involves concepts from various branches of mathematics including calculus. In this unit, we shall briefly review certain important theorems in calculus which are essential for the development and understanding of numerical methods. You are already familiar with some fundamental theorems about continuous functions from your calculus course. Here we shall review three theorems given in that course, namely, intermediate value theorem, Rolle’s Theorem and Lagrange’s mean value theorem. Then we state another important theorem in calculus due to B. Taylor and illustrate the theorem through various examples.

Most of the numerical methods give answers that are approximation to the desired solutions. In this situation, it is important to measure the accuracy of the approximate solution compared to the actual solution. To find the accuracy we must have an idea of the possible errors that can arise in computational procedures. In this unit we shall introduce you to different forms of errors which are common in numerical computations.

The basic ideas and result that we have illustrated in this unit will be used often throughout this course. So we suggest you go through this unit very carefully.

2.0 OBJECTIVES

After studying this unit you should be able to:

- apply
    - Intermediate value theorem
    - Rolle's Theorem
    - Lagrange's mean value theorem
    - Taylor's theorem;
- define the term 'error' in approximation
- distinguish between rounded-off error and truncation error and calculate these errors as the situation demands.

3.0 MAIN CONTENT

3.1 Three Fundamental Theorems

In this section we shall discuss three fundamental theorems, namely, intermediate value theorem, Rolle's Theorem and Lagrange's mean value theorem. All these theorems give properties of continuous functions defined on a closed interval [a, b]. we shall not prove them here, but we shall illustrate their utility with various examples. Let us take up these theorems one by one.

3.1.1 Intermediate Value Theorem

The intermediate value theorem says that a function that is continuous on a closed interval [a, b] takes on every intermediate value i.e., every value lying between f(a) and f(b) if $f(a) \neq f(b)$.

Formally, we can state the theorem a follows:

Theorem 1:

let f be a function defined on a closed interval [a, b]. let c be a number lying between f(a) and f(b) (i.e. f(a) < c < f(b) if f(a) < f(b) or f(b) < c < f(a) if f(b) < f(a)). Then there exists at least one point $x_0 \in [a, b]$ such that $f(x_0) = c$.

The following figure (Fig. 1) may help you to visualise the theorem more easily. It gives the graph of a function f.

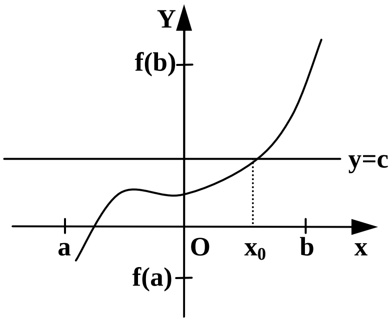

Fig. 1

In this figure f(a) < f(b). the condition f(a) < c < f(b) implies that the points (a, f(a)) and (b, f(b)) lie on opposite sides of the line y = c. This, together with the fact that f is continuous, implies that the graph crosses the line y = c at some point. In Fig. 1 you see that the graph crosses the line y = c at (x_0, c).

The importance of this theorem is as follows: If we have a continuous function f defined on a closed interval [a, b], then the theorem guarantees the existence of a solution of the equation f(x) = c, where c is as in Theorem 1. However, it does not say what the solution is. We shall illustrate this point with an example.

Example 1:

Find the value of x in $0 \le x \le \frac{\pi}{2}$ for which $\sin(x) = \frac{1}{2}$.

Solution: You know that the function f(x) = sin x is continuous on $\left(0, \frac{P}{2}\right)$. Since f(0) = 0 and $f\left(\frac{\pi}{2}\right) = 1$, we have $f(0) < \frac{1}{2} < f\left(\frac{\pi}{2}\right)$. thus f satisfies all the conditions of Theorem 1. Therefore, there exists at least one value of x, say x_0 such that $\sin(x_0) = \frac{1}{2}$, that is, the theorem guarantees that there exists a point x_0 such that sin $(x_0) = \frac{1}{2}$. Let us try to find this point from the graph of sin x in $\left(0, \frac{P}{2}\right)$ (see Fig. 2).

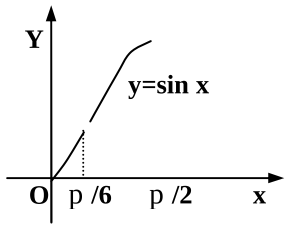

Fig. 2

From the figure, you can see that the line $x = \frac{1}{2}$ cuts the graph at the point $\left(\frac{p}{6}, \frac{1}{2}\right)$. Hence there exists a point $x_0 = \frac{p}{6}$ in $\left(0, \frac{P}{2}\right)$ such that sin $((x_0) = \frac{1}{2}$.

Let us consider another example.

Example 2:

Show that the equation $2x^3 + x^2 - x + 1 = 5$ has a solution in the interval [1, 2].

Solution:

Let $f(x) = 2x^3 + x^2 - x + 1$. Since f is a polynomial in x, f is continuous in [1, 2]. Also f(1) = 3, f(2) = 19 and 15 lies between f(1) and f(2). Thus f satisfied all conditions of Theorem 1. Therefore, there exists a number x_0 between 1 and 2 such that $f(x_0) = 5$. That is, the equation $2x^3 + x^2 - x + 1 = 5$ has solution in the interval [1, 2].

Thus we saw that the theorem enables us in establishing the existence of the solutions of certain equations of the type f(x) = 0 without actually solving them. In other words, if you want to find an interval in which a solution (or root) of f(x) = 0 exists, then find two numbers a, b such that f(a) f(b) < 0. Theorem 1, then states that the solution lies in]a, b[. We shall need some other numerical methods for finding the actual solution. We shall study the problem of finding solution of the equation f(x) = 0 more elaborately in Unit 2.

Let us now discuss another important theorem in calculus.

3.1.2 Rolle's Theorem

In this section we shall review the Rolle's Theorem. The theorem is named after the seventeenth century French mathematician Michel Rolle (1652 – 1719).

Theorem 2 (Rolle's Theorem):

Let f be a continuous function defined on [a, b] and differentiable on]a, b[. If f(a) = f(b), then there exists a number x_0 in]a, b[such that $f'(x_0) = 0$.

Geometrically, we can interpret the theorem easily. You know that since f is continuous, the graph of f is a smooth curve (see Fig. 3).

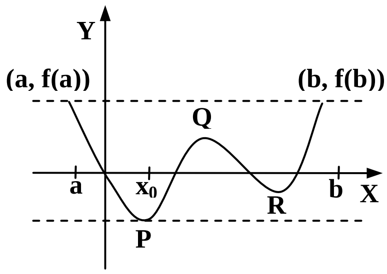

Fig. 3

You have already seen in your calculus course that the derivative $f'(x_0)$ at some point x_0 gives the slope of the tangent at $(x_0, f(x_0))$ to the curve y = f(x). Therefore the theorem states that if the end values f(a) and f(b) are equal, then there exists a point x_0 in]a, b[such that the slope of the tangent at the point $P(x_0, f(x_0))$ is zero, that is, the tangent is parallel to x-axis at the point (see Fig. 3). In fact we can have more than one point at which f'(x) = 0 as shown in Fig. 3. This shows that the number x_0 in Theorem 2 may not be unique.

The following example gives an application of Rolle's Theorem.

Example 3:

Use Rolle's Theorem to show that there is a solution of the equation cot x = x in $]0, \frac{\pi}{2}[$.

Solution: Here we have to solve the equation cot x – x = 0. We rewrite cot x – x as $\frac{\cos x \quad x\sin x}{\sin x}$. Solving the equation $\frac{\cos x \quad x\sin x}{\sin x} = 0$ in $]0, \frac{\pi}{2}[$ is same as solving the equation cos x – x sin x = 0. now we shall see whether we can find a function f which satisfies the conditions of

Rolle's Theorem and for which f'(x) = cos x – x sin x. Our experience in differentiation suggests that we try f(x) = x cos x. this function f is continuous in]0, $\frac{\pi}{2}$[, differentiable in]0, $\frac{\pi}{2}$[and the derivative f'(x) = cos x – x sin x. Also f(0) = 0 = f ($\frac{\pi}{2}$). Thus f satisfies all the requirements of Rolle's Theorem. Hence, there exists a point x_0 in]a, b[such that $f'(x_0) = \cos x_0 - x_0 \sin x_0 = 0$. This shows that a solution to the equation cot x – x = 0 exists in]0, $\frac{\pi}{2}$[.

Now, let us look at Fig. 3 carefully. We see that the line joining (a, f(a)) and (b, f(b)) is parallel to the tangent at $(x_0, f(x_0))$. Does this property hold when f(a) ≠ f(b) also? In other words, does there exists a point x_0 in]a, b[such that the tangent at $(x_0, f(x_0))$ is parallel to the line joining (a, f(a)) and (a, f(b))? The answer to this question is the content of the well-known theorem. "Lagrange's mean value theorem", which we discuss next.

3.1.3 Lagrange's Mean Value Theorem

This theorem was first proved by the French mathematician Count Joseph Louis Lagrange (1736 – 1813).

Theorem 3:

Let f be a continuous function defined on [a, b] and differentiable in]a, b[. Then there exists a number x_0 in]a, b[such that

$$f'(x_0) = \frac{f(b) - f(a)}{b - a} \qquad (1)$$

geometrically we can interpret this theorem as given in Fig. 4.

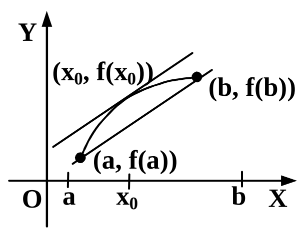

Fig. 4

In this figure you can see that the straight line connecting the end points (a, f(a)) and (b, f(b)) of the graph is parallel to some tangent to the curve at an intermediate point.

You may be wondering why this theorem is called ‘mean value theorem’. This is because of the following physical interpretation.

Suppose f(t) denotes the position of an object at time t. Then the average (mean) velocity during the interval [a, b] is given by

$$\frac{f(b) - f(a)}{b - a}$$

Now Theorem 3 states that this mean velocity during an interval [a, b] is equal to the velocity $f'(x_0)$ at some instant x_0 in]a, b[.

We shall illustrate the theorem with an example.

Example 4:

Apply the mean value theorem to the function $f(x) = \sqrt{x}$ in [0, 2] (see Fig. 5).

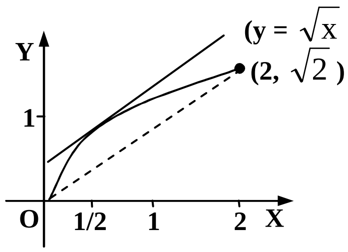

Fig. 5: Graph of f(x) = $\sqrt{x}$

Solution:

We first note that the function $f(x) = \sqrt{x}$ is continuous on [0, 2] and differentiable in]0, 2[and f’(x0 = $\frac{1}{2\sqrt{x}}$.

Therefore by Theorem 3, there exists a point x_0 in]0, 2[such the

$f(2) = \sqrt{2}$ and $f(0) = f'(x_0)\ (2 – 0)$

Now $f(2) = \sqrt{2}$ and $f(0) = 0$ and $f'(x_0) = \frac{1}{2x_0}$.

Therefore we have

$$\sqrt{2} = \frac{1}{x_0}$$

i.e. $\frac{1}{x_0} = \frac{1}{\sqrt{2}}$ and $x_0 = \frac{1}{2}$.

Thus we get that the line joining the end points (0, 0) and (2, $\sqrt{2}$) of the graph of f is parallel to the tangent to the curve at the point ($\frac{1}{2}$, $\frac{1}{\sqrt{2}}$).

We shall consider one more example.

Example 5:

Consider the function f(x) = (x – 1) (x – 2) (x – 3) in [0, 4]. Find a point x_0 in]0, 4[such that

$$f'(x_0) = \frac{f(4) - f(0)}{4 - 0}.$$

Solution: We rewrite the function f(x) as

$$f(x) = (x – 1) (x – 2) (x – 3) = x^3 – 6x^2 + 11x – 6$$

we know that f(x) is continuous on [0, 4], since f is a polynomial in x. Also the derivative

$$f'(x) = 3x^2 – 12x = 11$$

exists in]0, 4[. Thus f satisfies all conditions of the mean value theorem. Therefore, there exists a point x_0 in]0, 4[such that

$$f'(x_0) = \frac{f(4) - f(0)}{4 - 0}$$

i.e., $3x_0^2 - 12x_0 + 11 = \frac{6 + 6}{4 - 0} = 3$

i.e., $3x_0^2 - 12x_0 + 8 = 0$

This is a quadratic equation in x_0. The roots of this equation are

$$\frac{6 + 2\sqrt{3}}{8} \text{ and } \frac{6 - 2\sqrt{3}}{8}$$

Taking $\sqrt{3}$ = 1.732, we see that there are twp values for x_0 lying in the interval]0, 4[.

The above example shows that the number x_0 in Theorem 3 may not be unique. Again, as we mentioned in the case of theorems 1 and 2, the mean value theorem guarantees the existence of a point only.

So far we have used the mean value theorem to show the existence of a point satisfying Eqn. 1. Next we shall consider an example which shows another application of mean value theorem.

Example 6:

Find an approximate value of $\sqrt[3]{26}$ using the mean value theorem.

Solution:

Consider the function $f(x) = x^{1/3}$. Then $f(26) = \sqrt[3]{26}$. The number nearest to 26 for which the cube root is known is 27, i.e., $f(27) = \sqrt[3]{27} = 3$. Now we shall apply the mean value theorem to the function $f(x) = x^{1/3}$ in the interval $]26, 27[$. The function f is continuous in $[26, 27]$ and the derivative is

$$f'(x) = \frac{1}{3x^{2/3}}$$

Therefore, there exists a point x_0 between 26 and 27 such that

$$\sqrt[3]{27} - \sqrt[3]{26} = \frac{1}{3x_0^{2/3}}(27 - 26)$$

$$\text{i.e., } \sqrt[3]{26} = 3 - \frac{1}{3x_0^{2/3}} \qquad (2)$$

Since x_0 is close to 27, we approximate $\frac{1}{3x_0^{2/3}}$ by $\frac{1}{3(27)^{2/3}}$, i.e.;

$$\frac{1}{3x_0^{2/3}} \approx \frac{1}{27}$$

Substituting this value in Eqn. (2) we get

$$\sqrt[3]{26} = 3 - \frac{1}{27} = 2.963.$$

Note that in writing the value of we have rounded off the number after three decimal places. Using the calculator we find that the exact value of $\sqrt[3]{26}$ is 2.9624961.

We have given this example just to illustrate the usefulness of the theorem. The mean value theorem has got many other application which you will come across in later units.

Now we shall discuss another theorem in calculus.

3.2 Taylor's Theorem

You are already familiar with the name of the English mathematician Brook Taylor (1685 – 1731) from your calculus course. In this section we shall introduce you to a well-known theorem due to B. Taylor. Here we shall state the theorem without proof and discuss some of its applications.

You are familiar with polynomial equations of the form $f(x) = a_0 + a_1 x + ... + a_n x^n$ where $a_0, a_1 ....., a_n$ are real numbers. We can easily compute the value of a polynomial at any point x = a by using the four basic operation of addition, multiplication, subtraction and division. On the other hand there are function like e^x, cos x. In x etc. which occur frequently in all branches of mathematics which cannot be evaluated in the same manner. For example, evaluating the function f(x) = cos x at 0.524 is not so simple. Now, to evaluate such functions we try to approximate them by polynomial which are easier to evaluate. Taylor's theorem gives us a simple method for approximating functions f(x) by polynomials.

Let f(x) be a real-valued function defined on R which is n-times differentiable. Consider the function

$P_1(x) = f(x_0) + (x – x_0) f'(x_0)$
where x_0 is any given real number.

Now $P_1(x)$ is a polynomial in x of degree 1 and $P_1(x_0) = f(x_0)$ and $P'_1(x_0) = f'(x_0)$. The polynomial $P_1(x)$ is called the first Taylor polynomial of f(x) at x_0. Now consider another function

$$P_2(x_0) = f(x_0) + (x – x_0)f'(x_0) + \frac{(x - x_0)^2}{2!} f''(x_0).$$

Then $P_2(x)$ is a polynomial in x of degree 2 and $P_2(x_0) = f(x_0)$, $P'_2(x_0) = f'(x_0)$ and $P''_2(x_0) = f''(x_0)$. $P_2(x)$ is called the second Taylor polynomial of f(x) at x_0.

Similarly, we can define the r^{th} Taylor polynomial of f(x) at x_0 where $1 \leqslant r \leqslant n$. The r^{th} Taylor polynomial at x_0 is given by

$$P_r(x) = f(x_0) + (x – x_0) f'(x_0) + \dots + \frac{f^{(n)}(x_0)}{r!}(x – x_0)^r. \qquad (3)$$

You can check that $P_r(x_0) = f(x_0)$, $P'_r(x_0) = f'(x_0)$, ….

$P_r^{(r)}(x_0) = f^{(r)}(x_0)$ (see E6)
Let us consider an example.

Example 7:

Find the fourth Taylor polynomial of f(x) = In x about x_0=1.

Solution:

The fourth Taylor polynomial of f(x) is given by

$$P_4(x) = f(1) + (x - 1)f'(1) + \frac{(x-1)^2}{2!} f''(1) + \frac{(x-1)^3}{3!} f^{(3)}(1) + \frac{(x-1)^4}{4!} f^{(4)}(1).$$

Now, f(1) = In1 = 0

$f'(x) = \frac{1}{x}$; $f'(1) = 1$

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

$f^{(3)}(x) = \frac{2}{x^3}$; $f^{(3)}(1) = 2$

$f^{(4)}(x) = \frac{-6}{x^4}$; $f^{(4)}(1) = -6$

Therefore, $P_4(x) = (x - 1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4}$

We are now ready to state the Taylor's theorem.

Theorem 4 (Taylor's Theorem):

Let f be a real valued function having (n + 1) continuous derivatives on]a, b[for some n $\geqslant$ 0. Let x_0 be any point in the interval]a, b[. Then for any x $\in$]a, b[, we have

$$f(x) = f(x_0) + \frac{(x - x_0)}{1!} f'(x_0) + \frac{(x - x_0)}{2!} f^{(2)}(x_0) + \dots$$

$$+ \dots + \frac{(x - x_0)^n}{n!} f^{(n)}(x_0) + \frac{(x - x_0)^{n+1}}{n+1!} f^{n+1}(c). \qquad (4)$$

where c is point between x_0 and x.

The series given in Eqn. (4) is called the nth Taylor's expansion of f(x) at x_0.

We rewrite Eqn. (4) in the form
$f(x) = P_n(x) + R_{n+1}(x)$
where $P_n(x)$ is the nth Taylor polynomial of f(x) about x_0 and

$$R_{n+1}^{(x)} = \frac{(x - x_0)^{n+1}}{n+1!} f^{n+1}(c).$$

$R_{n+1}(x)$ depends on x, x_0 and n. $R_{n+1}(x)$ is called the remainder (or error) of the nth Taylor's expansion after n + 1 terms.

Suppose we put $x_0 = a$ and $x = a + h$ where $h > 0$, in Eqn (4). Then any point between a and a + h will be of the form $a +/\theta h$, $0 < \theta < 1$.

Therefore, Eqn (4) can be written as

$$f(a+h) = f(a)+h f'(a)+\frac{h^2}{2!} f''(a)+\ldots+\frac{h^n}{n!} f^{(n)}(a)+\frac{h^{n+1}}{n+1!} f^{(n+1)}(a+\theta h) \qquad (5)$$

Let us now make some remarks on the Taylor's theorem.

Remark 1: Suppose that the function f(x) in Theorem 4 is a polynomial of degree m. Then $f^{(r)}(x) = 0$ for all $r > m$. Therefore $R_{n+1}(x) = 0$ for all $n \geqslant m$. Thus, in this case, the m^{th} Taylor's expansion of f(x) about x_0 will be

$$f(x) = f(x_0) + \frac{(x - x_0)}{1!} f'(x_0) + \ldots + \frac{(x - x_0)^m}{m!} f^{(m)}(x_0).$$

Note that the right hand side of the above equation is simply a polynomial in $(x – x_0)$.

Therefore, finding Taylor's expansion of a polynomial function f(x) about x_0 is the same as expressing f(x) as a polynomial in $(x – x_0)$ with coefficients from R.

Remark 2:

Suppose we put $x_0 = a$, $x = b$ and $n = 0$ in Eqn. (4). Then Eqn (4) becomes
$f(b) = f(a) + f'(c)(b – a)$
or equivalently
$f(b) – f(a) = f'(c) (b – a)$
which is the Lagrange's mean value theorem. Therefore we can consider the mean value theorem as a special case of Taylor's theorem.

Let us consider some examples.

Example 8

Expand $f(x) = x^4 – 5x^3 + 5x^2 + x + 2$ in powers of $(x – 2)$.

Solution:

The function f(x) is a polynomial in x of degree 4. Hence, derivatives of all orders exists and are continuous. Therefore by Taylor's theorem, the 4^{th} Taylor expansion of f(x) about 2 is given by

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

Here $f(2) = 0$
$f'(x) = 4x^3 – 15x^2 + 10x + 1,$ $f'(2) = -7$
$f''(x) = 12x^2 – 30x + 10,$ $f''(2) = -2$
$f^{(3)}(x) = 24x – 30,$ $f^{(3)}(2) = 18$
$f^{(4)}(x) = 24,$ $f^{(4)}(2) = 24$

Hence the expansion is

$$f(x) = -7(x – 2) - \frac{2(x-2)^2}{2!} + \frac{18(x-2)^3}{3!} + \frac{24(x-2)^4}{4!}$$
$$= -7(x – 2) – (x – 2)^2 + 3(x – 2)^3 + (x – 2)^4$$

Example 9:

Find the nth Taylor expansion of 1n (1 + x) about x = 0 for $x \in]-1, 1[$.

Solution:

We first note that the point x = 0 lies in the given interval. Further; the function f(x) = 1n (1 + x) has continuous derivatives of all orders. The derivatives are given by

$f'(x) = \frac{1}{1+x},$ $f'(0) = 1$

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

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

.. .
.. .
.. .

$f^{(n)})x) = \frac{(-1)^{n-1}(n-1)!}{(1+x)^n}$, $f^{(n)}(0) = (-1)^{n-1}(n-1)!$

Therefore by applying Taylor's theorem we get that for any x $\in$]-1, 1[

$$1n\ (1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + \frac{(-1)^{n-1}x^n}{n} + \frac{(-1)^{n-1}n!\,x^{n+1}}{(n+1)!(1+c)^{n+1}}$$

where c is a point lying between 0 and x.

Now, let us consider the behaviour of the remainder in a small interval, say, [0, 0.5]. then for x in [0, 0.5], we have

$$|R_{n+1}(x)| = \left|\frac{(-1)^n n!\,x^{n+1}}{(n+1)!(1+c)^{n+1}}\right|$$

where 0 < c < x.

Since $|x| < 1$, $|x|^{n+1} < 1$ for any positive integer n.

Also since c > 0, $\frac{1}{(1+c)^{n+1}} < 1$. Therefore we have

$$|R_{n+1}(x)| < \frac{1}{n+1}$$

Now $\frac{1}{n+1}$ can be made as small as we like by choosing n sufficiently large i.e. $\lim_{n\to\infty} \frac{1}{n+1} = 0$. This shows that $\lim_{n\to\infty} |R_{n+1}(x)| = 0$.

The above example shows that if n is sufficient large, the value of the nth Taylor polynomial $P_n(x)$ at any x_0 will be approximately equal to the value of the given function $f(x_0)$. In fact, the remainder $R_{n+1}(x)$ tell(s) us how close the value $P_n(x_0)$ is to $f(x_0)$.

Now we shall make some general observations about the remainder $R_{n+1}(x)$ in the Taylor's expansion of a function f(x).

Remark 3: Consider the nth Taylor expansion of f about x_0 given by
$f(x) = P_n(x) + R_{n+1}(x)$.

Then $R_{n+1}(x) = f(x) - P_n(x)$. If $\lim_{n\to\infty} R_{n+1}(x) = 0$ for some x, then for that x we say that we can approximate f(x) by $P_n(x)$ and we write f(x) as the infinite series.

$$f(x) = f_0(x) + f'(x)(x-x_0) + \frac{f^{(2)}(x_0)}{2!}(x - x_0)^2 + \ldots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \ldots$$

$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} x^n$$

is called Maclaurin's series.

Remark 4: If the remainder $R_{n+1}(x)$ satisfies the condition that $|R_{n+1}(x)| < M$ for some n at some fixed point x = a, then M is called the bound of the error at x = a.

In this case we have

$|R_{n+1}(x)| = |f(x) - | < M$
That is, f(x) lies in the interval $]P_n(x) – M, P_n(x) + M[$.

Now if M is considerably small for some n, then this interval becomes very small. In this case we say that f(x) is approximately equal to the value of the nth Taylor polynomial with error M. Thus the remainder is used to determine a bound for the accuracy of the approximation.

We shall explain these concepts with an example.

Example 10:

Find the 2nd Taylor's expansion of $f(x) = \sqrt{1+x}$ in]-1, 1[about x = 0. find the bound of the error at x = 0.2.

Solution:

Since $f(x) = \sqrt{1+x}$, we have
$f(0) = 1$
$f'(x) = \frac{1}{2\sqrt{1+x}}$, $f(0) = \frac{1}{2}$
$f''(x) = -\frac{1}{4}(1 + x)^{-3/2}$, $f'(0) = -\frac{1}{4}$
$f^{(3)}(x) = \frac{3}{8}(1 + x)^{-5/2}$,

Applying Taylor's theorem to f(x), we get

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3(1 + c)^{-5/2}$$

where c is a point lying between 0 and x.

The error is given by $R_3(x) = \frac{x^3}{16}(1 + c)^{-5/2}$.

When x = 0.2, we have

$$R_3(0.2) = \frac{(0.2)^3}{16(1+c)^{5/2}}$$

Where 0 < c < 0.2. Since c > 0 we have

$$\left|\frac{1}{(1+c)^{5/2}}\right| < 1.$$

Hence,

$$|R_3(0.2)| \leqslant \frac{(0.2)^3}{16} = (0.5)\ 10^{-3}$$

Hence the bound of the error for n = 2 at x = 0.2 is (0.5) 10^{-3}.

There are some functions whose Taylor's expansion is used very often. We shall list their expansion here.

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!}e^c \ldots \tag{7}$$

$$\text{Sin } x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots + \frac{(-1)^{n-1}x^{2n-1}}{(2n-1)!} + \frac{(-1)^n x^{2n-1}}{(2n+1)!}\cos(c) \tag{8}$$

$$\text{Cos } x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots + \frac{(-1)^n x^{2n}}{(2n)!} + \frac{(-1)^{n+1}x^{2n+2}}{(2n+2)!}\cos(c). \tag{9}$$

$$\frac{1}{1-x} = 1 + x + x^2 + \ldots + x^n + \frac{x^{n+1}}{(1-c)^{n+2}} \tag{10}$$

where c, in each expansion, is as given in Taylor's theorem.

Now, let us consider some examples that illustrate the use of finding approximate values of some functions at certain points using truncated Taylor series.

Example 11:

Using Taylor's expansion for sin x about x = 0, find the approximate value of sin 10° with error less than 10^{-7}.

Solution:

The nth Taylor's expansion for sin x given in Eqn. (9) is

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots + \frac{(-1)^{n-1}x^{2n-1}}{(2n-1)!} + \frac{(-1)^n x^{2n+1}}{(2n+1)!}\cos(c). \qquad (11)$$

where x is the angle measured in radians.

Now, in radian measure , we have

$10^\circ = \frac{\pi}{18}$ radians.

Therefore, by putting $x = \frac{\pi}{18}$ in Eqn. (11) we get

$$\sin\frac{\pi}{18} = \frac{\pi}{18} - \frac{1}{3!}\left(\frac{\pi}{18}\right)^3 + \frac{1}{5!}\left(\frac{\pi}{18}\right)^5 + \ldots + R_{n+1}\left(\frac{\pi}{18}\right)$$

where $R_{n+1}(\frac{\pi}{18})$ is the remainder after (n + 1) terms.

Now

$$R_{n+1}\left(\frac{\pi}{18}\right) = \frac{(-1)^n}{(2n+1)!}\left(\frac{\pi}{18}\right)^{2n+1}\cos c.$$

If we approximate $\sin \frac{\pi}{18}$ by $P_n(\frac{\pi}{18})$, then the error introduced will be less than 10^{-7} if

$$\left|\sin\left(\frac{\pi}{18}\right) - P_n\left(\frac{\pi}{18}\right)\right| = \left|R_{n+1}\left(\frac{\pi}{18}\right)\right| = \left|\frac{(-1)^n}{(2n+1)!}\left(\frac{\pi}{18}\right)^{2n+1}\cos c\right| < 10^{-7}.$$

Maximizing cos c, we require that

$$\frac{1}{(2n+1)!}\left(\frac{\pi}{18}\right)^{2n+1} < 10^{-7} \qquad (12)$$

Using the calculator, we find that the value of left hand side of Eqn. (12) for various n is

n	1	2	3
Left hand side	89×10^{-3}	13×10^{-5}	99×10^{-9}

From the table we find that the inequality in (12) is satisfied for n = 3. Hence the required approximation is

$$\sin\left(\frac{\pi}{18}\right) \approx \frac{\pi}{18} - \frac{1}{3!}\left(\frac{\pi}{18}\right)^3 + \frac{1}{5!}\left(\frac{\pi}{18}\right)^5 = 0.1745445$$

with error less than 1.0×10^{-7}.

Let us now find the approximate value of e using Taylor’s theorem.

Example 12:

Using Maclaurin's series for e^x, show that $e \approx 2.71806$ with error less than 0.001. (Assume that $e < 3$).

Solution:

The Maclaurin's series for e^x is

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots$$

Putting x = 1 in the above series, we get

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \dots$$

Now we have to find n for which

$|e - P_n(1)| = |R_{n+1}(1)| < 0.001.$

Now $|R_{n+1}(1)| \leqslant e^c \dfrac{1}{(n+1)!}$

Since we have chosen $x_0 = 0$ and $x = 1$, the value c lies between 0 and 1 i.e. $0 < c < 1$. Since $e^c < c < 3$, we get

$$|R_{n+1}(1)| \leqslant e^c \frac{3}{(n+1)!}$$

The bound for $R_{n+1}(1)$ for different n is given in the following table.

n	1	2	3	4	5	6
Bounds for R_{n+1}	1.5	.5	.1	.125	.004	.0006

From this table, we see that

$R_{n+1} < .001$ if n = 6

Thus $P_6(1)$ is the desired approximation to e. i.e.

$$e \approx 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \frac{1}{720} + \frac{1957}{720} \approx 2.71806$$

In numerical analysis we are concerned with developing a sequence of calculations that will give a satisfactory answer to a problem. Since this process involves a lot of computations, there is a chance for the presence of some errors in these computations. In the next section we shall introduce you to the concept of 'errors' that arise in numerical computations.

3.3 Errors

In this section we shall discuss the concept of an 'error'. We consider two types of errors that are commonly encountered in numerical computations.

You are already familiar with the rounding off a number which has non-terminal decimal expansion from your school arithmetic. For example we use 3.1425 for 22/7. These rounded off numbers are approximations of the actual values. In any computational procedure we make use of these approximate values instead of the true values. Let x_T denote the true value and x_A denote the approximate value. How do we measure the goodness of an approximation x_A to x_T? The simplest measure which naturally comes to our mind is the difference between x_T and x_A. This measure is called the 'error'. Formally, we define error as a quantity which satisfies the identity.

True value x_T = Approximate value x_A + error.

Now if an 'error' in approximation is considered small (according to some criterion), then we say that 'x_A is a good approximation to x'.

Let us consider an example.

Example 13:

The true value of π is 3.14159265 … In some mensuration problems the value 22/7 is commonly used as an approximation to π. What is the error in this approximation?

Solution:

The true value of π is

$$\pi = 3.14159265 \qquad (13)$$

Now, we convert 22/7 to decimal form, so that we can find the difference between the approximate value and true value. Then the approximate value of π is

$$\frac{22}{7} = 3.14285714 \qquad (14)$$

Therefore,

$$\text{error} = \text{True value} - \text{approximate value} = -0.00126449 \qquad (15)$$

Note that in this case the error is negative. Error can be positive or negative. We shall in general be interested in absolute value of the error which is defined as

|error| = |True value – approximate value|
For example, the absolute Error in Example 13 is
|error | = |-0.00126449…| = 0.00126…
Sometimes, when the true value is very small we prefer to study the error by comparing it with the value. This is known as Relative error and we define this error as

$$|\text{Relative error}| = \left|\frac{\text{True value - approximate value}}{\text{True value}}\right|$$

In the case of Example 13,

$$|\text{Relative error}| = \frac{0.00126449...}{3.14159265...} = 0.00040249966\ldots$$

But note that in certain computations, the true value may to be available. In that case we replace the true value by the computed approximate value by the computed approximate value in the definition of relative error.

In numerical calculations, you will encounter mainly two types of errors: round-off error and truncation error. We shall discuss these errors in the next two subsections 1.4.1 and 1.4.2 respectively.

3.3.1 Round-off Error

Let us look at Example 13 again. You can see that the numbers appearing in Eqn. (13), (14) and (15) consists of 8 digits after the decimal point followed by dots. The line of dots indicates that the digits continue and we are not able to write all of them. That is, these numbers cannot be represented exactly by a terminating decimal expansion. Whenever we use much numbers in calculations we have to decide how many digits we are going to take into account. For example, consider again the approximate value of π. If we approximate π using 2 digits after the decimal point (say), chopping off the other digits, then we have

π = 3.14

The error in this approximation is
error = 0.00159265 (16)

If we use 3 digits after the decimal point, then using chopping we have
$\pi \approx 3.141$

In this case the error is given by
error = -0.00059265 (17)

Now suppose we consider the approximate value rounded-off to three decimal places. You already know how to round off a number which has non-terminal decimal expansion. Then the value of π rounded-off to 3 digits is 3.142. The error in this case is

error = -0.00040734…
which is smaller, in absolute value than 0.00059265…given in Eqn. (17). Therefore in general whenever we want to use only a certain number of digits after the decimal point, then it is always better to use the value rounded-off to that many digits because in this case the error is usually small. The error involved in a process where we use rounding-off method is called round-off error.

We now discuss the concept of floating point arithmetic.

In scientific computations a real number x is usually represented in the form

$x = \pm (. d_1 d_2 \dots d_n)\, 10^m$
where $d_1\ d_2 \dots d_n$ are natural numbers between 0 and 9 and m is an integer called exponent. Writing a number in this form is known as floating point representation. We denote this representation by fl(x). Such a floating point number is said to be normalized if $d_1 \neq 0$. To translate a number into floating point representation we adopt any of the two methods – rounding and chopping. For example, suppose we want to represent the number 537 in the normalized floating point representation with n = 1, then we get

fl (537) = $.5 \times 10^3$ chopped
= $.5 \times 10^3$ rounded

In this case we are getting the same representation in rounding and chopping. Now if we take n = 2, then we get

fl (537) = $.53 \times 10^3$ chopped
= $.54 \times 10^3$ rounded

In this case, the representations are different.

Now if we take n = 3, then we get

fl (537) = $.537 \times 10^3$ chopped

$= .537 \times 10^3$ rounded

The number n in the floating point representation is called precision.

The difference between the true value of a number x and rounded fl(x) is called round-off error. From the earlier discussion it is clear that the round-off error decreases when precision increases.

Mathematically, we define these concepts as follows:

Definition 2:

Let x be a real number and x* be a real number having non-terminal decimal expansion, then we say that x* represents x rounded to k decimal places if

$|x - x^*| \leqslant \frac{1}{2}\, 10^{-k}$, where $k > 0$ is a positive integer.

Next definition gives us a measure by which we can conclude that the round-off error occurring in an approximation process is negligible or not.

Definition 3:

Let x be a real number and x* be an approximation to x. Then we say that x* is accurate to k decimal places if

$$\frac{1}{2}10^{-(k+1)} \leqslant |x - x^*| \leqslant \frac{1}{2}10^{-k} \tag{18}$$

Let us consider an example.

Example 14:

Find out to how many decimal places the value of 22/7 obtained in Example 13 is accurate as an approximation to $\pi = 3.14159265$?

Solution:

We have already seen in Example 13 that

$$\left|\pi - \frac{22}{7}\right| = 0.00126449\ldots$$

Now $.0005 < .00126\ldots < 0.005$

or $\frac{1}{2}10^{-3} < .00126\ldots < \frac{1}{2}10^{-2}$

Therefore the inequality (18) is satisfied for k = 2.

Hence, by Definition 3, we conclude that the approximation is accurate to 2 decimal places.

Now we make an important remark.

Remark 5:

Round-off errors can create serious difficulties in lengthy computations. Suppose we have a problem which involves a long calculation. In the course of these computations many rounding errors (some positive, and some negative) may occur in a number of ways. At the end of the calculations these errors will get accumulated and we don't know the magnitude of this error. Theoretically it can be large. But, in reality some of these errors (between positive and negative errors) may get cancelled so that the accumulated error will be much smaller.

Let us now define another type of error called Truncation error.

3.3.2 Truncation Error

We shall first illustrate this error with a simple example. In Sec. 1.3. we have already discussed how to find approximate value of a certain function f(x) for a given value of x using Taylor's series expression. Let

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n$$

denote the Taylor's series of f(x) about x_0. In practical situations, we cannot, in general, find the sum of an infinite number of terms. So we must stop after a finite number of terms, say N. This means that we are taking

$$f(x) = \sum_{n=0}^{N} (x - x_0)^n$$

and ignoring the rest of the terms, that is, $\sum_{n=N+1}^{\infty} a_n (x - x_0)^n$

There is an error involved in this truncating process which arises from the terms which we exclude. This error is called the 'truncation error'. We denote this error by T E. Thus we have

$$T\,E = f(x) - \sum_{n=0}^{N} a_n\,(x - x_0)^n \quad \sum_{n=N+1}^{\infty} a_n\,(x - x_0)^n$$

You already know how to calculate this error from Sec. 1.3. There we saw that using Taylor's theorem we can estimate the error (or remainder) involved in a truncation process in some cases.

Let's see what happen if we apply Taylor's theorem to the function f(x) about the point $x_0 = 0$. We assume that f satisfies all conditions of Taylor's theorem. Then we have

$$f(x) = \sum_{n=0}^{N} a_n\,x^n + \frac{x^{N+1}}{N+1!}\,f^{N+1}(c) \qquad (19)$$

where $a_n = \dfrac{f^{(n)}(0)}{n!}$ and $0 < c < x$.

now, suppose that we want to approximate f(x) by $\sum_{n=0}^{N} a_n\,x^n$.

Then Eqn. (19) tells us that the truncation error in approximating f(x) by $\sum_{n=0}^{N} a_n\,x^n$ is given by

$$T\,E = R_{N+1}(x) = \frac{x^{N+1}}{N+1!}\,f^{N+1}(c) \qquad (20)$$

Theoretically we can use this formula for truncation error for any sufficiently differentiable function. But practically it is not easy to calculate the nth derivative of many functions. Because of the complexity in differentiation of such functions, it is better to obtain indirectly their Taylor polynomials by using one of the standard expansions we have listed in Sec. 1.3.

For example consider the function $f(x) = e^{x^2}$. It is difficult to calculate the nth derivative of this function. Therefore, for convenience, we obtain Taylor's expansion of e^{x^2} using Taylor's expansion of e^y by putting $y = x^2$. We shall illustrate this in the following example.

Example 15:

Calculate a bound for the truncation error in approximation e^{x^2} by

$$e^{x^2} \approx 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} \text{ for } x \in\,]-1, 1[.$$

Solution:

Put $u = x^2$. Then $e^{x^2} = e^u$. Now we apply the Taylor's theorem to function $f(u) = e^u$ about $u = 0$. Then, we have

$$e^u = 1 + u + \frac{u^2}{2!} + \frac{u^3}{3!} + \frac{u^4}{4!} + R_5(u) \text{ where}$$

$$R_5(u) = \frac{e^c u^5}{5!}$$

And $0 < c < u$. Since $|x| < 1$, $u = x^2 < 1$ i.e. $c < 1$. Therefore, $e^c < e < 3$. Thus

$$|R_5(u)| \leqslant \left|\frac{3x^{10}}{5!}\right| < \frac{3}{5!} = \frac{1}{40} = .025$$

Hence the truncation error in approximating e^{x^2} by the above expression is less than 25×10^{-1}.

If the absolute value of the TE is less, then we say that he approximation is good.

Now, in practical situations we should be able to find out the value of n for which the summation $\sum a_n x^n$ gives a good approximation to f(x). For this we always specify the accuracy (or error bound) required in advance. Then we find n using formula (20) such that the absolute error $|R_{n+1}(x)|$ is less than the specified accuracy. This gives the approximation within the prescribed accuracy.

Let us consider an example.

Example 16:

Find an approximate value of the integral

$$\int_0^1 e^{x^2}\, dx$$

with an error less than 0.025

Solution:

In Example 15 we observed that

$$e^{x^2} \approx 1 + \frac{x^2}{1!} + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}$$

with TE = $\frac{e^{x^2} x^{10}}{5!}$ dx.

Now we use this approximation to calculate the integral. We have

$$\int_0^1 e^{x^2} dx \approx \int_0^1 (1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!})dx \qquad (20)$$

with the truncation error

$$TE = \int_0^1 \frac{e^{x^2} x^{10}}{5!} dx.$$

We have

$$|TE| \int_0^1 \frac{e^{x^2} |x|^{10}}{5!} \leqslant \frac{3}{5!} = .25 \times 10^{-1}$$

Integrating the right hand side of (21), we get

$$\int_0^1 e^{x^2} \approx \int_0^1 (1 + x^2 + \frac{x^4}{2!} \left(+ \frac{x^6}{3!} + \frac{x^8}{4!})dx \right) =$$

$$x + \frac{x^3}{3} + \frac{x^5}{5 \times 2!} + \frac{x^7}{7 \times 3!} + \frac{x^9}{9 \times 4!} \Big|_0^1$$

$$= \left(x + \frac{x^3}{3} + \frac{x^5}{10} + \frac{x^7}{42} + \frac{x^9}{216} \right)_0^1$$

$$= 1 + \frac{1}{2} + \frac{1}{10} + \frac{1}{40} + \frac{1}{216}$$

= 0.0048

Here is an important remark.

Remark: The magnitude of the truncation error could be reduced within any prescribed accuracy by retaining sufficient large number of terms. Likewise the magnitude of the round-off error could be reduced by retaining additional digits.

You can now try the following self assessment exercises.

SELF ASSESSMENT EXERCISE

a) Calculate a bound for the truncation error in approximation f(x) = sin x by

$$\sin x \approx 1 - \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} \text{ where } -1 \leqslant x \leqslant 1.$$

b) Using the approximation in (a), calculate an approximate value of the integral

$$\int_0^1 \frac{\sin x}{x} dx$$

with an error 10^{-4}.

SELF ASSESSMENT EXERCISE

a) Calculate the truncation error in approximating

$$e^{-x^2} \text{ by } 1 - x^2 + \frac{x^4}{2}, -1 \leqslant x \leqslant 1.$$

b) Using the approximation in (a) calculate an approximate value of

$\int_0^1 e^{-x^2} dx$ within an error bound of 10^{-7}.

4.0 CONCLUSION

We end this unit by summarizing what we have learnt in this unit.

5.0 SUMMARY

In this unit we have:

- recalled three important theorems in calculus, namely

  i) Intermediate value theorem
  ii) Rolle's Theorem
  iii) Lagrange's mean value theorem

- State Taylor's theorem and demonstrated it with the help of examples.
  The nth Taylor's expansion:

$$f(x) = f(x_0) + \frac{(x \quad x_0)}{1!} f'(x_0) + \frac{(x \quad x_0)^2}{2!} f^{(2)}(x_0) + \ldots$$

$$\ldots + \frac{(x - x_0)^n}{n!} f^{(n)}(x_0) + \frac{(x - x_0)^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

- Defined the term 'error' occurring in numerical computations.
- Discussed two types of errors namely

  i) Round-off error: Error occurring in computations where we use rounding off method to represent a number is called round-off error.

  ii) Truncation error: Error occurring in computations where we use truncation process to represent the sum of an infinite number of terms.

- Explained how Taylor's theorem is used to calculate the truncation error.

6.0 TUTOR-MARKED ASSIGNMENT

1) Show that the following equations have a solution in the interval given alongside.

2) Using Rolle's Theorem show that there is a solution to the equation tan x – 1 + x = 0 in]0, 1[.

3) Let $f(x) = \frac{1}{3}x^3 + 2x$. Find a number x_0 in]0, 3[such that

   $f'(x_0) = \frac{f(3) - f(0)}{3 - 0}$

4) Find all numbers x_0 in the interval]-2, 1[for which the tangent to the graph of $f(x) = x^3 + 4$ is parallel to the line joining the end points (-2, f(-2)) and (1, f(1)).

5) Show that Rolle's Theorem is a special case of mean value theorem.

6) If P_r denotes the rth Taylor polynomial as given y Eqn (3), then show that $P_r(x_0) = f(x_0)$, $P'_r(x_0) = f'(x_0)$, ... $P_r^{(r)}(x_0) = f^{(r)}(x_0)$.

7) Obtain the third Taylor polynomial of $f(x) = e^x$ about x = 0.

8) Obtain the nth Taylor expansion of the function $f(x) = \frac{1}{1+x}$ in $]-\frac{1}{2}, 1[$ about $x_0 = 0$.

9) Does $f(x) = \sqrt{x}$ have a Taylor series expansion about $x = 0$? Justify your answer.

10) Obtain the 8^{th} Taylor expansion of the function $f(x) = \cos x$ in $[-\frac{\pi}{4}, \frac{\pi}{4}]$ about $x_0 = 0$. Obtain a bound for the error $R_9(x)$.

11) Using Maclaurin's expansion for cos x, find the approximate value of $\cos \frac{\pi}{4}$ with the error bound 10^{-5}.

12) How large should n be chosen in Maclaurin's expansion for e^x to have $|e^x - P_n(x)| \leqslant 10^{-5}$, $-1 \leqslant x \leqslant 1$.

13) In some approximation problems where graphic methods are used, the value $\frac{355}{133}$ is used as an approximation to π = 3.14159265….To how many decimal places the value $\frac{355}{133}$ is accurate as an approximation to π?

7.0 REFERENCES/FURTHER READINGS

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

UNIT 2 REVIEW OF CALCULUS

CONTENTS

1.0 INTRODUCTION

We often come across equation of the form $x^4 + 3x^2 + 2x + 1 = 0$ or $e^x = x - 2$ or tan h x = x etc. Finding one or more values of x which satisfy these equations is one of the important problems in Mathematics. From your elementary algebra course, you are already familiar with some methods of solving equations of degrees 1, 2, 3 and 4 equations of degree 1, 2, 3 and 4 are called linear, quadratic, cubic and biquadratic respectively. There you might have realized that it is very difficult to use the methods available for solving cubic and biquadratic equations. In fact no formula exists for solving equations of degree $n \geqslant 5$. In these cases we take recourse to approximate methods for the determination of the solution of equations of the form.

$$f(x) = 0 \qquad (1)$$

The problem of finding approximate values of roots of polynomial equations of higher degree was initiated by Chinese mathematicians. The methods of solution in various forms appeared in the 13th century work che' in kiu-shoo. The first noteworthy work in this direction was done in Euope by the English mathematician Fibonacci. Later in the year 1600 Vieta and Isaac Newton made significant contribution to the theory.

In this unit as well as in the next two units we shall discuss some numerical methods which gives an approximate solution of an equation f(x) = 0. We can classify the methods of solution into two types namely (i) Direct methods and (ii) Iteration methods. Direct methods produce solution by in finite number of steps whereas iteration methods give an approximate solution by repeated application of a numerical process. You will find later that for using iteration methods we have to start with an approximate solution. Iteration methods improve this approximate solution. We shall begin this unit by first discussing methods which enable us to determine an initial approximate solution and then discuss iteration methods to refine this approximate solution.

2.0 OBJECTIVES

After studying this unit you should be able to:

- find an initial approximation of the root using (1) tabulation method (2) graphical method.
- use bisection method for finding approximate roots.
- use fixed point iteration method for finding approximate roots.

3.0 MAIN BODY

3.1 Initial Approximation to a Root

You know that in many problems of engineering and physical sciences you come across equations in one variable of the form f(x) = 0.

For example, in Physical, the pressure-volume-temperature relationship of real gases can be described by the equation

$$PV = RT + \frac{\beta}{V} + \frac{r}{V^2} + \frac{s}{V^3} \qquad (2)$$

where P, V, T are pressure, volume and temperature respectively. R, β, r, s are constants. We can rewrite Eqn. (2) as

$$PV^4 - RTV^3 - \beta V^3 - rV - s = 0 \qquad (3)$$

Therefore the problem of finding the specific volume of a gas at a given temperature and pressure reduces to solving the biquadratic equation Eqn. (3) for the unknown variable V.

Consider another example in life sciences, the study of genetic problem of recombination of chromosomes can be described in the form

$$p(1 - p) = p^2 - p + k - 0,$$

where p stands for the recombination fraction with the limitation $0 \leqslant p \leqslant \frac{1}{2}$ and (1 – p) stands for the non-recombination fraction. The problem of finding the recombination fraction of a gene reduces to the problem of finding roots of the quadratic equation $p^2 - p + k = 0$.

In these problems we are concerned with finding value (or values) of the unknown variable x that satisfies the equation f(x) = 0. the function f(x) may be a polynomial of the form

$$f(x) = a_0 + a_1 x + ... + a_n x_n$$

or it may be a combination of polynomials, trigonometric, exponential or logarithmic functions. By a root of this equation we mean a number x_0 such that $f(x_0) = 0$. The root is also called a zero of f(x).

If f(x) is linear, then Eqn. (1) is of the form ax + b = 0, $a \neq 0$ and it has only one root given by $x = -\frac{b}{a}$. Any equation which is not linear is called a non-equation. In this unit we shall discuss some methods for finding roots of the equation f(x) = 0 where f(x) is a non linear function. You are already familiar with various methods for calculating roots of quadratic, cubic and biquadratic equations. But there is no such formula for solving polynomial equations of degree more than 4 or even for a simple equation like

$$x - \cos x = 0$$

Here we shall discuss some of the numerical approximation methods. These methods involve two steps:

Step 1: To find an initial approximation of a root.

Step 2: To improve this approximation to get a more accurate value.

We first consider step 1. Finding an initial approximation to a root means locating (or estimating) a root of an equation approximately. There are two ways for achieving this-tabulation method and graphical method.

Let us start with Tabulation method.

3.1.1 Tabulation Method

This method is based on the intermediate value theorem (IV Theorem), (see Theorem 1, Unit 1). Let us try to understand the various steps involved in the method through an example.

Suppose we want to find a root of the equation

$2x – \log_{10}x = 7$

We first compute value of $f(x) = 2x – \log_{10}x – 7$ for different value of x, say x = 1, 2, 3 and 4.

When x = 1, we have $f(1) = 2 – \log_{10}1 – 7 = -5$

Similarly, we have

$f(2) = 4 – \log_{10}2 – 7 = -3.301$
(Note that $\log_{10}2$ is computed using a scientific calculator.)

$f(3) = 6 – \log_{10}2 – 7 = -1.477$
$f(4) = 8 – \log_{10}4 – 7 = -0.3977$

These values are given in the following table:

Table 1

x	1	2	3	4
f(x)	-5	-3.301	-1.477	0.397

We find that f(3) is negative and f(4) is positive. Now we apply IV Theorem to the function $f(x) = 2x – \log_{10}x – 7$ in the interval $I_1 = [3, 4]$. Since f(3) and f(4) are of opposite signs, by IV theorem there exists a number x_0 lying between 3 and 4 such that $f(x_0) = 0$. That is, a root of the function lies in the interval]3, 4[. Note that this root is positive.

Let us now repeat the above computations for some values of x lying in]3, 4[say x = 3.5, 3.7 and 3.8. In the following table we report the values of f(x).

Table 2

x	3.5	3.7	3.8
f(x)	-0.544	-0.168	0.0202

We find that f(3.7) are of opposite signs. By applying IV theorem again to f(x) in the interval $I_2 = [3.7, 3.8]$, we find that the root of f(x) lies in the interval]3.7, 3.8[. Note that this interval is smaller than the previous interval. We call this interval a refinement of the previous interval. Let us repeat the above procedure once again for the interval I_2. In Table 3 we give the values of f(x) for some x between 3.7 and 3.8.

Table 3

x	3.75	3.78	3.79
f(x)	-0.074	-0.017	-0.00137

Table 3 shows that the root lies within the interval]3.78, 3.79[and this interval is much smaller compared to the original interval]3, 4[. The procedure is terminated by taking any value of x between 3. 78 and 3.79 as an approximate value of the root of the equation $f(x) = 2x – \log_{10}x – 7 = 0$.

The method illustrated above is known as Tabulation method. Let us write the steps involved in the method.

Step 1:

Select some numbers $x_1, x_2, ..., x_n$ and calculate $f(x_1)$ and $f(x_2), ..., f(x_n)$. If $f(x_i) = 0$ for some i, then x_i is a root of the equation. If none of the x_is are zero, then proceed to step 2.

Step 2:

Find values x_i and x_{i+1} such that $f(x_i)\ f(x_{i+1}) < 0$. Rename $x_i = a_1$ and $x_{i+1} = b_1$. Then by the IV Theorem a root lies in between a_1 and b_1. Test for all values of $f(x_j)$, j = 1, 2, ..., n and determine other intervals, if any, in which some more roots may lie.

Step 3:

Repeat Step 1 by taking some numbers between a_1 and b_1. Again, if $f(x_j) = 0$ for some x_j between a_1 then we have found the root x_j. Otherwise, continue step 2.

Continue the step 1, 2, 3 till we get a sufficiently small interval]a, b[in which the root lies. Then any value between]a, b[can be chosen as an initial approximation to the root. You may have noticed that the test values x_j, j = 1, 2, ..., n chosen are dependent on the nature of the function f(x).

We can always gather some information regarding the root either from the physical problem in which the equation f(x) = 0 occur, or it is specified in the problem. For example, we may ask for the smallest positive root or a root closest to a given number etc.

For a better understanding of the method let us consider one more example.

Example 1:

Find the approximate value of the real root of the equation

$2x - 3 \sin x - 5 = 0.$

Solution:

Let $f(x) = 2x - 3 \sin x - 5.$

Since $f(-x) = -2x + 3 \sin x - 5 < 0$ for $x > 0$, the function f(x) is negative for all negative real numbers x. Therefore the function has no negative real root. Hence the roots of this equation must lie in $[0, \propto[$. Now following step 1, we compute values of f(x), for x = 0, 1, 2, 3, 4, ...

We have

$f(0) = -5.0,$
$f(1) = 2 - 3 \sin 1 - 5 = 5.5224$

using the calculator. Note that x is in radians. The values f(0), f(1), f(2) and f(3) are given in Table 4.

Table 4

x	0	1	2	3
f(x)	-5.0	-5.51224	-3.7278	0.5766

Now we follow step 2. From the table we find that f(2) and f(3) are of opposite signs. Therefore a root lies between 2 and 3. Now, to get a more refined interval, we evaluate f(x) for some values between 2 and 3. The values are given in Table 5.

Table 5

x	2	2.5	2.8	2.9
f(x)	-3.7278	-1.7954	-0.4049	0.0822

This table of values shows that f(2.8) and f(2.9) are of opposite signs and hence the root lies between 2.8 and 2.9. We repeat the process once again for the interval [2.8, 2.9] by taking some values as given in Table 6.

Table 6

x	2.8	2.85	2.88	2.89
f(x)	-0.4049	-1.1624	-0.0159	0.0232

From Table 6 we find that the root lies between 2.88 and 2.89. This interval is small, therefore we take any value between 2.88 and 2.89 as an initial approximation of the root. Since f(2.88) is near to zero than f(2.89), we can take any number near to 2.88 as an initial approximation to the root.

You might have realized that the tabulation method is a lengthy process for finding an initial approximation of a root. However, since only a rough approximation to the root is required, we normally use only one application of the tabulation method. In the next sub-section we shall discuss the graphical method.

3.1.2 Graphical Method

In this method, we draw the approximate graph of $y = f(x)$. The points where the curve cuts the x-axis are taken as the required approximate values of the roots of the equation $f(x) = 0$. Let us consider an example.

Example 2: Find an approximate value of a root of the bi-quadratic equation

$x^4 + 4x^3 + 4x^2 – 2 = 0$
using graphical method.

Solution:

We first sketch the fourth degree polynomial $f(x) = x^4 + 4x^3 + 4x^2 – 2$. This graph is given in Fig. 1.

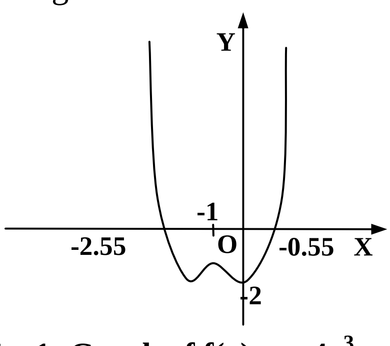

Fig. 1: Graph of $f(x) = + 4x^3 + 4x^2 – 2$

The figure shows that the graph cuts the x-axis at two points -2.55 and 0.55, approximately. Hence -2.55 and 0.55 are taken as the approximate roots of the equation

$$x^4 + 4x^3 + 4x^2 - 2 = 0$$

Now go back for a moment to Unit 1 and see Example 1 in Sec. 1.2. There we applied graphical method to find the roots of the equation sin x $= \frac{1}{2}$.

Let us consider another example.

Example 3:

Find the approximate value of a root of

$$x^2 - e^x = 0$$

using graphical method.

Solution:

First thing to do is to draw the graph of the function $f(x) = x^2 - e^x$. It is not easy to graph this function. Now if we split the function as

$$f(x) = f_1(x) - f_2(x)$$

where $f_1(x) = x^2$ and $f_2(x) = e^x$, then we can easily draw the graphs of the functions $f_1(x)$ and $f_2(x)$. The graphs are given in fig. 2.

The figure shows that the two curves $y = x^2$ and $y = e^x$ intersect at some point P. From the figure, we find that the approximate point of intersection of the two curves is -0.7. Thus we

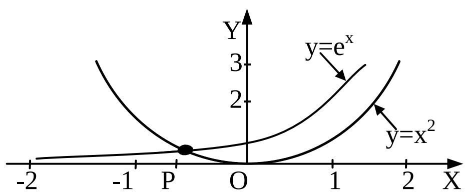

Fig. 2: Graphs of $f_1(x) = x^2$ and $f_2(x) = e^x$.

have $f_1(-0.7) - f_2(-0.7)$, and therefore $f(-0.7) = f_1(-0.7) - f_2(-0.7) \approx 0$. Hence -0.7 is an approximate value of the root of the equation $f(x) = 0$.

From the above example we observe the following: Suppose we want to apply the graphic method for finding an approximate root of f(x) = 0. Then we may try to simply the method by splitting the equation as

$$f(x) = f_1(x) - f_2(x) = 0 \qquad (4)$$

where the graphs of $f_1(x)$ and $f_2(x)$ are easy to draw. From Eqn. (4), we have $f_1(x) = f_2(x)$. The x-coordinate of the point at which the two curves $y_1 = f_1(x)$ and $y_2 = f_2(x)$ intersect gives an approximate value of the root of the equation f(x) = 0. Note that we are interested only in the x-coordinate, we don't have to worry about the point of intersection of the curves.

Often we can split the function f(x) in the form (4) in a number of ways. But we should choose that form which involves minimum calculations and the graphs of $f_1(x)$ and $f_2(x)$ are easy to draw. We illustrate this point in the following example.

Example 4:

Find an approximate value of the positive real root of 3x – cos x – 1 = 0 using graphic method.

Solution:

Since it is easy to plot 3x – 1 and cos x, we rewrite the equation as 3x – 1 = cos x. The graphs of $y = f_1(x) = 3x - 1$ and $y = f_2(x) = \cos x$ are given in Figure 3.

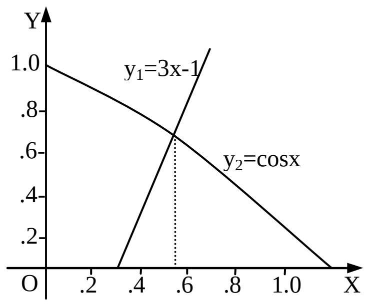

Fig. 3: Graphs of $f_1(x) = 3x - 1$ and $f_2(x)$ cos x

It is clear from the figure that the x-coordinate of the point of intersection is approximately 0.6. Hence x = 0.6 is an approximate value of the root of the equation 3x – cos x – 1 = 0.

We now make a remark.

Remark 1:

You should take some care while choosing the scale for graphing. A magnification of the scale may improve the accuracy of the approximate value.

We have discussed two methods, namely, tabulation method and graphical method which help us in finding an initial approximation to a root. But these two methods give only a rough approximation to a root. Now to obtain more accurate results, we need to improve these crude approximations. In the tabulation method we found that one way of improving the process is refining the intervals within which a root lies. A modification of this method is known as bisection method. In the next section we discuss this method.

3.2 Bisection Method

In the beginning of the previous section we have mentioned that there are two steps involved in finding an approximate solution. The first step has already been discussed. In this section we consider the second step which deals with refining an initial approximation to a root.

Once we know an interval in which a root lies, there are several procedures to refine it. The bisection method is one of the basic methods among them. We repeat the steps 1, 2, 3 of the tabulation method given in subsection 3.3.1 in a modified form. For convenience we write the method as an algorithm.

Suppose that we are given a continuous function f(x) defined on [a, b] and we want to find the roots of the equation f(x) = 0 by bisection method. We described the procedure in the following steps:

Step 1:

Find points x_1, x_2 in the interval [a, b] such that $f(x_1). f(x_2) < 0$. That is, those points x_1 and x_2 for which $f(x_1)$ and $f(x_2)$ are of opposite signs-(see Step 1 subsection 3.3.1). This process is called “finding an initial bisecting interval”. Then IV theorem a root lies in the interval $]x_1, x_2[$.

Step 2:

Find the middle point c of the interval $]x_1, x_2[$ i.e., $c = \frac{x_1 + x_2}{2}$. If f(c) = 0, then c is the required root of the equation and we can stop the procedure. Otherwise we go to Step 3.

Step 3:

Find out if

$f(x_1)\, f(c) < 0$

If it holds, then the root lies in $]x_1, c[$. Otherwise the root lies in $]c, x_2[$ (see Fig 4). Thus in either case we have found an interval half as wide as the original interval that contains the root.

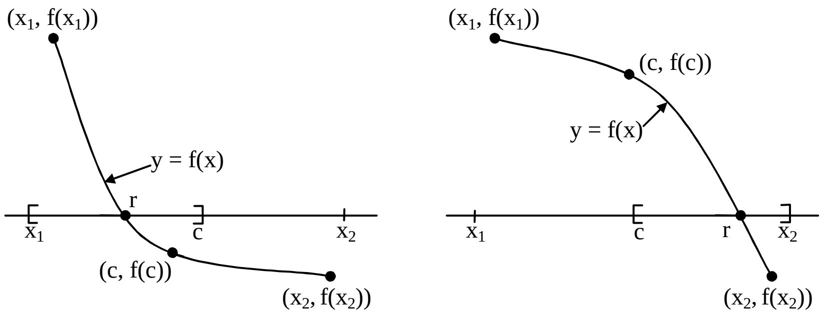

Fig. 4: The decision process for the bisection method

Step 4:

Repeat Step 2 and 3 with the new interval. This process either gives you the root or an interval having width ¼ of the original interval $]x_1, x_2[$ which contains the required root.

Step 5:

Repeat this procedure until the interval width is as small as we desire. Each bisection halves the length of the preceding interval. After N steps, the original interval length will be reduced by a factor $1/2^N$.

Now we shall see how this method helps in refining the initial intervals in some of the problems we have done in subsection 2.2.1.

Example 5:

Consider the equation $2x - \log_{10}x - 7$ lies in]3.78, 3.79[. Apply bisection method to find an approximate root of the equation correct to three decimal places.

Solution:

Let $f(x) = 2x - \log_{10}x - 7$. From Table 2 in subsection 3.3.1, we find that f(3.78) = -0.01749 and f(3.79) = 0.00136. Thus s root lies in the interval]3.78, 3.79[.

Then we find the middle point of the interval]3.78, 3.79[. The middle point is c = (3.78 + 3.79)/2 = 3.785 and f(c) = f3.785) = -0.0806 $\neq$ 0. Now, we check the condition in Step 3. Since f(3.78) f(3.785) > 0, the root does not lie in the interval]3.78. 3.78[. Hence the root lies in the interval]3.785, 3.9[. We have to refine this interval further to get better approximation. Further bisection are shown in the following Table.

Table 7

Number of Bisection	Bisected value x_i	$f(x_i)$	Improved Interval
1	3.785	-0.00806	]3.785, 3.79[
2	3.7875	-3.3525×10^{-3}	]3.7875, 3.79[
3	3.78875	9.9594×10^{-4}	]3.78875, 3.79[
4	3.789375	1.824×10^{-4}	]3.78875, 3.789375[
5	3.7890625	-4.068×10^{-4}	]3.78906, 3.7989375[

The table shows that the improved interval after 5 bisections is]3.78906, 3.789375[. The width of this interval in 3.789375 - 3.78906 = 0.000315. If we stop further bisections, the maximum absolute error would be 0.000315. The approximate root can therefore be taken as (3.78906 - 3.789375)/2 = 3.789218. Hence the desired approximate value of the root rounded off to three decimal places is 3.789.

Example 6:

Apply bisection method to find an approximation to the positive root of the equation.

$2x - 3 \sin x - 5 = 0$

rounded off to three decimal places.

Solution: Let f(x) = 2x – 3 sin x – 5.

In Example 1, we had shown that a positive root lies in the interval]2.8, 2.9[. Now we apply bisection method to this interval. The results are given in the following table.

Table 8

Number of Bisection	Bisected value x_i	$f(x_i)$	Improved Interval
1	2.85	-0.1624	]2.85, 2.79[
2	2.875	-0.0403	]2.875, 2.79[
3	2.8875	0.02089	]2.875, 2.8875[
4	2.88125	-9.735×10^{-3}	]2.88125, 2.8875[
5	2.884375	5.57781×10^{-3}	]2.88125, 2.884375[
6	2.8828125	-2.0795×10^{-3}	]2.8828125, 2.884375[
7	2.8835938	1.7489×10^{-3}	]2.8828125, 2.8835938[
8	2.8832031	-1.6539×10^{-4}	]2.8832031, 2.8835938[

After we bisection the width of the interval is 2.8835938 **-** 2.8832031 = 0.0003907. Hence, the maximum possible absolute error to the root is 0.0003907. Therefore the required approximation to the root is 2.883.

Now let us make some remarks.

Remark 2:

While applying bisection method we must be careful to check that f(x) is continuous. For example, we may come across functions like f(x) = $\frac{1}{x-1}$. If we consider the interval].5, 1.5[, then f(.5) f(1.5) < 0. In this case we may be tempted to use bisection method. But we cannot use the method here because f(x) is not defined at the middle point x = 1. We can overcome these difficulties by taking f(x) to be continuous throughout the initial bisecting interval. (Note that if f(x) is continuous by IV theorem f(x) assumes all values between the intervals.)

Therefore you should always examine the continuity of the function in the initial interval before attempting the bisection method.

Remark 3:

It may happen that a function has more than one root in an interval. The bisection method helps us in determining one root only. We can determine the other roots by properly choosing the initial intervals.

While applying bisection method we repeatedly apply steps 2, 3, 4 and 5. You recall that in the introduction we classified such a method as an Iteration method. As we mentioned in the beginning of Sec. 3.1, a numerical process starts with an initial approximation and iteration improves this approximation until we get the desired accurate value of the root.

Let us consider another iteration method now.

3.3 Fixed Point Iteration Method

The bisection method we have described earlier depends on our ability to find an interval in which the root lies. The task of finding such intervals is difficult in certain situations. In such cases we try an alternate method called Fixed Point Iteration Method. We shall discuss the advantage of this method later.

The first step in this method is to rewrite the equation f(x) = 0 as

$$x = g(x) \tag{5}$$

For example consider the equation $x^2 – 2x – 8 = 0$. We can write it as

$$x = \sqrt{2x+8} \tag{6}$$

$$x = \frac{2x+8}{x} \tag{7}$$

$$x = \frac{x^2 -8}{2} \tag{8}$$

We can choose the form (5) in several ways. Since f(x) = 0 is the same s x = g(x), finding a root of f(x) = 0 is the same as finding a root of x = g(x) i.e., a fixed point of g(x). Each such g(x) given in (6), (7) or (8) is called an iteration function for solving f(x) = 0.

Once an iteration function is chosen, our next step is to take a point x_0 close to the root as the initial approximation of the root.

Starting with x_0, we find the first approximation x_1 as

$$x_1 = g(x_0)$$

Then we find the next approximation as

$$x_2 = g(x_1)$$

Similarly we find the successive approximation x_2, x_3, x_4 ... as

$x_3 = g(x_2)$
$x_4 = g(x_3)$
. .
..
. .
$x_{n+1} = g(x_n)$

Each computation of the type $x_{n+1} = g(x_n)$ is called an iteration. Now, two questions arise (i) when do we stop these iterations? (ii) Does this procedure always give the required solution?

To ensure this we make the following assumptions on g(x):

Assumption*

The derivative g'(x) of g(x) exists g'(x) is continuous and satisfies $|g'(x)| < 1$ in an interval containing x_0. (That would mean that we require $|g'(x)| < 1$ at all iterates x_i.)

The iteration is usually stopped whenever $|x_{i+1}|$ is less than the accuracy required.

In Unit 3 you will prove that if g(x) satisfies the above conditions, then there exists a unique point α such that $g(\alpha) = \alpha$ and the sequence of iterates approach α, provided that the initial approximation is close to the point α.

Now we shall illustrate this method with the following example.

Example 7:

Find an approximate root of the equation

$x^2 - 2x - 8 = 0$
using fixed point iteration method, starting with $x_0 = 5$. Stop the iteration whenever

$|x_{i+1} - x_i| < 0.001$.

Solution:

Let $f(x) = x^2 - 2x - 8$. We saw that the equation $f(x) = 0$ can be written in three forms (6), (7) and (8). We shall take up the three forms one by one.

Case 1: Suppose we consider form (5). In this form the equation is written as

$x = (2x + 8)^{1/2}$

Here $g(x) = (2x + 8)^{1/2}$. Let's see whether Assumption (*) is satisfied for this g(x). We have

$$g'(x) = \frac{1}{(2x+8)^{1/2}}$$

Then $|g'(x)| < 1$ whenever $(2x + 8)^{1/2} > 1$. For any positive real number x, we see that the inequality $(2x + 8)^{1/2} > 1$ is satisfied. Therefore, we consider any interval on the positive side of x-axis. Since the starting point is $x_0 = 5$, we may consider the interval at I = [3, 6]. This contains the point 5. Now, g(x) satisfies the condition that g'(x) exists on I, g'(x) is continuous on I and $|g'(x)| < 1$ for every x in the interval [3, 6]. Now we apply fixed point iteration method to g(x).

We get
$x_1 = g(5) = \sqrt{18} = 4.243$
$x_2 = g(4.243) = 4.060$
$x_3 = 4.015$
$x_4 = 4.004$
$x_5 = 4.001$
$x_6 = 4.000$.

Since $|x_6 – x_5| = |-0.001| = 0.001$, we conclude that an approximate value of a root of f(x) = 0 is 4.

Case 2: Let us consider the second form,

$$x = \frac{2x+8}{x}$$

Here $g(x) = \frac{2x+8}{x}$ and $g'(x) = \frac{-8}{x^2}$. The $|g'(x)| < 1$ for any real number x $\geqslant 3$. Hence g(x) satisfies Assumption (*) in the interval [3, 6]. Now we leave is as an exercise for you to complete the computations (See TMA 6).

Case 3: Here we have $x = \frac{x^2-8}{2}$. Then $g(x) = \frac{x^2-8}{2}$ and $g'(x) = x$. In this case $|g'(x)| < 1$ only if $|x| < 1$ i.e. if x lies in the interval]-1, 1[. But this interval does not contain 5. Therefore g(x) does not satisfy the Assumption (*) in any interval containing the initial approximation. Hence, the iteration method cannot provide approximation to the desired root.

Note: This example may appear artificial to you. You are right because in this case we have got a formula for calculating the root. This example is taken to illustrate the method in a simple way.

Let us consider another example.

Example 8:

Use fixed point iteration procedure to find an approximate root of 2x = 3 sin x – 5 = 0 starting with the point $x_0 = 2.8$. Stop the iteration whenever $|x_{i+1} + x_i| < 10^{-5}$.

Solution: We can rewrite the equation in the form,

$$x = \frac{3}{2}\sin x + \frac{5}{2}.$$

Here $g(x) = \frac{3}{2}\sin x + \frac{5}{2}$ and $g'(x) = \frac{3}{2}\cos x$.

Now at $x_0 = 2.8$, we have

$|g'(2.8)| = 1.413$

which is greater than 1. Thus g(x) does not satisfy Assumption (*) and therefore in this form the iteration method fails.

Let us now rewrite the equation in another form. We write

$$x = x - \frac{2x - 3\sin x - 5}{2 - 3\cos x}$$

Then $g(x) = x - \frac{2x - 3\sin x - 5}{2 - 3\cos x}$

You may wonder how did we get this form. Note that here g(x) is of the form $g(x) = x - \frac{f(x)}{f'(x)}$. You will find later that the above equation is the iterated formula for another popular iteration method.

Then $g'(x) = 1 - \left(\frac{(2-3\cos x)(2-3\cos x) - (2x-3\sin x+5)3\sin x}{(2-3\cos x)^2}\right)$

$= \frac{2x-3\sin x+5}{(2-3\cos x)^2} 3 \sin x$

At $x_0 = 2.8$ $|g'(x_0)| = 0.0669315$ (or 0.02174691) < 1

Therefore g(x) satisfies the Assumption (*). Using the initial approximation as $x_0 = 2.8$, we get the successive approximation as

$x_1 = 2.8839015$
$x_2 = 2.8832369$
$x_3 = 2.8832369$

Since $|x_2 - x_3| < 10^{-5}$ we stop the iteration here and conclude that 2.88323 is an approximate value of the root.

Next we shall use another form

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

Here $g(x) = \sin^{-1}\left(\frac{2x-5}{3}\right)$ and $g'(x) = \frac{2}{\sqrt{9-(2x-5)^2}}$

At $x_0 = 2.8$, $g'(x_0) = 0.6804 < 1$. In fact, we can check that in any small interval containing 2.8 $|g'(x)| < 1$. Thus g(x) satisfies the Assumption (*). Applying the iteration method, we have

$x_1 = \sin^{-1}\left(\frac{2(2.8)-5}{3}\right) = 0.201358$

We find that there are two values which satisfy the above equation. One value is 0.201358 and the other is π - 0.201358 = 2.940235. In situations, we take a value close to the initial approximation. In this case the value close to the initial approximation is 2.940235. Therefore we take this value as the starting point of the next approximation.

$x_1 = 2.940235$

Next we calculate

$x_2 = \sin^{-1}\left(\frac{2(2.940235)-5}{3}\right)$
= 0.297876 or 2.843717

Continuing like this, it needed 17 iteration to obtain the value x_{17} = 2.88323, which we got from the previous form. This means that in this form the convergence is very slow.

From examples 7 and 8, we learn that if we choose the form x = g(x) properly, then we can get the approximate root provided that the initial approximation is sufficiently close to the root. The initial approximation is usually given in the problem or we can find using the IV theorem.

Now we shall make a remark here

Remark: The Assumption (*) we have given for an iteration function, is a stronger assumption. In actual practice there are a variety of assumptions which the iteration function g(x) must satisfy to ensure that the iterations approach the root. But, to use those assumptions you would require a lot of practice in the application of techniques in mathematical analysis. In this course, we will be restricting ourselves to functions that satisfies Assumption (*). If you would like to know about the other assumptions, you may refer to ‘Elementary Numerical Analysis’ by Samuel D Conte and Carl de Boor.

4.0 CONCLUSION

Let us now briefly recall what we have done in this unit.

5.0 SUMMARY

In this unit we have covered the following points:

- We have seen that the methods for finding an approximate solution of an equation involve two steps:

  i) Find an initial approximation to a root.

  ii) Improve the initial approximation to get a more accurate value of the root.

- We have described the following iteration methods for improving an initial approximation of a root.

  i) Bisection method

  ii) Fixed point iteration method.

6.0 TUTOR-MARKED ASSIGNMENT (TMA)

1) Find an initial approximation to a root of the equation $3x - \sqrt{1+\sin x} = 0$ using tabulation method.

2) Find a initial approximation to a positive root of the equation $2x - \tan x = 0$ using tabulation method.

3) Find the approximate location of the roots of the following equations in the regions given using graphic method.

   a) $f(x) = e^{-x} - x = 0$, in $0 \leqslant x \leqslant 1$

   b) $f(x) = e^{-0.4x} - 0.4x - 9 = 0$, in $0 < x \leqslant 7$

4) Starting with the interval $[a_0, b_0]$, apply bisection method to be the following equations and find an interval of width 0.05 that contains a solution of the equations

   a) $e^x - 2 - x = 0$, $[a_0, b_0] = [1.0, 1.8]$

   b) $1n\ x - 5 + x = 0$, $[a_0, b_0] = [3.2, 4.0]$

5) Using bisection method find an approximate root of the equation $x^3 - x - 4 = 0$ in the interval]1, 2[to two places of decimal.

6) Apply fixed point iteration method to the form $x = \frac{2x+8}{x}$ starting with $x_0 = 5$ to obtain a root of $x^2 - 2x - 8 = 0$.

7) a) Apply fixed point iteration method to the following equations with the initial approximation given alongside. In each case find an approximate root rounded off to 4 decimal places.

      i) $x = -45 + \frac{2}{x}$ $x_0 = 20$.

      ii) $x = \frac{1}{2} + \sin x$, $x_0 = 1$.

   b) Compute the exact roots of the equation $x^2 + 45x - 2 = 0$ using quadratic formula and compare with the approximate root obtained in (a) (i).

7.0 REFERENCES/FURTHER READINGS

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

UNIT 3 CHORD METHOD FOR FINDING ROOTS

CONTENTS

1.0 INTRODUCTION

In the last unit we introduced you to two iteration methods for finding roots of an equation f(x) = 0. There we have shown that a root of the equation f(x) = 0 can be obtained by writing the equation in the form x = g(x). Using this form we generate a sequence of approximations $x_{i+1} = g(x_i)$ for i = 0, 1, 2, ... We had also mentioned there that the success of the iteration methods depends upon the form of g(x) and the initial approximation x_0. In this unit, we shall discuss two iteration methods: regula-falsi and Newton-Raphson methods. These methods produce results faster than bisection method. The first two sections of this unit deal with derivations and the use of these two methods. You will be able to appreciate these iteration methods better if you can compare the efficiency of these methods. With this in view we introduce the concept of convergence criterion which helps us to check the efficiency of each method. Sec. 3.3 is devoted to the study of rate of convergence of different iterative methods.

2.0 OBJECTIVES

After studying the unit you should be able to:

- apply regula-falsi and secant methods for finding roots
- apply Newton-Raphson method for finding roots
- define 'order of convergence' of an iterative scheme
- obtain the order of convergence of the following four methods:
    - bisection method

- fixed point iteration method
- secant method
- Newton-Raphson method

3.0 MAIN BODY

3.1 Regula-Falsi Method (or Method of False Position)

In this section we shall discuss the 'regula-falsi method'. The Latin word 'Regula Falsi' means rule of falsehood. It does not mean that rule is a false statement. But it conveys that the roots that we get according to the rule are approximate roots and not necessarily exact roots. The method is also known as the method of false position. This method is similar to the bisection method you have learnt in Unit 3.

The bisection method for finding approximate roots has a drawback that it makes use of only the signs of f(a) and f(b). It does not use the values f(a), f(b) in the computations. For example, if f(a) = 700 and f(b) = -0.1, then by the bisection method the first approximate value of a root of f(x) is the mid value x_0 of the interval]a, b[. But at x_0, $f(x_0)$ is nowhere near 0. Therefore in this case it makes more sense to take a value near to -0.1 than the middle value as the approximation to the root. This drawback is to some extent overcome by the regula-falsi method. We shall first describe the method geometrically.

Suppose we want to find a root of the equation f(x) = 0 where f(x) is a continuous function. As in the bisection method, we first find an interval]a, b[such that f(a) f(b) < 0. Let us look at the graph of f(x) given in Fig. 1.

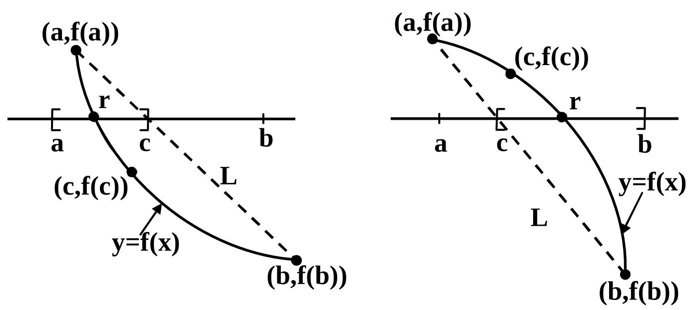

Fig 1: Regula-Falsi

The condition f(a) f(b) < 0 means that the points (a, f(a)) and (b, f(b)) lie on the opposite sides of the x-axis. Let bus consider the line joining (a, f(a)) and (b, f(b)). This line crosses the x-axis at some point (c, 0) [see Fig. 1]. Then we take the x-coordinate of that point as the first approximation. If f(c) = 0, then x = c is the required root. If f(a) f(c) < 0, then the root lies in]a, c[(see Fig. 1 (a)). In this case the graph of y = f(x) is concave near the root r). Otherwise, if f(a) f(c) > 0, the root lies in

]c, b[(see Fig. 1 (b)). In this case the graph of y = f(x) is convex near the root. Having fixed the interval in which the roots lies, we repeat the above procedure.

Let us now write the above procedure in the mathematical form. Recall the formula for the line joining two points in the Cartesian plane. The line joining (a, f(a)) and (b, f(b)) is given by

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

We can rewrite this in the form

$$\frac{y - f(a)}{f(b) - f(a)} = \frac{x - a}{b - a} \qquad (1)$$

Since the straight line intersects the x-axis at (c, 0), he point (c, 0) lies on the straight line. Putting x = c, y = 0 in Eqn. (1), we get

$$\frac{-f(a)}{f(b) - f(a)} = \frac{c - a}{b - a}$$

i.e. $\frac{c}{b - a} - \frac{a}{b - a} = \frac{-f(a)}{f(b) - f(a)}$

Thus $c = a \; \frac{f(a)}{f(b) - f(a)} \; (b - a)$. (2)

This expression for c gives an approximate value of a root of f(x). Simplifying (2), we can also write as

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

Now, examine the sign of f(c) and decide in which interval]a, c[or]c, b[, the root lies. We thus obtain a new interval such that f(x) is of opposite signs at the end points of this interval. By repeating this process, we get a sequence of intervals]a, b[,]a, a_1[,]a, a_2[, ... as shown in Fig. 2.

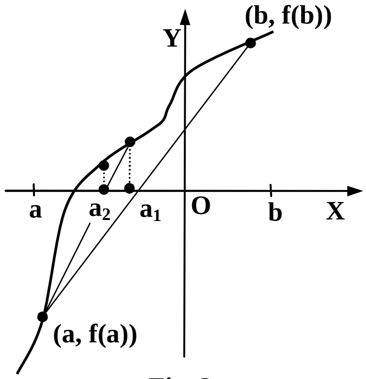

Fig. 2

We stop the process when either of the following holds.

i) The interval containing the zero of f(x) is of sufficiently small length or

ii) The difference between two successive approximation is negligible.

In the iteration format, the method is usually written as

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)}$$

where $]x_0, x_1[$ is the interval in which the root lies.

We now summarise this method in the algorithm form. This will enable you to solve problems easily.

Step 1: Find numbers x_0 and x_1 such that $f(x_0)\, f(x_1) < 0$, using the tabulation method.

Step 2: Set $x_2 = \dfrac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)}$. This gives the first approximation.

Step 3: If $f(x_2) = 0$ then x_2 is the required root. If $f(x_2) \neq 0$ and $f(x_0)\, f(x_2) < 0$, then the next approximation lies in $]x_0, x_2[$. Otherwise it lies in $]x_2, x_1[$.

Step 4: Repeat the process till the magnitude of the difference between two successive iterated values x_i and x_{i+1} is less than the accuracy required. (Note that $|x_{i+1} – x_i|$ gives the error after i^{th} iteration).

Let us now understand these steps through an example.

Example 1:

It is known that the equation $x^3 + 7x^2 + 9 = 0$ has a root between -8 and -7. Use the regula-falsi method to obtain the root rounded off to 3 decimal places. Stop the iteration when $|x_{i+1} - x_i| < 10^{-4}$.

Solution:

For convenience we rewrite the given function f(x) as

$$f(x) = x^3 + 7x^2 + 9$$
$$= x^2(x + 7) + 9$$

Since we are given that $x_0 = -8$ and $x_1 = -7$, we do not have to use Step 1. Now to get the first approximation, we apply the formula in Step 2.

Since, $f(x_0) = f(-8) = -55$ and $f(x_1) = f(-7) = 9$ we obtain

$$x_2 = \frac{(-8)9 - (-7)(-5)}{9 + 55} = -7.1406$$

Therefore our first approximation is -7.1406.

To find the next approximation we calculate $f(x_2)$ with the signs of $f(x_0)$ and $f(x_1)$. We can see that $f(x_0)$ and $f(x_2)$ are of opposite signs. Therefore a root lies in the interval]-8, -7.1406[. We apply the formula again by renaming the end points of the interval as $x_1 = -8$, $x_2 = -7.1406$. Then we get the second approximation as

$$x_3 = \frac{-8\ \ f(-7.1406) + 7.1406\ \ f(-8)}{1.862856 + 55} = -7.168174.$$

We repeat this process using Step 2 and 3 given above. The iterated values are given in the following table.

Table 1

Number of iterations	Interval	Iterated Values x_i	The function value $f(x_i)$
1	]-8,-7[	-7.1406	1.862856
2	]-8,-7.1406[	-7.168174	0.3587607
3	]-8,-7.168174[	-7.1735649	0.0683443
4	]-8,-7.1735649[	-7.1745906	0.012994
5	]-8,-7.1745906[	-7.1747855	0.00246959
6	]-8, -7.1747855[	-7.1748226	0.00046978

From the able, we see that the absolute value of the difference between the 5th and 6th iterated values is |7.1748226 – 7.1747855| = .0000371. Therefore we stop the iteration here. Further, the values of f(x) at 6th iterated value is .00046978 = 4.6978×10^{-4} which is close to zero. Hence we conclude that -7.175 is an approximate root of $x^3 + 7x^2 + 9 = 0$

Rounded off to three decimal places.

You note that in regula-falsi method, at each stage we find an interval $]x_0, x_1[$ which contains a root and then apply iteration formula (3). This procedure has a disadvantage. To overcome this, regula-falsi method is modified. The modified method is known as secant method. In this method we choose x_0 and x_1 as any two approximations of the root. The Interval $]x_0, x_1[$ need not contain the root. Then we supply formula (3) with x_0, x_1, $f(x_0)$ and $f(x_1)$.

The iterations are now defined as:

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x)}$$

$$x_3 = \frac{x_1 f(x_2) - x_2 f(x_1)}{f(x_2) - f(x_1)}$$

....

....

$$x_{n+1} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} \qquad (4)$$

Note: Geometrically, in secant Method, we replace the graph of f(x) in the interval $]x_n, x_{n+1}[$ by a straight line joining two points $(x_n, f(x_{n+1}), (x_{n+1}), f(x_{n+1}))$ on the curve and take the point of intersection with x-axis as the approximate value of the root. Any line joining two points on the curve is called a secant line. That is why this method is known as secant method. (see Fig. 3).

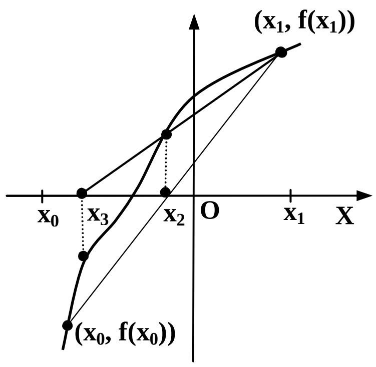

Fig. 3

Let us solve an example.

Example 2:

Determine an approximate root of the equation

$\cos x - x e^x = 0$

using
i) secant method starting with the two initial approximations as $x_0 = 1$ and $x_1 = 1$
and
ii) regula-falsi method.
(This example was considered in the book 'Numerical methods for scientific and engineering computation' by M. K. Jain, S. R. K. Iyengar and R. K. Jain).

Solution:

Let $f(x) = \cos x - x e^x$.

Then $f(0) = 1$ and $f(1) = \cos 1 - e = -2.177979523$. Now we apply formula (4) with $x_0 = 0$ and $x_1 = 1$. Then

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} = \frac{0(-2177979523 + (-1)1}{-2.177979523 - 1}$$

$$= \frac{-1}{-2.177979523 - 1} = \frac{1}{3.177979523} = 0.3146653378.$$

Therefore the first iterated value is 0.3146653378. to get the 2nd iterated value, we apply formula (4) with $x_1 = 1$, $x_2 = 0.3144653378$. Now $f(1) = -2.177979523$ and $f(0.\ 3144653378) = 0.519871175$.

Therefore

$$x_3 = \frac{x_1 f(x_2) - x_2 f(x_1)}{f(x_2) - f(x_1)}$$

$$= \frac{1(0.519871175) - 0.3146653378(-2.177979523)}{0.519871175 + 2.177979523}$$

$= 0.4467281466$

We continue this process. The iterated values are tabulated in the following table.

Table 2: Secant Method

Number of iterations	Iterated Values x_i	$f(x_i)$
1	0.3146653378	0.519871
2	0.4467281466	0.203545
3	0.5317058606	-0.0429311
4	0.5169044676	.00259276
5	0.5177474653	0.00003011
6	0.5177573708	-0.215132×10^{-7}
7	0.5177573637	0.178663×10^{-12}
8	0.5177573637	0.222045×10^{-15}

From the table we find that the iterated values for 7th and 8th iterations are the same. Also the value of the function at the 8th iteration is closed to zero. Therefore we conclude that 0.5177573637 is an approximate root of the equation.

ii) To apply regula-falsi method, let us first note that f(0) f(1) < 0. Therefore a root lies in the interval]0, 1[. Now we apply formula (3) with $x_0 = 0$ and $x_1 = 1$. then the first approximation is

$$x_2 = \frac{0(-2177979523 + (-1)1}{-2.177979523 - 1}$$

$$= 0.3146653378$$

You may have noticed that we have already calculated the expression on the right hand side of the above equation in part (i).

Now $f(x_2) = 0.51987 > 0$. This shows that the root lies in the interval]0.3146653378, 1[. To get the second approximation, we compute

$$x_3 = \frac{0.3146653378\ f(1)\ -1f(0.3146653378)}{f(1) - f(0.3146653378)} = 0.4467281446$$

which is same as x_3 obtained in (i). We find $f(x_2) = 0.203545 > 0$. Hence the root lies in]0.4467281446, 1[. To get the third approximation, we calculate

$$x_4 = \frac{0.4467281446\ f(1)\ -1f(0.4467281446)}{f(1) - f(0.4467281446)}$$

The above expression on the right hand side is different from the expression for x_4 in part (i). This is because when we use regula-falsi method, at each stage, we have to check the condition $f(x_1)\ f(x_{i-1}) < 0$.

The computed values of the rest of the approximations are given in Table 3.

Table 3: Regula-Falsi Method

No.	Interval	Iterated value x_i	$f(x_i)$
1	[0, 1[	0.3146653378	0.519871
2	].04467281446, 1[	0.4467281446	0.203545
3	]0.4940153366, 1[	0.4940153366	0.708023×10^{-1}
4	]0.5099461404, 1[	0.5099461404	0.236077×10^{-1}
5	]0.5152010099, 1[	0.5152010099	0.776011×10^{-2}
6	]0.5176683450, 1[	0.5177478783	0.288554×10^{-4}
7	]0.5177478783, 1[	0.5177573636	0.396288×10^{-9}

From the table, we observe that we have to perform 20 iterations using regula-falsi method to get the approximate value of the root 0.5177573637 which we obtained by secant method after 8 iterations. Note that the end point 1 is fixed in all iterations given in the table.

Next we shall discuss another iteration method.

3.2 Newton-Raphson Method

This method is one of the most useful methods for finding roots of an algebraic equation.

Suppose that we want to find an approximate root of the equation $f(x) = 0$. If $f(x)$ is continuous, then we can apply either bisection method or regula-falsi method to find approximate roots. Now if $f(x)$ and $f'(x)$ are continuous, then we can use a new iteration method called Newton-Raphson method. You will learn that this method gives the result more faster than the bisection or regula-falsi methods. The underlying idea of the method is due to mathematician Isac Newton. But the method as now used is due to the mathematician Raphson.

Let us begin with an equation $f(x) = 0$ where $f(x)$ and $f'(x)$ and are continuous. Let x_0 be an initial approximation and assume that x_0 is close to the exact root α and $f'(x) \neq 0$. Let $\alpha = x_0 + h$ where h is a small quantity in magnitude. Hence $f(\alpha) = f(x_0 + h) = 0$.

Now we expand $f(x_0 + h)$ using Taylor's theorem. Note that $f(x)$ satisfies all the requirements of Taylor's theorem. Therefore, we get

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \ldots = 0$$

Neglecting the terms containing h^2 and higher powers we get

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

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

This gives a new approximation to α as

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

Now the iteration can be defined by

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

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

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \qquad (5)$$

Eqn. (5) is called the Newton-Raphson formula. Before solving some examples we shall explain this method geometrically.

Geometrical Interpretation of Newton-Raphson Method

Let the graph of the function y = f(x) be as shown in Fig. 4.

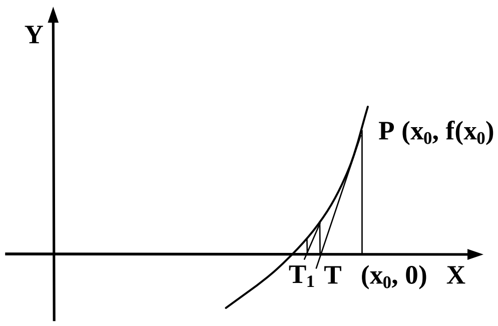

Fig. 4 Newton-Raphson Method

If x_0 is an initial approximation to the root, then the corresponding point on the graph is $P(x_0, f(x_0))$. We draw a tangent to the curve at P. Let it intersect the x-axis at T. (see Fig. 4). Let x_1 be the x-coordinate of T. Let $S(\alpha, 0)$ denote the point on the x-axis where the curve cuts the x-axis. We know that α is a root of the equation f(x) = 0. We take x_1 as the new approximation which may be closer to α than x_0. Now let us find the tangent at $P(x_0, f(x_0))$. The slope of the tangent at $P(x_0, f(x_0))$ is given by $f'(x_0)$. Therefore by the point-slope form of the expression for a tangent to a curve, we can write

$y – f(x_0) = f'(x_0) (x_1 – x_0)$

This tangent passes through the point $T(x_1, 0)$ (see fig. 4). Therefore we get

$0 – f(x_0) = f'(x_0) (x_1 – x_0)$

i.e. $x_1 f'(x_0) = x_0 f'(x_0) – f(x_0)$

i.e. $x_1 = x_0 – \frac{f(x_0)}{f'(x_0)}$

This is the first iterated value. To get the second iterated value we again consider a tangent at a point $P(x_1, f(x_1))$ on the curve (see Fig. 4) and repeat the process. Then we get a point $T_1(x_2, 0)$ on the x-axis. From the figure, we observe that T_1 is more closer to $S(\alpha, 0)$ than T. therefore after each iteration the approximation is coming closet and closer to the actual root. In practice we do not know the actual root of a given function.

Let us now take up some examples.

Example 3:

Find the smallest positive root of
$2x – \tan x = 0$
by Newton-Raphson method, correct to 5 decimal places.

Solution:

Let $f(x) = 2x – \tan x$. Then $f(x)$ is a continuous function and $f'(x) = 2 – \sec^2x$ is also a continuous function. Recall that the given equation has already appeared in an exercise in Unit 2 (see TMA in Unit 2). From that exercise we know that an initial approximation to the positive root of the equations is $x = 1$. Now we apply the Newton-Raphson iterated formula.

$x_1 = x_{i-1} - \frac{f(x_i)}{f'(x_i)}$, i = 1, 2, 3 ....

Here $x_0 = 1$. Then $f(x_0) = f(1) = 2 – \tan 1 = 0.4425922$

$$f'(x_0) = f'(1) = 2 – \sec^2 1 = 2 – (1 + \tan^2 1)$$
$$= 1 - \tan^2 1$$
$$= -1.425519$$

Therefore $x_1 = 1 - \frac{0.4425922}{-1.425519}$

$= 1.31048$

For i = 2, we get
$x_3 = 1.17605$
$x_4 = 1.165926$
$x_5 = 1.165562$
$x_6 = 1.165561$

Now x_5 and x_6 are correct to five decimal places. Hence we stop the iteration process here. The root correct to 5 decimal places is 1.16556.

Next we shall consider an application of Newton-Raphson formula. We know that finding the square root of a number is not easy unless we use a calculator. Calculators use some algorithm to obtain such an algorithm for calculating square roots. Let's consider an example.

Example 4:

Find an approximate value of $\sqrt{2}$ using the Newton-Raphson formula.

Solution:

Let x = $\sqrt{2}$. Then we have $x^2 = 2$ i.e. $x^2 - 2 = 0$. Hence we need to find the positive root of the equation $x^2 - 2 = 0$. Let

$f(x) = x^2 - 2$.

Then f(x) satisfies all the conditions for applying Newton-Raphson method. We choose $x_0 = 1$ as the initial approximation to the root. This is because we know that $\sqrt{2}$ lies between $\sqrt{1}$ and $\sqrt{4}$ and therefore we can assume that the root will be close to 1.

Now we compute the iterated values.

The iteration formula is

$$x_i = x_{i-1} - \frac{x_{i-1}^2 - 2}{2x_{i-1}}$$

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

Putting i = 1, 2, 3 ….. we get

$$x_1 = \frac{1}{2}\left(x_0 + \frac{2}{x_0}\right) = 1.5$$

$x_2 = \frac{1}{2}\left(1.5 + \frac{2}{1.5}\right) = 1.4166667$

$x_3 = \frac{1}{2}\left(1.4166667 + \frac{2}{1.416667}\right)$

$= 1.41242157$

Similarly

$x_4 = 1.4142136$

$x_5 = 1.4142136$

Thus the value of $\sqrt{2}$ correct to seven decimal places is 1.4142136. Now you can check this value with the calculator.

Note 1:

The method used in the above example is applicable for finding square root of nay positive real number. For example suppose we want to find an approximate value of $\sqrt{A}$ where A is a positive real number. Then we consider the equation $x^2 – A = 0$. The iterated formula in this case is

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

This formula involves only the basic arithmetic operations +, -, × and ÷ .

Note 2:

From examples (3) and (4), we find that Newton-Raphson method gives the root very fast. One reason for this is that the derivative $|f'(x)|$ is large compared to $|f(x)|$ for any $x = x_i$. The quantity $\left|\frac{f(x)}{f'(x)}\right|$ which is the difference between two iterated values is small in this case. In general we can say that if $|f'(x_i)|$ is large compared to $|f(x_i)|$, then we can obtain the desired root very fast by this method.

The Newton-Raphson method has some limitations. In the following remarks we mention some of the difficulties.

Remark 1:

Suppose $f'(x_i)$ is zero in a neighbourhood of the root, then it may happen that $f'(x_i) = 0$ for some x_i. In this case we cannot apply Newton-Raphson formula, since division by zero is not allowed.

Remark 2:

Another difficulty is that it may happen that f'(x) is zero only at the roots. This happens in either of the situations.

i) f(x) has multiple root at α. Recall that a polynomial function f(x) has a multiple root α of order N if we can write

$$f(x) = (x - \alpha)^N h(x)$$

where h(x) is a function such that $h(\alpha) \neq 0$. For a general function f(x), this means $f(\alpha) = 0 = f'(\alpha) = ... = f^{N-1}(\alpha)$ and $f^N(\alpha) \neq 0$.

ii) f(x) has a stationary point (point of maximum of minimum) point at the root [recall from your calculus course that if f'(x) = 0 at some point x then x is called a stationary point].

In such cases some modifications to the Newton-Raphson method are necessary to get an accurate result. We shall not discuss the modifications here as they are beyond the scope of this course.

You can try some exercise now. Whenever needed, should use a calculator for computation.

In the next section we shall discuss a criterion using which we can check the efficiency of an iteration process.

3.3 Convergence Criterion

In this section we shall introduce a new concept called 'convergence criterion' related to an iteration process. This criterion gives us an idea of how much successive iteration has to be carried out to obtain the root to the desired accuracy. We begin with a definition.

Definition 1:

Let x_0, x_1 …..x_n …. be the successive approximation of an iteration process. We denote the sequence of these approximation as $\{x_n\}_{n=0}^{\infty}$. We say that $\{x_n\}_{n=0}^{\infty}$ converges to a root α with order p $\geqslant 1$ if

$$|x_{n+1} - \alpha| \leqslant \lambda |x_n - \alpha|^P \qquad (6)$$

for some number $\lambda > 0$. p is called the order of convergence and λ is called the asymptotic error constant.

For each i. we denote by $\varepsilon_i = x_i - \alpha$. Then the above inequality be written as

$$|\varepsilon_{i+1}| \leqslant \lambda |\varepsilon_i|^P \qquad (7)$$

This inequality shows the relationship between the errors in successive approximations. For example, suppose p = 2 and $|\varepsilon_i| \approx 10^{-2}$ for some i. then we can expect that $|\varepsilon_{i+1}| \approx \lambda 10^{-4}$. Thus if p is large, the iteration converges rapidly. When p takes the integer values 1, 2, 3 then we say that the convergences are linear, quadratic and cubic respectively. In the case of linear convergence (i.e. p =1). Then we require that $\lambda < 1$. In this case we can write (6) as

$$|x_{n+1} - \alpha| \leqslant \lambda |x_n - \alpha| \text{ for all } n \geqslant 0 \qquad (8)$$

In this condition is satisfied for an iteration process then we say that the iteration process converges linearly.

Setting n = 0 in the inequality (8), we get

$$|x_1 - \alpha| \leqslant \lambda |x_0 - \alpha|$$

For n = 1, we get

$$|x_2 - \alpha| \leqslant \lambda |x_1 - \alpha| \leqslant \lambda^2 |x_0 - \alpha|$$

Similarly for n = 2, we get

$$|x_3 - \alpha| \leqslant \lambda |x_2 - \alpha| \leqslant \lambda^2 |x_1 - \alpha| \leqslant \lambda^3 |x_0 - \alpha|$$

Using induction on n, we get that

$$|x_n - \alpha| \leqslant \lambda^n |x_0 - \alpha| \text{ for } n \geqslant 0 \qquad (9)$$

If either of the inequality (8) or (9) is satisfied, then we conclude that $\{x_n\}_{n=0}^{\infty}$ converges to the root.

Now we shall find the order of convergence of the iteration methods which you have studied so far.

Let us first consider bisection method.

Convergence of bisection method

Suppose that we apply the bisection method on the interval $[a_0, b_0]$ for the equation $f(x) = 0$. In this method you have seen that we construct intervals $[a_0, b_0]$ $[a_1, b_1]$ $[a_2, b_2]$ … each of which contains the required root of the given equation.

Recall that in each step the interval width is reduced by $\frac{1}{2}$ i.e.

$$b_1 - a_1 = \frac{b_0 - a_0}{2}$$

$$b_2 = a_2 = \frac{b_1 - a_1}{2} = \frac{b_0 - a_0}{2^2}$$

$$\vdots \qquad \vdots$$

$$\text{and } b_n - a_n = \frac{b_0 - a_0}{2^n} \tag{10}$$

We know that the equation $f(x) = 0$ has a root in $[a_0, b_0]$. Let α be the root of the equation. Then α lies in all intervals $[a_i, b_i]$, $i = 0, 1, 2, \ldots$. For any n, let $c_n = \frac{a_n - b_n}{2}$ denote the middle point of the interval $[a_n, b_n]$. Then $c_0, c_1, c_2, \ldots$ are taken as successive approximations to the root α. Let's check the inequality (8) for $\{c_n\}_{n=0}^{\infty}$ converges to the root α. Hence we can say the bisection method always converges.

For practical purposes, we should be able to decide at what stage we can stop the iteration to have an acceptably good approximate value of α. The number of iterations required to achieve a given accuracy for the bisection method can be obtained. Suppose that we want an approximate solution within an error bound of 10^{-M} (Recall that you have studied error bounds in Unit 1, Sec. 3.4). Taking logarithms on both sides of Eqn. (10), we find that the number of iteration required, say n, is approximately given by

$$n = \text{int}\left[\frac{\text{In}(b_0 - a_0) - \text{In}10^{-M}}{\text{In}2}\right] \tag{11}$$

where the symbol 'int' stands for the integral part of the number in the bracket and $]a_0, b_0[$ is the initial interval in which a root lies.

Let us work out an example.

Example 5:

Suppose that the bisection method is used to find a zero of f(x) in the interval [0, 1]. How many times this interval be bisected to guarantee that we have an approximate root with absolute error less than or equal to 10^{-5}.

Solution:

Let n denote the required number. To calculate n, we apply the formula in Eqn. (11) with $b_0 = 1$, $a_0 = 0$ and M = 5.

Then

$$n = \text{int}\left(\frac{\text{In}1 - \text{In}10^{-5}}{\text{In}2}\right)$$

Using a calculator, we find

$$n = \text{int}\left(\frac{11.51292547}{0.69314718}\right)$$
$$= \text{int}\,[16.60964047] = 17$$

The following table gives the minimum number of iterations required to find an approximate root in the interval]0, 1[for various acceptable errors.

E	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
n	7	10	14	17	20	24

This table shows that for getting an approximate value with an absolute error bounded by 10^{-5}, we have to perform 17 iterations. Thus even though the bisection method is simple to use, it requires a large number of iterations to obtain a reasonably good approximate root. This is one of the disadvantages of thee bisection method.

Note: The formula given in Eqn. (11) shows that, given an acceptable error, the number of iterations depends upon the initial interval and thereby depends upon the initial approximation of the root and not directly on the values of f(x) at these approximations.

Next we shall obtain the convergence criteria for the secant method.

Convergence criteria for Secant Method

Let f(x) = 0 be the given equation. Let α denote a simple root of the equation f(x) = 0. Then we have $f'(\alpha) \neq 0$. The iteration scheme for the secant method is

$$x_{i+1} = x_i - \frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})} \tag{12}$$

For each i, set $\varepsilon_i = x_i - \alpha$. Then $x_i + \alpha$. Substituting in Eqn. (12) we get

$$\varepsilon_{i+1} + \alpha = \varepsilon_i + \alpha - \frac{\varepsilon_i - \varepsilon_{i-1}}{f(\varepsilon_i + \alpha) - f(\varepsilon_{i+1} + \alpha)} f(\varepsilon_i + \alpha)$$

$$\varepsilon_{i+1} = \varepsilon_i - \frac{\varepsilon_i - \varepsilon_{i-1}}{f(\varepsilon_i + \alpha) - f(\varepsilon_{i-1} + \alpha)} f(\varepsilon_i + \alpha) \tag{13}$$

Now we expand $f(\varepsilon_i + \alpha)$ and $f(\varepsilon_i - \alpha)$ using Taylor's theorem about the point x = α.

We get $f(\varepsilon_i + \alpha) = f(\alpha) + \frac{f'(\alpha)}{1} \varepsilon_i + \frac{f''(\alpha)}{2} \varepsilon_i^2 + \ldots$

i.e. $f(\varepsilon_i + \alpha) = f'(\alpha) \left[\varepsilon_i + \frac{f''(\alpha)}{2f'(\alpha)} \varepsilon_i^2 + \ldots\right]$ (14)

since $f'(\alpha) = 0$.

Similarly,

$$f(\varepsilon_{i-1} + \alpha) = f'(\alpha) \left[\varepsilon_{i-1} + \frac{f''(\alpha)}{2f'(\alpha)} \varepsilon_{i-1}^2 + \ldots\right] \tag{15}$$

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

$$= f'(\alpha)(\varepsilon_i - \varepsilon_{i-1}) \left[1 + (\varepsilon_i + \varepsilon_{i-1})\frac{f''(\alpha)}{2f'(\alpha)} + \ldots\right] \tag{16}$$

Substituting Eqn. (14) and Eqn. (13), we get

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

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

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

By neglecting the terms involving $\varepsilon_i\,\varepsilon_{i-1}^2 + \varepsilon_i^2\,\varepsilon'_{i-1}$ the above expression, we get

$$\varepsilon_{i+1} \approx \varepsilon_i\,\varepsilon_{i-1}\left[\frac{f''(\alpha)}{2f'(\alpha)}\right] \tag{17}$$

This relationship between the errors is called the error equation. Note that this relationship holds only if α is a simple root. Now using Eqn. (17) we will find a number p and λ such that

$$\varepsilon_{i+1} = \lambda\,\varepsilon_i^p \quad i = 0, 1, 2, ... \tag{18}$$

Setting i = j – 1, we obtain

$$\varepsilon_j = \lambda\,\varepsilon_{j-1}^p$$

or

$$\varepsilon_i = \lambda\,\varepsilon_{i-1}^p$$

Taking p^th root on both sides, we get

$$\varepsilon_i^{1/p} = \lambda^{1/p}\,\varepsilon_{i-1}$$

$$\text{i.e. } \varepsilon_{i-1} = \lambda^{-1/p}\,\varepsilon_i^{1/p} \tag{19}$$

Combining Eqns. (17) and (18). We get

$$\lambda\,\varepsilon_i^p = \varepsilon_i\,\varepsilon_{i-1}\frac{f''(\alpha)}{2f'(\alpha)}$$

Substituting the expression for ε_{i-1} from Eqn. (19) in the above expression we get

$$\lambda\,\varepsilon_i^p \approx \frac{f''(\alpha)}{2f'(\alpha)}\,\varepsilon_i\,\lambda^{-1/p}\,\varepsilon_i^{1/p}$$

$$\text{i.e. } \lambda\,\varepsilon_i^p \approx \frac{f''(\alpha)}{2f'(\alpha)}\,\lambda^{-1/p}\,\varepsilon_i^{1+1/p} \tag{20}$$

Equating the powers of ε_i on both sides of Eqn. (20) we get

$$p = 1 + \frac{1}{p} \text{ or } p^2 - p - 1 = 0.$$

This is a quadratic equation in p. The roots are given by

$$p = \frac{1+\sqrt{5}}{2} \approx 1.618.$$

Now, to get the number λ, we equate the constant terms on both sides of Eqn. (20). Then we get

$$\lambda = \left[\frac{f''(\alpha)}{2f'(\alpha)}\right]^{P/1+p}$$

Hence the order of convergence of the secant method is p = 1.62 and the asymptotic error constant is $\left[\frac{f''(\alpha)}{2f'(\alpha)}\right]^{P/1+p}$

Example 6:

The following are the five successive iterations obtained by secant method to find the root α = -2 of the equation $x^3 - 3x + 2 = 0$.

$x_1 = -2.6$, $x_2 = -2.4$, $x_3 = -2.106598985$.
$x_4 = -2.022641412$ and $x_5 = -2.000022537$.

Compute the asymptotic error constant and show that $\varepsilon_5 \approx \frac{2}{3}\varepsilon_4$.

Solution:

Let $f(x) = x^3 - 3x + 2$

Then
$f'(x) = 3x^2 - 3$, $f'(-2) = 9$
$f''(x) = 6x$, $f(-2) = -12$

Therefore $\lambda = \left[-\frac{12}{18}\right]^{.618}$

$= \left[-\frac{2}{3}\right]^{.618} = -0.778351205$

Now
$\varepsilon_5 = | x_5 - \alpha | = | -2.000022537 + 2$
$= 0.000022537$

and

ε_4 = | -2.022641412 + 2 | = 0.022641412.

Then $\lambda\, \varepsilon_4$ = 0.778351205 × 2.022641412
= 0.000021246
$\approx$0.00002253
Hence we get that $\lambda\, \varepsilon_4 \approx \varepsilon_5$

Convergence criterion for fixed point iteration method

Recall that in this method we write the equation in the form

x = g(x)

Let α denote a root of the equation. Let x_0 be an initial approximation to the root. The iteration formula is

$x_{i+1} = g(x_i)$, i = 0, 1, 2, ... (21)

We assume that g'(x) exists and is continuous and | g'(x) | < 1 in an interval containing the root α. We also assume that x_0, x_1, ... lie in this interval.

Since g'(x) is continuous near the root and | g'(x) | < 1, there exists an interval] α - h, α + h[, where h > 0, such that | g'(x) | $\leqslant$ k for some k, where 0 < k < 1.

Since α is a root of the equation, we have

$\alpha = g(\alpha)$. (22)

Subtracting (22) from (21) we get

$x_{i+1} - \alpha = g(x_i) = g(\alpha)$

Now the function g(x) is continuous in the interval]x_i, α[and g'(x) exists in this interval. Hence g(x) satisfies all the conditions of the mean value theorem [see Unit 1]. Then, by the mean value theorem there exists a ξ between x_i and α such that

$| x_{i+1} - \alpha | \leqslant | g(x_i) - g(\alpha) | \leqslant | g'(\xi) | \, | (x_i - \alpha |$

Note that ξ lies in]α - h, α + h[and therefore | g'(ξ) | < k and hence

$| x_{i+1} - \alpha | \leqslant | x_i - \alpha |$

Setting i = 0, 1, 2, ..., n we get

$| x_1 - \alpha | \leqslant k | x_0 - \alpha |$

$| x_2 - \alpha | \leqslant k | x_1 - \alpha | \leqslant k^2 | x_0 - \alpha |$

$$\vdots \qquad \vdots$$

$| x_n - \alpha | \leqslant k^n | x_0 - \alpha |$

This shows that the sequence of approximation $| x_i |$ converges to α provided that the initial approximation is close to the root.

We summarise the result obtained for this iteration process in the following Theorem.

Theorem 1:

If g(x) and g'(x) are continuous in an interval about a root α of the equation x = g(x), and if $| g'(x) | < 1$ for all x in the interval, then the successive approximations $x_1, x_2, \ldots$ given by

$x_i = g(x_{i-1}), i = 1, 2, 3, \ldots$

converges to the root α provided that the initial approximation x_0 is chosen in the above interval.

We shall now discuss the order of convergence of this method. From the previous discussions we have the result.

$| x_{i+1} - \alpha | \leqslant g'(\xi) | (x_i - \alpha) |$

Note that ξ is dependent on each x_i. Now we wish to determine the constant λ and p independent of x_i such that

$| x_{i+1} - \alpha | \leqslant c | (x_i - \alpha) |^P$

Note that as the approximations x_i get closer to the root α, $g'(\xi)$ approaches a constant value $g'(\alpha)$. Therefore, in the limiting case, as $i \to \propto$,the approximation satisfy the relation

$| x_{i+1} - \alpha | \leqslant g'(\alpha) | (x_i - \alpha) |$

Therefore, we conclude that if $g'(\alpha) \neq 0$, then the convergence of the method is linear.

If $g'(\alpha) = 0$, then we have

$_{i+1} - \alpha = g(x_i) - \alpha$
$= g(x_i - \alpha) + \alpha - \alpha$
$= g(\alpha) + (x_i - \alpha)\, g'(\alpha) + \frac{(x_i \alpha)^2}{2} g''(\xi) - \alpha$
$= \frac{(x_i \alpha)^2}{2}\, g''(\xi)$
since $g(\alpha) = \alpha$ and $g'(\alpha) = 0$ and ξ lies between x_i and α.

Therefore, in the limiting case we have

$$| x_{i+1} - \alpha| \leqslant \frac{1}{2}\, | g''(\alpha) |\, | (x_i - \alpha) |^2$$

Hence, if $f'(\alpha) = 0$ and $g'(\alpha) \neq 0$, then this iteration method is of order 2.

Example 7:

Suppose α and β are the roots of the equation $x^2 + ax + b = 0$. Consider a rearrangement of this equation as

$$x = -\frac{(ax + b)}{x}$$

Show that the iteration $x_{i+1} = -\frac{(ax_i + b)}{x_i}$ will converge near $x = \alpha$ when $|\alpha| > |\beta|$

Solution:

The iteration are given by

$$x_{i+1} = g(x_i) = -\frac{(ax_i + b)}{x_i}, \; i = 0, 1, 2, ...$$

By Theorem 1, these iterations converge to α if $|g'(x)| < 1$ near α i.e. if $|g'(x)| = \left|-\frac{b}{x^2}\right| < 1$. Note that $g'(x)$ is continuous near α. If the iterations converge to $x = \alpha$, then we require $|g'(x)| = \left|-\frac{b}{\alpha^2}\right| < 1$.

Thus $| b | < | \alpha |^2$
i.e. $| \alpha |^2 > | b |$. (23)

Now you recall from your elementary algebra course that if α and β are the roots, then

$\alpha + \beta = -a$ and $\alpha\beta = b$
Therefore $|b| = |\alpha||\beta|$. Substituting in Eqn. (23), we get
$|\alpha|^2 > |b| = |\alpha||\beta|$.

Hence $|\alpha| > |\beta|$

Finally, we shall discuss the convergence of the Newton-Raphson method.

Convergence of Newton-Raphson Method

Newton-Raphson iteration formula is given by

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \qquad (24)$$

To obtain the order of the method we proceed as in the secant method. We assume that α is a simple root of $f(x) = 0$. Let

$x_i - \alpha = \varepsilon_i$, $i = 0, 1, 2, \ldots$

Then we have

$$\varepsilon_{i+1} + \alpha = \varepsilon_i + \alpha - \frac{f(\varepsilon_i + \alpha)}{f'(\varepsilon_i + \alpha)}$$

i.e. $$\varepsilon_{i+1} = \frac{\varepsilon_i f'(\varepsilon_i + \alpha) - f(\varepsilon_i + \alpha)}{f'(\varepsilon_i + \alpha)}$$

Now we expand $f(\varepsilon_i + \alpha)$ and $f'(\varepsilon_i + \alpha)$, using Taylor's theorem about the point α. We have

$$\left[\varepsilon_i \left\{f'(\alpha + \varepsilon_i f''(\alpha) + \frac{\varepsilon_i^2}{2} f''(\alpha) + \ldots\right\}\right.$$

$$\varepsilon_{i+1} = \frac{\left.- \left\{f(\alpha)\varepsilon_i f'(\alpha) + \frac{\varepsilon_i^2}{2} f''(\alpha) + \ldots\right\}\right]}{f'(\alpha) + \varepsilon_i f''(\alpha) + \varepsilon_1^2 f''(\alpha) + \ldots}$$

But $f(\alpha) = 0$ and $f'(\alpha) \neq 0$. Therefore

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

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

Hence, by neglecting higher powers of ε_i, we get

$$\varepsilon_{i+1} \approx \frac{f''(\alpha)}{2f'(\alpha)}\varepsilon_i^2$$

This shows that the errors satisfy Eqn. (6) with p = and $\lambda = \frac{f''(\alpha)}{2f'(\alpha)}$. Hence, Newton-Raphson method is of order 2. That is at each step, the error is proportional to the square of the previous error.

Now, we shall discuss an alternate method for showing that the order is 2. Note that we can write (24) in the form x = g(x) where

$$g(x) = x \quad \frac{f(x)}{f'(x)}$$

$$g'(x) = \frac{d}{dx}\left[x - \frac{f(x)}{f(x)}\right] = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2}$$

$$= \frac{f(x)f''(x)}{[f'(x)]^2}$$

Now, $g'(\alpha) = \frac{f(\alpha)f''(\alpha)}{[f'(\alpha)]^2} = 0$, since $f(\alpha) = 0$ and $f'(\alpha) \neq 0$.

Hence by the conclusion drawn just above Example 7, the method is of order 2. Note that this is true only if α is a simple root. If α is a multiple root i.e. if $g'(\alpha) = 0$, then the convergence is not quadratic, but only linear. We shall not prove this result, but we shall illustrate this with an example.

Let us consider an example.

Example 8:

Let $f(x) = (x - 2)^4 = 0$. Starting with the initial approximation $x_0 = 2.1$, compute the iterations x_1, x_2, x_3 and x_4 using Newton-Raphson method. Is the sequence conveying quadratically or linearly?

Solution:

The given function has multiple roots at x = 2 and is of order 4.

Newton-Raphson iteration formula for the given equation is

$$x_{i+1} = x_i - \frac{(x_i - 2)^4}{4(x_i - 2)^3} = x_i - \frac{1}{4}(x_i - 2)$$

$$= \frac{1}{4}(3x_i - 2) \qquad (25)$$

Starting with $x_0 = 2.1$, the iteration are given by

$$x_1 = \frac{1}{4}(6.3 + 2) = \frac{8.3}{2} = 2.075$$

Similarly,

$x_2 = 2.05625$
$x_3 = 2.0421875$
$x_4 = 2.031640625$

Now $\varepsilon_0 = x_0 - 2 = 0.1$, $\varepsilon_i = x_1 - 2 = 0.075$, $\varepsilon_2 = 0.05625$, $\varepsilon_3 = 0.0421875$, $\varepsilon_4 = 0.031640625$.

Then

$$\varepsilon_i = .075 = \frac{3}{4} \times 0.1 = \frac{3}{4}\varepsilon_0$$

and

$$\varepsilon_2 = \frac{3}{4}\varepsilon_i$$

$$\varepsilon_3 = \frac{3}{4}\varepsilon_2$$

$$\varepsilon_4 = \frac{3}{4}\varepsilon_3$$

Thus the convergence is linear in this case. The error is reduced by a factor of $\frac{3}{4}$ with each iteration. This result can also be obtained directly from Eqn. (25).

4.0 CONCLUSION

Same as in Summary

5.0 SUMMARY

In this unit we have

- described the following methods for finding a root of an equation f(x) = 0

  i) Regula-Falsi method:
  The formula is
  $$c = \frac{a\,f(b) - b\,f(a)}{f(b) - f(a)}$$
  where]a, b[is an interval such that f(a) f(b) < 0.

  ii) Secant method:
  The iteration formula is
  $$x_{i+1} = \frac{x_{i-1}f(x_i) - x_i f(x_{i-1})}{f(x_i) - f(x_{i-1})} \quad i = 0, 1, 2, \ldots.$$
  where x_0 and x_1 are any two given approximation of the root.

  iii) Newton-Raphson method:
  The iteration formula is
  $$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)},\ i = 0, 1, 2, \ldots$$
  where x_0 is an initial approximation to the root.

- introduced the concept called convergence criterion of an iteration process.

- discussed the convergence of the following iterative methods
  i) Bisection method.
  ii) Fixed point iteration method.
  iii) Secant method.
  iv) Newton-Raphson method.

6.0 TUTOR-MARKED ASSIGNMENT (TMA)

1.) Obtain an approximate root for the following equations rounded off to three decimal places, using regula-falsi method

a) $x \log_{10}x - 1.2 = 0$

b) $x \sin x - 1 = 0$

2.) Use secant method to find an approximate root to the equation $x^2 - 2x + 1 = 0$, rounded off to 5 decimal places, starting with $x_0 = 2.6$ and $x_1 = 2.5$. Compare the result with the exact root $1 + \sqrt{2}$.

3.) Find an approximate root of the cubic equation $x^3 + x^2 + 3x - 3 = 0$ using

a) i) regula-falsi method, correct to three decimal places.

ii) secant method starting with a = 1, b = 2, rounded-off to three decimal places.

b) compare the results obtained by (i) and (ii) in part (a).

4.) Starting with $x_0 = 0$ find an approximate root of the equation $x^3 - 4x + 1 = 0$, rounded off to five decimal places using Newton-Raphson method.

5.) The motion of a planet in the orbit is governed by an equation of the form y = x – e sin x where e stands for the eccentricity. Let y = 1 and $e = \frac{1}{2}$. Then find a approximate root of $2x - 2 - \sin x = 0$ in the interval $[0, \pi]$ with error less than 10^{-5}. Start with $x_0 = 1.5$.

6.) Using Newton-Raphson square root algorithm, find the following roots within an accuracy of 10^{-4}.

i) $8^{1/2}$ starting with $x_0 = 3$

ii) $91^{1/2}$ starting with $x_0 = 10$

7.) Can Newton-Raphson iteration method be used to solve the equation $x^{1/3} = 0$? Give reasons for your answer.

8.) For the problem given in Example 5, Unit 2, find the number n of bisection required to have an approximate root with absolute error less than or equal to 10^{-7}.

9.) For the equation given in Example 7, show that the iteration $x_{i+1} = \frac{b}{x_i + a}$ will converge to the root $x = \alpha$, when $|\alpha| < |\beta|$.

7.0 REFERENCES/FURTHER READINGS

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.

UNIT 4 APPROXIMATE ROOTS OF POLYNOMIAL EQUATION

CONTENTS

1.0 INTRODUCTION

In the last two units we discussed methods for finding approximate roots of the equation f(x) = 0. In this unit we restrict our attention to polynomial equations. Recall that a polynomial equation is an equation of the form f(x) = 0 where f(x) is a polynomial in x. Polynomial equation arise very frequently in all branches of science especially in physical applications. For example, the stability of electrical of mechanical systems is related to the real part of one of the complex roots of a certain polynomial equation. Thus there is a need to find all roots, real and complex, of a polynomial equation. The four iteration methods, we have discussed so far, applies to polynomial equations also. But you have seen that all those methods are time consuming. Thus it is necessary to find some efficient methods for obtaining roots of polynomial equations.

The sixteenth century French mathematician Francois Vieta was the pioneer to develop methods for finding approximate roots of polynomial equations. Later, several other methods were developed for solving polynomial equations. In this unit we shall discuss two simple methods: Birge-Vieta's and Graeffe's root squaring methods. To apply these methods we should have some prior knowledge of location and nature of roots of a polynomial equation. You are already familiar with some results regarding location and nature of roots from the elementary algebra course. We shall begin this unit by listing some of the important result about the roots of polynomial equations.

2.0 OBJECTIVES

After reading this unit you should be able to:

- apply the following methods for finding approximate roots of polynomial equations
  - o Birge-Vieta method
  - o Graeffe's root squaring method
- list the advantages of the above methods over the methods discussed in the earlier units.

3.0 MAIN BODY

3.1 Some Results on Roots of Polynomial Equations

The main contribution in the study of polynomial equations due to the French mathematician Rene Descartes' The results appeared in the third part of his famous paper 'La geometric' which means 'The geometry'.

Consider a polynomial equation of degree n

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + ... +a_1x + a_0 \qquad (1)$$

where a_0, a_1, ... a_n are real numbers and $a_n \neq 0$. You know that the roots of a polynomial equation need not be real numbers, it can be complex numbers, that is numbers of the form z = a + ib where a and b are real numbers. The following results are basic to the study of roots of polynomial equations.

Theorem 1:

(Fundamental Theorem of Algebra): Let p(x) be a polynomial of degree $n \geqslant 1$ given by Eqn. (1). Then p(x) = 0 has at least one root: that is there exists a number $\alpha \in C$ such that $p(\alpha) = 0$. In fact p(x) has n complex roots which may not be distinct.

Theorem 2:

Let p(x) be a polynomial of degree n and α is a real number. Then

$$p(x) = (x - \alpha)\, q_0(x) + r_0 \qquad (2)$$

for some polynomial $q_0(x)$ of degree n – 1 and some constant number r_0 . $q_0(x)$ and r_0 are called the quotient polynomial and the remainder respectively.

In particular, if α is a root of the equation p(x) = 0, then $r_0 = 0$: that is (x - α) divides p(x).

Then we get

$$p(x) = (x - \alpha)\, q_0(x)$$

How do we determine $q_0(x)$ and r_0? We can find them by the method of synthetic division of a polynomial p(x). Let us now discuss the synthetic division procedure.

Consider the polynomial p(x) as given in Eqn. (1)

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$$

Dividing p(x) by x - α we get

$$p(x) = q_0(x)\,(x - \alpha) + r_0 \qquad (3)$$

where $q_0(x)$ is a polynomial of degree n – 1 and r_0 is a constant.

Let $q_0(x)$ be represented as

$$q_0(x) = b_n x^{n-1} + b_{n-1} x^{n-2} + ... + b_2 x + b_1$$

(Note that for convenience we are denoting the coefficient by b_1, ..., b_n instead of b_0, b_1, ... b_{n-1}). Set $b_0 = r_0$. Substituting the expressions for $q_0(x)$ and r_0 in Eqn. (3) we get

$$p(x) = (x - \alpha)\,(b_n x^{n-1} + b_{n-1} x^{n-2} + ... + b_2 x + b_1) + b_0 \qquad (4)$$

Now, to find b_0, b_1 ..... b_n we simplify the right hand side of Eqn. (4) and compare the coefficients of x^i, i = 0, 1, ... n on both sides. Note that $p(\alpha) = b_0$. Comparing the coefficient we get

Coefficient of x^n	: $a_n = b_n$	$b_n = a_n$
Coefficient of x^{n-1}	: $a_{n-1} = b_{n-1} - \alpha b_n$,	$b_{n-1} = a_{n-1} + \alpha b_n$
	.	
	.	
	.	
Coefficient of x^k	: $a_k – b_k - \alpha b_{k+1}$,	$b_k = a_k + \alpha b_{k+1}$
	.	
	.	
	.	
Coefficient of x^0	: $a_0 = b_0 - \alpha$,	$b_0 = a_0 + \alpha b_1$

It is easy to perform the calculations if we write the coefficient of p(x) on a line and perform the calculation $b_k = a_k + \alpha b_{k+1}$ below a_k as given in the table below.

Table 1: Horner's table for synthetic division procedure

α	a_n	a_{n-1}	a_{n-2}	...	a_k	...	a_2	a_1	a_0
		αb_n	αb_{n-1}	...	αb_{k+1}	...	αb_3	αb_2	αb_1
	b_n	b_{n-1}	b_{n-2}		b_k		b_2	b_1	$b_0=p_0(\alpha)$

We shall illustrate this procedure with an example.

Example 1:

Divide the polynomial

$p(x) = x^5 – 6x^4 + 8x^3 + 8x^2 + 4x – 40$
by x – 3 by the synthetic division method and find the remainder.

Solution:

Here p(x) is a polynomial of degree 5. If a_5, a_4, a_3, a_2, a_1, a_0 are the coefficients of p(x), then the Horner's table in this case is

Table 2

a_5	a_4	a_3	a_2	a_1	a_0
1	-6	8	8	4	-40
	3	-9	-3	15	57
1	-3	-1	5	19	17
b_5	b_4	b_3	b_2	b_1	b_1

Hence the quotient polynomial $q_0(x)$ is

$q_0(x) = x^4 – 3x^3 – x^2 + 5x + 19$
and the remainder is $r_0 = b_0 = 17$. thus we have p(3) = b =17.

Theorem 3:

Suppose that z = a + ib is a root of the polynomial equation p(x) = 0. Then the conjugate of z, namely $\overline{z}$, = a – ib is also a root of the equation p(x) = 0, i.e. complex roots occur in pairs.

We denote by p(-x) the polynomial obtained by replacing x by –x in p(x). We next give an important Theorem due to Rene Descartes.

Theorem 4:

(Descartes' Rule of signs): A polynomial equation p(x) = 0 cannot have more positive roots than the number of changes in sign of its coefficients. Similarly $p(x) = 0$ cannot have more negative roots than the number of changes in sign of the coefficients of $p(-x)$.

For example, let us consider the polynomial equation
$p(x) = x^4 – 15x^2 + 7x – 11 = 0$
$= 1x^4 – 15x^2 + 7x – 11 = 0$
We count the changes in the sign of the coefficients. Going from left to right there are changes between 1 and -15, between -15 and 7 and between 7 and -11. The total number of changes is 3 and hence it can have at most 3 positive roots. Now we consider
$p(-x) = (-x)^4 – 15(-x)^2 + 7(-x) – 11 = 0$
$= x^4 – 15x^2 – 7x – 11$

Here there is only one change between 1 and -15 and hence the equation cannot have more than one negative root.

We now give another theorem which helps us in locating the real roots.

Theorem 5:

Let $p(x) = 0$ be a polynomial equation of degree $n \geqslant 1$. Let a and b be two real numbers with $a < b$. Suppose further that $p(a) \neq 0$ and $p(b) \neq 0$. Then,

i) if p(a) and p(b) have opposite signs, the equation $p(x) = 0$ has an odd number of roots between a and b.

ii) if p(a) and p(b) have like signs, then $p(x) = 0$ either has no root or an even number of roots between a and b.

Note: In this theorem multiplicity of the root is taken into consideration i.e. if a is a root of multiplicity k it has to be counted k times.

As a corollary of Theorem 5, we have the following results.

Corollary 1: An equation of odd degree with real coefficients has at least one real root whose sign is opposite to that of the last term.

Corollary 2: An equation of even degree whose constant term has the sign opposite to that of the leading coefficient, has at least two real roots one positive and the other negative.

Corollary 3: the result given in Theorem 5(i) is the generalization of the Intermediate value theorem.

The relationship between roots and coefficients of a polynomial equation is given below.

Theorem 6: Let $\alpha_1, \alpha_2, ..., \alpha_n$ be a roots ($n \geqslant 1$) of the polynomial equation

$p(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0 = 0$

Then $\alpha_1 + \alpha_2 + ... + \alpha_n = \frac{-a_{n-1}}{a_n}$

$$\alpha_1 \alpha_2 + \alpha_2 \alpha_3 + ... + \alpha_{n-1} \alpha_n = \frac{a_{n-2}}{a_n}$$

....

....

$$\alpha_1 \alpha_2 ... \alpha_n = (-1)^n \frac{a_0}{a_n}$$

In the next section we shall discuss one of the simple methods for solving polynomial equations.

3.2 Birge-Vieta Method

We shall now discuss the Birge-Vieta method for finding the real roots of a polynomial equation. This method is based on an original method due to two English mathematicians Birge-Vieta. This method is a modified form of Newton-Raphson method.

Consider now, a polynomial equation of degree n, say

$$p_n(x) = a_n x^n + ... + a_1 x + a_0 = 0. \quad (5)$$

Let x_0 be an initial approximation to the root α. The Newton-Raphson iterated formula for improving this approximation is

$$x_i = x_{i-1} - \frac{p_n(x_{i-1})}{p'_n(x_{i-1})}, \ i = 1, 2, ... \quad (6)$$

To apply this formula we should be able to evaluate both $p_n(x)$ and $p'_n(x_i)$ at any x_i. The most natural way is to evaluate

$$p_n(x_i) = a_n x_i^n + a_{n-1}x^{n-1} + ... + a_2 x_i^2 + a_1 x_i + a_0$$

$$p'_n(x_i) = n\, a_n x_i^{n-1} + (n - 1)a_{n-1}x_i^{n-2} + ... + 2a_2 x_i + a_1$$

However, this is the most inefficient way of polynomial because of the amount of computations involved and also due to the possible growth of round off errors. Thus there is a need to look for some efficient method for evaluating $p_n(x_i)$ and $p'_n(x_i)$.

Let us consider the evaluation of $p_n(x_i)$ and $p'_n(x_i)$ at x_0 using Horner's method as discussed in the previous section.

We have

$p_n(x_i) = (x – x_0)\, q_{n-1}(x) + r_0$ (7)
where
$q_{n-1}(x) = b_nx^{n-1} + b_{n-2}x^{n-2} + \ldots + b_2x + b_1$
and $b_0 = p_n(x_0) = r_0$ (8)

We have already discussed in the previous section how to find b_1, I = 1, 2, …, n.

Next we shall find the derivative $p'_n(x_0)$ using Horner's method. We divide $q_{n-1}(x)$ by $(x – x_0)$ using Horner's method. That is, we write

$q_{n-1}(x) = (x – x_0)\, q_{n-2}(x) + r_1$
$q_{n-1}(x) = c_nx^{n-2} + c_{n-1}x^{n-3} + \ldots + c_3x + c_2.$

Comparing the coefficients, we get c_i as given in the following table

Table 3

	b_n	b_{n-1}	...	b_k	...	b_2	b_1
x_0		x_0c_n	...	x_0c_{k+1}	...	x_0c_3	x_0c_2
	$c_n=b_n$	c_{n-1}		c_k		c_2	c_1

As observed in Sec. 1, we have
$c_1 = q_{n-1}(x_0)$ (9)

Now, from Eqn. (7) and (8), we have
$p_n(x) = (x – x_0)\, q_{n-1}(x) + p_n(x_0)$ (10)

Differentiating both sides of Eqn. (10) w.r.t.x, we get
$p'_n(x) = q_{n-1}(x) + (x – x_0)\, q'_{n-1}(x)$ (11)

Putting $x = x_0$ in Eqn. (11), we get
$p'_n(x_0) = q_{n-1}(x_0)$ (12)

Comparing (9) and (12), we get
$p'_n(x_0) = q_{n-1}(x_0) = c_1$

Hence the Newton-Raphson method (Eqn. (6)) simplies to

$$x_i = x_{i-1} - \frac{b_0}{c_1} \quad (13)$$

We summarise the evaluation of b_i and c_i in the following table.

Table 4

	a_n	a_{n-1}	...	a_k	...	a_2	a_1	a_0
x_0		x_0b_n	...	x_0b_{k+1}	...	x_0b_3	x_0b_2	x_0b_1
	$a_n=b_n$	b_{n-1}		b_k		b_2	b_1	$b_0=p_n(x_0)$
x_0		x_0c_n	...	x_0c_{k+1}	...	x_0c_3	x_0c_2	
	$c_n=b_n$	c_{n-1}		c_k		c_2		$c_1=p'_n(x_0)$

Let us consider an example.

Example 2:

Evaluate p'(3) for the polynomial
$p(x) = x^5 - 6x^4 + 8x^3 + 8x^2 + 4x - 40$.

Solution:

Here the coefficients are $a_0 = -40$, $a_1 = 4$, $a_2 = 8$, $a_3 = 8$, $a_4 = -6$ and $a_5 = 1$. To compute b_0, we form the following table.

Table 5

3	1	-6	8	8	4	-40
		3	-9	-3	15	57
3	1	-3	-1	5	19	$17=p(3)=b_0$
		3	0	-3	6	
	1	0	-1	2	$25 = p'(3)=c_1$	

Therefore p'(3) = 25

Now we shall illustrate why this method is more efficient than the direct method. Let us consider an example. Suppose we want to evaluate the polynomial

$$p(x) = -8x^5 + 7x^4 - 6x^3 + 5x^2 - 4x + 3$$

for any given x.

When we evaluate by direct method, we compute each power of x by multiplying with x the preceding power of x as

$x^3 = x(x^2)$, $x^4 = x(x^3)$ etc.

Thus each term c^k takes two multiplications for k > 1. Then the total number of multiplications involved in the evaluation pf p(x) is 1 + 2 + 2 + 2 + 2 = 9.

When we use Horner's method the total number of multiplications in 5. The number of additions in both cases are the same. This shows that less computation is involved while using Horner's method and therapy reduces the error in computation.

Let us now solve some problems using Birge-Vieta method.

Example 3:

Use Birge-Vieta method to find all the positive real roots, rounded off to three decimal places of the equation

$x^4 + 7x^3 + 24x^2 + x – 15 = 0$

Stop the iteration whenever $| x_{i+1} – x_i | < 0.0001$

Solution:

We first note that the given equation
$p4(x) = x^4 + 7x^3 + 24x^2 + x – 15 = 0$
is of degree 4. Therefore, by Theorem 1, this equation has 4 roots. Since there is only one change of sign in the coefficients of this equation, Descartes' rule of signs (see Theorem 4), states that the equation can have at most one positive real root.

Now let us examine whether the equation has a positive real root.

Since p4(0) = -15 and p4(1) = 19, by Intermediate value theorem, the equation has a root lying in]0, 1[.

We take x_0 = 0.5 as the initial approximation to the root. The first iteration is given by

$$x_1 = x_0 - \frac{p_4(x_0)}{p'_4(x_0)}$$

$$= 0.5 - \frac{p_4(0.5)}{p'_4(0.5)}$$

Now we evaluate $p_4(0.5)$ and $p'_4(0.5)$ using Horner's method. The results are given in the following table.

Table 6

	1	7	24	1	-15
0.5		0.5	3.75	13.875	7.4375
	1	7.5	27.75	14.875	-75625=$p_4(0.5)$
0.5		0.5	4.00	15.875	
	1	8.0	31.75	30.750 = $p'_4(0.5)$	

Therefore $x_1 = 0.5 - \dfrac{-7.5625}{30.75} = 0.7459$

The second iteration is given by

$x_2 = x_1 - \dfrac{p_4(x_1)}{p'_4(x_1)} = 0.7459 - \dfrac{p_4(0.7459)}{p'_4(0.7459)}$

Uisng synthetic division, we form the following table of values

Table 7

	1	7	24	1	-15
0.7459		0.7459	5.7777	22.2119	17.3138
	1	7.7459	29.7777	23.2119	2.3138
0.7459		0.7459	6.3340336		26.935717
	1	8.4918	36.111701		50.146879

Therefore $x_2 = 0.7459 - \dfrac{2.3132}{50.1469} = 0.6998$

Third iteration is given by

$x_3 = x_2 - \dfrac{p_4(0.6998)}{p'_4(0.6998)}$

Table 8

	1	7	24	1	-15
0.6998		0.6998	5.3881	20.5649	15.0905
	1	7.6998	29.3881	21.5649	0.0905
0.6998		.6998	5.8778	24.6780	
	1	8.3996	35.2659	46.2429	

$x_3 = 0.6998 - \dfrac{0.0905}{46.2429} = 0.6978$

For the fourth iteration we have

$x_4 = x_3 - \dfrac{p_4(0.6978)}{p'_4(0.6978)}$

Table 9

	1	7	24	1	-15
0.6978		0.6978	5.3715248	20.495459	14.999525
	1	7.6978	29.3715248	21.495459	0.0905
0.6978		.6978	5.8584497	24.583476	
	1	8.3956	35.229975	46.078926	

$$x_4 = 0.6978 - \frac{0.0005}{46.0789} = 0.6978$$

Since x_3 and x_4 are the same, we get $| x_4 – x_3 | < 0.0001$ and therefore we stop the iteration here. Hence the approximate value of the root rounded off to three decimal places is 0.698.

Next we shall illustrate how Birge-Vieta's method helps us to find all real roots of a polynomial equation.

Consider Eqn. (4)

$$p(x) = (x - \alpha)(b_n x^{n-1} + b_{n-1}x^{n-2} + ... + b_2 x + b_1) + b_0$$

If α is a root of the equation $p(x) = 0$, then $p(x)$ is exactly divisible by $x - \alpha$, that is, $b_0 = 0$. In finding the approximations to the root by the Birge-Vieta method, we find that b_0 approaches zero ($b_0 \rightarrow 0$) as x_i approaches α ($x_i \rightarrow \alpha$). Hence, if x_n is taken as the final approximation, to the root satisfying the criterion $| x_n – x_{n-1} | < \varepsilon$, then to this approximation, the required quotient is

$$q_{n-1}(x) = b_n x^{n-1} + b_{n-1}x^{n-2} + ... + b_1$$

where b'_1 are obtained by using x_n and the Horner's method. This polynomial is called the deflated polynomial or reduced polynomial. The next root is now obtained using $q_{n-1}(x)$ and not $p_n(x)$. Continuing this process, we can successively reduce the degree of the polynomial and find one real root at a time.

Let us consider an example.

Example 4:

Find all the roots of the polynomial equation $p_3(x) = x^3 + x – 3 = 0$ rounded off to three decimal places. Stop the iteration whenever $|x_{i+1} – x_i| < 0.0001$.

Solution:

The equation $p_3(x) = 0$ has three root. Since there is only one change in the sign of the coefficients, by Descartes' rule of signs the equation can have at most one positive real root. The equation has no negative real root since $p_3(-x) = 0$ has no change of sign of coefficients. Since $p_3(x) = 0$ is of odd degree it has at least one real root. Hence the given equation $x^3 + x – 3 = 0$ has one positive real root and a complex pair. Since $p(1) = -1$ and $p(2) = 7$, by intermediate value theorem the equation has a real root lying in the interval]1, 2[. Let us find the real root using Birge-Vieta Method. Let the initial approximation be 1.1.

First iteration

Table 10

	1	0	14	-3
1.1		1.1	1.21	2.431
	1	1.1	2.21	0.0905
1.1		1.1	2.42	
	1	2.2	4.63	

Therefore $x_1 = 1.1 - \dfrac{-0.569}{4.63} = 1.22289$

Similarly, we obtain
$x_2 = 1.21347$
$x_3 = 1.21341$

Since $| x_2 – x_3 | < 0.0001$, we stop the iteration here. Hence the required value of the root is 1.213, rounded off to three decimal places. Next let us obtain the deflated polynomial of $p_3(x)$. To get the deflated polynomial of, we have to find the polynomial $q_2(x)$ by using the final approximation $x_3 = 1.213$ (see Table 11).

Table 11

	1	0	1	-3
1.213		1.213	1.4714	2.9978
	1	1.213	2.4714	-0.0022

Note that $p_3(1.213) = -0.0022$. That is, the magnitude of the error in satisfying $p_3(x_3) = 0$ is 0.0022.
We find $q_2(x) = x^2 + 1.213x + 2.4714 = 0$

This is a quadratic equation and its roots are given by

$$x = \frac{-1.213 \pm \sqrt{(1.213)^2 - 4 \times 2.4714}}{2}$$

$= \dfrac{-1.213 \pm 2.9009i}{2}$

$= 0.\ 6065 \pm 1.4505\ i$

Hence the three roots of the equation rounded off to three decimal places are 1.213, 0.6065 + 1.4505 i and -0.6065 – 1.4505 i.

Remark: We now know that we can determine all the real roots of a polynomial equation using deflated polynomials. This procedure reduces the amount of computations also. But this method has certain limitations. The computations using deflated polynomial can cause unexpected errors. If the roots are determined only approximately, the coefficients of the deflated polynomials will contain some errors due to rounding off. Therefore we can expect loss of accuracy in the remaining roots. There are some ways of minimizing this error. We shall not be going into the details of these refinements.

3.3 Graeffe's Root Squaring Method

In the last section we have discussed a method for finding real roots of polynomial equations. Here we shall discuss a direct method for solving polynomial equations. This method was developed independently by three mathematicians Dandelin, Lobachesky and Graeffe. But Graeffe's name is usually associated with this method. The advantage of this method is that it finds all roots of a polynomial equation simultaneously: the roots may be real and distinct, real and equal (multiple) or complex roots.

The underlying idea of the method is based on the following fact: Suppose $\beta_1, \beta_2, ..., \beta_n$ are the n real and distinct roots of a polynomial equation of degree n such that they are widely separated, that is,

$|\beta_1| >> |\beta_2| >> |\beta_3| >> ... >> |\beta_n|$

where >> stands for 'much greater than'. Then we can obtain the roots approximately from the coefficients of the polynomial equation as follows:

Let the polynomial equation whose roots are $\beta_1, \beta_2, ..., \beta_n$ be

$a_0 + a_1x + a_2x^2 + ... + a_nx^n = 0,\ a_n \neq 0.$

Using the relations between the roots and the coefficients of the polynomial as given in Sec. 4.2, we get

$$\beta_1 + \beta_2 + \ldots + \beta_n = -\frac{a_{n-1}}{a_n}$$

$$\beta_1, \beta_2 + \beta_1\beta_3 + \ldots + \beta_{n-1}\beta_n = \frac{a_{n-2}}{a_n}$$

$$\beta_1\beta_2\beta_3 + \ldots + \beta_{n-2}\beta_{n-1}\beta_n = -\frac{a_{n-3}}{a_n} \tag{14}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\beta_1\beta_2 \ldots \beta_n = (-1)^n \frac{a_0}{a_n}$$

Since $|\beta_1| >> |\beta_2| >> |\beta_3| >> \ldots >> |\beta_n|$, we have from (14) the approximations

$$\left.\begin{aligned}
&\beta_1 \approx -\frac{a_{n-1}}{a_n} \\
&\beta_1\beta_2 \approx \frac{a_{n-2}}{a_n} \\
&\beta_1\beta_2\beta_3 \approx -\frac{a_{n-3}}{a_n} \\
&\ldots \\
&\ldots \\
&\beta_1\beta_2 \ldots \beta_n \approx (-1)^n \frac{a_0}{a_n}
\end{aligned}\right\} \tag{15}$$

These approximations can be simplified as

$$\left.\begin{aligned}
&|\beta_1| \approx -\frac{a_{n-1}}{a_n} \\
&|\beta_2| \approx \frac{a_{n-2}}{a_n}\frac{a_n}{a_{n-1}} \approx \frac{a_{n-2}}{a_{n-1}} \\
&|\beta_3| \approx -\frac{a_{n-3}}{a_n}\frac{a_{n-1}}{a_{n-2}}\frac{a_n}{a_{n-1}} = \frac{a_{n-3}}{a_{n-2}} \\
&. \\
&. \\
&.
\end{aligned}\right\} \tag{16}$$

$$|\beta_n| \approx \frac{a_0}{a_1}$$

So the problem now is to find from the given polynomial equation, a polynomial equation whose roots are widely separated. This can be done by the method which we shall describe now.

In the present course we shall discuss the application of the method to a polynomial equation whose roots are real and distinct.

Let α_1, α_2, ..., α_n be the n real and distinct roots of the polynomial equation of degree n given by

$$a_0 + a_1x + a_2x^2 + ... + a_nx^n = 0. \qquad (17)$$

where $a_0, a_1, a_2, ..., a_{n-1}, a_n$ are real numbers and $a_n \neq 0$. We rewrite Eqn. (17) by collecting all even terms on one side and all odd terms on the other side, i.e.

$$a_0 + a_2x^2 + a_4x^4 + ... = -(a_1x + a_3x^3 + a_5x^5 + ...) \qquad (18)$$

Squaring both sides of Eqn. (18), we get

$$(a_0 + a_2x^2 + a_4x^4 + ...)^2 = (a_1x + a_3x^3 + a_5x^5 + ...)^2$$

Now we expand both the right and left sides and simplify by collecting the coefficients. We get

$$a_0^2 - (a_1^2 - 2a_0a_2)x^2 + (a_2^2 - 2a_1a_3 + 2a_0a_4)x^4 –$$
$$(a_3^2 - 2a_2a_4 + 2a_1a_5 - 2a_0a_6)x^6 + ... + (-1)^n a_n^2 x^{2n} = 0 \qquad (19)$$

Putting $x^2 = -y$ in Eqn. (19), we obtain a new equation given by

$$b_0 + b_1y + b_2y^2 + ... + b_n = 0 \qquad (20)$$

where

$b_0 = a_0^2$

$b_1 = a_1^2 - 2a_0a_2$

$b_2 = a_2^2 - 2a_1a_3 + 2a_0a_4$

$b_n = a_n^2$

The following table helps us to compute the coefficients $b_0, b_1, ..., b_n$ of Eqn. (20) directly from Eqn. (17).

Table 12

a_0	a_1	a_2	$a_{3...}$	a_n
a_0^2	a_1^2	a_2^2	a_3^2	a_n^2
0	$-2a_0a_2$	$-2a_1a_3$	$-2a_2a_4$	0
0	0	$-2a_0a_4$	$-2a_1a_5$	0
0	0	0	$-2a_0a_6$	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
b_0	b_1	b_2	b_3...	b_n

To form Table 12 we first write the coefficients a_0, a_1, a_2, ..., a_n as the first row. Then we form (n + 1) columns as follows.

The terms in each column alternate in sign starting with a positive sign. The first term in each column is the square of the coefficients a_k, k = 0, 1, 2, ..., n. The second term in each column is twice the product of the nearest neighbouring coefficients, if there are nay with negative sign: otherwise put it as zero. For example, the second term in the first column is zero and second term in the second column is $-2a_0\ a_2$. Likewise the second term of the $(k + 1)^{th}$ column is $2a_{k-1}\ a_{k+1}$. The third term in the $(k + 1)^{th}$ column is twice the product of the next neighbouring coefficients a_{k-2} and a_{k+2}, if there are nay, otherwise put it as zero. This procedure is continued until there are no coefficients available to form the cross products. Then we add all the term in each column. The sum gives the coefficients b_k for k = 0, 1, 2, ..., n which are listed as the last term in each column. Since the substitution $x^2 = -y$ is used, it is easy to see that if α_1, α_2, ..., α_n are the n roots of Eqn. (17), then $-\alpha_1^2$, α_2^2, ..., α_n^2 are the roots of Eqn. (20).

Thus, starting with a given polynomial equation, we obtained another polynomial equation whose roots are the squares of the roots of the original equation with negative sign.

We repeat the procedure for Eqn. (20) and obtain another equation

$c_0 + c_1x + ... + c_nx^n = 0.$

Whose roots are the squares of the roots of Eqn. (20) with a negative sign i.e. they are fourth powers of the roots of the original equation with a negative sign. Let this procedure be repeated n times. Then, we obtain an equation

$$q_0 + q_1x + ... + q_nx^n = 0 \qquad (21)$$

whose roots γ_1, γ_2, ..., γ_n are given by

γ_i =, i = 0, 1, 2, ..., n. (22)

Now, since all the roots of Eqn. (17) are real and distinct, we have

$|\alpha_1| > |\alpha_2| > \dots\dots\dots\dots > |\alpha_n|$

Hence $|\gamma_1| = |\alpha_1^{2^m}| = \left|\frac{q_{n-1}}{q_n}\right|$

$|\gamma_2| = |\alpha_2^{2^m}| = \left|\frac{q_{n-2}}{q_{n-1}}\right|$

. . .

. . .

. . .

$|\gamma_n| = |\alpha_n^{2^m}| = \left|\frac{q_0}{q_1}\right|$

The magnitude of the roots of the original equations are therefore given by

$|\alpha_1| = \sqrt[2^m]{\frac{q_{n-1}}{q_n}}$

$|\alpha_2| = \sqrt[2^m]{\frac{q_{n-2}}{q_{n-1}}}$

.

.

.

$|\alpha_n| = \sqrt[2^m]{\frac{q_0}{q_1}}$

This gives the magnitude of the roots. To determine the sign of the roots, we substitute these approximations in the original equation and verify whether positive or negative value satisfies it.

We shall now illustrate this method with an example.

Example 5:

Find the roots of the cubic equation $x^3 - 15x^2 + 62x - 72 = 0$ by Graeffe's method using three squaring.

Solution:

Let $P_3(x) = x^3 - 15x^2 + 62x - 72 = 0$.

The equation has no negative real roots. Let us now apply the root squaring method successively. The get the following results:

First Squaring

Table 13

a_0 -72	a_1 62	a_2 -15	a_3 1
$a_0^2=5184$	$a_1^2=3844$	$a_2^2=225$	$a_3^2=1$
0	$-2a_0a_2=-2160$	$-2a_1a_3=-124$	0
5184	1684	101	1
b_0	b_1	b_2	b_3

Therefore the new equation is

$x^3 + 101x^2 + 168x + 5184 = 0$.

Applying the squaring method to the new equation we get the following results.

Second Squaring

Table 14

5184	1684	101	1
26873856	2835856	10201	1
0	-1047168	-3368	0
26873856	1788688	6833	1

Thus the new equation is

$x^3 + 6833x^2 + 1788688x + 26873856 = 0$.

For the third squaring, we have the following results.

Third Squaring

Table 15

26873856	1788688	6833	1
7.2220414×10^{14}	3.1994048×10^{12}	46689889	1
0	-3672581×10^{12}	-3577376	0
7.2220414×10^{14}	2.83214×10^{12}	43112513	1
q_0	q_1	q_2	q_3

Hence the new equation is

$$x^3 + 43112513x^2 + (2.83214 \times 10^{12})x + (7.2220414 \times 10^{14}) = 0$$

After three squaring, the roots γ_1, γ_2, and γ_3 of this equation are given by

$$|\gamma_1| = \left|\frac{q_2}{q_3}\right| = 43112513$$

$$|\gamma_2| = \left|\frac{q_1}{q_2}\right| = \frac{2.83214 \times 10^{12}}{43112513}$$

$$|\gamma_3| = \left|\frac{q_0}{q_1}\right| = \frac{7.22204 \times 10^{14}}{2.83214 \times 10^{12}}$$

Hence, the roots

$$|\alpha_1| = \sqrt[8]{443112513} = 9.0017$$

$$|\alpha_2| = \sqrt[8]{\frac{2.83214 \times 10^{12}}{43112513}} = 4.0011$$

$$|\alpha_3| = \sqrt[8]{\frac{7.22204 \times 10^{14}}{2.83214 \times 10^{12}}} = 1.9990$$

Since the equation has no negative real roots, all the roots are positive. Hence the roots can be taken as 9.0017, 4.0011 and 1.9990. If the approximations are rounded to 2 decimal places, we have the roots as 9, 4 and 2. Alternately, we can substitute the approximate roots in the given equation and find their sign.

4.0 CONCLUSION

We have seen that Graeffe's root squaring method obtain all real roots simultaneously. There is considerable saving in time also. The method can be extended to find multiple and complex roots also. However the method is not efficient to find these roots. We shall not discuss these extensions.

We shall end this block by summarizing what we have covered in this unit.

5.0 SUMMARY

In this unit we have

- discussed the following methods for finding approximate roots of polynomial equations.

    i) Birge-Vieta method.

    ii) Graeffe's root squaring method.

- Mentioned the advantage and disadvantages of the above methods.

6.0 TUTOR-MARKED ASSIGMENT (TMA)

1) Find the quotient and the remainder when $2x^3 – 5x^2 + 3x – 1$ is divided by $x – 2$.

2) Using synthetic division check whether $\alpha_0 = 3$ is a root of the polynomial equation $x^4 + x^3 – 13x^2 – x + 12 = 0$ and find the quotient polynomial.

3) How many negative roots does the equation $3x^7 + 5x^5 + 4x^3 + 10x – 6 = 0$ have? Also determine the number of positive roots, if any.

4) Show that the biquadratic equation
$p(x) = x^4 + x^3 – 2x^2 + 4x – 24 = 0$ has at least two real roots one positive and the other negative.

5) Using synthetic division, show that 2 is a simple root of the equation
$p(x) = x^4 – 2x^3 – 7x^2 + 8x + 12 = 0$.

6) Evaluate p(0.5) and p'(0.5) for
$p(x) = -8x^5 + 7x^4 – 6x^3 + 5x^2 – 4x + 3$

7) Find an approximation to one of the roots of the equation
$p(x) = 2x^4 – 3x^2 + 3x – 4 = 0$
using Birge-Vieta method starting with the initial approximation $x_0 = -2$. Stop the iteration whenever $| x_{i+1} – x_i | < 0.4 \times 10^{-2}$.

8) Find all the roots of the equation $x^3 – 2x – 5 = 0$ using Birge-Vieta method.

9) Find the real root rounded off to two decimal places of the equation $x^4 - 4x^3 - 3x + 23 = 0$ lying in the interval]2, 3[by Birge-Vieta method.

10) Determine all roots of the following equations by Graeffe's root squaring method using three squaring.

i) $x^3 + 6x^2 - 36x + 40 = 0$

ii) $x^3 - 2x^2 - 5x + 6 = 0$

iii) $x^3 - 5x^2 - 17x + 20 = 0$

7.0 REFERENCES/FURTHER READINGS

Engineering Mathematics P.D.S. Verma.

Generalized Functions in Mathematical Physics by V.S. Viadimirov.

Fundamentals of the Finite Element Method. Hartley Grandin, Fr.