SHIVAJI UNIVERSITY, KOLHAPUR

CENTRE FOR DISTANCE AND ONLINE EDUCATION

Numerical Analysis

(Mathematics)

For

M. Sc.-I, Sem. II

(In accordance with National Education Policy 2020)

(Academic Year 2022-23 onwards)

First Edition 2014
Second Edition 2015
Third Edition 2016
Fourth Edition 2019
Revised Edition 2023

Prescribed for **M. Sc. Part-I**

Copies : 500

Published by:
Dr. V. N. Shinde
Ag. Registrar,
Shivaji University,
Kolhapur-416 004

Printed by :

Shri. B. P. Patil
Superintendent,
Shivaji University Press,
Kolhapur-416 004

ISBN-978-81-8486-537-0

★ Further information about the Centre for Distance and Online Education & Shivaji University may be obtained from the University Office at Vidyanagar, Kolhapur-416 004, India.

Centre for Distance and Online Education
Shivaji University, Kolhapur

ADVISORY COMMITTEE

Prof. (Dr.) D. T. Shirke
Honourable Vice Chancellor,
Shivaji University, Kolhapur

Prof. (Dr.) P. S. Patil
Honourable Pro-Vice Chancellor,
Shivaji University, Kolhapur

Prof. (Dr.) Prakash Pawar
Department of Political Science,
Shivaji University, Kolhapur

Prof. (Dr.) S. Vidyashankar
Vice-Chancellor, KSOU,
Mukthagangotri, Mysuru, Karnataka

Dr. Rajendra Kankariya
G-2/121, Indira Park,
Chinchwadgaon, Pune

Prof. (Dr.) Smt. Cima Yeole
Git Govind, Flat No. 2,
1139 Sykes Extension, Kolhapur

Dr. Sanjay Ratnaparkhi
D-16, Teachers Colony,
Vidhyanagari, Mumbai University,
Santacruz (E), Mumbai

Prof. (Dr.) Smt. Kavita Oza
Department of Computer Science,
Shivaji University, Kolhapur

Prof. (Dr.) Chetan Awati
Department of Technology,
Shivaji University, Kolhapur

Prof. (Dr.) M. S. Deshmukh
Dean, Faculty of Humanities,
Shivaji University, Kolhapur

Prof. (Dr.) S. S. Mahajan
Dean, Faculty of Commerce and
Management, Shivaji University, Kolhapur

Prof. (Dr.) Smt. S. H. Thakar
I/c. Dean, Faculty of Science and
Technology, Shivaji University, Kolhapur

Prin. (Dr.) Smt. M. V. Gulavani
I/c. Dean, Faculty of Inter-disciplinary
Studies, Shivaji University, Kolhapur

Dr. V. N. Shinde
Ag. Registrar,
Shivaji University, Kolhapur

Dr. A. N. Jadhav
Director, Board of Examinations and
Evaluation, Shivaji University, Kolhapur

Shri. A. B. Chougule
I/c. Finance and Accounts Officer,
Shivaji University, Kolhapur

Prof. (Dr.) D. K. More
(Member Secretary) Director,
Centre for Distance and Online Education,
Shivaji University, Kolhapur.

Centre for Distance and Online Education
Shivaji University, Kolhapur

MEMBERS OF B.O.S. IN MATHEMATICS

Chairman**- Prof. Dr. Kishor Kucche**
Department of Mathematics, Shivaji University, Kolhapur

- **Dr. Mrs. S. H. Thakar**
  Head, Dept. of Mathematics,
  Shivaji University, Kolhapur.
- **Dr. Girish Dhondiram Shelake**
  Willingdon College, Sangli
- **Dr. Smt. Bebitai Annaso Sajane**
  Smt. Kasturbai Walchand College of Arts & Science, Sangli
- **Dr. Dadasaheb Rajaram Phadatare**
  Balasaheb Desai College, Patan,
  Tal. Karad, Dist. Satara
- **Dr. Santosh Bhaurao Joshi**
  Walchand College of Engineering,
  Sangli
- **Dr. Sanjay Anant Morye**
  Rajaram College, Vidyanagar,
  Kolhapur
- **Dr. Hambirrao Tatyasaheb Dinde**
  Shri Shiv Shahu Mahavidyalaya,
  Sarud, Tal. Shahuwadi, Dist.
  Kolhapur
- **Dr. Santaji Shrikant Khopade**
  Karmaveer Hire Arts, Science,
  Commerce and Education College,
  Gargoti, Tal. Bhudaragad, Dist.
  Kolhapur

Preface

The Shivaji University, Kolhapur has established the Distance and Online Education Centre for external students from the year 2022-23, with the goal that, those students who are not able to complete their studies regularly, due to unavoidable circumstances, they must be involved in the main stream by appearing externally. The centre is trying hard to provide notes to those aspirants by entrusting the task to experts in the subjects to prepare the Self Instructional Material (SIM). Today we are extremely happy to present a book on Numerical Analysis for M. Sc. Mathematics students as SIM prepared by us. The SIM is prepared strictly according to syllabus NEP 2020 and we hope that the exposition of the material in the book will meet the needs of all students.

This book has grown from the lectures we deliver in the Department of Mathematics at Shivaji University, Kolhapur. The book is based on the curriculum recommended for M. Sc. Mathematics at Shivaji University, Kolhapur.

This book has four units. Unit 1 provides an introduction to error analysis and methods to estimate roots of polynomial and Transcendental equations. This unit deals with direct and iterative method for finding the roots of transcendentel and polynomial equations. In unit 2, the direct and iterative methods for the solution of a system of linear algebraic equations are discussed. The error analysis and convergence of iterative methods are also discussed. Various methods for finding eigenvalues and corresponding eigen vectors are explained. Unit 3 gives the numerical methods of differentiation and integration. Lagrange's interpolation and Newton's divided difference formula is derived that approximates a function by a polynomial of given degree. Uniqueness of interpolating polymial is proved. Error analysis for Lagvage's interpolation is carried out. Various methods for numerical differentiation and numerical integration are discussed along with their error analysis. Unit 4 deals with numerical solutions of ordinary differential Equations. Various methods used to determine the numerical solutions of ordinary differential Equations are discussed. Error analysis is for all the methods is given.

All the units are followed by solved problems. A good number of examples have been solved at the end of each unit to enable the student to understand the concepts described in the text. Good number of excercises are given at the end of each unit.

We hope that the content of the SIM will be helpful for the students having their education in distance mode.

Editor

Writing Team	Unit
Prof. Dr. (Mrs.) S. H. Thakar Department of Mathematics, Shivaji University, Kolhapur. (Maharashtra)	**1, 2, 3**
Dr. M. T. Gophane Dept. of Mathematics, Shivaji University, Kolhapur. (Maharashtra)	**4**

■ **Editor** ■

Prof. Dr. (Mrs.) S. H. Thakar
Department of Mathematics,
Shivaji University, Kolhapur.
(Maharashtra)

M. Sc. (Mathematics)

Numerical Analysis

Contents

Each Unit begins with the section objectives -

Objectives are directive and indicative of :

1. what has been presented in the unit and
2. what is expected from you
3. what you are expected to know pertaining to the specific unit, once you have completed working on the unit.

The exercises at the end of each unit are not to be submitted to us for evaluation. They have been provided to you as study tools to keep you in the right track as you study the unit.

Dear Students

The SIM is simply a supporting material for the study of this paper. It is also advised to see the new syllabus 2022-23 and study the reference books & other related material for the detailed study of the paper.

INTRODUCTION

Numerical analysis involves the study, development and analysis of algorithms for obtaining numerical solutions to various mathematical problems. Frequently numerical analysis is called the mathematics of scientific computing. Numerical analysis is the development and study of procedures for solving problems with computer. The art and science of preparing and solving scientific and engineering problems have undergone considerable changes due to the available digital computing systems.

Digital computers are the principal means of calculation in numerical analysis and consequently it is very important to understand how they operate. A computer has a finite word length and so only a fixed number of digits are stored and used during computation.

1. Errors

Even in storing an exact decimal number in its converted form in the computer memory, an error is introduced. This error is machine dependent. Also at the end of computation of aparticular problem, the final result in the computer should be converted into a form understandable to the user. Therefore an additional error is committed at this stage too. This error is called local round off error. Thus we define

Error = True Value – Computed Value

In order to determine the accuracy of an approximate solution, errors are measured in different ways.

Definition 2 : Absolute error = | error |

Definition 3 : Relative error = $\dfrac{|\text{Error}|}{|\text{True Value}|}$

Definition 4 : Round Off Error : is the quantity R which must be added to the finite representation of a computed number in order to make it the true representation of that number.

When a number N is written in floating point form with t digits, say, in base 10 as,

$$N = \left(0 \cdot d_1 d_2 d_3 .... d_t\right) 10^e, \; d_1 \neq 0$$

We say that the number N has t significant digits. For example, 0.3 agrees with $\frac{1}{3}$ to one significant digit. The round off error for this representation will be $\frac{1}{3} - 0.3$.

All the errors defined above are machine errors and can be minimized by using computing aids of higher precision.

Methematically, in numerical analysis we usually come across two types of errors.

(i) Inherent Errors

It is that quantity of error which is present in the statement of the problem itself, before finding its solution. It arises due to the simplified assumptions made in the mathematical modelling of a problem. It can also arise when the data is obtained from certain physical measurements of the parameters of the problem.

(ii) Truncation Errors

These are errors caused by using approximate formulae in computations. e.g. when a function $f(x)$ is evaluated from an infinite series, if we use only first few terms of the series to compute value of function $f(x)$, we get an approximate answer. Here, the error is due to truncating the series.

Suppose $f(x) = \cos x$. Then

$$f(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} ..... + (-1)^n \frac{x^{2n}}{(2n)!} + ...$$

If we retain the first n terms, the truncation error (TE) is

$$TE = (-1)^{n+1} \cdot \frac{x^{2n+2}}{(2n+2)!} + (-1)^{n+2} \frac{x^{2n+4}}{(2n+4)!} + (-1)^{n+3} \frac{x^{2n+6}}{(2n+6)!} .....$$

The study of this type of error is associated with the problem of convergence. Some special terminology is used to describe the rapidity with which a sequence converges.

Big O and Little o Notation :

Let $\{x_n\}$ and $\{\alpha_n\}$ be two different sequences.

The equation $x_n = o(\alpha_n)$ (we say x_n is "little oh" of α_n)

$$\lim_{n \to \infty} \frac{x_n}{\alpha_n} = 0$$

To avoid division by zero, we say that

$$x_n = o(\alpha_n) \text{ if } |x_n| \le \varepsilon_n |\alpha_n| \text{ and } \varepsilon_n \to 0 \text{ as } n \to \infty .$$

We write $x_n = o(\alpha_n)$ (we say x_n is "big oh" of α_n) if there is a constant C and number n_0 such that $|x_n| \le C|\alpha_n|$ when $n \ge n_0$.

These two notations give a coarse method of comparing two sequences. They are frequently used when both sequences converge to 0. If $x_n \to 0$, $\alpha_n \to 0$ and $x_n = o(\alpha_n)$, then x_n converges to 0 at least as rapidly as α_n does. If $x_n \to 0$, $\alpha_n \to 0$ and $x_n = o(\alpha_n)$ then x_n converges to 0 more rapidly than α_n does.

Definition : The truncation error is the quantity T which must be added to the true representation of the quantity in order that the result is exactly equal to the quantity we are seeking to generate.

2. Stability in Numerical Analysis

A number of mathematical problems have solutions that are quite sensitive to small computational errors, for example round off error. To deal with this phenomenon, we introduce the concept of stability. A numerical method for solving mathematical problem is considered stable if the sensitivity of the numerical answer to the data is no greater than in the original mathematical problem.

A numerical method is said to be stable if the effect of any single fixed round off error is bounded.

3. Problem Solving Using Computers

In order to solve a given problem using computer the major steps involved are -

(i) Choosing an appropriate numerical method

(ii) Designing an algorithm

(iii) Programming

(iv) Computer Execution.

In Unit1 to 4 we discuss various numerical methods (and their analysis) for solving transcendental and polynomial equations, system of linear equations, differential equations, numerical methods available to interpolate and approximate functions, integration and evaluation of eigen values and eigen vectors of symmetric matrices.

TRANSCENDENTAL AND POLYNOMIAL EQUATIONS

Introduction :

One of the basic problems in science and engineering is the computation of roots of an equation $f(x) = 0$. The equation $f(x) = 0$ is called algebraic or polynomial equation if it is purely a polynomial in x. It is called a transcendental equation if $f(x)$ contains trignometric, exponential or logarithmic functions.

Definition 1.0.2 :

A number ξ is a solution of $f(x) = 0$ if $f(\xi) = 0$. Such ξ is called root or zero of $f(x) = 0$.

Geometrically, a root of $f(x) = 0$ is the value of x at which the graph of $y = f(x)$ intersects the x-axis.

Definition 1.0.2 :

If we can write $f(x) = (x - \xi)^m g(x)$ where $g(x)$ is bounded and $g(\xi) \neq 0$ then ξ is called a multiple root of multiplicity m. In this case $f(\xi) = f'(\xi) = f''(\xi) = ... = f^{(m-1)}(\xi) = 0$. and $f^{|m|}(\xi) \neq 0$ for m = 1 the root ξ is called simple root.

The following are the basic properties of polynomial equation.

(i) Every polynomial equation of n^{th} degree, where n is positive integer has exactly n roots.

(ii) Complex roots occur in pairs i.e. if a + ib is a root of $f(x) = 0$, so is a – ib.

(iii) If $x = a$ is a root of $f(x) = 0$, a polynomial of degree n then $f(x) = (x - a) g(x)$ where $g(x)$ is a polynomial of degree $(n - 1)$.

(iv) Descartes Rule of Signs :

The number of positive roots of a polynomial equation $f(x) = 0$ with real coefficients cannot exceed the number of changes in sign of the coefficients in the polynomial $f(x) = 0$. Similarly, the number of negative roots of $f(x) = 0$ cannot exceed the number of changes in the sign of the coefficients

of $f(-x) = 0$. For example, consider $f(x) = x^3 - 3x^2 + 4x - 5 = 0$. The coefficients of this equation are (1, –3, 4, –5). As there are three changes in sign, the given equation will have at the most three positive roots.

(v) Intermediate Value Property :

If $f(x)$ is a real valued continuous function in the closed interval $a \le x \le b$ then a function takes each value between $f(a)$ and $f(b)$. In particular if $f(a)$ and $f(b)$ have opposite signs, then the graph of a function $y = f(x)$ crosses the x-axis at least once. i.e. $f(x) = 0$ has at least one root between *a* and *b*.

i.e. $f(\xi) = 0$, $a < \xi < b$

There are generally two types of methods used to find roots of $f(x) = 0$.

(i) Direct Methods :

These methods give the exact value of the roots in a finite number of steps. Further the methods give all the roots at the same time. These methods require no knowledge of the initial approximation of a root of the equation $f(x) = 0$.

e.g. solutions of polynomial equation are known for polynomials of degree upto cubic. i.e.

for $a_0 x + a_1 = 0$, $x = -\dfrac{a_1}{a_0}$

for $a_0 x^2 + a_1 x + a_2 = 0$, $x = \dfrac{-a_1 \pm \sqrt{a_1^{\,2} - 4a_0 a_2}}{2a_0}$

(ii) Interrative Methods

These methods are based on the idea of successive approximations i.e. starting with one or more initial approximations to the root, we obtain a sequence of approximate solutions which converges to a root of given equation. In the next section we describe some numerical methods for the solution of equation $f(x) = 0$.

1.2 Bisection Method

This method is due to Bolzano.

Step 1 : Choose x_0 and x_1 such that $f(x_0)f(x_1)<0$ suppose $x_0<x_1$.

(By intermediate value principle root lies between x_0 and x_1)

Define $I_0=(x_0,x_1)$.

Step 2 : The desired root is approximately defined by $x_2=\frac{x_0+x_1}{2}$.

If $f(x_2)=0$ then x_2 is the desired root of $f(x)=0$.

If $f(x_2)\neq 0$, calculate $f(x_0)f(x_2)$.

If $f(x_0)f(x_2)<0$ then define $I_1=(x_0,x_2)$.

Otherwise define $I_1=(x_2,x_1)$ and $f(x_2)f(x_1)<0$.

Step 3 : Define $I_0=I_1$ and go to step 1. Thus at each iteration we either find the desired root to the required accuracy or narrow the length of interval to half of the length of interval at previous step. This process is continued to determine a smaller and smaller interval within which the desired root lies. If he permissible error is ε, then the approximate no. of iterations (n) required may be determined from the relation

$$\frac{x_1-x_0}{2^n}\leq\varepsilon$$

Note : The no. of iterations required to achieve required accuracy depends upon the initial interval I_0. If the length of I_0 is sufficiently small we will reach at the solution in less no. of iterations.

EXAMPLES .....

1.2.1 : Find a real root of the equation, $f(x)=x^3-x_1-1=0$

Answer :

Step 1 : Since $f(1)=-1<0$ and $f(2)=5>0$, the root lies between 1 and 2.

i.e. $I_0=(x_0,x_1)=(1,2)$.

Step 2 : $x_2=\frac{x_0+x_1}{2}=1.5$, $f(x_2)=f(1.5)=$

1.3 Iteration Methods Based on First Degree Equation

Although the bisection method is easy to compute it is not very efficient. For most functions we can improve the speed at which the root is approach through different schemes. Almost every functions can be approximated by a straight line over a small interval. We begin from a value that is near to a root. This initial value can be obtained by looking at the graph of a function or from few iterations of bisection method.

Iteration methods are obtained by approximating $f(x)$ by a polynomial of degree one in the neighbourhood of root. Thus

$$f(x) = a_0 x + a_1 = 0 \Rightarrow x = -\frac{a_1}{a_0},\ a_0 \neq 0 \qquad \text{..... (1.3.1.)}$$

The parameters a_0 and a_1 are to be determined by prescribing two approximate conditions on $f(x)$ and / or its derivatives.

1.3.1 Secant Method

Suppose x_{k-1} and x_k are two approximations to the root, then we determine a_0 and a_1 by using linear approximation.

$$f\left(x_k\right) = a_0 x_k + a_1$$

$$f\left(x_{k-1}\right) = a_0 x_{k-1} + a_1$$

On solving above two equations simultaneously for a_0 and a_1 we get

$$a_0 = \frac{f\left(x_k\right) - f\left(x_{k-1}\right)}{x_k - x_{k-1}} \quad \text{and} \quad a_1 = \frac{x_k f\left(x_{k-1}\right) - x_{k-1} f\left(x_k\right)}{x_k - x_{k-1}}$$

from equation (1.3.1) we get, the next approximate root

$$x_{k+1} = \frac{x_{k-1} f\left(x_k\right) - x_k f\left(x_{k-1}\right)}{f\left(x_k\right) - f\left(x_{k-1}\right)}$$

which may be written as

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f_k - f_{k-1}} f_k,\ k = 1, 2, 3, ... \qquad .... (1.3.2)$$

where $f_k = f\left(x_k\right)$ and $f_{k-1} = f\left(x_{k-1}\right)$.

This is called the Secant or the Chord Method.

Geometrically, in this method we choose two points on the curve and plot the line passing through these two points. The point of intersection of the straight line with the x-axis is the next approximation to the root (Fig. 1.1).

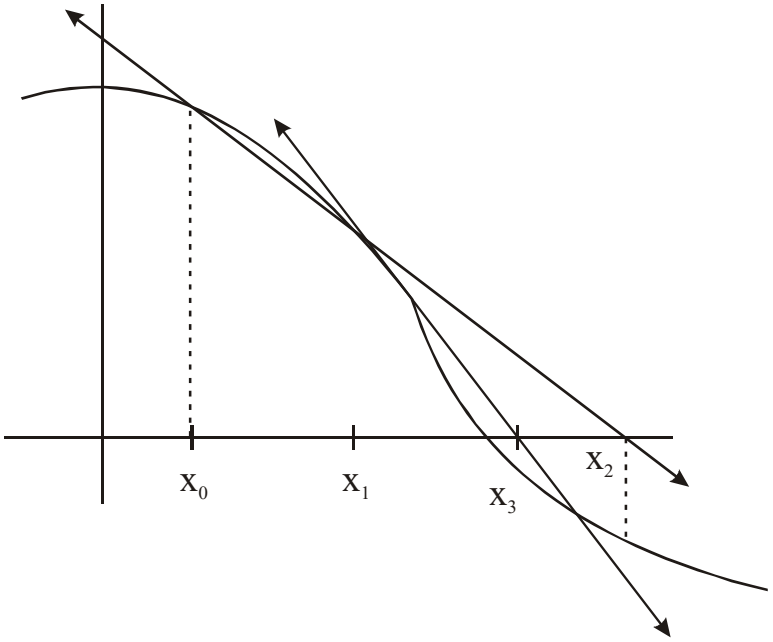


Fig. 1.1

1.3.2 Regula Falsi Method

This is the oldest method for finding the real root of an equation $f(x) = 0$ and closely resembles bisection method. This method is also called as method of false position. In this method we choose two points x_0 and x_1 such that $f(x_0)$ and $f(x_1)$ are of opposite signs. Since the graph of $y = f(x)$ crosses the x-axis between these two points, a root must lie in between these points. Now the equation of the Chord joining the two points $\left(x_0, f\left(x_0\right)\right), \left(x_1, f\left(x_1\right)\right)$ is

$$y - f\left(x_0\right) = \left[\frac{f\left(x_1\right) - f\left(x_0\right)}{x_1 - x_0}\right]\left(x - x_0\right)$$

The point of intersection of the chord with the x-axis is given by putting $y = 0$. Thus we get,

$$-f\left(x_0\right) = \frac{f\left(x_1\right) - f\left(x_0\right)}{x_1 - x_0}\left(x - x_0\right)$$

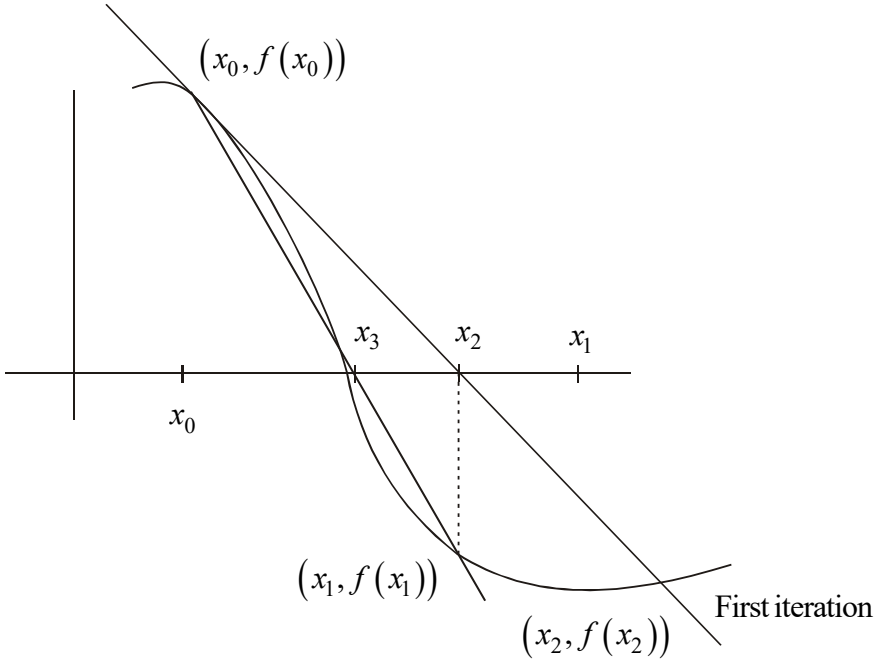
On solving above equation for x we obtain

$$x = x_0 - \frac{f\left(x_0\right)}{f\left(x_1\right) - f\left(x_0\right)}\left(x_1 - x_0\right)$$

Hence the second approximation to the root of $f(x)=0$ is given by

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1)-f(x_0)}(x_1 - x_0) \qquad \text{.... (1.3.2.1)}$$

If now $f(x_2)$ and $f(x_0)$ are of opposite signs then the root lies between x_0 and x_2 and we replace x_1 by x_2 in (1.3.2.1) and obtain the next approximation. Otherwise we replace x_0 by x_2 and generate next approximation. The procedure is repeated till the root is obtained to the desired accuracy. The repeated application of this procedure generates a sequence.



Suppose the approximate solution after $(k-1)$ iterations is denoted by x_k. Then the sequence $\{x_k\}$ approaches to the root ξ as $k \to \infty$ i.e. $f(\xi)=0$.

1.3.3 Newton Raahson Method

This method is generally used to improve the result obtained by one of the previous methods. Firstly we derive this method by using linear approximation. Suppose x_k is a point in the neighbourhood of the root of $f(x)=0$. If we approximate $f(x)$ by a polynomial of degree one we get $f(x)=a_0x+a_1$.

$$\therefore f(x_k) = a_0x_k + a_1$$

and $$f'(x_k) = a_0 \qquad \text{... (1.3.3.1)}$$

where a prime denotes differentiation with respect to x. On solving for as and a_1 we get

$$a_0 = f'(x_k) \text{ and } a_1 = f(x_k) - f'(x_k)x_k$$

From equation (1.3.1) we get,

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

$$= x_k - \frac{f(x_k)}{f'(x_k)}$$

Thus we get the next approximate root as

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, 3, \ldots \qquad \ldots (13.3.2)$$

This method is called the Newton Raphson method. The method (1.3.3.2) may also be obtained directly from Secant method (1.3.2.2) by taking the limit $x_{k-1} \to x_k$. In the limiting process i.e. when $x_{k-1} \to x_k$ the chord passing through the points $(x_k, f(x_k))$ and $(x_{k-1}, f(x_{k-1}))$ converges to the tangent at point $(x_k, f(x_k))$. Thus in this case the problem of finding a root of equation $f(x) = 0$ is equivalent to finding the point of inteersection of the tangent to the curve $y = f(x)$ at the point $(x_k, f(x_k))$ with the x-axis. The Newton Raphson method requires two values $f(x_k)$ and $f'(x_k)$. The method is applicable only when $f'(x_k) \neq 0$ i.e. root is a simple root.

The method can also be derived by using Taylor series representation. Let x_0 be an approximate root of $f(x) = 0$ and let $x_1 = x_0 + h$ be the correct root so that $f(x_1) = 0$. Expanding $f(x_1) = f(x_0 + h)$ by Taylor series about x_0, we obtain

$$f(x_1) = f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{x} f''(x_0) + \ldots = 0$$

Neglecting the second and higher order derivatives we have

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

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

Therefore $$x_1 = x_0 + h = x_0 - \frac{f(x_0)}{f'(x_0)}$$

is the better approximation than x_0.

Successive approximations are given by x_2, x_3, ... where

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \; n = 0, 1, 2, 3, \ldots$$

This is same as the formula (1.3.3.2).

1.4 Rate of Convergence

All the methods described in section 1.3 are iterative methods and repeatative application of these methods generate a sequence of approximate solutions. Convergence of this sequence is an important subject that we will discuss now.

1.4.1 Orders of Convergence

Some special terminology is used todescribe the rapidity with which a sequence convergences. Let $\{x_n\}$ be a sequence of real numbers tending to a limit x^*. We say tha the rate of convergence is at least linear if there is a constant $c < 1$ and an integer N such that

$$|x_{n+1} - x^*| \le C|x_n - x^*| \qquad (n \ge N) \qquad \ldots (1.4.1.1)$$

The convergence is atleast quadratri C if there are a constant C (not necessary less tnhan 1) and an integer N such that

$$|x_{n+1} - x^*| \le C|x_n - x^*|^2 \qquad (n \ge N) \qquad \ldots (1.4.1.2)$$

In general if there are positive constants C, largest α and an integer N such that

$$|x_{n+1} - x^*| \le C|x_n - x^*|^{\alpha} \qquad (n \ge N) \qquad \ldots (1.4.1.3)$$

We say that the rate of convergence is atleast α. The constant C is called the asymptotic error constant.

If x_k is an approximate root of $f(x) = 0$ and ξ is a solution of equation $f(x) = 0$ then $\varepsilon_k = x_k - \xi$ is the error in the solution. If the sequence $\varepsilon_k \to 0$ as $k \to \infty$, we say that iterative methods discussed in section 1.3 are convergent. We assume that ξ is a simple root of $f(x) = 0$ so that $f'(\xi) \ne 0$.

1.4.2 Rate of convergence of Secant Method

Suppose ξ is a simple root of $f(x)=0$ i.e. $f(\xi)=0$.

$\varepsilon_k = x_k - \xi$ is an error. On substituting $x_k = \xi + \varepsilon_k$ in (13.2.2) we get,

$$\varepsilon_{k+1} = \varepsilon_k - \frac{(\varepsilon_k - \varepsilon_{k-1}) f(\xi+\varepsilon_k)}{f(\xi+\varepsilon_k) - f(\xi+\varepsilon_{k-1})}$$

Expanding $f(\xi+\varepsilon_k)$ and $f(\xi+\varepsilon_{k-1})$ in Taylor series about the root ξ and observing that $f(\xi)=0$ we get,

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

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

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

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

$$\left(\because \frac{1}{1-x} = 1 + x + x^2 + x^3 + \ldots\right)$$

$$= \varepsilon_k - \left\{\varepsilon_k - \frac{1}{2}\varepsilon_k(\varepsilon_k + \varepsilon_{k-1})\frac{f''(\xi)}{f'(\xi)} + \frac{\varepsilon_k^{\,2}}{2!}\frac{f''(\xi)}{f'(\xi)} + \ldots\right.$$

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

$$= \frac{1}{2}\varepsilon_k\varepsilon_{k-1}\frac{f''(\xi)}{f'(\xi)} + 0\left(\varepsilon_k^{\,2}\varepsilon_{k-1} + \varepsilon_k^{\,3}\right)$$

Thus we write $\varepsilon_{k+1} = C \cdot \varepsilon_k \varepsilon_{k-1}$ where $C = \frac{1}{2}\frac{f''(\xi)}{f'(\xi)}$ and we ignore higher powers of ε_k.

The equation $\varepsilon_{k+1} = C \cdot \varepsilon_k \varepsilon_{k-1}$ ... (1.4.2.1)

is called error equation.

To determine the order of converence discussed in section 1.4.1, we have to determine the number α such that $\varepsilon_{k+1} = A \cdot \varepsilon_k{}^{\alpha}$ where A and α are to be determined. If we replace k by $k-1$ we get,

$$\varepsilon_k = A\varepsilon_{k-1}{}^{\alpha}$$

i.e. $\varepsilon_{k-1} = A^{-1/\alpha}\varepsilon_k{}^{1/\alpha}$

Substituting the values of ε_{k-1} and ε_{k+1} in equation (1.4.2.1) we get,

$$A\varepsilon_k{}^{\alpha} = C \cdot \varepsilon_k A^{-1/\alpha}\varepsilon_k{}^{1/\alpha}$$

$$\varepsilon_k{}^{\alpha} = CA^{-1-1/\alpha}\varepsilon_k{}^{1+1/\alpha}$$

Comparing the powers of ε_k on both sides of above equation we get,

$$\alpha = 1 + \frac{1}{\alpha}$$

which is quadratic equation in α and we get $\alpha = \frac{1}{2}(1 \pm \sqrt{5})$. The highest value of $\alpha = \frac{1}{2}(1+\sqrt{5})$ and we find that the rate of convergence of secant method is $\alpha = 1.618$ and $A = C^{\alpha/1+\alpha}$.

1.4.3 Rate of Convergence of Regula Falsi Method

If the function $f(x)$ in the equation $f(x) = 0$ is convex in the interval (x_0, x_1) that contains the root, then one of the points x_0 or x_1 is always fixed and the other point varies with k. If the point x_0 is fixed, then the function $f(x)$ is approximated by the straight line passing through the points $(x_0, f(x_0))$ and $(x_k, f(x_k))$, $k = 1, 2, 3, \ldots$.

Suppose ξ is simple root of $f(x)=0$ i.e. $f(\xi)=0$ and $\varepsilon_k = x_k - \xi$ is an error in approximate solution x_k. Since the point $(x_0, f(x_0))$ is fixed we can write

$$x_{k+1} = x_0 - \frac{f(x_0)(x_k - x_0)}{f(x_k) - f(x_0)}$$

$$\therefore \xi + \varepsilon_{k+1} = \xi + \varepsilon_0 - \frac{f(\xi+\varepsilon_0)\left[\xi + \varepsilon_k - (\xi+\varepsilon_0)\right]}{f(\xi+\varepsilon_k) - f(\xi+\varepsilon_0)}$$

$$\therefore \varepsilon_{k+1} = \varepsilon_0 - \frac{\left[\varepsilon_0 f'(\xi) + \frac{\varepsilon_0^{\,2}}{2!} f''(\xi) + \ldots\right](\varepsilon_k - \varepsilon_0)}{\left[\varepsilon_k f'(\xi) + \frac{\varepsilon_k^{\,2}}{2!} f''(\xi) + \ldots\right] - \left[\varepsilon_0 f'(\xi) + \frac{\varepsilon_0^{\,2}}{2!} f''(\xi) + \ldots\right]}$$

$$= \varepsilon_0 - \frac{(\varepsilon_k - \varepsilon_0) f'(\xi)\left[\varepsilon_0 + \frac{\varepsilon_0^{\,2}}{2!}\frac{f''(\xi)}{f'(\xi)} + \ldots\right]}{(\varepsilon_k - \varepsilon_0) f'(\xi)\left[1 + \frac{\varepsilon_k + \varepsilon_0}{2!}\frac{f''(\xi)}{f'(\xi)} + \ldots\right]}$$

$$= \varepsilon_0 - \left[\varepsilon_0 + \frac{\varepsilon_0^{\,2}}{2!}\frac{f''(\xi)}{f'(\xi)} + \ldots\right]\left[1 - \frac{\varepsilon_k + \varepsilon_0}{2!}\frac{f''(\xi)}{f'(\xi)} + \ldots\right]$$

$$= \frac{\varepsilon_0(\varepsilon_k + \varepsilon_0)}{2!}\frac{f''(\xi)}{f'(\xi)} - \frac{\varepsilon_0^{\,2}}{2!}\frac{f''(\xi)}{f'(\xi)} + \ldots$$

$$= \frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\varepsilon_0\varepsilon_k + 0\left(\varepsilon_0\varepsilon_k^{\,2}, \varepsilon_k^{\,3}\right)$$

Since $\varepsilon_0 = x_0 - \xi$ is the error in the first approximation and is independent of *k*, we can write

$$C = \frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\varepsilon_0$$

and we get error equation

$$\varepsilon_{k+1} = C\varepsilon_k$$

Here C is asymptotic error constant and by equation (1.4.1.2), we observe that the Regular Falsi method has at least linear rate of convergence.

1.4.4 Rate of Convergence of Newton Raphson Method

Suppose ξ is a simple root of $f(x)=0$ i.e. $f(\xi)=0$ but $f'(\xi)\neq 0$. Suppose $\varepsilon_k = x_k - \xi$ is an error in the approximate solution x_k. On substituting $x_k = \xi + \varepsilon_k$ in equation (1.3.3.2) and expanding $f(\xi+\varepsilon_k)$ and $f'(\xi+\varepsilon_k)$ in Taylor series about the point ξ and using the fact that $f(\xi)=0$ and $f'(\xi)\neq 0$, we obtain

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

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

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

$$= \varepsilon_k - \varepsilon_k + \frac{\varepsilon_k^{\,2}}{2!}\frac{f''(\xi)}{f'(\xi)} + \ldots. + \varepsilon_k^{\,2}\frac{f''(\xi)}{f'(\xi)} + \frac{\varepsilon_k^{\,3}}{2!}\left(\frac{f''(\xi)}{f'(\xi)}\right)^2$$

$$= \frac{\varepsilon_k^{\,2}}{2!}\frac{f''(\xi)}{f'(\xi)} + 0(\varepsilon_k^{\,3})$$

On neglecting terms containing $\varepsilon_k^{\,3}$ and higher powers of ε_k we get an error equation

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

Thus by equation (1.4.1.2) we obsrve that Newton Raphson method has second order convergence.

Note : If the root ξ of $f(x)=0$ is a root of multiplicity two or more then the rate of convergence for Newton Raphson method is one.

1.5 Interation Methods

To describe this method for finding the roots of equation

$$f(x) = 0 \qquad \text{... (1.5.1)}$$

We rewrite this equation in the form

$$x = \phi(x) \qquad \text{... (1.5.2)}$$

Let x_0 be an approximate value of the desired root ξ. Substituting it for x on the right hand side of equation (1.5.2) we obtain the first approximation

$$x_1 = \phi(x_0)$$

The successive approximations are then given by

$$x_2 = \phi(x_1)$$

$$x_3 = \phi(x_2)$$

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

Thus we get a sequence of approximate solutions $\{x_k\}$. The convergence of this sequence depends on the suitable choice of function $\phi(x)$ and initial approximation x_0. The function $\phi(x)$ is called an iteration function. If the function ϕ is continuous and the sequence $\{x_n\}$ converges to x^* then

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

$$\Rightarrow x^* = \lim_{n\to\infty} x_{n+1} = \underset{n\to\infty}{Lt}\ \phi(x_n) = \phi\left(\underset{n\to\infty}{Lt}\ x_n\right) = \phi(x^*)$$

Thus x^* is a root of equation (1.5.2) if the iteration function ϕ is continuous function.

The following theorem gives a necessary and sufficient condition for the convergence of sequence $\{x_n\}$.

Theorem : If $\phi(x)$ is a continuous function in some interval [a, b] that contains the root and $|\phi'(x)| \le C < 1$ in this interval, then for every choice of $x_0 \in [a, b]$, the sequence $\{x_k\}$ determined from

$$x_{k+1} = \phi(x_k) \quad , k = 0, 1, 2, 3, ...$$

Converges to the root ξ of $x = \phi(x)$.

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

$$\xi = \phi(\xi) \qquad \text{... (1.5.3)}$$

$$x_{k+1} = \phi(x_k) \qquad \text{... (1.5.4)}$$

From equation (1.5.3) and (1.5.4) we get,

$$\xi - x_{k+1} = \phi(\xi) - \phi(x_k), \quad k = 0, 1, 2, 3, \ldots$$

Using the mean value theorem, we get

$$\xi - x_{k+1} = \phi'(\theta_k)(\xi - x_k) \qquad \theta_k \in L(\xi, x_k)$$

where $L(\xi, x_k)$ represents a line segment joining ξ and x_k.

Similarly, we obtain

$$\xi - x_k = \phi'(\theta_{k-1})(\xi - x_{k-1}) \qquad \theta_{k-1} \in L(\xi, x_{k-1})$$

$$\xi - x_{k-1} = \phi'(\theta_{k-2})(\xi - x_{k-2}) \qquad \theta_{k-2} \in L(\xi, x_{k-2})$$

$$\xi - x_1 = \phi'(\theta_0)(\xi - x_0) \qquad \theta_0 \in L(\xi, x_0)$$

Since each $\theta_i \in [a, b]$, $|\phi'(\theta_i)| \le C < 1$ and $|\xi - x_{k+1}| \le C^{k+1}(\xi - x_0)$

Thus, $|\varepsilon_{k+1}| \le C^{k+1} |\varepsilon_0|$

Since C < 1, the right hand side of above inequality goes to zero as *k* becomes large and it follows that the sequence of approximations $\{x_k\}$. Converges to the root ξ if C < 1.

Note : The root obtained by this method is unique. Suppose ξ_1 and ξ_2 are two distinct roots of equation (1.5.2). i.e. $\xi_1 = \phi(\xi_1)$ and $\xi_2 = \phi(\xi_2)$.

Then we get

$$\xi_1 - \xi_2 = \phi(\xi_1) - \phi(\xi_2)$$

$$= \phi'(\theta)(\xi_1 - \xi_2)$$

$$\therefore (\xi_1 - \xi_2)[1 - \phi'(\theta)] = 0$$

But, $|\phi'(\theta)| \le C < 1 \qquad \therefore 1 - \phi'(\theta) \ne 0 \qquad \therefore \xi_1 - \xi_2 = 0$

and therefore $\xi_1 = \xi_2$, hence the root is unique.

In general the speed of convergence depends on the value of C; the smaller the value of C, the faster would be the convergence. Therefore, the speed of convergence dependent upon the choice of $\phi(x)$. There are many ways of rewriting $f(x) = 0$ in the form $x = \phi(x)$. For example the equation $f(x) = x^3 + x^2 - 1 = 0$ can be expressed as

$$x = (1+x)^{-1/2} = \phi_1(x) \qquad \text{(say)}$$

$$x = (1-x^3)^{1/2} = \phi_2(x) \qquad \text{(say)}$$

$$x = (1-x^2)^{1/3} = \phi_3(x) \qquad \text{(say)}$$

$$x = 1 + x - x^2 - x^3 = \phi_4(x) \qquad \text{(say)}$$

We have to choose that function $\phi_i(x)$ for which $|\phi_i'(x)| < 1$. Since $f(0) = -1$ and $f(1) = 1$, we know that root lies between 0 and 1.

$$|\phi_4'(x)| = |1 - 2x - 3x^2| \not< 1 \qquad \text{for } x \in [0,1]$$

$$|\phi_2'(x)| = \left|\frac{1}{2}(1-x^3)^{-1/2}(-3x^2)\right|$$

$$= \left|\frac{3}{2}x^2(1-x^3)^{-1/2}\right| \to \infty \qquad \text{as } x \to 1$$

Observe that the functions ϕ_2, ϕ_3, ϕ_4 are **not** the expected choices of iteration function, as $|\phi_i'(x)| \not< 1$ for $i = 2, 3, 4$ in the interval [0, 1].

If we choose $\phi(x) = (1+x)^{-1/2}$ then $\phi'(x) = -\frac{1}{2}(1+x)^{-3/2}$

and $\qquad |\phi'(x)| = \frac{1}{2}(1+x)^{-3/2} < 1 \qquad \forall x \in [0,1]$

$$\max_{0 \le x \le 1} |\phi'(x)| = \frac{1}{2\sqrt{8}} = 0.17678 < 1$$

and the iteration method $x_{k+1} = \phi(x_k)$ converges to the root as $|\phi'(x)| < 1$.

1.6 Polynomial Equations

Polynomial functions are of special importance. They are everywhere continuous, they are smooth, their derivatives are also continuous and smooth and they are readily evaluated. Descarte's rule of signs predict the number of positive roots. Polynomials are particularly well adapted to computers because the only mathematical operations they require for evaluation are addition, subtraction and multiplication.

To determine the roots of polynomial equation it is necessary to have the following information

(i) The exact number of real and complex roots along with their multiplicity.

(ii) The interval in which each real root lies.

By fundamental theorem of algebra we know that a polynomial of degree *n* has exactly *n* roots. Decarte's rule of signs gives only upper limit of no. of + ve and – ve real roots. This rule does not give the exact number of positive and negative real roots. The exact number of real roots of a polynomial can be found by Sturms theorem.

Let $f(x)$ be a polynomial of degree *n*. Let $f_1(x)$ represent its first order derivative. The remainder of $f(x)$ divided by $f_1(x)$ taken with the reverse sign is denoted by $f_2(x)$. Let $f_3(x)$ denotes the remainder of $f_1(x)$ divided by $f_2(x)$ with the reverse sign. Continue this process till we arrive at a constant. We thus obtain a sequence of functions

$$f(x), f_1(x), f_2(x), ..., f_k(x)$$

This sequence is called Sturm sequence.

Sturm Theorem : The number of real roots of the equation $f(x) = 0$ on [a, b] equals the difference between the number of changes of sign in the Sturm sequence at $x = a$ and $x = b$, provided that $f(a) \neq 0$, $f(b) \neq 0$.

Since a polynomial of degree *n* has exactly *n* roots, the number of complex roots equals (n - number of real roots), where a real root of multiplicity *r* is counted *r* times.

If $x_1, x_2, x_3, ..., x_n$ are real roots of $f(x)$ then

$$f(x) = a_0 (x - x_1)(x - x_2)(x - x_3)...(x - x_n)$$

Complex roots occur in pair. If x_1, x_2 are complex roots then $(x - x_1)(x - x_2)$ is a polynomial of degree two with real co-efficients and in this case

$$f(x) = a_0 (x^2 + px + q)(x - x_3)...(x - x_n)$$

Thus it is obvious that the methods of finding roots of polynomial equation should include the determination of either linear factor $(x-p)$ or quadratic factor x^2+px+q.

In this section two methods are presented. Birge Vieta method is used to determine the linear factor $(x-p)$ whereas Bairstow method is used to determine the quadratic factor x^2+px+q.

1.6.1 Birge-Vieta Method

This method is used to determine real root of polynomial equation

$$P_n(x)=a_0x^n+a_1x^{n-1}+a_2x^{n-1}+.....+a_{n-1}x+a_n=0 \qquad \text{... (1.6.1.1)}$$

If P is a root of polynomial $P_n(x)$ then $(x-p)$ is factor of polynomial $P_n(x)$. Suppose p is an approximate root of $P_n(x)$. If we divide $P_n(x)$ by a factor $(x-p)$ then we get a quotient Q_{n-1} a polynomial of degree (n – 1) and a remainder. The remainder R depends on choice of p. i.e. if we change the value of p, R will get change. Birge Vieta method gives a procedure to make R zero.

Suppose p is an approximate root of $P_n(x)$.

$$P_n(x)=(x-p)Q_{n-1}(x)+R \qquad \text{... (1.6.1.2)}$$

where $Q_{n-1}(x)=b_0x^{n-1}+b_1x^{n-2}+b_3x^{n-3}+....+b_{n-2}x+b_{n-1}$ .... (1.6.1.3)

a polynomial of degree (n – 1) and R is remainder.

The coefficients in polynomial Q_{n-1} i.e. b_i and remainder R are functions of p. Birge Vieta method gives a procedure to determine p such that $R(p)=0$.

$$P_n(p)=(p-p)Q_{n-1}(p)+R(p)=R(p) \qquad \text{... (1.6.1.4)}$$

$$R(p)=0$$

Equation (1.6.1.4) is the equation in variable p and any iterative method discussed in section 1.3 can be used to determine the root p. Application of Newton Raphson method discussed in section 1.3.3 for equation (1.6.1.4) gives

$$p_{k+1}=p_k-\frac{P_n(p_k)}{P_n'(p_k)} \qquad \text{... (1.6.1.5)}$$

For a polynomial equation, the computation of P_n and P_n' can be obtained with the help of synthetic division. On comparing the coefficients of like powers of x on both sides of equation (1.6.1.2) and using equations (1.6.1.1) and (1.6.1.3) we get,

$$a_0x^n+a_1x^{n-1}+a_2x^{n-2}+.....+a_{n-1}x+a_n=(x-p)\left(b_0x^{n-1}+b_1x^{n-2}+.....+b_{n-1}\right)+R$$

Thus

$$a_0 = b_0 \Rightarrow b_0 = a_0$$

$$a_1 = b_1 - pp_0 \Rightarrow b_1 = a_1 + pb_0$$

$$a_2 = b_2 - pp_1 \Rightarrow b_2 = a_2 + pb_1$$

$$\vdots$$

$$a_k = b_k - pb_{k-1} \Rightarrow b_k = a_k + pb_{k-1}$$

$$\vdots$$

$$a_n = R - pb_{n-1} \Rightarrow R = a_n + pb_{n-1}$$

In general

$$b_k = a_k + pb_{k-1},\ k = 1, 2, 3, \ldots, n \qquad \ldots (1.6.1.6)$$

with $b_0 = a_0$ and $b_n = R$.

From equation (1.6.1.4) we have

$$P_n(p) = R = b_n \qquad \ldots (1.6.1.7)$$

To determine $P_n'(p)$, differentiate (1.6.1.6) with respected p.

$$\frac{db_k}{dp} = b_{k-1} + p\frac{db_{k-1}}{dp} \qquad \ldots (1.6.1.8)$$

Equation (1.6.1.8) can be written as

$$C_{k-1} = b_{k-1} + pC_{k-2}$$

where $\frac{db_k}{dp} = C_{k-1},\ k = 1, 2, 3, \ldots, n$

Above equation can also be represented by

$$C_k = b_k + pC_{k-1},\ k = 1, 2, , n - 1 \qquad \ldots (1.6.1.9)$$

and $$C_0 = \frac{db_1}{dp} = \frac{d}{dp}\left(a_1 + pb_0\right) = b_0$$

(Since $b_0 = a_0$ and a_0 is independent of p, differentiation of b_0 with respect to p is zero).

On differentiating (1.6.1.7) with respect to p we get

$$\frac{dP_n(p)}{dp} = \frac{dR}{dp} = \frac{db_n}{dp} = C_{n-1} \qquad \text{... (1.6.1.10)}$$

By substituting the values of $P_n(p)$ and $P_n'(p)$ from equation (1.6.1.7) and (1.6.1.10) in equation (1.6.1.5) we get

$$p_{k+1} = p_k - \frac{b_n}{C_{n-1}},\ k = 0,\ 1,\ 2,\ .... \qquad \text{.... (1.6.1.11)}$$

where p_0 is initial approximation of factor $(x - p)$.

The method (1.6.1.11) is called the **Birge Vieta Method**.

The calculations of the coefficients b_k and C_k are carried out by using synthetic division.

p	a_0	a_1	a_2	a_3	...	a_{n-2}	a_{n-1}	a_n
		pb_0	pb_1	pb_2	...	pb_{n-3}	pb_{n-2}	pb_{n-1}
p	a_0	b_1	b_2	b_3	...	b_{n-2}	b_{n-1}	$b_n = R = P_n(p)$
		pC_0	pC_1	pC_2	...	pC_{n-3}	pC_{n-2}	
	C_0	C_1	C_2	C_3	...	C_{n-2}	$C_{n-1} = \frac{dR}{dp} = \frac{dP_n(p)}{dp}$	

Note : If the polynomial $P_n(x)$ do not contain the term x^k, write $a_k = 0$.

Once p is calculate with desired accuracy then repeat the procedure for $Q_{n-1}(x)$ to determine second factor of $P_n(x)$. The continuous application of Birge-Vieta method produces all real roots.

1.6.2 Bairstow Method

This method is used to extract quadratic factor from polynomial $P_n(x)$, which may give a pair of complex roots or pair of real roots. If we divide the polynomial $P_n(x)$ defined by equation (1.6.1.1) by the quadratic factor $x^2 + px + q$ then the quotient is a polynomial of degree $(n-2)$ and a remainder is a polynomial of degree one .

Thus $P_n(x) = (x^2 + px + q)Q_{n-1}(x) + Rx + S$ ... (1.6.2.1)

where $Q_{n-2}(x) = b_0x^{n-2} + b_1x^{n-3} + b_2x^{n-4} + ... + b_{n-3}x + b_{n-2}$ .... (1.6.2.2)

The coefficient b_0, b_1, b_2, ... b_{n-3}, b_{n-2}, R and S are functions of p and q. $x^2 + px + q$ is a factor of $P_n(x)$, if

$$R(p,q) = S(p,q) = 0 \qquad \text{.... (1.6.2.3)}$$

Suppose (p_0, q_0) is an initial approximation for equation (1.6.2.3) and $(p_0 + \Delta p, q_0 + \Delta q)$ is the true solution of equation (1.6.2.3). Then

$$R(p_0 + \Delta p, q_0 + \Delta q) = R(p_0, q_0) + \frac{\partial R}{\partial p}\Delta p + \frac{\partial R}{\partial q}\Delta q = 0$$

$$S(p_0 + \Delta p, q_0 + \Delta q) = S(p_0, q_0) + \frac{\partial S}{\partial p}\Delta p + \frac{\partial S}{\partial q}\Delta q = 0$$

On solving above equations simultaneously for Δp and Δq we get

$$\Delta p = -\frac{RS_q - SR_q}{R_p S_q - R_q S_p}, \quad \Delta q = -\frac{R_p S - RS_p}{R_p S_q - R_q S_p} \qquad \text{... (1.6.2.4)}$$

where R_p, R_q, S_p, S_q are partial derivatives of R and S with respect to p and q respectively evaluated at (p_0, q_0). Functions R and S are also evaluated at (p_0, q_0).

Thus to determine the true solution it is necessary to calculate Δp and Δq. The increaments Δp and Δq are known in terms of R, S, R_p, R_q, S_p, S_q evaluated at initial approximation (p_0, q_0). These functions R, S and their partial derivatives are obtained by comparing the coefficients of equal powers of x in equation (1.6.2.1). From equation (1.6.1.1) and (1.6.2.1) we get

$$x^n : a_0 = b_0 \Rightarrow b_0 = a_0$$

$$x^{n-1} : a_1 = b_1 + pb_0 \Rightarrow b_1 = a_1 - pb_0$$

$$x^{n-2} : a_2 = b_2 + pb_1 + qb_0 \Rightarrow b_2 = a_2 - pb_1 - b_0$$

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

$$x^{n-k} : a_k = b_k + pb_{k-1} + qb_{k-2} \Rightarrow b_k = a_k - pb_{k-1} - qb_{k-2}$$

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

$$x : a_{n-1} = R + pb_{n-2} + qb_{n-3} \Rightarrow R = a_{n-1} - pb_{n-2} - qb_{n-3}$$

$$x^0 : a_n = S + qb_{n-2} \Rightarrow S = a_n - qb_{n-2}$$

In general we write

$$b_k = a_k - pb_{k-1} - qb_{k-2},\ k = 1, 2, 3, ..., n \qquad .... (1.6.2.5)$$

where $b_0 = a_0$ and $b_{-1} = 0$.

From last two equations we get

$$R = b_{n-1} \text{ and } S = b_n + pb_{n-1} \qquad ... (1.6.2.6)$$

The partial derivatives R_p, R_q, S_p, S_q can be determined by differentiating (1.6.2.5) with respect to p and q. From equation (1.6.2.5) we get,

$$-\frac{\partial b_k}{\partial p} = b_{k-1} + p\frac{\partial b_{k-1}}{\partial p} + q\frac{\partial b_{k-2}}{\partial p};\ \ \frac{\partial b_0}{\partial p} = \frac{\partial b_{-1}}{\partial p} = 0 \qquad .... (1.6.2.7)$$

$$-\frac{\partial b_k}{\partial q} = p\frac{\partial b_{k-1}}{\partial q} + b_{k-2} + q\frac{\partial b_{k-2}}{\partial q};\ \ \frac{\partial b_0}{\partial q} = \frac{\partial b_{-1}}{\partial q} = 0 \qquad .... (1.6.2.8)$$

Substitution $\frac{\partial b_k}{\partial p} = -C_{k-1}, k = 1, 2, 3, ..., n$ converts equation (1.6.2.7) into

$$C_{k-1} = b_{k-1} - pC_{k-2} - qC_{k-3} \qquad .... (1.6.2.9)$$

If we write $C_{k-2} = -\frac{\partial b_k}{\partial q}$ then equation (1.6.2.8) becomes

$$C_{k-2} = b_{k-2} - pC_{k-3} - qC_{k-4} \qquad .... (1.6.2.10)$$

From equation (1.6.2.9) and (1.6.2.10) we get a recurrence relation

$$C_k = b_k - pC_{k-1} - qC_{k-2},\ k = 1, 2, 3, ..., n - 1$$

where $C_{-1} = 0$ and $C_0 = -\frac{\partial b_1}{\partial p} = -\frac{\partial}{\partial p}(a_1 - pb_0) = b_0$

(Since a_1 and $b_0 = a_0$ are independent of p and q)

From equation (1.6.2.6), (1.6.2.7) we get

$$R_p = \frac{\partial R}{\partial p} = \frac{\partial b_{n-1}}{\partial p} = -C_{n-2}$$

$$S_p = \frac{\partial b_n}{\partial p} + b_{n-1} + p\frac{\partial b_{n-1}}{\partial p} = b_{n-1} - C_{n-1} - pC_{n-2}$$

From equation (1.6.2.6) and (16.2.8) we get

$$R_q = \frac{\partial R}{\partial q} = \frac{\partial b_{n-1}}{\partial q} = -C_{n-3}$$

$$S_q = \frac{\partial S}{\partial q} = \frac{\partial b_n}{\partial q} + p\frac{\partial b_{n-1}}{\partial q} = -C_{n-2} - pC_{n-3}$$

On substituting the above values of $R, S,\ R_p\ ,\ R_q\ ,\ S_p\ ,\ S_q$ in equation (1.6.2.4) and using equation (1.6.2.6) we get,

$$\Delta p = -\frac{b_{n-1}\left(-C_{n-2} - pC_{n-3}\right) - \left(b_n + pb_{n-1}\right)\left(-C_{n-3}\right)}{\left(-C_{n-2}\right)\left(-C_{n-2} - pC_{n-3}\right) - \left(-C_{n-3}\right)\left(b_{n-1} - C_{n-1} - pC_{n-2}\right)}$$

$$= -\frac{b_n C_{n-3} - b_{n-1}C_{n-2}}{C_{n-2}^2 - C_{n-3}\left(C_{n-1} - b_{n-1}\right)}$$

and $$\Delta q = -\frac{\left(-C_{n-2}\right)\left(b_n + pb_{n-1}\right) - \left(b_{n-1}\right)\left(b_{n-1} - C_{n-1} - pC_{n-2}\right)}{\left(-C_{n-2}\right)\left(-C_{n-2} - pC_{n-3}\right) - \left(-C_{n-3}\right)\left(b_{n-1} - C_{n-1} - pC_{n-2}\right)}$$

$$= -\frac{b_{n-1}\left(C_{n-1} - b_{n-1}\right) - b_n C_{n-2}}{C_{n-2}^2 - C_{n-3}\left(C_{n-1} - b_{n-1}\right)}$$

The improved values of p and q are now

$$p_1 = p_0 + \Delta p \ \text{ and } \ q_1 = q_0 + \Delta q$$

Repeat the procedure by replacing the initial approximation $\left(p_0, q_0\right)$ by $\left(p_1, q_1\right)$, till we get the required accuracy of p and q.

For computing b_k's and C_k's we use the following representation

$-p$	a_0	a_1	a_2	...	a_{n-2}	a_{n-1}	a_n
$-q$		$-pb_0$	$-pb_1$	...	$-pb_{n-3}$	$-pb_{n-2}$	$-pb_{n-1}$
			$-qb_0$	...	$-qb_{n-4}$	$-qb_{n-3}$	$-qb_{n-2}$
$-p$	b_0	b_1	b_2	...	b_{n-2}	b_{n-1}	b_n
$-q$		$-pC_0$	$-pC_1$	...	$-pC_{n-3}$	$-pC_{n-2}$	
			$-qC_0$	...	$-qC_{n-4}$	$-qC_{n-3}$	
	C_0	C_1	C_2	...	C_{n-2}	C_{n-1}	

When p and q are obtained to the desired accuracy the polynomial $Q_{n-2}(x) = \dfrac{P_n(x)}{x^2 + px + q}$ is obtained from the synthetic division procedure. The next quadratic factor of Q_{n-2} (and hence of $P_n(x)$) is obtained by applying Bairstow method to $Q_{n-2}(x)$.

ILLUSTRATIVE EXAMPLES

1. Obtain a root, correct to three decimal places for each of the following equations using the bisection method

(a) $x^3 - x - 1 = 0$ (b) $x^3 + x^2 + x + 7 = 0$

(c) $x^3 - 2x - 5 = 0$ (d) $xe^x - 1 = 0$

Answer (a) : $f(x) = x^3 - x - 1 = 0$

Since $f(1) = -1$ and $f(2) = 8 - 2 - 1 = 5$, $f(1) \cdot f(2) < 0$. Therefore root lies between 1 and 2. Take $x_0 = \dfrac{1+2}{2} = \dfrac{3}{2} = 1.5$.

$$f\left(x_0\right) = \left(\frac{3}{2}\right)^3 - \left(\frac{3}{2}\right) - 1 = 0.875 > 0$$

$f(1) = -1$, $f\left(\dfrac{3}{2}\right) = 0.875$. $f(1) f\left(\dfrac{3}{2}\right) < 0$. We therefore conclude that root lies between 1 and $\dfrac{3}{2}$.

$$x_1 = \frac{x_0 + 1}{2} = \frac{\frac{3}{2} + 1}{2} = 1.25; \quad f(1.25) = (1.25)^3 - (1.25) - 1 = -0.2968 < 0$$

Root lies between 1.25 and 1.5 ($\because f\left(x_0\right) \cdot f\left(x_1\right) < 0$)

$$x_2 = \frac{x_0 + x_1}{2} = \frac{1.5 + 1.25}{2} = 1.375; \; f\left(x_2\right) = 0.224609375 > 0$$

Put $x_0 = x_2 = 1.375$

$$x_3 = \frac{x_0 + x_1}{2} = \frac{1.375 + 1.25}{2} = 1.3125; \; f\left(x_3\right) = -0.051513671 < 0$$

Put $x_1 = 1.3125$

$$x_4 = \frac{x_0 + x_1}{2} = \frac{1.375 + 1.3125}{2} = 1.34375,\ f(x_4) = 0.082611083 > 0$$

Put $x_0 = x_4 = 1.34375$

$$x_5 = \frac{x_0 + x_1}{2} = \frac{1.34375 + 1.3125}{2} = 1.328125;\ f(x_5) = 0.0145... > 0$$

$$\therefore x_0 = x_5 = 1.328125$$

$$x_6 = \frac{x_0 + x_1}{2} = \frac{1.328125 + 1.3125}{2} = 1.3203125,\ f(x_6) = -0.018710613 < 0$$

$$x_1 = x_6 = 1.3203125$$

$$x_7 = \frac{x_0 + x_1}{2} = \frac{1.328125 + 1.3203125}{2} = 1.32421875,\ f(x_7) = -0.002127... < 0$$

$$x_1 = x_7 = 1.32421875$$

$$x_8 = \frac{x_0 + x_1}{2} = \frac{1.328125 + 1.32421875}{2} = 1.326171875,\ f(x_8) = -0.00620... < 0$$

$$x_1 = x_8 = 1.326171875$$

$$x_9 = \frac{x_0 + x_1}{2} = \frac{1.328125 + 1326171875}{2} = 1.327158438,\ f(x_9) = 0.01038... < 0$$

$$\therefore x_0 = 1.327148438$$

$$x_{10} = \frac{x_0 + x_1}{2} = \frac{1.327148438 + 1.326171875}{2} = 1.326660157,\ f(x_{10}) = 0.00829... < 0$$

$$x_0 = 1.326660157$$

$$x_{11} = \frac{x_0 + x_1}{2} = \frac{1.326660157 + 1.326171875}{2} = 1.326416016,\ f(x_{11}) = 0.00725... > 0$$

$$x_0 = 1.326416016$$

$$x_{12} = \frac{x_0 + x_1}{2} = \frac{1.326416016 + 1.326171875}{2} = 1.326293946$$

$$|x_{11} - x_{12}| = |1.326416016 - 1.326293946| = 1.22 \times 10^{-4}$$

Therefore root is correct upto three decimal places.

Thus $x = 1.326293946$ is a root of $x^3 - x - 1 = 0$ which is correct upto three decimal places.

Answer (b) : $f(x) = x^3 + x^2 + x + 7 = 0$

$f(-2) = (-2)^3 + (-2)^2 + (-2) + 7 = -8 + 4 - 2 + 7 = 1 > 0$

$f(-3) = (-3)^3 + (-3)^2 + (-3) + 7 = -27 + 9 - 3 + 7 = -14 < 0$

Root lies between – 2 and – 3.

Let $x_0 = -2$, $x_1 = -3$, $f(x_0) = 1 > 0$, $f(x_1) = -14 < 0$.

$x_2 = \frac{x_0 + x_1}{2} = -2.5$, $f(x_2) = -4.875 < 0$

Put $x_1 = x_2 = -2.5$

$x_3 = \frac{x_0 + x_1}{2} = \frac{-2 - 2.5}{2} = -2.25$, $f(-2.25) = -1.578 < 0$

Put $x_1 = x_3 = -2.25$

$x_4 = \frac{x_0 + x_1}{2} = \frac{-2 - 2.25}{2} = -2.125$, $f(x_4) = -0.205078... < 0$

Put $x_1 = x_4 = -2.125$

$x_5 = \frac{x_0 + x_1}{2} = \frac{-2 - 2.125}{2} = -2.0625$, $f(x_5) = 0.417724609 > 0$

Put $x_0 = x_5 = -2.0625$

$x_6 = \frac{x_0 + x_1}{2} = \frac{-2.0625 - 2.125}{2} = -2.09375$, $f(x_6) = 0.11148... > 0$

Put $x_0 = x_6 = -2.09375$

$x_7 = \frac{x_0 + x_1}{2} = \frac{-2.09375 - 2.125}{2} = -2.109375$, $f(x_7) = -0.0454... < 0$

$x_1 = x_7 = -2.109375$

$x_8 = \frac{x_0 + x_1}{2} = \frac{-2.09375 - 2.109375}{2} = -2.1015625$, $f(x_8) = 0.0333... > 0$

Put $x_0 = x_8 = -2.1015625$

$x_9 = \frac{x_0 + x_1}{2} = \frac{-2.1015625 - 2.109375}{2} = -2.10546875$, $f(x_9) = -0.00601... < 0$

$x_1 = x_9 = -2.10546875$

$$x_{10} = \frac{x_0 + x_1}{2} = \frac{-2.1015625 - 2.10546875}{2} = -2.103516\,,\ f(x_{10}) = 0.0136... > 0$$

$$x_0 = x_{10} = -2.103516$$

$$x_{11} = \frac{x_0 + x_1}{2} = \frac{-2.103516 - 2.10546875}{2} = -2.104492\,,\quad f(x_4) = 0.0038... > 0$$

$$x_0 = x_{11} = -2.104492$$

$$x_{12} = \frac{x_0 + x_1}{2} = \frac{-2.104492 - 2.10546875}{2} = -2.104980375$$

$$|x_{11} - x_{12}| = |-2.104492 + 2.104980375| = 4.88 \times 10^{-4}$$

$\therefore$ Root correct upto three decimal paces is $x_{12} = -2.104980375$.

Answer (c) : $f(x) = x^3 - 2x - 5 = 0$

$$f(2) = 2^3 - 4 - 5 = -1 < 0\,,\ f(3) = 27 - 6 - 5 = 16 > 0$$

$$f(2)\, f(3) < 0$$

$\therefore$ Root lies between 2 and 3.

$$x_0 = 2\,,\ x_1 = 3\,,\ f(x_0) = -1 < 0\,,\ f(x_1) = 16 > 0$$

$$x_2 = \frac{x_0 + x_1}{2} = 2.5\,,\quad f(x_2) = 5.62 > 0$$

Put $x_1 = x_2 = 2.5$

$$x_3 = \frac{x_0 + x_1}{2} = \frac{2 + 2.5}{2} = 2.125\,,\ f(x_3) = 0.3457 > 0$$

Put $x_1 = x_3 = 2.125$

$$x_4 = \frac{x_0 + x_1}{2} = \frac{2 + 2.125}{2} = 2.0625\,,\ f(x_4) = -0.3513 < 0$$

Put $x_0 = 2.0625$

$$x_5 = \frac{x_0 + x_1}{2} = \frac{2.0625 + 2.125}{2} = 2.09375\,,\ f(x_5) = -0.0089 < 0$$

Put $x_0 = 2.09375$

$$x_6 = \frac{x_0 + x_1}{2} = \frac{2.09375 + 2.125}{2} = 2.109385\,,\ f(x_6) = 0.1668 > 0$$

Put $x_1 = x_6 = 2.10938$

$$x_7 = \frac{x_0 + x_1}{2} = \frac{2.09375 + 2.10938}{2} = 2.101565,\ f(x_7) = 0.07856 > 0$$

Put $x_1 = x_7 = 2.101565$

$$x_8 = \frac{x_0 + x_1}{2} = \frac{2.09375 + 2.101565}{2} = 2.09766,\ f(x_8) = 0.034... > 0$$

Put $x_1 = x_8 = 2.09766$

$$x_9 = \frac{x_0 + x_1}{2} = \frac{2.09375 + 2.09766}{2} = 2.09570,\ f(x_9) = 0.01286... > 0$$

Put $x_1 = x_9 = 2.09570$

$$x_{10} = \frac{x_0 + x_1}{2} = \frac{2.09375 + 2.09870}{2} = 2.09473,\ f(x_{10}) = 0.00195 > 0$$

Put $x_1 = x_{10} = 2.09473$

$$x_{11} = \frac{x_0 + x_1}{2} = \frac{2.09375 + 2.09473}{2} = 2.09424$$

$$|x_{10} - x_{11}| = |2.09473 - 2.09424| = 4.9 \times 10^{-4}$$

Root is correct upto three decimal places and $x_{11} = 2.09424$ is root of $f(x) = 0$.

Answer (d) : $f(0) = 0 - 1 = -1 < 0$, $f(1) = e - 1 = 1.718 > 0$

$x_0 = 0$, $x_1 = 1$, $f(x_0) = -1 < 0$, $f(x_1) = 1.718 > 0$

Root lies between 0 and 1.

$$x_2 = \frac{x_0 + x_1}{2} = 0.5,\ f(0.5) = -0.1756 < 0$$

Put $x_0 = 0.5$

$$x_3 = \frac{x_0 + x_1}{2} = \frac{0.5 + 1}{2} = 0.75,\ f(0.75) = 0.5877... > 0$$

Put $x_1 = 0.75$

$$x_4 = \frac{x_0 + x_1}{2} = \frac{0.5 + 0.75}{2} = 0.625,\ f(x_4) = 0.1676... > 0$$

Put $x_1 = x_4 = 0.625$

$$x_5 = \frac{x_0 + x_1}{2} = \frac{0.5 + 0.625}{2} = 0.5625,\ f(x_5) = -0.01278... < 0$$

Put $x_0 = x_5 = 0.5625$

$$x_6 = \frac{x_0 + x_1}{2} = \frac{0.5625 + 0.625}{2} = 0.59375,\ f(x_6) = 0.07514... > 0$$

Put $x_1 = 0.59375$

$$x_7 = \frac{x_0 + x_1}{2} = \frac{0.5625 + 0.59375}{2} = 0.578125,\ f(x_7) = 0.0306... > 0$$

Put $x_1 = 0.578125$

$$x_8 = \frac{x_0 + x_1}{2} = \frac{0.5625 + 0.578125}{2} = 0.5703125,\ f(x_8) = 0.008779... > 0$$

Put $x_1 = 0.5703125$

$$x_9 = \frac{x_0 + x_1}{2} = \frac{0.5625 + 0.5703125}{2} = 0.56640625,\ f(x_9) = -0.002... < 0$$

Put $x_0 = 0.56640625$

$$x_{10} = \frac{x_0 + x_1}{2} = \frac{0.56640625 + 0.5703125}{2} = 0.56836,\ f(x_{10}) = 0.0033... > 0$$

Put $x_1 = 0.56836$

$$x_{11} = \frac{x_0 + x_1}{2} = \frac{0.56640625 + 0.56836}{2} = 0.56738,\ f(x_{11}) = 0.0006... > 0$$

Put $x_1 = 0.56738$

$$x_{12} = \frac{x_0 + x_1}{2} = \frac{0.56640625 + 0.56738}{2} = 0.566893,\ f(x_{12}) = -0.0006... < 0$$

Put $x_0 = 0.566893$

$$x_{13} = \frac{x_0 + x_1}{2} = \frac{0.566893 + 0.56738}{2} = 0.5671365$$

$$|x_{12} - x_{13}| = |0.566893 - 0.5671365| = 2.4 \times 10^{-4}$$

The required root is 0.5671365 which is correct upto 3 decimal places.

2. A real root of following functions lies in the interval (0, 1). Perform four iterations of secant method to obtain the root.

(a) $x^3 - 5x + 1 = 0$ (b) $\cos x - xe^x = 0$ (c) $xe^x - 1 = 0$

Answer (a) : $f(x) = x^3 - 5x + 1 = 0$

Here $x_0 = 0$, $x_1 = 1$. $f(x_0) = 1$, $f(x_1) = -3$.

By Secant method (Chord method)

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \cdot f(x_k)$$

For $k = 1$, we have

$$x_2 = 1 - \frac{1-0}{(-3)-(1)} \cdot (-3) = 0.25 ; \; f(x_2) = -0.234375$$

For $k = 2$, we have

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} \cdot f(x_2)$$

$$= 0.25 - \frac{0.25 - 1}{(-0.234375) - (-3)} \cdot (-0.234375)$$

$$= 0.186441, \quad f(x_3) = 0.074276$$

For $k = 3$, we have

$$x_4 = x_3 - \frac{x_3 - x_2}{f(x_3) - f(x_2)} \cdot f(x_3)$$

$$= 0.186441 - \frac{0.186441 - 0.25}{0.074276 - (-0.234375)} \cdot (0.074276)$$

$$= 0.201736, \quad f(x_4) = -0.00047$$

For $k = 4$, we have

$$x_5 = x_4 - \frac{x_4 - x_3}{f(x_4) - f(x_3)} \cdot f(x_4)$$

$$= 0.201736 - \frac{0.201736 - 0.186441}{(-0.00047) - (0.074276)} \cdot (-0.00047)$$

$$= 0.201640$$

Thus after four iterations approximate root is 0.20164.

Answer (b) : $f(x) = \cos x - xe^x = 0$

By Secant method

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \cdot f(x_k)$$

Here $x_0 = 0$, $x_1 = 1$, $f(x_0) = 1$, $f(x_1) = -2.177979523$,

For $k = 1$, we have

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} \cdot f(x_1)$$

$$= 1 - \frac{1 - 0}{(-2.177979523) - 1} \cdot (-2.177979523)$$

$$= 0.3146653378, \quad f(x_2) = 0.519871175$$

For $k = 2$, we have

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} \cdot f(x_2)$$

$$= 0.3146653378 - \frac{0.3146653378 - 1}{0.519871175 - (-2.177979523)} \cdot (0.519871175)$$

$$= 0.4467281466, \quad f(x_3) = 0.203544710$$

For $k = 3$, we have

$$x_4 = x_3 - \frac{x_3 - x_2}{f(x_3) - f(x_2)} \cdot f(x_3)$$

$$= 0.4467281466 - \frac{0.4467281466 - 0.3146653378}{0.203544710 - 0.519871175} \cdot 0.20354471$$

$$= 0.5317058606, \quad f(x_4) = 0.0950824$$

For $k = 4$, we have

$$x_5 = x_4 - \frac{x_4 - x_3}{f(x_4) - f(x_3)} \cdot f(x_4)$$

$$= 0.5317058606 - \frac{0.5317058606 - 0.4467281466}{0.0950824 - 0.203544710} \cdot 0.0950824$$

$$= 0.606200724$$

After four iterations approximate root is 0.606200724.

Answer (c) : $f(x) = xe^x - 1$

By Secant method

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \cdot f(x_k)$$

Here $x_0 = 0$, $x_1 = 1$, $f(x_0) = -1$, $f(x_1) = 1.718281828$.

For $k = 1$,

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} \cdot f(x_1)$$

$$= 1 - \frac{1 - 0}{1.718281828 - (-1)} \cdot (1.718281828)$$

$$= 0036787944, \quad f(x_2) = -0.468536395$$

For $k = 2$,

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} \cdot f(x_2)$$

$$= 0.36787944 - \frac{0.36787944 - 1}{(-0.468536395) - (1.718281828)} \cdot (-0.468536395)$$

$$= 0.503314365, \quad f(x_3) = -0.167419994$$

For $k = 3$,

$$x_4 = x_3 - \frac{x_3 - x_2}{f(x_3) - f(x_2)} \cdot f(x_3)$$

$$= 0.503314365 - \frac{0.503314365 - 0.3146653378}{0.203544710 - 0.519871175} \cdot 0.20354471$$

$$= 0.624703233, \quad f(x_4) = 0.166752984$$

For $k = 4$,

$$x_5 = x_4 - \frac{x_4 - x_3}{f(x_4) - f(x_3)} \cdot f(x_4)$$

$$= 0.624703233 - \frac{0.624703233 - 0.503314365}{0.166752984 + 0.167419994} \cdot (0.166752984)$$

$$= 0.564129945, \quad f(x_5) = -0.008306022475$$

The approximate root of $f(x) = xe^x - 1 = 0$ is 0.564129945.

3. A real root of following functions lies in the internal (0, 1). Perform four iterations of Regula falsi method to obtain the approximate root.

(a) $x^3 - 5x + 1 = 0$ (b) $\cos x - xe^x = 0$ (c) $xe^x - 1 = 0$

Answer (a) : $f(x) = x^3 - 5x + 1 = 0$

We have $x_0 = 0$, $x_1 = 1$, $f(x_0) = 1$, $f(x_1) = -3$

Since $f(x_0) \cdot f(x_1) < 0$, root lies between 0 and 1.

By Regula falsi method,

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \cdot f(x_k)$$

For $k = 1$,

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} \cdot f(x_1)$$

$$= 0.25, \quad f(x_2) = -0.234375$$

Since $f(x_0) \cdot f(x_2) < 0$, root lies between x_0 and x_2.

Put $x_1 = x_0 = 0$

For $k = 2$,

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} \cdot f(x_2)$$

$$= 0.25 - \frac{0.25 - 0}{-0.234375 - 1} \cdot (-0.234375)$$

$$= 0.202532, \quad f(x_3) = -0.004352$$

Since $f(x_0) \cdot f(x_3) < 0$, root lies between x_0 and x_3.

Put $x_2 = x_0 = 0$

For $k = 3$,

$$x_4 = x_3 - \frac{x_3 - x_2}{f(x_3) - f(x_2)} \cdot f(x_3)$$

$$= 0.202532 - \frac{0.202532 - 0}{-0.004352 - 1} \cdot (-0.004352)$$

$$= 0.201654, \quad f(x_4) = -0.00007$$

Since $f(x_0) \cdot f(x_4) < 0$, root lies between x_0 and x_4.

Put $x_3 = x_0 = 0$.

For $k = 4$,

$$x_5 = x_4 - \frac{x_4 - x_3}{f(x_4) - f(x_3)} \cdot f(x_4)$$

$$= 0.201654 - \frac{0.201654 - 0}{-0.00007 - 1} \cdot (-0.00007)$$

$$= 0.201640$$

$$|x_5 - x_4| = |0.201640 - 0.201654| = 0.000014 = 1.4 \times 10^{-5}$$

The root $x = 0.201640$ is correct upto 4 decimal places.

Answer (b) : $f(x) = \cos x - xe^x = 0$.

We have $x_0 = 0$, $x_1 = 1$, $f(x_0) = 1$, $f(x_1) = -2.177979523$.

Since $f(x_0) \cdot f(x_1) < 0$, root lies between 0 and 1.

For $k = 1$,

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} \cdot f(x_1)$$

$$= 0.3146653378, \quad f(x_2) = 0.519871175$$

Since $f(x_1) \cdot f(x_2) < 0$, root lies between x_1 and x_2.

For $k = 2$,

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} \cdot f(x_2)$$

$$= 0.3146653378 - \frac{0.3146653378 - 1}{0.519871175 + 2.177979523} \cdot 0.519871175$$

$$= 0.4467281466, \quad f(x_3) = 0.203544710$$

Since $f(x_1) \cdot f(x_3) < 0$, root lies between x_1 and x_3.

Put $x_2 = x_1$ then for $k = 3$,

$$x_4 = x_3 - \frac{x_3 - x_2}{f(x_3) - f(x_2)} \cdot f(x_3)$$

$$= 0.4467281466 - \frac{0.4467281466 - 1}{0.203544710 + 2.177979523} \cdot (0.203544710)$$

$$= 0.4940153366, \quad f(x_4) = 0.0708023$$

Since $f(x_1) \cdot f(x_4) < 0$, root lies between x_1 and x_4.

Put $x_3 = x_1$ then for $k = 4$,

$$x_5 = x_4 - \frac{x_4 - x_3}{f(x_4) - f(x_3)} \cdot f(x_4)$$

$$= 0.4940153366 - \frac{0.49401533 - 1}{0.0708023 + 2.177979523} \cdot (0.0708023)$$

$$= 0.5099461404$$

The approximate rooy after four iteration is 0.5099461404.

Answer (c) : $f(x) = xe^x - 1 = 0$

We have $x_0 = 0$, $x_1 = 1$, $f(x_0) = -1$, $f(x_1) = 1.718281828$.

Since $f(x_0) \cdot f(x_1) < 0$, root lies between x_0 and x_1.

By Rehula falsi method

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \cdot f(x_k)$$

For $k = 1$,

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} \cdot f(x_1)$$

$$= 1 - \frac{1 - 0}{1.718281828 + 1} \cdot (1.718281828)$$

$$= 0.36787944, \quad f(x_2) = -0.468536395$$

Since $f(x_1) \cdot f(x_2) < 0$, root lies between x_1 and x_2.

For $k = 2$,

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} \cdot f(x_2)$$

$$= 0.36787944 - \frac{0.36787944 - 1}{-0.46853695 - 1.718281828} \cdot (-0.46853695)$$

$$= 0.503314365, \quad f(x_3) = -0.167419994$$

Since $f(x_1) \cdot f(x_3) < 0$, root lies between x_1 and x_3.

Put $x_2 = x_1$ then for $k = 3$,.

$$x_4 = x_3 - \frac{x_3 - x_2}{f(x_3) - f(x_2)} \cdot f(x_3)$$

$$= 0.503314365 - \frac{0.503314365 - 1}{-0.167419994 - 1.718281828} \cdot (-0.167419994)$$

$$= 0.547412061, \quad f(x_4) = -0.053648664$$

Since $f(x_1) \cdot f(x_4) < 0$, root lies between x_1 and x_4.

Put $x_3 = x_1$ then for $k = 4$, we have

$$x_5 = x_4 - \frac{x_4 - x_3}{f(x_4) - f(x_3)} \cdot f(x_4)$$

$$= 0.547412061 - \frac{0.547412061 - 1}{-0.053648664 - 1.718281828} \cdot (-0.053648664)$$

$$= 0.561115046$$

The approximate root after four iterations is 0.561115046.

4. (a) Perform four iterations of the Newton Raphson method to find the smallest positive root of the equation $f(x) = x^3 - 5x + 1 = 0$.

(b) Apply Newton Raphson method to determine a root of the equation $f(x) = \cos x - xe^x = 0$, correct upto thnree decimal places.

(c) Perform four iterations on Newton-Raphson method to obtain approximate value of $(77)^{1/3}$ starting with initial approximation $x_0 = 2$.

(d) To get the convergent Newton-Raphson method, show that the initial approximation x_0 for finding $\frac{1}{N}$, where N is positive integer, must satisfy $0 < x_0 < \frac{2}{N}$.

(e) Perform four iterations of Newton Raphson method to find approximate root of $f(x) = xe^x - 1$.

Answer (a) : The smallest positive root of equation $f(x) = x^3 - 5x + 1 = 0$ lies between 0 and 1.

Let initial approximation $x_0 = 0.5$.

In Newton Raphson method $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$.

$f(x) = x^3 - 5x + 1$ and $f'(x) = 3x^2 - 5$

For $k = 0$, $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$

$$= 0.5 - \frac{f(0.5)}{f'(0.5)} = 0.176470588$$

For $k = 1$,

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

$$= 0.176470588 - \frac{f(0.176470588)}{f'(0.176470588)}$$

$$= 0.201568074$$

For $k = 2$,

$$x_3 = 0.201568074 - \frac{f(0.201568074)}{f'(0.201568074)}$$

$$= 0.201639675$$

For $k = 3$,

$$x_4 = 0.201639675 - \frac{f(0.201639675)}{f'(0.201639675)}$$

$$= 0.201639678$$

For $k = 4$,

$$x_5 = 0.201639678 - \frac{f(0.201639610)}{f'(0.201639678)}$$

$$= 0.201639675$$

Since $|x_5 - x_4| = |0.201639675 - 0.201639678|$

$$= 2.252 \times 10^{-9}$$

The root $x = 0.201639675$ is correct upto 8 decimal places.

Answer (b) : The root of equation $f(x) = \cos x - xe^x = 0$ lies between 0 and 1. Let initial approximation $x_0 = 0.5$.

By Newton Raphson method $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$.

Here $f(x) = \cos x - xe^x$ then $f'(x) = -\sin x - (x+1)e^x$.

For $k = 0$,

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

$$= 0.5 - \frac{\cos(0.5) - (0.5)e^{0.5}}{-\sin(0.5) - (0.5+1)e^{0.5}}$$

$$= 0.518026009$$

For $k = 1$,

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

$$= 0.517757424$$

For $k = 2$,

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}$$

$$= 0.517757363$$

$$|x_2 - x_3| = |0.517757363 - 0.517757424|$$

$$= 6.03 \times 10^{-8}$$

The root $x = 0.567125668$ is correct upto 6 decimal places.

Answer (c) : Let $x = 17^{1/3}$ then $x^3 = 17$ and $f(x) = x^3 - 17 = 0$.

By Newton Raphson Method,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

$$= x_k - \frac{{x_k}^3 - 17}{3{x_k}^2}$$

$x_0 = 2$. For $k = 0$,

$$x_1 = x_0 - \frac{{x_0}^3 - 17}{3{x_0}^2} = 2.75$$

For $k = 1$,

$$x_2 = x_1 - \frac{x_1^3 - 17}{3x_1^2} = 2.582644628$$

For $k = 2$,

$$x_3 = x_2 - \frac{x_2^3 - 17}{3x_2^2} = 2.571331512$$

For $k = 3$,

$$x_4 = x_3 - \frac{x_3^3 - 17}{3x_3^2} = 2.571281592$$

Since $|x_3 - x_4| = |2.571331512 - 2.571281592| = 4.9 \times 10^{-5}$, the root is correct upto 4 decimal places. The exact value of $17^{1/3}$ correct upto four decimal places is 2.571281592.

Answer (d) : To find $\frac{1}{N}$, let $x = \frac{1}{N}$ then $\frac{1}{x} = N$ i.e. $\frac{1}{x} - N = 0$.

We write $f(x) = \frac{1}{x} - N = 0$. By Newton Raphson method

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{\frac{1}{x_k} - N}{-\frac{1}{x_k^2}} = x_k + x_k^2\left(\frac{1}{x_k} - N\right) = 2x_k - Nx_k^2$$

Now plot the graphs of $y = x$ and $y = 2x - Nx^2$.

$$y = 2x - Nx^2 = -N\left(x^2 - \frac{2}{N}x\right) = -N\left(x - \frac{1}{N}\right)^2 + \frac{1}{N}$$

i.e. $\left(x - \frac{1}{N}\right)^2 = -\frac{1}{N}\left(y - \frac{1}{N}\right)$ which is parabola. From the graph of this function we find that for any initial approximation outside the range $0 < x_0 < \frac{2}{N}$ the method diverges. If $x_0 = 0$, the iterations do not converge to $\frac{1}{N}$ but remains zero always.

Answer (e) : The root of equation $f(x) = xe^x - 1$ lies between 0 and 1. Let $x_0 = 0.5$.

By Newton Raphson Method

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

$f(x) = xe^x - 1$ and $f'(x) = (x+1)e^x$

For $k = 0$,

$$x_1 = x_0 - \frac{f(x_2)}{f'(x_0)} = 0.5 - \frac{f(0.5)}{f'(0.5)} = 0.571020439$$

For $k = 1$,

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 0.567155569$$

For $k = 2$,

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = 0.567143291$$

For $k = 3$,

$$x_4 = x_3 - \frac{f(x_3)}{f'(x_3)} = 0.56714329$$

$|x_4 - x_3| = 5 \times 10^{-10}$, the root $x = 0.56714329$ is correct upto 9 decimal places.

5. Use the iteration method to find, correct to four significant figures, a real root of each of the following equation.

(a) $x^3 + x^2 - 1 = 0$ (b) $2x - \cos x - 3 = 0$ (c) $xe^x - 1 = 0$

Answer (a) : $f(x) = x^3 + x^2 - 1 = 0$

$$f(x) = x^2(x+1) - 1 = 0 \Rightarrow x^2(x+1) - 1 \text{ i.e. } x = \frac{1}{\sqrt{1+x}}$$

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

Observe that $f(0) = -1$ and $f(1) = 1$. Therefore root lies between 0 and 1. On the interval (0, 1), $|\phi'(x)| < \frac{1}{2} < 1$. Thus iteration method.

$$x_{k+1} = \phi(x_k) = \frac{1}{\sqrt{1+x_k}} \text{ converges}$$

Put $x_0 = 1$, $x_1 = \frac{1}{\sqrt{2}} = 0.707106781$

$$x_2 = \frac{1}{\sqrt{1+x_1}} = 0.765366864$$

$$x_3 = \frac{1}{\sqrt{1+x_2}} = 0.752993979$$

$$x_4 = \frac{1}{\sqrt{1+x_3}} = 0.755283135$$

$$x_5 = \frac{1}{\sqrt{1+x_4}} = 0.754790473$$

$$x_6 = \frac{1}{\sqrt{1+x_5}} = 0.75489642$$

$$x_7 = \frac{1}{\sqrt{1+x_6}} = 0.754873632$$

$|x_7 - x_6| = 2.2 \times 10^{-5}$. The approximate root 0.754873632 is correct upto 4 significant figure.

Answer (b) : $2x - \cos x - 3 = 0 \Rightarrow x = \frac{1}{2}(\cos x + 3)$

On comparing this equation with $x = \phi(x)$ we have

$$\phi(x) = \frac{1}{2}(\cos x + 3), \quad |\phi'(x)| = \left|\frac{\sin x}{2}\right| < 1$$

Hence the iteration method is convergent.

Consider $x_{k+1} = \frac{1}{2}(\cos x_k + 3)$

Since $f(x) = 2x - \cos x - 3$ and $f(1) < 0$ and $f(2) > 0$, the root lies between 1 and 2. Let the initial approximation

$x_0 = 1.5$

$x_1 = \frac{1}{2}(\cos x_0 + 3) = 1.535368601$

$x_2 = \frac{1}{2}(\cos x_1 + 3) = 1.517710158$

$x_3 = \frac{1}{2}(\cos x_2 + 3) = 1.526530619$

$x_4 = \frac{1}{2}(\cos x_3 + 3) = 1.522125627$

$x_5 = \frac{1}{2}(\cos x_4 + 3) = 1.524325743$

$x_6 = \frac{1}{2}(\cos x_5 + 3) = 1.52322693$

$x_7 = \frac{1}{2}(\cos x_6 + 3) = 1.523775729$

$x_8 = \frac{1}{2}(\cos x_7 + 3) = 1.523501637$

$x_9 = \frac{1}{2}(\cos x_8 + 3) = 1.52363853$

$x_{10} = \frac{1}{2}(\cos x_9 + 3) = 1.52357016$

$x_{11} = \frac{1}{2}(\cos x_{10} + 3) = 1.523604307$

$|x_{10} - x_{11}| = 3.4 \times 10^{-5}$. The approximate root 1.523604307 is correct upto four significant figures.

Answer (c) : $f(x) = xe^x - 1$. Since $f(0) = -1$ and $f(1) = e - 1 > 0$, root lies between 0 and 1. Let $x_0 = 0.5$.

$xe^x - 1 = 0 \Rightarrow x = e^{-x}$, $\phi(x) = e^{-x}$, $|\phi'(x)| = e^{-x} < 1$ for $x \in (0,1)$

$\therefore x_{k+1} = e^{-x_k}$ is the iteration formula.

$x_0 = 0.5$. $x_1 = e^{-0.5} = 0.606530659$

$x_2 = e^{-x_1} = 0.545239212$

$x_3 = e^{-x_2} = 0.579703094$

$x_4 = e^{-x_3} = 0.560064628$

$x_5 = e^{-x_4} = 0.571172148$

$x_6 = e^{-x_5} = 0.564862947$

$x_7 = e^{-x_6} = 0.568438047$

$x_8 = e^{-x_7} = 0.566409453$

$x_9 = e^{-x_8} = 0.567559634$

$x_{10} = e^{-x_9} = 0.566907213$

$x_{11} = e^{-x_{10}} = 0.567277195$

$x_{12} = e^{-x_{11}} = 0.567067352$

$x_{13} = e^{-x_{12}} = 0.56718636$

$x_{14} = e^{-x_{13}} = 0.567118864$

$x_{15} = e^{-x_{14}} = 0.567157143$

$|x_{14} - x_{15}| = 3.8 \times 10^{-5}$, root $x = 0.567157143$ is correct upto four significant figures.

6. Perform two iterations of Birge Vieta method to find the root of polynomial $P_3(x) = 2x^3 - 5x + 1 = 0$. Use the initial approximation $p_0 = 0.5$. Also obtain the deflated polynomial.

Answer : $p_3(x) = 2x^3 - 5x + 1$

$= 2x^3 + 0x^2 - 5x + 1$ and $p_0 = 0.5$.

0.5	2	0	– 5	1
		1	0.5	– 2.25
	2	1	– 4.5	$-1.25 = b_3$
		1	1	
	2	2	$-3.5 = c_2$	

$$p_1 = p_0 - \frac{b_3}{c_2} = 0.5 - \frac{(-1.25)}{(-3.5)} = 0.142857$$

0.142857	2	0	– 5	1
		0.285714	0.040816	– 0.708454
	2	0.285714	– 4.959184	$0.291546 = b_3$
		0.285714	0.081632	
	2	0.571428	$-4.877552 = c_2$	

$$p_2 = p_1 - \frac{b_3}{c_2} = 0.142857 - \frac{(0.291546)}{(-4.877552)} = 0.202630$$

Thus 0.202630 is root after two iterations. To find the deflated polynomial we use synthetic division

0.202630	2	0	– 5	1
		0.405260	0.082118	– 0.996510
	2	0.405260	– 4.917882	0.003490

Observe that $b_3 = 0.003490$, is the error in factorization.

The deflated polynomial $Q_2(x) = 2x^2 + 0.405260x - 4.917882$.

7. Perform two iterations of Birge Vieta method to find the root of polynomial $x^4 - 3x^3 + 3x^2 - 3x + 2 = 0$. Use the initial approximation $p_0 = 0.5$.

Answer :

0.5	1	–3	3	– 3	2
		0.5	– 1.25	0.875	– 1.0625
	1	– 2.5	1.75	– 2.125	$0.9375 = b_4$
		0.5	– 1.0	0.375	
	1	– 2.0	0.75	$-1.750 = c_3$	

$$p_1 = p_0 - \frac{b_4}{c_3} = 0.5 - \frac{0.9375}{(-1.750)} = 1.0356$$

Second iteration $p_1 = 1.0356$.

1.0356	1	– 3	3	– 3	2
		1.0356	– 2.0343	1.0001	– 2.0711
	1	– 1.9644	0.9657	– 1.9999	$-0.0711 = b_4$
		1.0356	– 0.9619	0.0039	
	1	– 0.9288	0.0038	$-1.9960 = c_3$	

$$p_2 = p_1 - \frac{b_4}{c_3} = 1.0356 - \frac{(-0.0711)}{(-1.9960)} = 0.999979$$

8. Perform two iterations of the Bainstow method to extract a quadratic factor $x^2 + p \times + q$ from the polynomial $p_3(x) = x^3 + x^2 - x + 2 = 0$. Use initial approximation $p_0 = -0.9$ and $q_0 = 0.9$.

Answer :

$-p_0 = 0.9$	1	1	– 1	2
$-q_0 = -0.9$		0.9	1.71	– 0.171
			– 0.9	– 1.71
	1	1.9	$-0.19 = b_2$	$0.119 = b_3$
		0.9	2.52	
			– 0.9	
	$1 = c_0$	$2.8 = c_1$	$1.43 = c_2$	

$$\Delta p = -\frac{b_3 c_0 - b_2 c_1}{{c_1}^2 - c_0\left(c_2 - b_2\right)} = -\frac{0.651}{6.22} = -0.1047$$

$$\Delta q = -\frac{b_2\left(c_2 - b_2\right) - b_3 c_1}{{c_1}^2 - c_0\left(c_2 - b_2\right)} = \frac{0.6410}{6.22} = 0.1031$$

$$p_1 = p_0 + \Delta p = -0.9 - 0.1047 = -1.0047$$

$$q_1 = q_0 + \Delta q = 0.9 + 0.1031 = 1.0031$$

Second Iteration

$-p_1 = 1.0047$	1	1	− 1	2
		1.0047	2.0141	0.0111
$-q_1 = -1.0031$			− 1.0031	− 2.0109
	1	2.0047	$0.0110 = b_2$	$0.0002 = b_3$
		1.0047	3.0235	
			− 1.0031	
	$1 = c_0$	$3.0094 = c_1$	$2.0314 = c_2$	

$$\Delta p = -\frac{b_3 c_0 - b_2 c_1}{{c_1}^2 - c_0\left(c_2 - b_2\right)} = \frac{0.0329}{7.0361} = 0.0047$$

$$\Delta q = -\frac{b_2\left(c_2 - b_2\right) - b_3 c_1}{{c_1}^2 - c_0\left(c_2 - b_2\right)} = -\frac{0.0216}{7.0361} = -0.0031$$

$$p_2 = p_1 + \Delta p = -1.0047 + 0.0047 = -1$$

$$q_2 = q_1 + \Delta q = -1.0031 - 0.0031 = 1.0$$

Hence the quadratic factor is $x^2 + px + q = x^2 - x + 1$.

9. Perform one iteration of the Bairstow method to find the quadratic factor of the polynomial $x^4 + x^3 + 2x^2 + x + 1 = 0$. Use $p_0 = 0.5$ and $q_0 = 0.5$.

Answer :

$-p_0 = -0.5$	1	1	2	1	1
$-q_0 = -0.5$		– 0.5	– 0.25	– 0.625	– 0.0625
			– 0.5	– 0.25	– 0.625
	1	0.5	1.25	$0.125 = b_3$	$0.3125 = b_4$
		– 0.5	0.0	– 0.375	
			– 0.5	0	
	1	$0 = c_1$	$0.75 = c_2$	$– 0.25 = c_3$	

$$\Delta p = -\frac{b_4 c_1 - b_3 c_2}{{c_2}^2 - c_1 (c_3 - b_3)} = 0.1667$$

$$\Delta q = -\frac{b_3 (c_3 - b_3) - b_4 c_2}{{c_2}^2 - c_1 (c_3 - b_3)} = 0.5$$

$$p_1 = p_0 + \Delta p = 0.6667$$

$$q_1 = q_0 + \Delta q = 1.0$$

Quadratic factor is $x^2 + px + q = x^2 + 0.6667x + 1$.

UNIT - II

SYSTEM OF LINEAR ALGEBRAIC EQUATIONS AND EIGEN VALUE PROBLEMS

2.1 Introduction :

System of linear equations arise in a large number of areas, both directly in modeling physical situations and indirectly in the numerical solution of other mathematical models. These applications occur in virtually all areas of the physical, biological and social sciences. Linear systems are involved in optimization theory, numerical solutions of boundry value problems, partial differential equations, integral equations and numerous other problems.

The present chapter deals with simultaneous linear algebraic equations which can be represented generally as,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + ... + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + ... + a_{nn}x_n &= b_n \end{aligned} \qquad \text{... (2.1.1)}$$

where a_{ij} $(i, j = 1, 2, ..., n)$ are the known coefficients, b_i $(i = 1, 2, ..., n)$ are the known values and x_i $(i = 1, 2, ..., n)$ are the unknowns to be determined.

In the matrix notation, the above system of simultaneous linear algebraic equations can be written as

$$Ax = b \qquad \text{.... (2.2.2)}$$

where A is square matrix of order n, x is column vector with elements x_i, $i = 1, 2, ..., n$ and b is column vector with elements b_i, $i = 1, 2, ..., n$.

2.2 Iteration Methods

Many linear systems are too large to be solved by dirct methods based on Gauss elimination or matrix inversion. For these systems, iteration methods are often the only possible method of solution, as well as being faster than elimination in many cases. In this section we discuss two iterative methods. *viz.* Jacobi iteration method and Gauss Seidel iteration method.

A general linear iterative method for the solution of the system of equations $Ax = b$ may be defined in the form

$$\overline{X}^{(k+1)} = H\overline{X}^{(k)} + \overline{c}, \; k = 0, 1, 2, ... \qquad .... (2.2.1)$$

where $\overline{X}^{(k+1)}$ and $\overline{X}^{(k)}$ are the approximations for $\overline{X}$ at the $(k + 1)^{th}$ and k^{th} iterations respectively. H is called the iteration matrix and $\overline{c}$ is column vector. In the limiting case $\overline{X}^{(k)}$ converges to the exact solution

$$\overline{X} = \overline{A}^{-1}\overline{b} \qquad ... (2.2.2)$$

When the system of equatons can be ordered so that each diagonal entry of the coefficient matrix is larger in magnitude than the sum of the magnitudes of the other coefficients in that row - such a system is called diagonally dominant. For such system the iteration will converge for any starting values. Formally we say that an *nxn* matrix A is diagonally dominant if and only if for each $i = 1, 2, 3, ..., n$

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

For iterative methods we rearrange the system of equations so that the diagonal entries of the coefficient matrix A become diagonally dominant. If not, we rearrange the system of equations in such a way that the diagonal entries of matrix A are non-zero and possibly large in magnitude. Such a rearrangement is called pivoting.

1. Partial Pivoting

In the first state, the first column is searched for the largest element in magnitude and brought as the first diagonal element by interchanging the first equation with the equation having the largest element in magnitude. In the second stage, the second column is searched for the largest element in magnitude among the (n – 1) elements leaving the first element, and this element is brought as the second diagonal entry (a_{22}) by an interchange of the second equation with the equation having the largest element in magnitude. This procedure is continued until we arrive at the last equation.

Example 1 : Consider the system of equations

$$2x + 2y + z + 2u = 7$$

$$x - 2y - u = 2$$

$$3x - y - 2z - u = 3$$

$$x - 2u = 0$$

Ans. :

$$\begin{bmatrix} 2 & 2 & 1 & 2 \\ 1 & -2 & 0 & -1 \\ 3 & -1 & -2 & -1 \\ 1 & 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ u \end{bmatrix} = \begin{bmatrix} 7 \\ 2 \\ 3 \\ 0 \end{bmatrix}$$

$\max.\{|2|,|1|,|3|,|1|\} = |3|$ appears in third equation. Therefore interchange first and third equation

$$3x - y - 2z - u = 3$$

$$x - 2y - u = 2$$

$$2x + 2y + z + 2u = 7$$

$$x - 2u = 0$$

$$\begin{bmatrix} 3 & -1 & -2 & -1 \\ 1 & -2 & 0 & -1 \\ 2 & 2 & 1 & -2 \\ 1 & 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ u \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 7 \\ 0 \end{bmatrix}$$

Consider second column excluding (a_{12}) entry.

$\max\{|-2|, 2, 0\} = 2$ we can keep second equation as it is, since the $|-2|$ also gives the maximum value.

Consider third column excluding first two entries and calculate $\max\{|1|, 0\} = 1$. There is no need to interchange third and fourth equation and partial pivoting is complete.

The rearrangement of system is generally carries out on the augmented matrix [A, b].

Complete Pivoting :

In this procedure we search the matrix A for the largest element in magnitude and bring it as the first pivot. This requires not only an interchange of equations but also an interchange of position of the variables.

2.2.1 Jacobi Iteration Method

This method is an iteration method and is used to determine the solution of system of linear equations. In the system $\overline{A}\overline{x} = \overline{b}$, we assume that the quantities a_{ii} are non-zero and sufficiently large. This can be done by partial or complete pivoting. The system of equations (2.2.1) may be written as

$$\begin{aligned} a_{11}x_1 &= -\left(a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n\right) + b_1 \\ a_{22}x_2 &= -\left(a_{21}x_1 + a_{23}x_3 + ... + a_{2n}x_n\right) + b_2 \qquad .... (2.2.1.1) \\ &\vdots \\ a_{nn}x_n &= -\left(a_{n1}x_1 + a_{n2}x_2 + ... + a_{n(n-1)}x_{n-1}\right) + b_n \end{aligned}$$

From equation (2.2.1.1) we have an iteration method

$$\begin{aligned} x_1^{(k+1)} &= -\frac{1}{a_{11}}\left(a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + ... + a_{1n}x_n^{(k)}\right) + \frac{1}{a_{11}}b_1 \\ x_2^{(k+1)} &= -\frac{1}{a_{22}}\left(a_{21}x_1^{(k)} + a_{23}x_3^{(k)} + ... + a_{2n}x_n^{(k)}\right) + \frac{1}{a_{22}}b_2 \qquad .... (2.2.1.2) \\ &\vdots \\ x_n^{(k+1)} &= -\frac{1}{a_{nn}}\left(a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + ... + a_{n(n-1)}x_{n-1}^{(k)}\right) + \frac{1}{a_{nn}}b_n \end{aligned}$$

Initially we can assume that $x_1^{(0)} = b_1$, $x_2^{(0)} = b_2 ... x_n^{(0)} = b_n$.

Since we replace the complete vector $\overline{x}^{(k)}$ in the right side of (2.2.1.2) at the end of each iteration, this method is called the method of simultaneous displacement.

In the matrix form equation (2.2.1.1) can be written as

$$D\overline{x} = -(L+U)\overline{x} + \overline{b}$$

where L and U are respectively lower and upper traingular matries with zero diagonal entries, D is diagonal matrix such that A = L + D + U.

The matrix form of equation (2.2.1.1) is used to write an iteration method in the form,

$$\overline{x}^{(k+1)} = -D^{-1}(L+U)\overline{x}^{(k)} + D^{-1}\overline{b} \text{ , } k = 0, 1, 2, 3, ... \qquad ... (2.2.1.3)$$

$\overline{x}^{(0)} = \overline{b}$ is the initial approximation.

Alternatively equation (2.2.1.3) can be written as

$$\overline{x}^{(k+1)} = \overline{x}^{(k)} - \overline{x}^{(k)} - D^{-1}(L+U)\overline{x}^{(k)} + D^{-1}\overline{b}$$

$$= \bar{x}^{(k)} - D^{-1}(D+L+U)\bar{x}^{(k)} + D^{-1}\bar{b}$$

$$= \bar{x}^{(k)} - D^{-1}\left[A\bar{x}^{(k)} - \bar{b}\right]$$

$$= \bar{x}^{(k)} + D^{-1}\left[\bar{b} - A\bar{x}^{(k)}\right]$$

Define $\bar{V}^{(k)} = D^{-1}\bar{r}^{(k)} = \bar{x}^{(k+1)} - \bar{x}^{(k)}$ is the error in the approximation and $\bar{r}^{(k)} = \bar{b} - A\bar{x}^{(k)}$ is the residual vector.

We solve $D\bar{V}^{(k)} = \bar{r}^{(k)}$ for $\bar{V}^{(k)}$ and find

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \bar{V}^{(k)} \qquad \text{.... (2.2.1.4)}$$

These equations describe the Jacobi iteration method in an error format. Thus to solve the system of equations by Jacobi iteration method in an error form we have the following procedure

$$\bar{r}^{(k)} = \bar{b} - A\bar{x}^{(k)}$$

$$\bar{V}^{(k)} = D^{-1}\bar{r}^{(k)} \qquad \text{.... (2.2.1.5)}$$

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \bar{V}^{(k)}$$

2.2.2 Gauss Seidel Iteration Method

We know that every matrix A can be uniquely represented as the sum of lower and upper traingular matrix with zero diagonal entries and a diagonal matrix. The system of equations $A\bar{x} = \bar{b}$ can be represented by

$$(L+D+U)\bar{x} = \bar{b}$$

$$\therefore D\bar{x} = -(L+U)\bar{x} + \bar{b}$$

$$\therefore D\bar{x} = -L\bar{x} - U\bar{x} + \bar{b}$$

From above equation we have an iteration method

$$D\bar{x}^{(k+1)} = -L\bar{x}^{(k+1)} - U\bar{x}^{(k)} + \bar{b}\,,\ k = 0,\ 1,\ 2,\ 3,\ .... \qquad \text{... (2.2.2.1)}$$

Initially, we assume that $x_1^{(0)} = b_1$.

Method (2.2.2.1) is called Gauss Seidel iteration method. In the explicit form equation (2.2.2.1) can be written as

$$(L+D)\bar{x}^{(k+1)} = -U\bar{x}^{(k)} + \bar{b}$$

$$\bar{x}^{(k+1)} = -(L+D)^{-1}U\bar{x}^{(k)} + (L+D)^{-1}\bar{b}$$

$$= \bar{x}^{(k)} - \bar{x}^{(k)} - (L+D)^{-1} U\bar{x}^{(k)} + (L+D)^{-1}\bar{b}$$

$$= \bar{x}^{(k)} - (L+D)^{-1}\left[(L+D+U)\bar{x}^{(k)}\right] + (L+D)^{-1}\bar{b}$$

$$= \bar{x}^{(k)} - (L+D)^{-1}\left[A\bar{x}^{(k)} - \bar{b}\right]$$

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + (L+D)^{-1}\left[\bar{b} - A\bar{x}^{(k)}\right]$$

Thus to solve the system of equations by Gauss Seidel iteration method in an error form we have the following procedure.

$$\bar{r}^{(k)} = \bar{b} - A\bar{x}^{(k)}$$

Solve $(L+D)\bar{V}^{(k)} = \bar{r}^{(k)}$ for $\bar{V}^{(k)}$ by forward substitution

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \bar{V}^{(k)}, \quad k = 0, 1, ..., n \qquad .... (2.2.2.2)$$

System (2.2.2.2) describe the Gauss Seidel method in an error form.

2.2.3 Convergence Analysis of Iterative Methods

Iterative methods are methods of successive approximations. Convergence of iterative methods is studied through error analysis. To discuss the convergence of iterative method.

$$\bar{x}^{(k+1)} = H\bar{x}^{(k)} + \bar{c}, \quad k = 0, 1, 2,3, ... \qquad ... (2.2.3.1)$$

where $\bar{x}^{(k+1)}$ and $\bar{x}^{(k)}$ are the approximations for $\bar{x}$ at the $(k+1)^{th}$ and k^{th} iterations respectively, we study the behaviour of the difference between exact solution $\bar{x}$ and an approximation $\bar{x}^{(k)}$.

The exact solution of iterative method will satisfy

$$\bar{x} = H\bar{x} + \bar{c} \qquad .... (2.2.3.2)$$

Sibtracting (2.2.3.1) from (2.2.3.2) and substituting $\varepsilon^{-(k)} = \bar{x}^{(k+1)} - \bar{x}$ we get

$$\bar{\varepsilon}^{(k+1)} = H\bar{\varepsilon}^{(k)}, \quad k = 0, 1, 2, ... \qquad ... (2.2.3.3)$$

Repeatative application of (2.2.3.3) for $\bar{\varepsilon}^{(k)}, k = 1, 2, 3, ...$ gives

$$\bar{\varepsilon}^{(k+1)} = H^k\bar{\varepsilon}^{(0)}, \quad k = 0, 1, 2, 3, ...$$

For Jacobi iterative method $H = -D^{-1}(L+U)$ and $\bar{c} = D^{-1}\bar{b}$ whereas for Gauss Seidel iterative methods $H = (L+D)^{-1}U$ and $\bar{c} = (L+D)^{-1}\bar{b}$. For both the methods iteration matrix H remains constant for each iteration.

If the error sequence $\{\bar{\varepsilon}^{(k+1)}\}$ converges to zero as $k \to \infty$, we say that the iterativ method converges. To study the convergence of error sequence we use the following theorems.

Theorem 2.1 : Let A be a squar matrix. Then

$$\lim_{m\to\infty} A^m = 0 \text{ if } \|A\| < 1 \text{ or iff } \rho(A) < 1.$$

Before proving this theorem we explain the notations and definitions used in the statement of the theorem.

Definition : Matrix Norm

The matrix norm $\|A\|$ is a non-negative number which satisfies the properties

(i) $\|A\| \geq 0$ if $A \neq 0$ and $\|O\| = 0$ where O is zero matrix.

(ii) $\|cA\| = |c|\|A\|$ for arbitrary complex number c.

(iii) $\|A + B\| \leq \|A\| + \|B\|$

(iv) $\|AB\| \leq \|A\|\|B\|$

The most commonly used norms are

(i) Euclidean norm or frobenius norm

$$F(A) = \left[\sum_{ij=1}^{n} |a_{ij}|^2\right]^{\frac{1}{2}} \text{ where } A = \left[a_{ij}\right]_{i,j=1}^{n}$$

(ii) Maximum norm

$$\|A\| = \|A\|_\infty = \max_i \sum_{k=1}^{n} |a_{ik}| \qquad \text{(maximum absolute row sum)}$$

$$\|A\|_1 = \max_k \sum_{i=1}^{n} |a_{ik}| \qquad \text{(maximum absolute column sum)}$$

(iii) Hilbert norm or Spectral norm

The largest eigen value in modulus of a matrix A is called the spectral radius of the matrix A and is denoted by $\rho(A)$. The spectral radius is defined only for square matrices.

$$\|A\|_2 = \sqrt{\lambda} \text{ where } \lambda = \rho(A^* A) \text{ and } A^* = (\overline{A})^T,$$

$\overline{A}$ is the complex conjugate of A.

Proof of Theorem 2.1 :

If $\|A\| < 1$ then by definition of norm of matrix.

$$\|A^m\| \le \|A\|^m \qquad (\because \|AB\| \le \|A\|\|B\|)$$

and since norm is a continuous function,

$$\lim_{m\to\infty} \|A^m\| = \left\| \lim_{m\to\infty} A^m \right\| \le \lim_{m\to\infty} \|A\|^m = 0 \qquad (\because \|A\| < 1)$$

For simplicity, we assume that all the eigen values of the matrix A are distinct. Then there exist a similarity transformation S such that

$$A = S^{-1}DS$$

where D is diagonal matrix having eigen values of A on the diagonal. Therefore

$$A^m = (S^{-1}DS)(S^{-1}DS)...(S^{-1}DS)_{(m \text{ times})}$$

$$= S^{-1}D(SS^{-1})D...(SS^{-1})DS$$

$$= S^{-1}D^mS$$

where $D^m = \begin{bmatrix} \lambda_1^m & 0 & 0 & 0\ldots 0 \\ 0 & \lambda_2^m & 0 & 0\ldots 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0\ldots \lambda_n^m \end{bmatrix}$

$\lim_{m\to\infty} A^m = S^{-1}\left(\lim_{m\to\infty} D^m\right)S = 0$ iff all the eigen values satisfy $|\lambda_i| < 1$, i.e. $\rho(A) < 1$.

Theorem 2.2 : The infinite series $+A + A^2 + A^3 + ...$ converges iff $\lim_{m\to\infty} A^m = 0$. The series converges to $(I - A)^{-1}$.

Proof : From the definition of convergent series we have

$$\lim_{m\to\infty} A^m = 0 .$$

Suppose $\lim_{m\to\infty} A^m = 0$ then by theorem 2.1 we have

$$\rho(A) < 1$$

Since the magnitude of largest eigen value of matrix A is strictly less than 1, $|I - A| \neq 0$ and therefore $(I - A)^{-1}$ exists.

We know the identity

$$(I + A + A^2 + A^3 + ... + A^m)(I - A) = I - A^{m+1}$$

$$\therefore (I + A + A^2 + A^3 + ... + A^m) = (I - A^{m+1})(I - A)^{-1}$$

Since $\lim_{m \to \infty} A^m = 0$ (by theorem 2.1), we have

$$I + A + A^2 + A^3 + ... + A^m = (I - A)^{-1}$$

Theorem 2.3 : No eigen value of matrix A exceeds the norm of a matrix A. i.e. $\|A\| \geq \varrho(A)$

Proof : For eigen value λ of a matrix A we have

$$A\bar{x} = \lambda\bar{x}$$

where $\bar{x}$ is a non-zero eigen vector corresponding to eigen value λ.

$$\therefore \|\lambda\bar{x}\| = |\lambda|\|\bar{x}\| = \|A\bar{x}\| \leq \|A\|\|\bar{x}\|$$

Thus we have $|\lambda|\|\bar{x}\| \leq \|A\|\|\bar{x}\|$ $\quad (\|\bar{x}\| \neq 0)$

i.e. $|\lambda| \leq \|A\|$ (where λ is any eigen value)

i.e. $\varrho(A) \leq \|A\|$

Theorem 2.4 : The iteration method of the form

$$\bar{x}^{(k+1)} = H\bar{x}^{(k)} + \bar{c}$$

for the system of equation $A\bar{x} = \bar{b}$ converges to the exact solution for any initial vector $\bar{x}^{(0)}$ if $\|H\| < 1$.

Proof : We take initial vector $\bar{x}^{(0)} = \bar{0}$. Then the repeated application of iteration method gives

$$\bar{x}^{(1)} = \bar{c}$$

$$\bar{x}^{(2)} = H\bar{x}^{(1)} + \bar{c} = H\bar{c} + \bar{c} = (H + I)\bar{c}$$

$$\bar{x}^{(3)} = H\bar{x}^{(2)} + \bar{c} = H(H + I)\bar{c} + \bar{c} = (H^2 + H + I)\bar{c}$$

$$\vdots$$

$$\bar{x}^{(k+1)} = (H^k + H^{k-1} + ... + H^2 + H + I)\bar{c}$$

$$\lim_{k\to\infty} \bar{x}^{(k+1)} = \lim_{k\to\infty} \left(H^k + H^{k-1} + H^{k-2} + ... + H^2 + H + I\right)\bar{c}$$

$$= (I-H)^{-1}\bar{c} \qquad \text{(if } \|H\| < 1\text{)}$$

Thus $\lim_{k\to\infty} \bar{x}^{(k+1)} = (I-H)^{-1}\bar{c}$

In case of Jacobi iteration method we have

$$H = -D^{-1}(L+D) \text{ and } \bar{c} = D^{-1}\bar{b}$$

$$\therefore (I-H)^{-1}\bar{c} = \left[I + D^{-1}(L+U)\right]^{-1} D^{-1}\bar{b}$$

$$= \left[D^{-1}D + D^{-1}(L+U)\right]^{-1} D^{-1}\bar{b}$$

$$= \left[D^{-1}(D+L+U)\right]^{-1} D^{-1}\bar{b}$$

$$= (D+L+U)^{-1}\left(D^{-1}\right)^{-1} D^{-1}\bar{b}$$

$$= A^{-1}\bar{b}$$

$$= \bar{x}$$

Thus for Jacobi iteration method we have

$$\lim_{k\to\infty} \bar{x}^{(k+1)} = (I-H)^{-1}\bar{c} = \bar{x}$$

In case of Gauss Seidel iteration method we have

$$H = -(L+D)^{-1}U \text{ and } \bar{c} = (L+D)^{-1}\bar{b}$$

$$\therefore (I-H)^{-1}\bar{c} = \left[I + (L+D)^{-1}U\right]^{-1} (L+D)^{-1}\bar{b}$$

$$= \left[(L+D)^{-1}(L+D) + (L+D)^{-1}U\right]^{-1} (L+D)^{-1}\bar{b}$$

$$= \left[(L+D)^{-1}(L+D+U)\right]^{-1} (L+D)^{-1}\bar{b}$$

$$= A^{-1}\left[(L+D)^{-1}\right]^{-1} (L+D)^{-1}\bar{b} \qquad \left(\because (AB)^{-1} = B^{-1}A^{-1}\right)$$

$$= A^{-1}\bar{b} = \bar{x}$$

Thus for Gauss Seidel iteration method we have

$$\lim_{k\to\infty} \bar{x}^{(k+1)} = (I-H)^{-1}\bar{c} = A^{-1}\bar{b} = \bar{x}$$

From theorem 2.3 we observe that Jacobi iterative method converges if

$$\rho(H_J) = \rho\left(-D^{-1}(L+U)\right) = \rho\left(D^{-1}(L+U)\right) < 1$$

and Gauss Seidel iterative method converges if

$$\rho(H_G) = \rho\left(-(L+D)^{-1}U\right) = \rho\left((L+D)^{-1}U\right) < 1$$

Theorem 2.5 : A necessary and sufficient condition for convergence of an iterative method $\bar{x}^{(k+1)} = H\bar{x}^{(k)} + \bar{c}$ is that the eigen values of the iteration matrix H satisfy $|\lambda_i(H)| < 1, i = 1, 2, ..., n$ where $\lambda_i(H)$ are eigen values of matrix H.

Proof : Suppose $\lambda_1, \lambda_2, \lambda_3, ..., \lambda_n$ are the eigen values of matrix H and $\bar{x}_1, \bar{x}_2, \bar{x}_3, ..., \bar{x}_n$ be the corresponding independent eigen vectors. $\bar{\varepsilon}^{(0)}$ is an n-vector. We write

$$\bar{\varepsilon}^{(0)} = c_1\bar{x}_1 + c_2\bar{x}_2 + ... + c_n\bar{x}_n$$

$$\therefore H^k\bar{\varepsilon}^{(0)} = c_1H^k\bar{x}_1 + c_2H^k\bar{x}_2 + ... + c_nH^k\bar{x}_n$$

$$= c_1\lambda_1{}^k\bar{x}_1 + c_2\lambda_2{}^k\bar{x}_2 + ... + c_n\lambda_n{}^k\bar{x}_n$$

$$\lim_{k\to\infty}\bar{\varepsilon}^{(k)} = H^k\bar{\varepsilon}^{(0)} = \lim_{k\to\infty}\left(\sum_{i=1}^{n} c_i\lambda_i{}^k\bar{x}_i\right)$$

$\bar{\varepsilon}^{(k)} \to \bar{0}$ iff $\lambda_i{}^k \to 0$ as $k \to \infty$.

i.e. $\bar{\varepsilon}^{(k)} \to \bar{0}$ iff $|\lambda_i(H)| < 1$.

Definition 2.2 : The rate of convergence of an iterative method is given by

$v = -\log_{10}[\rho(H)]$ where $\rho(H)$ is the spectral radius of matrix H.

Theorem 2.6 : If A is a strictly diagonally dominant matrix, then the Jacobi iteration scheme converges for any initial starting vector.

Proof : The Jacobi iteration scheme is given by

$$\bar{x}^{(k+1)} = -D^{-1}(L+U)\bar{x}^{(k)} + D^{-1}\bar{b}$$

$$= -D^{-1}(A-D)\bar{x}^{(k)} + D^{-1}\bar{b}$$

$$= \left(I - D^{-1}A\right)\bar{x}^{(k)} + D^{-1}\bar{b}$$

The iteration scheme will be convergent if $\|I - D^{-1}A\| < 1$.

Using absolute row sum norm we have

$$\|I - D^{-1}A\| = \max\left\{1 - \frac{1}{a_{ii}}\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\right\}$$

Since A is strictly diagonal dominant, $\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}| < |a_{ii}|$ for all i = i, 1, 2, 3, ..., n and therefore $\|I - D^{-1}A\| < 1$ and therefor the Jacobi iteration scheme converges for any initial starting vector.

Theorem 2.7 : If A is strictly diagonally dominant matrix, then the Gauss-Seidel iteration scheme converges for any initial starting vector.

Proof : The Gauss Seidel iteration scheme is given by

$$\begin{aligned}\bar{x}^{(k+1)} &= -(D+L)^{-1}U\bar{x}^{(k)} + (D+L)^{-1}\bar{b}\\ &= -(D+L)^{-1}[A-(D+L)]\bar{x}^{(k)} + (D+L)^{-1}\bar{b}\\ &= \left[I-(D+L)^{-1}A\right]\bar{x}^{(k)} + (D+L)^{-1}\bar{b}\end{aligned}$$

The iteration scheme will be convergent if

$$\rho\left(I-(D+L)^{-1}A\right) < 1$$

Let λ be an eigen value of $I-(D+L)^{-1}A$.

$$\left[I-(D+L)^{-1}A\right]\bar{x} = \lambda\bar{x}$$

$$\left[(D+L)^{-1}(D+L)-(D+L)^{-1}A\right]\bar{x} = \lambda\bar{x}$$

$$(D+L)^{-1}[D+L-A]\bar{x} = \lambda\bar{x}$$

$$\therefore -U\bar{x} = \lambda(D+L)\bar{x} \qquad (\because A = L+D+U)$$

i.e. $$-\sum_{j=i+1}^{n}a_{ij}x_j = \lambda\sum_{j=1}^{i}a_{ij}x_j \qquad 1 \le i \le n$$

i.e. $$-\sum_{j=i+1}^{n}a_{ij}x_j = \lambda a_{ii}x_i + \lambda\sum_{j=i}^{i-1}a_{ij}x_j$$

i.e. $$\lambda a_{ii}x_i = -\lambda\sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j$$

$$|\lambda a_{ii}x_i| \le |\lambda|\sum_{j=i}^{i-1}|a_{ij}||x_j| + \sum_{j=i+1}^{n}|a_{ij}||x_j| \quad \text{.... (2.2.3.1)}$$

Since $\overline{x}$ is an eigen vector, $\overline{x} \neq \overline{0}$. We assume that $\|\overline{x}\| = 1$.

Choose an index *i* such that $|x_i| = 1$ and $|x_j| \le 1$, $\forall j \neq i$.

From equation (2.2.3.4) we get

$$|\lambda||a_{ii}| \le |\lambda|\sum_{j=i}^{i-1}|a_{ij}| + \sum_{j=i+1}^{n}|a_{ij}|$$

$$\therefore |\lambda|\left[|a_{ii}| - \sum_{j=i}^{i-1}|a_{ij}|\right] \le \sum_{j=i+1}^{n}|a_{ij}|$$

$$\therefore |\lambda| \le \frac{\sum_{j=i+1}^{n}|a_{ij}|}{|a_{ii}| - \sum_{j=i}^{i-1}|a_{ij}|}$$

Since matrix A is diagonally dominant $|a_{ii}| > \sum_{\substack{j=i \\ j\neq i}}^{n}|a_{ij}|$.

$$\therefore |a_{ii}| > \sum_{j=i}^{i-1}|a_{ij}| + \sum_{j=i+1}^{n}|a_{ij}|$$

or $$\therefore |a_{ii}| - \sum_{j=i}^{i-1}|a_{ij}| > \sum_{j=i+1}^{n}|a_{ij}|$$

$$\therefore \frac{\sum_{j=i+1}^{n}|a_{ij}|}{|a_{ii}| - \sum_{j=i}^{i-1}|a_{ij}|} < 1 \text{ and therefore } |\lambda| < 1.$$

But λ is any eigen value of $I-(D+L)^{-1}A$.

$$\therefore \rho\left(I-(D+L)^{-1}A\right)<1$$

Note : The rate of convergence of Gauss Seidel scheme is twice that of the Jacobi scheme. It may happen that for the system $A\bar{x}=\bar{b}$, $\rho(H_J)<1$ but $\rho(H_G)>1$. Similarly it is possible to have the system of equation $A\bar{x}=\bar{b}$ for $\rho(H_G)<1$ which but $\rho(H_J)>1$. For these systems matrix A is not diagonally dominant.

2.3 Matrix Factorization Method

The system of equation $A\bar{x}=\bar{b}$ can be directly solved in the following cases.

Case (i) : A = D

The system of equations is

$$a_{ii}x_i=b_i \Rightarrow x_i=\frac{b_i}{a_{ii}} \quad \text{for } i = 1, 2, 3, ..., n ;\ a_{ii}\neq 0 \quad \forall i$$

Case (ii) : A = L (Lower Traingular Matrix)

The system of equations is of the form

$$a_{11}x_1=b_1$$

$$a_{21}x_1+a_{22}x_2=b_2$$

$$a_{31}x_1+a_{32}x_2+a_{33}x_3=b_3$$

$$\vdots$$

$$a_{n1}x_1+a_{n2}x_2+a_{n3}x_3+.....+a_{nn}x_n=b_n$$

Solving first equation we get x_1. If we substitute this value of x_1 in second equation we get x_2 and so on. Since unknowns are determined by forward substitution, the method is called forward substitution method.

Case (iii) : A = U (Upper Traingular Matrix)

In this case system of equation is of the form

$$a_{11}x_1+a_{12}x_2+a_{13}x_3+.....+a_{1n}x_n=b_1$$

$$a_{22}x_2+a_{23}x_3+.....+a_{2n}x_n=b_2$$

$$a_{33}x_3 + \ldots\ldots + a_{3n}x_n = b_3$$

$$\vdots$$

$$a_{(n-1)(n-1)}x_{n-1} + a_{(n-1)n}x_n = b_{n-1}$$

$$a_{nn}x_n = b_n$$

From last equation we get x_n from second last equation on substituting the value of x_n we get x_{n-1} and so on.

Since the unknown are determined from back substitution, this method is called the back substitution method.

Thus the equation $A\bar{x} = \bar{b}$ is directly solvable if the matrix A can be transformed into one of the three cases discussed above.

2.3.1 Triagulization Method

This method is also known as the decomposition method or factorization method. In this method, the coefficient matrix A of the system of equations $A\bar{x} = \bar{b}$ is decomposed or factorized into the product of a lower traingular matrix L and an upper traingular matrix U. We write the matrix A = LU.

where, $L = \begin{bmatrix} \ell_{11} & 0 & 0 & \ldots & 0 \\ \ell_{21} & \ell_{22} & 0 & \ldots & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \ldots & \ell_{nn} \end{bmatrix}$ and $U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \ldots & u_{1n} \\ 0 & u_{22} & u_{23} & \ldots & u_{2n} \\ 0 & 0 & u_{33} & \ldots & u_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & u_{nn} \end{bmatrix}$

Then the system of equation $A\bar{x} = \bar{b}$ becomes

$$LU\bar{x} = \bar{b}$$

We write above equation as the following two systems of equations

$$\begin{aligned} U\bar{x} &= \bar{z} \\ L\bar{z} &= \bar{b} \end{aligned} \qquad \ldots. (2.3.1)$$

From Case (ii) we determine $\bar{z}$ and the form Case (iii) we solve $U\bar{x} = \bar{z}$ to calculate $\bar{x}$.

2.3.1 (a) Doolittle's Method

In this method we write A = LU where the diagonal elements of matrix L are 1. We write

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & u_{33} & \dots & a_{3n} \\ \vdots & & & & \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \ell_{21} & 1 & 0 & \dots & 0 \\ \ell_{31} & \ell_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \dots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & 0 & u_{33} & \dots & u_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & u_{nn} \end{bmatrix}$$

These are n^2 equations in $\left(\frac{n(n+1)}{2} - n\right) + \frac{n(n+1)}{2} = n^2$ unknowns comparing left hand side with right hand side product. We get componentwise equations.

$$a_{11} = u_{11},\ a_{12} = u_{12},\ a_{13} = u_{13},\ ....,\ a_{1n} = u_{1n}$$

$$a_{21} = \ell_{21}u_{11},\ a_{22} = \ell_{21}u_{12} + u_{22},\ ...,\ a_{2n} = \ell_{21}u_{1n} + u_{2n}$$

By using forward substitution we calculate

$$u_{11},\ u_{12},\ u_{13},\ ....,\ u_{1n},\ \ell_{21},\ u_{22},\ u_{23},\ ....,\ u_{2n},\ .....$$

Once the matrices L and U are known the solution is obtained by representing the system $A\bar{x} = \bar{b}$ in the form of equation (2.3.1)

2.3.1 (b) Crout's Method

In this method we write A = LU where the diagonal elements of matrix U are all 1. We write

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & & & & \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} \ell_{11} & 0 & 0 & \dots & 0 \\ \ell_{21} & \ell_{22} & 0 & \dots & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \dots & \ell_{nn} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & 1 & u_{23} & \dots & u_{2n} \\ 0 & 0 & 1 & \dots & u_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Again these are n^2 equations in n^2 unknowns. Equating the componentwise elements of l.h.s. and r.h.s. we get

$$a_{11} = \ell_{11},\ a_{12} = \ell_{11}u_{12},\ a_{13} = \ell_{11}u_{13},\ ....,\ a_{1n} = \ell_{11}u_{1n}$$

$$a_{21} = \ell_{21},\ a_{22} = \ell_{21}u_{12} + \ell_{22},\ a_{23} = \ell_{21}u_{13} + \ell_{22}u_{23},\ ....,\ a_{2n} = \ell_{21}u_{1n} + \ell_{22}u_{2n}$$

------- ------ ------ ------

By using forward substitution we get,

$$\ell_{11} = a_{11},\ u_{12} = \frac{a_{12}}{\ell_{11}},\ u_{13} = \frac{a_{13}}{\ell_{11}},\ ....,\ u_{1n} = \frac{a_{1n}}{\ell_{11}}$$

$$\ell_{21} = a_{21},\ \ell_{22} = a_{22} - \ell_{21}u_{12},\ ....,\ u_{2n} = a_{2n} - \ell_{21}u_{1n}$$

Once the matrices L and U are determined the system of equations $A\bar{x} = \bar{b}$ is represented in the form of equation (2.3.1) and system of equations (2.3.1) is solved from Case (iii) and from Case (i) respectively.

2.4 Eigen Values and Eigen Vectors

In section 2.2 and 2.3 iterative methods for linear system of equations are discussed. Consider the system of equation

$$A\bar{x} = \lambda\bar{x} \qquad \text{..... (2.4.1)}$$

Equation (2.4.1) is called eigen value problem. The eigen values of A are given by the roots of the characteristics equation

$$\det(A - \lambda I) = 0 \qquad \text{..... (2.4.2)}$$

If A is square matrix of order n, equation (2.4.2) gives a polynomial equation of degree n. The roots of this polynomial equation are called eigen values and may be determined by the methods given in Unit 1. Once the roots λ_i of polynomial (2.4.2) are known then a non-zero vector $\bar{x}_i$ such that

$$A\bar{x}_i = \lambda_i\bar{x}_i \qquad \text{.... (2.4.3)}$$

is called the eigen vector or characteristic vector corresponding to λ_i. On multiplying equation (2.4.3) by a constant c we get

$$Ac\bar{x}_i = \lambda_i c\bar{x}_i \Rightarrow A\bar{y} = \lambda_i\bar{y}$$

where $\bar{y} = c\bar{x}_i$ i.e. $\bar{y}$ is also a characteristic vector of A corresponding to eigen value λ_i. This shows that an eigen vector is determined only to within an arbitrary multiplicative constant. On premultiplying equation (2.4.1) (m – 1) times by A we obtain

$$A^m\bar{x} = \lambda A^m\bar{x} = \lambda A^{m-1}(A\bar{x}) = \lambda A^{m-1}\lambda\bar{x} = \lambda^2 A^{m-1}\bar{x}...$$

$$= \lambda^m\bar{x}$$

$$A^m \bar{x} = \lambda^m \bar{x} \quad \text{for } m = 1, 2, 3, 4, \ldots\ldots \qquad \ldots\ldots (2.4.4)$$

Equation (2.4.4) shows that λ^m is an eigen value of A^m if λ is an eigen value of A.

Since $\det\left(A^T - \lambda I\right) = \det\left(A - \lambda I\right)$, A and A^T have the same eigen values. If $\bar{u}_i$ is an eigen vector corresponding to the eigen value λ, then $A\bar{u}_i = \lambda_i \bar{u}_i$. Premultiplication by $\bar{u}_i^T$ gives $\bar{u}_i^T A\bar{u}_i = \lambda_i \bar{u}_i^T \bar{u}_i$ and we get

$$\lambda_i = \frac{\bar{u}_i^T A\bar{u}_i}{\bar{u}_i^T \bar{u}_i}$$

If $\bar{u}_i$ is an eigen vector of a matrix A then λ_i is the eigen value of matrix A. Thus given eigen value λ we find eigen vector by solving $A\bar{x} = \lambda\bar{x}$ and given eigen vector $\bar{x}$ we find the corresponding eigen vector as $\frac{\bar{x}^T A\bar{x}}{\bar{x}^T \bar{x}}$. For arbitrary $\bar{u}$, the ratio $\frac{\bar{u}^T A\bar{u}}{\bar{u}^T \bar{u}}$ is called the Rayleigh quotient.

Let A and B be two square matrices of same order. If a non-singular matrix S can be determined such that

$$B = S^{-1}AS, \qquad \ldots\ldots (2.4.5)$$

then the matrices A and B are said to be similar and the matrix S is called similarly matrix and the transformation is called similarly transformation. From equation (2.4.5) we write

$$A = SBS^{-1}$$

If λ_i is an eigen value of A and $\bar{u}_i$ is the corresponding eigen vector then

$$A\bar{u}_i = \lambda_i \bar{u}_i$$

$$S^{-1}A\bar{u}_i = \lambda S^{-1}\bar{u}_i$$

Put $\bar{u}_i = S\bar{v}_i$ then $S^{-1}AS\bar{v}_i = \lambda S^{-1}S\bar{v}_i = \lambda\bar{v}_i$.

i.e. $B\bar{v}_i = \lambda\bar{v}_i$. But then eigen values of A and B are same and given eigen vectors $\bar{u}_i$ of matrix A, $S^{-1}\bar{u}_i$ are the eigen vectors of the matrix B. A similarity transformation, where S is the matrix of eigen vectors reduces a matrix A to its diagonal form. The eigen values of A are the diagonal elements. If eigen vectors of A are linearly independent then S^{-1} exists and the matrix A is said to be diagonalizable.

2.4.1 Bounds on Eigen Values

The bounds on the eigen values of the matrix A are given by the theorems by Gerschgorin and Brauer.

Theorem (Gerschgorin)

The largest eigne value in modulus of a square matrix A cannot exceed the largest sum of the moduli of the elements along any row or any column.

Proof : Let λ_i we an eigen values of A and $\bar{x}_i$ be the corresponding eigen vector. Suppose $\bar{x}_i^T = [x_{i1}, x_{i2}, ..., x_{in}]$. Since λ_i is an eigen value of A and $\bar{x}_i$ is corresponding eigen vector,

$$A\bar{x}_i = \lambda_i \bar{x}_i$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix} = \lambda_i \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix}$$

Let $|x_{ik}| = \max_r |x_{ir}|$. Select the kth equation and divide it by x_{ik}. The kth equation is

$$a_{k1}x_{i1} + a_{k2}x_{i2} + ... + a_{kk}x_{ik} + ... + a_{kn}x_{in} = \lambda_i x_{ik}$$

Then $$\lambda_i = a_{k1}\frac{x_{i1}}{x_{ik}} + a_{k2}\frac{x_{i2}}{x_{ik}} + ... + a_{kk}\frac{x_{ik}}{x_{ik}} + ... + a_{kn}\frac{x_{in}}{x_{ik}}$$

Since $|x_{ik}| = \max_r |x_{ir}|$, $\left|\frac{x_{ir}}{x_{ik}}\right| \le 1$, $r = 1, 2, ..., n$

and $$|\lambda_i| \le |a_{k1}| + |a_{k2}| + |a_{k3}| + ... + |a_{kk}| + ... + |a_{kn}|$$

Thus if λ is an eigen value then

$$|\lambda| \le \sum_{i=1}^{n} |a_{ki}| \quad \text{for some } k.$$

$$\therefore |\lambda| \le \max_k \sum_{i=1}^{n} |a_{ki}|$$

Thus each eigen value and therefore the largest eigen value in modulus of a square matrix A cannot exceed the largest sum of the moduli of the elements along any row.

Since A and A^T have same eigen values, the theorem is also true for columns. (Repeat the procedure for A^T)

Theorem (Brauer) : Let P_k be the sum of the moduli of the elements along the k[th] row excluding the diagonal element a_{kk} of a square matrix A. Every eigen value of A lies inside or on the boundary of at least one of the circles $|\lambda - a_{kk}| = P_k$, k = 1, 2, 3, ..., n.

Proof : Let λ_i be an eigen value of A and $\bar{x}_i$ be the corresponding eigenvector. Suppose

$$\bar{x}_i^T = [x_{i1}, x_{i2}, ..., x_{in}]$$

Then $A\bar{x}_i = \lambda_i \bar{x}_i$ can be written as

$$a_{11}x_{i1} + a_{12}x_{i2} + a_{13}x_{i3} + ... + a_{1n}x_{in} = \lambda_i x_{i1}$$

$$a_{21}x_{i1} + a_{22}x_{i2} + a_{23}x_{i3} + ... + a_{2n}x_{in} = \lambda_i x_{i2}$$

$$\vdots$$

$$a_{n1}x_{i1} + a_{n2}x_{i2} + a_{n3}x_{i3} + ... + a_{nn}x_{in} = \lambda_i x_{in}$$

Let $|x_{ik}| = \max_r |x_{ir}|$. Select k[th] equation from above n equations. The k[th] equation is

$$a_{k1}x_{i1} + a_{k2}x_{i2} + a_{k3}x_{i3} + ... + a_{kk}x_{ik} + ... + a_{kn}x_{in} = \lambda_i x_{ik}$$

Divide above equation by x_{ik} and rearrange the terms.

$$\lambda_i - a_{kk} = a_{k1}\frac{x_{i1}}{x_{ik}} + a_{k2}\frac{x_{i2}}{x_{ik}} + ... + a_{k(k-1)}\frac{x_{i(k-1)}}{x_{ik}} + a_{k(k+1)}\frac{x_{i(k+1)}}{x_{ik}} + ... + a_{kn}\frac{x_{in}}{x_{ik}}$$

Since $|x_{ik}| = \max_r |x_{ir}|$, $\left|\frac{x_{ir}}{x_{ik}}\right| \le 1$, $\forall 1 \le r \le n$.

and $$|\lambda_i - a_{kk}| \le |a_{k1}| + |a_{k2}| + ... + |a_{k(k+1)}| + |a_{k(k-1)}| + ... + |a_{kn}|$$

Thus $$|\lambda_i - a_{kk}| \le P_k$$

Therefore all the eigenvalues of A lie inside or on the union of the above circles.

Since A and A^T have same eigenvalues, theorem holds for column sum also i.e.

$$|\lambda_i - a_{kk}| \le |a_{1k}| + |a_{2k}| + ... + |a_{(k-1)k}| + |a_{(k+1)k}| + ... + |a_{nk}|, \; k = 1, 2, 3, ..., n.$$

The bounds obtained for rows and columns are independent. Hence all the eigen values of A must lie in the intersection of these bounds. These circles are called the Gerschgorin circles and the bounds are called the Gerschgorin bounds.

Example 2.4.1 : Estimate the eigenvalues of the matrix

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 1 & 1 & 1 \\ 1 & 3 & -1 \end{bmatrix}$$

using the Gerschgorin bounds and Brauer theorem.

Answer : By Gerschgorin theorem corresponding to row we have

$$|\lambda| \le \max\{|1|+|2|+|-1|, |1|+|1|+|1|, |1|+|3|+|-1|\}$$

$$\le \max\{4, 3, 5\} = 5$$

i.e. $|\lambda| \le 5$

Similarly by considering column sum we get,

$$|\lambda| \le \max\{|1|+|1|+|1|, |2|+|1|+|3|, |-1|+|1|+|-1|\}$$

$$\le \max\{3, 6, 3\} = 6$$

By Brauer's theorem every eigenvalue of A lies inside or on the boundary of atleast one of the circles $|\lambda - a_{kk}| \le A_k$.

Corresponding to rows we have

$$\therefore |\lambda - 1| \le 3,\ |\lambda - 1| \le 2,\ |\lambda + 1| \le 4$$

Corresponding to columns we have

$$|\lambda - 1| \le 2,\ |\lambda - 1| \le 5,\ |\lambda + 1| \le 2$$

The union corresponding to row sum gives

$$\{|\lambda - 1| \le 3 \cup |\lambda - 1| \le 2 \cup |\lambda + 1| \le 4\} = \{|\lambda - 1| \le 3 \cup |\lambda + 1| \le 4\}$$

The union corresponding to column sum gives

$$\{|\lambda - 1| \le 2 \cup |\lambda - 1| \le 5 \cup |\lambda + 1| \le 2\} = \{|\lambda - 1| \le 5 \cup |\lambda + 1| \le 2\}$$

Thus the required region is given by

$$\{|\lambda| \le 5\} \cap \{|\lambda| \le 6\} \cap \{|\lambda - 1| \le 3 \cup |\lambda + 1| \le 4\} \cap \{|\lambda - 1| \le 5, |\lambda + 1| \le 2\}$$

Example 2.4.2 : Estimate the eigen value region of the matrix

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

By Gerschgorin theorem $|\lambda| \le \max\{7,9,9\} = 9$

Corresponding to column we have $|\lambda| \le \max\{7,9,7\} = 9$

Since matrix A is symmetric all eigenvalues are real and $|\lambda| \le 9$ gives the interval [–9, 9]. Thus all eigenvalues lie in the interval [–9, 9].

By Brauer theorem the eigenvalues lie in the region $\{|\lambda - 3| \le 4 \cup |\lambda - 5| \le 4 \cup |\lambda - 3| \le 4\}$ corresponding to row sum and corresponding to coulmn sum we have

$$\{|\lambda - 3| \le 4 \cup |\lambda - 5| \le 4 \cup |\lambda - 3| \le 4\}$$

Thus we have the region

$$\{|\lambda - 3| \le 4 \cup |\lambda - 5| \le 4 \cup |\lambda - 3| \le 4\}$$

$$= \{-4 \le \lambda - 3 \le 4 \cup -4 \le \lambda - 5 \le 4\}$$

$$\{-1 \le \lambda \le 7 \cup 1 \le \lambda \le 9\} = [-1, 9]$$

Thus all roots lie in the interval $[-1, 9]$.

2.5 Jacobi Method for Symmetric Matrices

For real symmetric matrix all eigenvalues are real and there exist a real orthogonal matrix S such that $S^{-1}AS$ is a diagonal matrix D. The diagonal entries of D are all eigenvalues of matrix A. The diagonalization is achieved by applying series of orthogonal transformations.

Computational Procedure

Let $|a_{ik}| = \max\{|a_{ij}| : i \ne j, i, j = 1,2,3,...,n\}$. Consider 2 × 2 matrix formed by the intersection of i^{th} & k^{th} row and i^{th} & k^{th} column. Then we get a matrix

$$A_1 = \begin{bmatrix} a_{ii} & a_{ik} \\ a_{ik} & a_{kk} \end{bmatrix} \qquad (a_{ik} = a_{ki} \text{ since A is symmetric matrix})$$

Choose $S_1^* = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ and find θ such that $S_1^{*-1} A_1 S_1^*$ is diagonal matrix.

$$S_1^{*-1} A_1 S_1^* = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} a_{ii} & a_{ik} \\ a_{ik} & a_{kk} \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$= \begin{bmatrix} a_{ii}\cos^2\theta + a_{ik}\sin 2\theta + a_{kk}\sin^2\theta & (a_{kk} - a_{ii})\dfrac{\sin 2\theta}{2} + a_{ik}\cos 2\theta \\ (a_{kk} - a_{ii})\dfrac{\sin 2\theta}{2} + a_{ik}\cos 2\theta & a_{ii}\sin^2\theta + a_{kk}\cos^2\theta - a_{ik}\sin 2\theta \end{bmatrix}$$

Now we choose θ such that the off diagonal entry of matrix $S_1^{*-1} A_1 S_1^*$ becomes zero so that $S_1^{*-1} A_1 S_1^*$ becomes a diagonal matrix. Thus we choose θ such that

$$(a_{kk} - a_{ii})\frac{\sin 2\theta}{2} + a_{ik}\cos 2\theta = 0$$

$$\tan 2\theta = \frac{-2a_{ik}}{a_{kk} - a_{ii}} = \frac{2a_{ik}}{a_{ii} - a_{kk}}$$

This equation produces four values of θ. We choose θ between $-\frac{\pi}{4}$ and $\frac{\pi}{4}$ and we get,

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2a_{ik}}{a_{ii} - a_{kk}}\right) \qquad \text{if } a_{ii} \neq a_{kk}$$

$$= \frac{\pi}{4} \qquad \text{if } a_{ii} = a_{kk},\ a_{ik} > 0$$

$$= -\frac{\pi}{4} \qquad \text{if } a_{ii} = a_{kk},\ a_{ik} < 0$$

With this choice of θ construct n $\times$ n orthogonal matrix S_1 as follows. Write $\cos\theta$, $-\sin\theta$, $\sin\theta$, $\cos\theta$ at (i,i), (i,k), (k,i), (k,k) positions of matrix S_1 respectively. Write remaining diagonal entries to be 1 and rest of the offdiagonal entries 0. Thus we get a matrix S_1 as

$$S_1 = \begin{array}{c} \\ \\ i^{th} \\ \\ j^{th} \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} & & i^{th} & & j^{th} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cos\theta & 0 & -\sin\theta & \cdots \\ 0 & 0 & \cdots & 1 & \cdots \\ \cdots & \sin\theta & \cdots & \cos\theta & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \cdots & 1 \end{bmatrix} \end{array}$$

Define $B_1 = S_1^{-1} A S_1$

In $B_1, (i,k)$ and (k,i) entries are zero. Repeat procedure for B_1. We get S_2. Define

$$B_2 = S_2^{-1} B_1 S_2 = S_2^{-1} S_1^{-1} A S_1 S_2$$

After making r transformations we get,

$$B_r = S_r^{-1} S_{r-1}^{-1} \ldots S_2^{-1} S_1^{-1} A S_1 S_2 S_3 \ldots S_r$$

$$= \left(S_1 S_2 S_3 \ldots S_r\right)^{-1} A \left(S_1 S_2 S_3 \ldots S_r\right)$$

$$= S^{-1} A S$$

where $S = S_1 S_2 S_3 \ldots S_r$

As $r \to \infty$, B_r approaches a diagonal matrix with the eigenvalues on the leading diagonal. This procedure is called Jacobi method.

The convergence to a diagonal matrix takes place even if the maximum of offdiagonal elements are not selected and we make any offdiagonal entry zero. This modification is called the special cyclic Jacobi method. In this method there is no search for maximum offdiagonal entry.

2.6 Householder's Method for Symmetric Matrices

In Jacobi method symmetric matrix is converted into a diagonal matrix through similarity transformation. In this method a symmetric matrix A is reduced to the tridiagonal form by orthogonal transformations. The orthogonal transformations are of the form

$$P = I - 2\bar{w}\bar{w}^T \qquad \ldots. (2.6.1)$$

where $\bar{w}$ is a column vector such that $\bar{w}^T \bar{w} = 1$.

Observe that P is symmetric and orthogonal.

$$P^T = \left(I - 2\bar{w}\bar{w}^T\right)^T = I^T - 2\left(\bar{w}\bar{w}^T\right)^T = I - 2\bar{w}\bar{w}^T = P$$

$$P^T P = \left(I - 2\bar{w}\bar{w}^T\right)\left(I - 2\bar{w}\bar{w}^T\right)$$

$$= I - 2\bar{w}\bar{w}^T - 2\bar{w}\bar{w}^T + 4\bar{w}\bar{w}^T\bar{w}\bar{w}^T = I \qquad \left(\because \bar{w}^T\bar{w} = 1\right)$$

At the first transformation we find x_i's such that we get zero in the position (1, 3), (1, 4),....., (1, n) and zero in the corresponding positions in the first column i.e. (3, 1), (4, 1), (5, 1) .... (n, 1). Thus one transformation $P_2^{-1}AP_2 = A_2$, bring (n – 2) zeros in the first row and first column. In the second transformation $P_3^{-1}A_2P_3$, we get (n – 3) zeros in the second column and second row namely (2, 4), (2, 5), (2, 6) .... (2, n) and (4, 2), (5, 2) ..... (n, 2) positions. The final matrix is tridiagonal. The tridiagonalization is completed with exactly (n – 2) Householder transformation.

The matrix P_r is constructed as follows.

The vector $\bar{w}_r$ is constructed with the first (r–1) components as zeros.

$$\bar{w}_r{}^T = \left[0, 0, 0, \ldots 0, x_r, x_{r+1}, \ldots, x_n\right]$$

Since $\bar{w}_r{}^T\bar{w}_r = 1$, $x_r{}^2 + x_{r+1}{}^2 + x_{r+1}{}^2 + \ldots + x_n{}^2 = 1$

with this choice of $\bar{w}_r$, $P_r = I - \bar{w}_r\bar{w}_r{}^T$.

Let us illustrate this procedure for 3 $\times$ 3 and 4 $\times$ 4 matrics.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

$$\bar{w}_2{}^T = \left[0, \bar{w}_2, \bar{w}_3\right], \quad \bar{w}_2{}^T\bar{w}_2 = 1 \Rightarrow w_2{}^2 + w_3{}^2 = 1$$

$$P_2 = I - \bar{w}_2\bar{w}_2{}^T$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 2\begin{bmatrix} o \\ w_2 \\ w_3 \end{bmatrix}\left(0, w_2, w_3\right)$$

$$= \begin{bmatrix} 1-0 & 0-0 & 0-0 \\ 0-0 & 1-2w_2{}^2 & -2w_2w_3 \\ 0-0 & -2w_2w_3 & 1-2w_3{}^2 \end{bmatrix}$$

$P_2^T P_2 = I$ therefore P_2 is orthogonal and $P_2^T = P_2^{-1}$.

$$AP_2 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-2w_2^2 & -2w_2w_3 \\ 0 & -2w_2w_3 & 1-2w_3^2 \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{12}\left(1-2w_2^2\right)-2a_{13}w_2w_3 & -2w_2w_3a_{12}+a_{13}\left(1-2w_3^2\right) \\ a_{12} & a_{22}\left(1-2w_2^2\right)-2a_{23}w_2w_3 & -2w_2w_3a_{22}+a_{23}\left(1-2w_3^2\right) \\ a_{13} & a_{23}\left(1-2w_2^2\right)-2a_{33}w_2w_3 & -2w_2w_3a_{23}+a_{33}\left(1-2w_3^2\right) \end{bmatrix}$$

$$\therefore P_2^T AP_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-2w_2^2 & -2w_2w_3 \\ 0 & -2w_2w_3 & 1-w_3^2 \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12}\left(1-2w_2^2\right)-2a_{13}w_2w_3 & -2w_2w_3a_{12}+a_{13}\left(1-2w_3^2\right) \\ a_{12} & a_{22}\left(1-2w_2^2\right)-2a_{23}w_2w_3 & -2w_2w_3a_{22}+a_{23}\left(1-2w_3^2\right) \\ a_{13} & a_{23}\left(1-2w_2^2\right)-2a_{33}w_2w_3 & -2w_2w_3a_{23}+a_{33}\left(1-2w_3^2\right) \end{bmatrix}$$

$P_2^T AP_2$ is tridiagonal if (1, 3) entry of the matrix $P_2^T AP_2$ is zero. But (1, 3) entry of $P_2^T AP_2$ is

$$-2w_2w_3a_{12}+a_{13}\left(1-2w_3^2\right)=0$$

$$a_{13}-2w_3\left(a_{12}w_2+a_{13}w_3\right)=0$$

i.e. $a_{13}-2w_3r=0$ where $r=a_{12}w_2+a_{13}w_3$

(1, 2) entry of $P_2^T AP_2$ denoted by $a_{12}^{'}$ is

$$a_{12}^{'} = a_{12}\left(1-2w_2^2\right)-a_{13}w_2w_3$$

$$= a_{12}-2w_2r$$

$$\therefore a_{12}^{'2} = \left(a_{12}-2w_2r\right)^2+\left(a_{13}-2w_3r\right)^2 \qquad \left[a_{13}-2w_3r=0\right]$$

$$= a_{12}^2+a_{13}^2+4r^2\left(w_2^2+w_3^2\right)-4r\left(w_2a_{12}+w_3a_{13}\right)$$

$$= a_{12}^2+a_{13}^2+4r^2-4r^2$$

Thus $a_{12}^{'2} = (a_{12} - 2w_2 r)^2 + (a_{13} - 2w_3 r)^2$

$= (a_{12} - 2w_2 r)^2 + 0$

$a_{12}^{'2} = a_{12}^2 + a_{13}^2$

Therefore, $a_{12}^{'} = \pm\sqrt{a_{12}^2 + a_{13}^2} = a_{12} - 2w_2 r = \pm S$ (say)

Now we have two equations

$$a_{13} - 2rw_3 = 0 \quad \text{... (i)}$$

$$a_{12} - 2rw_2 = \pm S \quad \text{... (ii)}$$

Multiply equation (i) by w_3 and (ii) by w_2 and add.

Since $a_{12}w_2 + a_{13}w_3 = r$, we have

$r - 2r = \pm Sw_2$ i.e. $r = \pm Sw_2$

Now from equation (ii) we have

$$a_{12} \pm 2Sw_2^2 = \pm S$$

Thus $w_2^2 = \pm\dfrac{\pm S - a_{12}}{2S}$

$$w_2^2 = \frac{1}{2}\left[1 \mp \frac{a_{12}}{S}\right], \quad w_3 = \mp\frac{1}{2}\frac{a_{13}}{2w_2\sqrt{a_{12}^2 + a_{13}^2}}$$

In computing w_3 from w_2, we choose w_2 as large as possible.

We demonstrate the reduction in the following example.

Example 2.6.1 : Reduce the matrix

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

to tridiagonal form.

Answer : Here $\sqrt{a_{12}^2 + a_{13}^2} = \sqrt{9+16} = \pm 5$.

$$w_2^2 = \frac{1}{2}\left[1 \mp \frac{a_{12}}{S}\right] = \frac{1}{2}\left[1 + \frac{3}{5}\right]$$ (w_2 is max. for S = 5 since a_{12} is +ve choose S +ve)

$$\therefore w_2^{\,2} = \frac{4}{5} \qquad \text{i.e. } w_2 = \frac{2}{\sqrt{5}}$$

$$w_3 = \frac{a_{13}}{2w_2\sqrt{a_{12}^{\,2} + a_{13}^{\,2}}} = \frac{4}{2 \cdot \frac{2}{\sqrt{5}} \cdot 5} = \frac{1}{\sqrt{5}}$$

$$\therefore \bar{w}_2^{\,T} = \left[0, \frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right]$$

and $P_2 = I - 2\bar{w}_2\bar{w}_2^{\,T}$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix}$$

$$A_2 = P_2^T A P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix} \begin{bmatrix} 1 & 3 & 4 \\ 3 & 2 & -1 \\ 4 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix} \begin{bmatrix} 1 & -5 & 0 \\ 3 & -\frac{2}{5} & -\frac{11}{5} \\ 4 & -\frac{1}{5} & \frac{7}{5} \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 1 & -5 & 0 \\ -5 & +\frac{2}{5} & \frac{1}{5} \\ 0 & \frac{1}{5} & \frac{13}{5} \end{bmatrix}$$

To illustrate the procedure for 4×4 matrix let

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{12} & a_{22} & a_{23} & a_{24} \\ a_{13} & a_{23} & a_{33} & a_{34} \\ a_{14} & a_{24} & a_{34} & a_{44} \end{bmatrix}$$

Since the transformations $P_r^T AP_r$ are orthogonal, the sum of squares of the elements in any row are invariant. We will use the fact that sum of squares of elements in any row of matrix A is same as the sum of squares of elements in corresponding row of matrix $P_r^T AP_r$.

Choose $\bar{w}_2^T = [0, x_2, x_3, x_4]$ and $\bar{w}_2^T \bar{w}_2 = x_2^2 + x_3^2 + x_4^2 = 1$.

At the first transformation we find x_2, x_3, x_4 such that we get zero in the position (1, 3), (1, 4) and (3, 1), (4, 1).

In the matrix P_2 first row is a unit vector and therefore the position (1, 3), (1, 4) have zero entry if the corresponding elements in AP_2 are zero. The first row of AP_2 is given by the following product.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{12} & a_{22} & a_{23} & a_{24} \\ a_{13} & a_{23} & a_{33} & a_{34} \\ a_{14} & a_{24} & a_{34} & a_{44} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2x_2^2 & -2x_2x_3 & -2x_2x_4 \\ 0 & -2x_2x_3 & 1-2x_3^2 & -2x_3x_4 \\ 0 & -2x_2x_4 & -2x_3x_4 & 1-2x_4^2 \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12}(1-2x_2^2) - 2a_{13}x_2x_3 & -2a_{14}x_2x_4 - 2a_{12}x_2x_3 + a_{13}(1-2x_3^2) - 2a_{14}x_3x_4 \\ & & -2a_{12}x_2x_4 - 2a_{13}x_3x_4 + a_{14}(1-2x_4^2) \end{bmatrix}$$

$$= [a_{11}a_{12} \quad -2p_1x_2, \quad a_{13} - 2p_1x_3, \quad a_{14} - 2p_1x_4]$$

where $p_1 = a_{12}x_2 + a_{13}x_3 + a_{14}x_4$

Now we need to find x_2, x_3, x_4 such that $a_{13} - 2p_1x_3 = 0$ and $a_{14} - 2p_1x_4 = 0$. So that (1, 3) and (1, 4) position of P_2AP_2 will become zero. Since the sum of the squares of the elements in any row is invariant underthe orthogonal transformation we have,

$$a_{11}^2 + a_{12}^2 + a_{13}^2 + a_{14}^2 = a_{11}^2 + (a_{12} - 2p_1x_2)^2 + (a_{13} - 2p_1x_3)^2 + (a_{14} - 2p_1x_4)^2$$

$$= a_{11}^2 + (a_{12} - 2p_1x_2)^2 + 0 + 0$$

$$\therefore a_{12} - 2p_1x_2 = \pm\sqrt{a_{12}^2 + a_{13}^2 + a_{14}^2} = \pm S_1 \quad \text{(say)}$$

Thus we have three equations

$$a_{13} - 2p_1x_3 = 0 \qquad \text{.... (2.6.1)}$$

$$a_{14} - 2p_1 x_4 = 0 \qquad \text{.... (2.6.2)}$$

$$a_{12} - 2p_1 x_2 = \pm S_1 \qquad \text{... (2.6.3)}$$

Since matrix A is given matrix, S_1 is known quantity. Multiply (2.6.3) by x_2, (2.6.1) by x_3 and (2.6.2) by x_4 and add. We get,

$$p_1 - 2p_1\left({x_2}^2 + {x_3}^2 + {x_4}^2\right) = \pm S_1 x_2$$

But ${x_2}^2 + {x_3}^2 + {x_4}^2 = 1$ and therefore

$$p_1 = \mp S_1 x_2$$

Now if we put this value of p_1 in equation (2.6.3) then equation becomes a quadratic equation in x_2 and can be solved for x_2.

$$a_{12} - 2\left(\mp S_1 x_2\right) x_2 = \pm S_1$$

$${x_2}^2 = \frac{1}{2}\left(1 \mp \frac{a_{12}}{S_1}\right) \qquad \text{.... (2.6.4)}$$

From equation (2.6.1) and (2.6.2) we get

$$x_3 = \frac{a_{13}}{2p_1} = \frac{a_{13}}{\mp 2S_1 x_2} \quad \text{and} \quad x_4 = \frac{a_{14}}{2p_1} = \frac{a_{14}}{\mp 2S_1 x_2} \qquad \text{.... (2.6.5)}$$

From equation (2.6.4) we observe that x_2 and therefore p_1 posses two values. Since x_3 and x_4 contains x_2 in the denominator, we choose the large root by x_2. This is done by taking suitable sign in equation (2.6.4).

$${x_2}^2 = \frac{1}{2}\left[1 + \frac{a_{12}\ \text{sign}\left(a_{12}\right)}{S_1}\right]$$

$$x_3 = \frac{a_{13}\ \text{sign}\left(a_{12}\right)}{2S_1 x_2}, \quad x_4 = \frac{a_{14}\ \text{sign}\left(a_{12}\right)}{2S_1 x_2}$$

where sign $\left(a_{12}\right)$ is sign function which takes value – 1 if $a_{12} < 0$ and value 1 if $a_{12} > 0$.

Thus the transformation $P_2 A P_2$ produces zero value in (1, 3), (1, 4) and therefore (3, 1), (4, 1) positions. One more transformation discussed for 3×3 matrix produces zeros in (2, 4) and (4, 2) positions. The resulting matrix will be a tridiagonal matrix.

Example 2.6.21 : Use the Householder's method to reduce the give matrix A into the tridiagonal form.

$$A=\begin{bmatrix} 4 & -1 & -2 & 2 \\ -1 & 4 & -1 & -2 \\ -2 & -1 & 4 & -1 \\ 2 & -2 & -1 & 4 \end{bmatrix}$$

Answer :

First iteration : Let $\bar{w}_2 = [0, x_2, x_3, x_4]^T$

$$S_1 = \sqrt{(-1)^2 + (-2)^2 + (2)^2} = 3$$

$$x_2{}^2 = \frac{1}{2}\left[1 + \frac{(-1)(-1)}{3}\right] = \frac{2}{3} \qquad \therefore x_2 = \sqrt{\frac{2}{3}}$$

$$x_3 = \frac{(-2)(-1)}{2(3)\sqrt{2/3}} = \frac{1}{\sqrt{6}}, \; x_4 = \frac{2(-1)}{2(3)\sqrt{2/3}} = -\frac{1}{\sqrt{6}}$$

$$P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2x_2{}^2 & -2x_2x_3 & -2x_2x_4 \\ 0 & -2x_2x_3 & 1-2x_3{}^2 & -2x_3x_4 \\ 0 & -2x_2x_4 & -2x_3x_4 & 1-2x_4{}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

$$A_2 = P_2 A P_2$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 4 & -1 & -2 & 2 \\ -1 & 4 & -1 & -2 \\ -2 & -1 & 4 & -1 \\ 2 & -2 & -1 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 4 & 3 & 0 & 0 \\ -1 & -2 & -4 & 1 \\ -2 & -3 & 3 & 0 \\ 2 & 4 & 2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 3 & 0 & 0 \\ 3 & \frac{16}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{16}{3} & -\frac{1}{3} \\ 0 & \frac{1}{3} & -\frac{1}{3} & \frac{4}{3} \end{bmatrix}$$

Second Iteration : $P_3 = [0, 0, x_3, x_4]^T$; $x_3^2 + x_4^2 = 1$

$$S_1 = \sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2} = \frac{\sqrt{5}}{3}; \; x_3^2 = \frac{1}{2}\left(1 \mp \frac{2/3}{\sqrt{5}/3}\right) = \frac{1}{2}\left(1 + \frac{2}{\sqrt{5}}\right)$$

$$x_3^2 = \frac{1}{2}\left(\frac{\sqrt{5}+2}{\sqrt{5}}\right) \text{ and } x_4^2 = 1 - x_3^2 = 1 - \frac{\sqrt{5}+2}{2\sqrt{5}} = \frac{\sqrt{5}-2}{2\sqrt{5}}$$

Suppose $x_3^2 = a$ then $x_4^2 = \frac{1}{20a}$ and we have

$$P_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-2a & -\frac{1}{\sqrt{5}} \\ 0 & 0 & -\frac{1}{\sqrt{5}} & 1-\frac{1}{10a} \end{bmatrix}$$

$$A_3 = P_3 A_2 P_3$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-2a & -\frac{1}{\sqrt{5}} \\ 0 & 0 & -\frac{1}{\sqrt{5}} & 1-\frac{1}{10a} \end{bmatrix} \begin{bmatrix} 4 & 3 & 0 & 0 \\ 3 & \frac{16}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{16}{3} & -\frac{1}{3} \\ 0 & \frac{1}{3} & -\frac{1}{3} & \frac{4}{3} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-2a & -\frac{1}{\sqrt{5}} \\ 0 & 0 & -\frac{1}{\sqrt{5}} & 1-\frac{1}{10a} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-2a & -\frac{1}{\sqrt{5}} \\ 0 & 0 & -\frac{1}{\sqrt{5}} & 1-\frac{1}{10a} \end{bmatrix} \begin{bmatrix} 4 & 3 & 0 & 0 \\ 3 & \frac{16}{3} & \frac{2\sqrt{5}(1-2a)-1}{3\sqrt{5}} & -\frac{2}{3\sqrt{5}}+\frac{1}{3}\left(1-\frac{1}{10a}\right) \\ 0 & \frac{2}{3} & \frac{16}{3}(1-2a)+\frac{1}{3\sqrt{5}} & -\frac{16}{2\sqrt{5}}-\frac{1}{3}\left(1-\frac{1}{10a}\right) \\ 0 & \frac{1}{3} & -\frac{1}{3}(1-2a)-\frac{4}{3\sqrt{5}} & \frac{1}{3\sqrt{5}}+\frac{4}{3}\left(1-\frac{1}{10a}\right) \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 3 & 0 & 0 \\ 3 & \frac{16}{3} & -\frac{5}{3\sqrt{5}} & 0 \\ 0 & -\frac{5}{3\sqrt{5}} & \frac{16}{3} & \frac{9}{5} \\ 0 & 0 & \frac{9}{5} & \frac{12}{5} \end{bmatrix}$$

2.7 Power Method

In section 2.2.3 we have seen that convergence of iterative method depends upon the spectral radius of iteration matrix H. Spectral radius of a matrix is the largest eigen value in modulus. Therefore, it is necessary to calculate the largest eigen value (in magnitude). Power method is normally used to determine the largest eigen value (in magnitude) of the given matrix (spectral radius of a matrix).

Suppose we want to determine the largest eigen value of a square matrix A of order n. Let λ_1, λ_2, λ_3, ..., λ_n be the distinct eigen values of matrix A arranged in decreasing order (in magnitude).

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > ..... > |\lambda_n| \qquad .... (2.7.1)$$

and $\bar{v}_1$, $\bar{v}_2$, $\bar{v}_3$, ..., $\bar{v}_n$ be the corresponding linearly independent eigen vectors.

$$A\bar{v}_i = \lambda_i \bar{v}_i, \quad i = 1, 2, 3, ..., n \qquad ..... (2.7.2)$$

Since eigen vectors are linearly independent, any eigen vector $\bar{v}$ in the eigen space, spanned by the eigen vectors $\bar{v}_1, \bar{v}_2, \bar{v}_3, ..., \bar{v}_n$, can be written as

$$\bar{v} = c_1\bar{v}_1 + c_2\bar{v}_2 + ... + c_n\bar{v}_n \qquad .... (2.7.3)$$

[Since $\bar{v}_1, \bar{v}_2, \bar{v}_3, ..., \bar{v}_n$ are linearly independent vectors, this set is a basis of eigen space]

Premultiplying equation (2.7.3) by A we get,

$$A\bar{v} = A\left(c_1\bar{v}_1 + c_2\bar{v}_2 + ... + c_n\bar{v}_n\right)$$

$$= c_1A\bar{v}_1 + c_2A\bar{v}_2 + c_3A\bar{v}_3 + ... + c_nA\bar{v}_n$$

From equation (2.7.2) we have

$$A\bar{v} = c_1\lambda_1\bar{v}_1 + c_2\lambda_2\bar{v}_2 + c_3\lambda_3\bar{v}_3 + ... + c_n\lambda_n\bar{v}_n$$

$$= \lambda_1\left[c_1\bar{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)c_2\bar{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)c_3\bar{v}_3 + ... + \left(\frac{\lambda_n}{\lambda_1}\right)c_n\bar{v}_n\right]$$

$$A(A\bar{v}) = A\left\{\lambda_1\left[c_1\bar{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)c_2\bar{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)c_3\bar{v}_3 + ... + \left(\frac{\lambda_n}{\lambda_1}\right)c_n\bar{v}_n\right]\right\}$$

$$= \lambda_1\left[c_1A\bar{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)c_2A\bar{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)c_3A\bar{v}_3 + ... + \left(\frac{\lambda_n}{\lambda_1}\right)c_nA\bar{v}_n\right]$$

$$= \lambda_1\left[c_1\lambda_1\bar{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)c_2\lambda_2\bar{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)c_3\lambda_3\bar{v}_3 + ... + \left(\frac{\lambda_n}{\lambda_1}\right)c_n\lambda_n\bar{v}_n\right] \quad \left(\because A\bar{v}_i = \lambda_i\bar{v}_i\right)$$

$$= \lambda_1^2\left[c_1\bar{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^2 c_2\bar{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)^2 c_3\bar{v}_3 + ... + \left(\frac{\lambda_n}{\lambda_1}\right)^2 c_n\bar{v}_n\right]$$

Repeatative premultiplication of A gives,

$$A^k\bar{v} = \lambda_1^k\left[c_1\bar{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^k c_2\bar{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)^k c_3\bar{v}_3 + ... + \left(\frac{\lambda_n}{\lambda_1}\right)^k c_n\bar{v}_n\right] \qquad .... (2.7.4)$$

$$A^{k+1}\bar{v} = \lambda_1^{k+1}\left[c_1\bar{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^{k+1} c_2\bar{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)^{k+1} c_3\bar{v}_3 + ... + \left(\frac{\lambda_n}{\lambda_1}\right)^{k+1} c_n\bar{v}_n\right] .... (2.7.5)$$

As k becomes very large, right hand side of equation (2.7.4) and (2.7.5) will be dominated by ${\lambda_1}^k c_1 \overline{v}_1$ and ${\lambda_1}^{k+1} c_1 \overline{v}_1$ respectively.

[Since $|\lambda_1| > |\lambda_i|$, $i =$ 2, 3, 4, ... n, $\left(\frac{\lambda_i}{\lambda_1}\right)^k \to 0$ as $k \to \infty$]

Thus $A^k \overline{v} \cong {\lambda_1}^k c_1 \overline{v}_1$

and $A^{k+1} \overline{v} \cong {\lambda_1}^{k+1} c_1 \overline{v}_1$

Now the eigen value λ_1 is obtained as the ratio of the corresponding components of $A^{k+1}\overline{v}$ and $A^k \overline{v}$.

$$\lambda_1 \cong \frac{{\lambda_1}^{k+1} c_1 (\overline{v}_1)_r}{{\lambda_1}^k c_1 (\overline{v}_1)_r} \quad \text{and}$$

$$\lambda_1 = \underset{k\to\infty}{Lt} \frac{(A^{k+1}\overline{v})_r}{(A^k \overline{v})_r}, \quad r = 1, 2, 3, ..., n$$

where the suffix r denotes the rth component of the vector.

The iteration is stoped when the magnitudes of the differences of the ratios are less than the given error tolerance.

In order to keep roundoff error in control, we normalize the vector before premultiplying by A.

For computation purpose we follow the following procedure.

Let $\overline{v}_0$ be a non-zero arbitrary initial vector. (we choose $\overline{v}_0$ in such a way that ${\overline{v}_0}^T \overline{v}_1 \neq 0$)

Define $\overline{y}_{k+1} = A\overline{v}_k$

Suppose m_{k+1} is the largest element in magnitude of $\overline{y}_{k+1}$,

Define $\overline{v}_{k+1} = \frac{\overline{y}_{k+1}}{m_{k+1}}$

Calculate $\frac{(\overline{y}_{k+1})_r}{(\overline{v}_k)_r}$, $r = 1, 2, 3, ..., n$

If all the ratios are less than the given error tolerance

i.e. if $\left|\frac{(\overline{y}_{k+1})_r}{(\overline{v}_k)_r} - \frac{(\overline{y}_{k+1})_s}{(\overline{v}_k)_s}\right| < \varepsilon \qquad \forall r, s = 1, 2, ..., n$

Then $\dfrac{(\bar{y}_{k+1})_r}{(\bar{v}_k)_r} \to \lambda_1$ as $k \to \infty$.

The vector $\bar{v}_{k+1}$ is the required eigen vector.

The initial vector $\bar{v}_0$ is usually chosen as a vector with all components equal to unity if no suitable approximation is available.

ILLUSTRATIVE EXAMPLES

1. Solve the system of equations.

$4x_1 + x_2 + x_3 = 2$

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

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

Using Jacobi iteration method. Take the approximation as $\bar{x}^0 = [0.5, -0.5, -0.5]^T$ and perform two iterations.

Answer : We have the system of equation $A\bar{x} = \bar{b}$ where

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 1 & 5 & 2 \\ 1 & 2 & 3 \end{bmatrix}, \quad \bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \text{and} \quad \bar{b} = \begin{bmatrix} 2 \\ -6 \\ -4 \end{bmatrix}$$

For Jacobi iteration method,

$$\bar{x}^{(k+1)} = -D^{-1}(L+U))\bar{x}^{(k)} + D^{-1}b$$

$$\because A = \begin{bmatrix} 4 & 1 & 1 \\ 1 & 5 & 2 \\ 1 & 2 & 3 \end{bmatrix}, \; L = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 2 & 0 \end{bmatrix}, \; U = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } D = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

$$\therefore D^{-1}(L+U) = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1/4 & 1/4 \\ 1/5 & 0 & 2/5 \\ 1/3 & 2/3 & 0 \end{bmatrix}$$

$$D^{-1}b = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 2 \\ -6 \\ -4 \end{bmatrix} = \begin{bmatrix} 1/2 \\ -6/5 \\ -4/3 \end{bmatrix}$$

Thus Jacobi iteration method becomes

$$\bar{x}^{(k+1)} = \begin{bmatrix} 0 & -1/4 & -1/4 \\ -1/5 & 0 & -2/5 \\ -1/3 & -2/3 & 0 \end{bmatrix} \bar{x}^{(k)} + \begin{bmatrix} 1/2 \\ -6/5 \\ -4/3 \end{bmatrix}$$

$$\bar{x}^{0} = \begin{bmatrix} 0.5 \\ -0.5 \\ -0.5 \end{bmatrix}$$

$$\bar{x}^{(1)} = \begin{bmatrix} 0 & -1/4 & -1/4 \\ -1/5 & 0 & -2/5 \\ -1/3 & -2/3 & 0 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.5 \\ -0.5 \end{bmatrix} + \begin{bmatrix} 1/2 \\ -6/5 \\ -4/3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.25 \\ 0.1 \\ 0.16666 \end{bmatrix} + \begin{bmatrix} 1/2 \\ -6/5 \\ -4/3 \end{bmatrix} = \begin{bmatrix} 0.75 \\ -1.1 \\ -1.16667 \end{bmatrix}$$

$$\bar{x}^{(2)} = \begin{bmatrix} 0 & -1/4 & -1/4 \\ -1/5 & 0 & -2/5 \\ -1/3 & -2/3 & 0 \end{bmatrix} \begin{bmatrix} 0.75 \\ -1.1 \\ -1.16667 \end{bmatrix} + \begin{bmatrix} 1/2 \\ -6/5 \\ -4/3 \end{bmatrix}$$

$$= \begin{bmatrix} 1.0666675 \\ -0.883332 \\ -0.85 \end{bmatrix}$$

2. For the following system of equations

$4x + y + 2z = 4$

$3x + 5y + z = 7$

$x + y + 3z = 3$

(a) Show that Jacobi iteration scheme converges.

(b) Obtain the Jacobi iteration scheme in matrix form.

(c) Starting with $\overline{x}^{(0)} = \overline{0}$, itrate two times.

Answer : We have the system of equations $A\overline{x} = \overline{b}$ where

$$A = \begin{bmatrix} 4 & 1 & 2 \\ 3 & 5 & 1 \\ 1 & 1 & 3 \end{bmatrix}, \quad \overline{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \overline{b} = \begin{bmatrix} 4 \\ 7 \\ 3 \end{bmatrix}$$

Jacobi iteration method is

$$\overline{x}^{(k+1)} = H_J \overline{x}^{(k)} + D^{-1} b$$

where, $H_J = -D^{-1}(L+U)$

Here $L = \begin{bmatrix} 0 & 0 & 0 \\ 3 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

Therefore, $H_J = -D^{-1}(L+U)$

$$= -\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 \\ 3 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$= -\begin{bmatrix} 0 & 1/4 & 1/2 \\ 3/5 & 0 & 1/5 \\ 1/3 & 1/3 & 0 \end{bmatrix}$$

$$D^{-1}b = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 4 \\ 7 \\ 3 \end{bmatrix}$$

Thus we have Jacobi iteration scheme.

$$\bar{x}^{(k+1)} = \begin{bmatrix} 0 & -1/4 & -1/2 \\ -3/5 & 0 & -1/5 \\ -1/3 & -1/3 & 0 \end{bmatrix} \bar{x}^{(k)} + \begin{bmatrix} 1 \\ 7/5 \\ 1 \end{bmatrix} \qquad \text{.... (i)}$$

To check the convergence of numerical scheme, let us analyze the eigenvalues of the matrix H_J.

Bounds on the eigenvalues of the matrix H_J are calculated by using Gerschgorin theorem.

By Grsungorin theorem

$$|\lambda| < \max\left\{\frac{3}{4}, \frac{4}{5}, \frac{2}{3}\right\} = \frac{4}{5}$$

$$|\lambda| < \max\left\{\frac{14}{15}, \frac{7}{12}, \frac{7}{10}\right\} < 1$$

Thus $|\lambda| < 1$ and $\rho(H_J) < 1$ therefore Jacobi iteration scheme is convergent scheme.

$$\bar{x}^{(k+1)} = \begin{bmatrix} 0 & -1/4 & -1/2 \\ -3/5 & 0 & -1/5 \\ -1/3 & -1/3 & 0 \end{bmatrix} \bar{x}^{(k)} + \begin{bmatrix} 1 \\ 7/5 \\ 1 \end{bmatrix}$$

$$\bar{x}^{(0)} = 0$$

$$\therefore \bar{x}^{(1)} = \begin{bmatrix} 1 \\ 7/5 \\ 1 \end{bmatrix}$$

$$\bar{x}^{(2)} = \begin{bmatrix} 0 & -1/4 & -1/2 \\ -3/5 & 0 & -1/5 \\ -1/3 & -1/3 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 7/5 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 7/5 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.15 \\ 0.6 \\ 0.2 \end{bmatrix}$$

3. For the following systemof equations.

$2x - y = 1$

$-x + 2y - z = 0$

$-y + 2z - w = 0$

$-z + 2w = 1$

Find the rate of convergence of Jacobi iteration method.

Answer : We have the system $A\bar{x} = b$ where

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}, \quad \bar{x} = \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix} \text{ and } \quad b = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

For Jacobi iteration method

$$H_J = -D^{-1}(L+U)$$

Here $L = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}$, $D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$, $U = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

Thus $H_J = -D^{-1}(L+U)$

$$= -\begin{bmatrix} 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

To determine the rate of convergence, let us calculate $\rho(H_J)$, consider $\det(H_J - \lambda I) = 0$.

$$\therefore \begin{vmatrix} -\lambda & 1/2 & 0 & 0 \\ 1/2 & -\lambda & 1/2 & 0 \\ 0 & 1/2 & -\lambda & 1/2 \\ 0 & 0 & 1/2 & -\lambda \end{vmatrix} = 0$$

$$\Rightarrow -\lambda \begin{vmatrix} -\lambda & 1/2 & 0 \\ 1/2 & -\lambda & 1/2 \\ 0 & 1/2 & -\lambda \end{vmatrix} - \frac{1}{2} \begin{vmatrix} 1/2 & 1/2 & 0 \\ 0 & -\lambda & 1/2 \\ 0 & 1/2 & -\lambda \end{vmatrix} = 0$$

$$\Rightarrow -\lambda \left\{ -\lambda \left(\lambda^2 - \frac{1}{4} \right) - \frac{1}{2} \left(-\frac{\lambda}{2} - 0 \right) + 0 \right\} - \frac{1}{2} \left\{ \frac{1}{2} \left(\lambda^2 - \frac{1}{4} \right) - \frac{1}{2}(0) \right\} = 0$$

$$\Rightarrow -\lambda \left\{ -\lambda^3 + \frac{1}{4}\lambda + \frac{1}{4}\lambda \right\} - \frac{1}{2} \left\{ \frac{1}{2}\lambda^2 - \frac{1}{8} \right\} = 0$$

$$\Rightarrow \lambda^4 - \frac{3}{4}\lambda^2 + \frac{1}{16} = 0$$

$$\lambda^2 = \frac{3/4 \pm \sqrt{\left(3/4\right)^2 - 4/16}}{2}$$

$$= \frac{3 \pm \sqrt{5}}{8}$$

$\lambda = \pm 0.654508497$, $\lambda = \pm 0.095491502$

$\therefore \rho(H_J) = 0.654508497$

The rate of convergence,

$$\gamma = -\log_{10} \rho(H_J)$$

$$= -\log(0.654508497)$$

$$= 0.1841$$

4. Solve the system of equations.

$2x_1 - x_2 = 7$

$-x_1 + 2x_2 - x_3 = 1$

$-x_2 + 2x_3 = 1$

Using Gauss-Seidel method. Take the initial approximation as $\bar{x}^{(0)} = \bar{0}$ and perform two iterations.

Answer : We have the system of equations $A\bar{x} = \bar{b}$ where

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \quad \bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} 7 \\ 1 \\ 1 \end{bmatrix}$$

By Gauss-Seidel method

$$\bar{x}^{(k+1)} = -(D+L)^{-1} U\, \bar{x}^{(k)} + (D+L)^{-1} b$$

Here $U = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}$ and $(D+L) = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 2 & 0 \\ 0 & -1 & 2 \end{bmatrix}$

$$(D+L)^{-1} = \frac{1}{8}\begin{bmatrix} 4 & 2 & 1 \\ 0 & 4 & 2 \\ 0 & 0 & 4 \end{bmatrix}^T$$

$$= \frac{1}{8}\begin{bmatrix} 4 & 0 & 0 \\ 2 & 4 & 0 \\ 1 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/4 & 1/2 & 0 \\ 1/8 & 1/4 & 1/2 \end{bmatrix}$$

$$(D+L)^{-1} U = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/4 & 1/2 & 0 \\ 1/8 & 1/4 & 1/2 \end{bmatrix}\begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1/2 & 0 \\ 0 & -1/4 & -1/2 \\ 0 & -1/8 & -1/4 \end{bmatrix}$$

$$(D+L)^{-1}b=\begin{bmatrix}1/2 & 0 & 0\\ 1/4 & 1/2 & 0\\ 1/8 & 1/4 & 1/2\end{bmatrix}\begin{bmatrix}7\\1\\1\end{bmatrix}=\begin{bmatrix}\frac{7}{2}\\ \frac{9}{4}\\ \frac{13}{8}\end{bmatrix}$$

Thus we have the iteration scheme

$$\bar{x}^{(k+1)}=\begin{bmatrix}0 & 1/2 & 0\\ 0 & 1/4 & 1/2\\ 0 & 1/8 & 1/4\end{bmatrix}\bar{x}^{(k)}+\begin{bmatrix}7/2\\ 9/4\\ 13/8\end{bmatrix}$$

$$\bar{x}^{0}=\begin{bmatrix}0\\0\\0\end{bmatrix}$$

$$\bar{x}^{(1)}=\begin{bmatrix}7/2\\ 9/4\\ 13/8\end{bmatrix}$$

$$\bar{x}^{(2)}=\begin{bmatrix}0 & 1/2 & 0\\ 0 & 1/4 & 1/2\\ 0 & 1/8 & 1/4\end{bmatrix}\begin{bmatrix}7/2\\ 9/4\\ 13/8\end{bmatrix}+\begin{bmatrix}7/2\\ 9/4\\ 13/8\end{bmatrix}$$

$$=\begin{bmatrix}4.625\\ 3.625\\ 2.3125\end{bmatrix}$$

5. Determine the convergence factor for the Jacobi and Gauss Seidel methods for the system.

$$\begin{bmatrix}4 & 0 & 2\\ 0 & 5 & 2\\ 5 & 4 & 10\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}4\\-3\\2\end{bmatrix}$$

Answer :

(i) $H_J = -D^{-1}(L+U)$

$$= -\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/10 \end{bmatrix}\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 2 \\ 5 & 4 & 0 \end{bmatrix}$$

$$= -\begin{bmatrix} 0 & 0 & 1/2 \\ 0 & 0 & 2/5 \\ 1/2 & 2/5 & 0 \end{bmatrix}$$

$\det(H_J - \lambda I) = 0$

$$\Rightarrow \begin{vmatrix} 0-\lambda & 0 & -1/2 \\ 0 & -\lambda & -2/5 \\ -1/2 & -2/5 & -\lambda \end{vmatrix} = -\lambda\left[\lambda^2 - \frac{4}{25}\right] - \frac{1}{2}\left[0 - \frac{\lambda}{2}\right] = 0$$

$$\Rightarrow -\lambda^3 + \left(\frac{4}{25} + \frac{1}{4}\right)\lambda = 0$$

$$\Rightarrow (-\lambda)\left[\lambda^2 - \frac{41}{100}\right] = 0$$

$$\Rightarrow \lambda = 0,\ \lambda = +\sqrt{\frac{41}{100}},\ \lambda = -\sqrt{\frac{41}{100}}$$

Thus $\rho(H_J) = \sqrt{0.41}$ is the convergence factor of Jacobi iteration method.

(ii) For Gauss Seidel iteration method

$H_G = -(D+L)^{-1}U$

$$U = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } D+L = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 5 & 4 & 10 \end{bmatrix}$$

$$(D+L)^{-1} = \frac{1}{200}\begin{bmatrix} 50 & 0 & -25 \\ 0 & 40 & -16 \\ 0 & 0 & 20 \end{bmatrix}^T$$

$$= \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ -1/8 & -2/25 & 1/10 \end{bmatrix}$$

$$H_G = -(D+L)^{-1}U$$

$$= -\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ -1/8 & -2/25 & 1/10 \end{bmatrix}\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= -\begin{bmatrix} 0 & 0 & 1/2 \\ 0 & 0 & 2/5 \\ 0 & 0 & -41/100 \end{bmatrix}$$

$\det\left(H_G - \lambda I\right) = 0$ gives

$$\begin{vmatrix} -\lambda & 0 & -1/2 \\ 0 & -\lambda & -2/5 \\ 0 & 0 & 41/100 - \lambda \end{vmatrix} = (-\lambda)\left[-\lambda\left(\frac{41}{100} - \lambda\right) - 0\right] - \frac{1}{2}[0-0] = 0$$

Thus $\lambda = 0$, $\lambda = 0$, $\lambda = \frac{41}{100}$ are eigen values of H_G. and $\rho(H_G) = \frac{41}{100} = 0.41$ is convergence factor of Gauss Seidel iteration method.

6. For the following system of equations

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

(a) Set up the Gauss Seidel iteration scheme in matrix form.

(b) Show that iteration scheme is convergent and hence find its rate of convergence.

(c) Starting with $\bar{x}^{(0)} = \bar{0}$, iterate two times.

Answer :

(a) $H_G = -(D+L)^{-1}U$

In this example

$$D+L = \begin{bmatrix} -3 & 0 & 0 \\ 2 & -3 & 0 \\ 0 & 2 & -3 \end{bmatrix} \text{ and } U = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$(D+L)^{-1} = \frac{1}{27}\begin{bmatrix} 9 & 6 & 4 \\ 0 & 9 & 6 \\ 0 & 0 & 9 \end{bmatrix}^T$$

$$= \begin{bmatrix} -1/3 & -2/9 & -4/27 \\ 0 & -1/3 & -2/9 \\ 0 & 0 & -1/3 \end{bmatrix}^T = \begin{bmatrix} -1/3 & 0 & 0 \\ -2/9 & -1/3 & 0 \\ -4/27 & -2/9 & -1/3 \end{bmatrix}$$

Thus, $H_G = -(D+L)^{-1}U$

$$= \begin{bmatrix} 1/3 & 0 & 0 \\ 2/9 & 1/3 & 0 \\ 4/27 & 2/9 & 1/3 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1/3 & 0 \\ 0 & 2/9 & 1/3 \\ 0 & 4/27 & 2/9 \end{bmatrix}$$

$$(D+L)^{-1} b = \begin{bmatrix} -1/3 & 0 & 0 \\ -2/9 & -1/3 & 0 \\ -4/27 & -2/9 & -1/3 \end{bmatrix} \begin{bmatrix} -2 \\ 0 \\ -1 \end{bmatrix}$$

$$= \begin{bmatrix} 2/3 \\ 4/9 \\ +17/27 \end{bmatrix}$$

Thus, $\bar{x}^{(k+1)} = \begin{bmatrix} 0 & 1/3 & 0 \\ 0 & 2/9 & 1/3 \\ 0 & 4/27 & 2/9 \end{bmatrix} \bar{x}^{(k)} + \begin{bmatrix} 2/3 \\ 4/9 \\ 17/27 \end{bmatrix}$

Answer (b) : To determine the rate of convergence we calculate spectral radius of matrix H_G.

$$\det\left(H_G - \lambda\tau\right) = 0$$

$$\Rightarrow \begin{vmatrix} -\lambda & 1/3 & 0 \\ 0 & 2/9 - \lambda & 1/3 \\ 0 & 4/27 & 2/9 - \lambda \end{vmatrix} = 0$$

$$\Rightarrow -\lambda\left[\left(\frac{2}{9}-\lambda\right)\left(\frac{2}{9}-\lambda\right)-\frac{4}{81}\right]-\frac{1}{3}[0,0]=0$$

$$\therefore -\lambda\left[\lambda^2-\frac{4}{9}\lambda+\frac{4}{81}-\frac{4}{81}\right]=0$$

$$\Rightarrow -\lambda^2\left(\lambda-\frac{4}{9}\right)=0$$

$\Rightarrow \lambda = 0, 0, \frac{4}{9}$ are eigen values of H_G.

Thus, $\rho(H_G) = \frac{4}{9}$

Since $\rho(H_G) = \frac{4}{9} < 1$, the method is convergent.

The rate of convergence $\nu = -\log\frac{4}{9} = 0.3522$.

Answer (c) : $\bar{x}^{(0)} = \bar{0}$

$$\bar{x}^{(1)} = \begin{bmatrix} 2/3 \\ 4/9 \\ 17/27 \end{bmatrix} = \begin{bmatrix} 0.6667 \\ 0.4445 \\ 0.6297 \end{bmatrix}$$

$$\bar{x}^{(2)} = \begin{bmatrix} 0 & 1/3 & 0 \\ 0 & 2/9 & 1/3 \\ 0 & 4/27 & 2/9 \end{bmatrix} \begin{bmatrix} 2/3 \\ 4/9 \\ 17/27 \end{bmatrix} + \begin{bmatrix} 2/3 \\ 4/9 \\ 17/27 \end{bmatrix}$$

$$= \begin{bmatrix} 22/27 \\ 61/81 \\ 203/243 \end{bmatrix} = \begin{bmatrix} 0.8148 \\ 0.7531 \\ 0.8354 \end{bmatrix}$$

7. Find the inverse of matrix $A = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$.

Using LU decomposition method. Take $U_{ii} = 1$.

Answer : We write

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

$$= \begin{bmatrix} \ell_{11} & \ell_{11}u_{12} & \ell_{11}u_{13} \\ \ell_{21} & \ell_{21}u_{12} + \ell_{22} & \ell_{21}u_{13} + \ell_{22}u_{23} \\ \ell_{31} & \ell_{31}u_{12} + \ell_{32} & \ell_{31}u_{13} + \ell_{32}u_{23} + \ell_{33} \end{bmatrix}$$

On comparing the corresponding elements we have from first column $\ell_{11} = 3$, $\ell_{21} = 2$, $\ell_{31} = 1$.

From first row $\quad \ell_{11}u_{12} = 2 \Rightarrow u_{12} = \frac{2}{3}$

$$\ell_{11}u_{13} = 1 \Rightarrow u_{13} = \frac{1}{3}$$

From second column $\quad \ell_{21}u_{12} + \ell_{22} = 3$

$$\ell_{22} = 3 - \frac{4}{3} = \frac{5}{3}$$

$$\ell_{31}u_{12} + \ell_{32} = 2$$

$$\therefore \ell_{32} = 2 - \frac{2}{3} = \frac{4}{3}$$

From third column we have

$$\ell_{21}u_{13} + \ell_{22}u_{23} = 2$$

$$(2)\left(\frac{1}{3}\right) + \left(\frac{5}{3}\right)u_{23} = 2$$

$$\therefore u_{23} = \frac{4}{5}$$

$$\ell_{31}u_{13} + \ell_{32}u_{23} + \ell_{33} = 2$$

$$\therefore \frac{1}{3} + \frac{16}{15} + \ell_{33} = 2$$

$$\ell_{33} = 2 - \frac{1}{3} - \frac{16}{15} = \frac{3}{5}$$

Thus we have, $L = \begin{bmatrix} 3 & 0 & 0 \\ 2 & 5/3 & 0 \\ 1 & 4/3 & 3/5 \end{bmatrix}$ and $U = \begin{bmatrix} 1 & 2/3 & 1/3 \\ 0 & 1 & 4/5 \\ 0 & 0 & 1 \end{bmatrix}$

Now, $L^{-1} = \frac{1}{3}\begin{bmatrix} 1 & -6/5 & 1 \\ 0 & 9/5 & -4 \\ 0 & 0 & 5 \end{bmatrix}^T$

$$= \begin{bmatrix} 1/3 & 0 & 0 \\ -2/5 & 3/5 & 0 \\ 1/3 & -4/3 & 5/3 \end{bmatrix}$$

$$U^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2/3 & 1 & 0 \\ 1/5 & -4/5 & 1 \end{bmatrix}^{\mathrm{T}}$$

$$= \begin{bmatrix} 1 & -2/3 & 1/5 \\ 0 & 1 & -4/5 \\ 0 & 0 & 1 \end{bmatrix}$$

Hence, $A^{-1} = (LU)^{-1} = U^{-1}L^{-1}$

$$= \begin{bmatrix} 1 & -2/3 & 1/5 \\ 0 & 1 & -4/5 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 0 & 0 \\ -2/5 & 3/5 & 0 \\ 1/3 & -4/3 & 5/3 \end{bmatrix}$$

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

8. Solve the following system of equations by Doolittie's method.

$$\begin{bmatrix} 2 & 1 & +1 & -2 \\ 4 & 0 & 2 & 1 \\ 3 & 2 & 2 & 0 \\ 1 & 3 & 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -10 \\ 8 \\ 7 \\ -5 \end{bmatrix}$$

Answer : We write the system of equations

$A\bar{x} = \bar{b}$ as $LU\bar{x} = \bar{b}$ or $L\bar{y} = \bar{b}$ and $U\bar{x} = \bar{y}$.

From $L\bar{y} = \bar{b}$ calculate $\bar{y}$ by forward substitution and calculate $\bar{x}$ from $U\bar{x} = \bar{y}$ by backward substitution. For Doolittle method diagonal elements of matrix L are 1. Thus we have,

$$\begin{bmatrix} 2 & 1 & 1 & -2 \\ 4 & 0 & 2 & 1 \\ 3 & 2 & 2 & 0 \\ 1 & 3 & 2 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \ell_{21} & 1 & 0 & 0 \\ \ell_{31} & \ell_{32} & 1 & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}$$

Comparing the corresponding elements on both sides, from first row we have,

$u_{11} = 2$, $u_{12} = 1$, $u_{13} = 1$, $u_{14} = -2$.

From first column we have,

$$u_{11}\ell_{21} = 4 \Rightarrow \ell_{21} = 2, \ell_{31} = \frac{3}{2}, \ell_{41} = \frac{1}{2}$$

From second row we have $\ell_{21}u_{12} + u_{22} = 0$ i.e. $u_{22} = -2$

$\ell_{21}u_{13} + u_{23} = 2$ i.e. $u_{23} = 0$

$\ell_{21}u_{14} + u_{24} = 1$ i.e. $u_{24} = 1 - (2)(-2) = 5$

From second column we have,

$$\ell_{31}u_{12} + \ell_{32}u_{22} = 2 \text{ i.e. } \ell_{32} = -\frac{1}{4}$$

$$\ell_{41}u_{12} + \ell_{42}u_{22} = 3 \text{ i.e. } \ell_{42} = -\frac{5}{4}$$

From third row,

$$\ell_{31}u_{13} + \ell_{32}u_{23} + u_{33} = 2 \text{ i.e. } u_{33} = \frac{1}{2}$$

$$\ell_{31}u_{14} + \ell_{32}u_{24} + u_{34} = 0 \text{ i.e } u_{34} = \frac{17}{4}$$

From third column we get,

$$\ell_{41}u_{13} + \ell_{42}u_{23} + \ell_{43}u_{33} = 2 \text{ i.e. } \ell_{43} = 3$$

Lastly $\ell_{41}u_{14} + \ell_{42}u_{24} + \ell_{43}u_{34} + u_{44} = -1$ i.e. $u_{44} = -\frac{13}{2}$

Thus we have

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3/2 & -1/4 & 1 & 0 \\ 1/2 & -5/4 & 3 & 1 \end{bmatrix} \text{ and } U = \begin{bmatrix} 2 & 1 & 1 & -2 \\ 0 & -2 & 0 & 5 \\ 0 & 0 & 1/2 & 17/4 \\ 0 & 0 & 0 & -13/2 \end{bmatrix}$$

From $L\bar{y} = \bar{b}$ we get

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3/2 & -1/4 & 1 & 0 \\ 1/2 & -5/4 & 3 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} -10 \\ 8 \\ 7 \\ -5 \end{bmatrix}$$

Using forward substitution we have

$y_1 = -10$, $y_2 = 8 - 2y = 28$

$$y_3 = 7 - \frac{3}{2}y_1 + \frac{1}{4}y_2 = 7 + \frac{30}{2} + \frac{28}{4} = 29$$

$$y_4 = -5 - \frac{1}{2}y_1 + \frac{5}{4}y_2 - 3y_3 = -5 + \frac{10}{2} + \frac{5}{4}(28) - 3(29) = -52$$

From $U\bar{x} = \bar{y}$ we get

$$\begin{bmatrix} 2 & 1 & 1 & -2 \\ 0 & -2 & 0 & 5 \\ 0 & 0 & 1/2 & 17/4 \\ 0 & 0 & 0 & -13/2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -10 \\ 28 \\ 29 \\ -52 \end{bmatrix}$$

By backward substitution we have,

$$-\frac{13}{2}x_4 = -52 \Rightarrow x_4 = 8$$

$$\frac{1}{2}x_3 + \frac{17}{4}x_4 = 29 \Rightarrow x_3 = -10$$

$$-2x_2 + 5x_4 = 28 \Rightarrow x_2 = 6$$

$$2x_1 + x_2 + x_3 - 2x_4 = -10 \Rightarrow x_1 = 5$$

Thus the solution is,

$$x_1 = 5,\ x_2 = 6,\ x_3 = -10,\ x_4 = 8.$$

9. Solve the following system of equations $A\bar{x} = \bar{b}$ by Crout method.

$$\begin{bmatrix} 1 & 1 & 1 \\ 4 & 3 & -1 \\ 3 & 5 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 6 \\ 4 \end{bmatrix}$$

Answer : We write A = LU

$$\begin{bmatrix} 1 & 1 & 1 \\ 4 & 3 & -1 \\ 3 & 5 & 3 \end{bmatrix} = \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

On comparing the corresponding elements we obtain,

First column, $\ell_{11} = 1,\ \ell_{21} = 4,\ \ell_{31} = 3.$

First row, $\ell_{11}u_{12} = 1$ i.e. $u_{12} = 1,\ u_{13} = 1.$

Second column, $\ell_{21}u_{12} + \ell_{22} = 3$ i.e. $\ell_{22} = 3 - 4 = -1$

$\ell_{31}u_{12} + \ell_{32} = 5$ i.e. $\ell_{32} = 5 - 3 = 2$

Second row, $\ell_{21}u_{13} + \ell_{22}u_{23} = -1$ i.e. $u_{23} = 5$

Second row, $\ell_{21}u_{13} + \ell_{22}u_{23} = -1$ i.e. $u_{23} = 5$

Third row, $\ell_{31}u_{13} + \ell_{32}u_{23} + \ell_{33} = 3$ i.e. $\ell_{33} = 3 - 3 - 10 = -10$

Thus we have,

$$L=\begin{bmatrix}1 & 0 & 0\\ 4 & -1 & 0\\ 3 & 2 & -10\end{bmatrix},\ U=\begin{bmatrix}1 & 1 & 1\\ 0 & 1 & 5\\ 0 & 0 & 1\end{bmatrix}$$

Now $(LU)\bar{x}=b$,

Put $U\bar{x}=\bar{y}$ then $L\bar{y}=\bar{b}$,

i.e. $$\begin{bmatrix}1 & 0 & 0\\ 4 & -1 & 0\\ 3 & 2 & -10\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix}=\begin{bmatrix}1\\ 6\\ 4\end{bmatrix}$$

Using forward substitution we have $y_1=1$.

$$4y_1-y_2=6 \text{ i.e. } y_2=4-6=-2$$

$$3y_1+2y_2-10y_3=4 \text{ i.e. } y_3=\frac{1}{10}[(3)(1)+2(-2)-4]=-\frac{1}{2}$$

Thus $y_1=1$, $y_2=-2$ and $y_3=-\frac{1}{2}$.

Now $U\bar{x}=\bar{y}$,

$$\begin{bmatrix}1 & 1 & 1\\ 0 & 1 & 5\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}1\\ -2\\ -\frac{1}{2}\end{bmatrix}$$

By back substitution we get $x_3=-\frac{1}{2}$,

$$x_2+5x_3=-2 \text{ i.e. } x_2=-2+5\left(\frac{1}{2}\right)=\frac{1}{2}$$

and $$x_1+x_2+x_3=1 \text{ i.e. } x_1=1-x_2-x_3=1-\frac{1}{2}+\frac{1}{2}=1$$

Thus $x_1=1$, $x_2=\frac{1}{2}$, $x_3=-\frac{1}{2}$ is the solution.

10. For the matrix $A=\begin{bmatrix}2 & -1 & 1\\ -1 & 2 & -1\\ 1 & -1 & 2\end{bmatrix}$, find all eigen values and the corresponding eigen vectors.

Answer : $\det(A-\lambda I)=0$

$$\Rightarrow\begin{vmatrix}2-\lambda & -1 & 1\\ -1 & 2-\lambda & -1\\ 1 & -1 & 2-\lambda\end{vmatrix}=(2-\lambda)\left[(2-\lambda)^2-1\right]+1\left[-(2-\lambda)+1\right]+\left(1-(2-\lambda)\right)=0$$

$$\Rightarrow(2-\lambda)\left(4-4\lambda+\lambda^2-1\right)+(\lambda-1)+(-1+\lambda)=0$$

$$\Rightarrow(2-\lambda)\left(\lambda^2-4\lambda+3\right)+2(\lambda-1)=0$$

$$\Rightarrow-\lambda^3+6\lambda^2-9\lambda+4=0$$

$$\Rightarrow(-\lambda+1)\left(\lambda^2-5\lambda+4\right)=0$$

$$\Rightarrow(-\lambda+1)(\lambda-4)(\lambda-1)=0$$

Thus $\lambda=1$, $\lambda=1$, $\lambda=4$ are eigenvalues.

The eigenvectors corresponding to $\lambda=1$ is solution of the system $(A-I)\bar{x}=0$.

$$\begin{bmatrix}1 & -1 & 1\\ -1 & 1 & -1\\ 1 & -1 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}$$

The solutions are $\bar{x}=\begin{bmatrix}1\\ 1\\ 0\end{bmatrix}$ and $\bar{x}=\begin{bmatrix}1\\ 0\\ -1\end{bmatrix}$.

The eigenvector corresponding to $\lambda=4$ is solution of $(A-4I)\bar{x}=0$.

$$\begin{bmatrix}-2 & -1 & 1\\ -1 & -2 & -1\\ 1 & -1 & -2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}$$

The solution is $\bar{x}=\begin{bmatrix}1\\ -1\\ 1\end{bmatrix}$.

11. Find the intervals which contain all the eigenvalues of the following matrix.

$$\begin{bmatrix} 1 & 2 & -3 \\ 2 & 1 & -1 \\ -3 & -1 & 2 \end{bmatrix}$$

Answer : By Gerschgorin bounds the eigen values lie in the region.

$$|\lambda| < \max\{6,4,6\} = 6$$

$$|\lambda| < \max\{6,4,6\} = 6$$

i.e. $|\lambda| < 6$

By Brauer theorem, all the eigen values lie in the union of circles $|\lambda - 1| < 5$, $|\lambda - 1| < 3$, $|\lambda - 2| < 4$ and union of circles $|\lambda - 1| < 5$, $|\lambda - 1| < 3$, $|\lambda - 2| < 4$.

Thus all the eigenvalues lie in the region

$$\{|\lambda| < 6\} \cap \{|\lambda - 1| < 5 \cup |\lambda - 1| < 3 \cup |\lambda - 2| < 4\}$$

Since A is summetric all eigen values are real.

$$(-6,6) \cap \{\{-4 < \lambda < 6\} \cup \{-2 < \lambda < 4\} \cup \{-2 < \lambda < 6\}\}$$

$$(-6,6) \cap \{(-4,6) \cup (-2,4) \cup (-2,6)\}$$

$$(-6,6) \cap (-4,6) = (-4,6)$$

12. Compute $\begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}^{10}$ exactly.

Answer : Let $S = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

$$S^T AS = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta + (0.1)\sin\theta & -\sin\theta + (0.1)\cos\theta \\ (0.1)\cos\theta + \sin\theta & -(0.1)\sin\theta + \cos\theta \end{bmatrix}$$

$$= \begin{bmatrix} \cos^2\theta + (0.2)\sin\theta\cos\theta + \sin^2\theta & -\sin\theta\cos\theta + (0.1)\cos^2\theta - (0.1)\sin^2\theta + \sin\theta\cos\theta \\ (0.1)(\cos^2\theta - \sin^2\theta) & \sin^2\theta - (0.1)\sin\theta\cos\theta - (0.1)\sin\theta\cos\theta + \cos^2\theta \end{bmatrix}$$

$$= \begin{bmatrix} 1+(0.2)\sin\theta\cos\theta & (0.1)\left(\cos^2\theta-\sin^2\theta\right) \\ (0.1)\left(\cos^2\theta-\sin^2\theta\right) & 1-(0.2)\sin\theta\cos\theta \end{bmatrix}$$

This matrix reduces to diagonal matrix if $\cos^2\theta-\sin^2\theta=0$.

$$\cos^2\theta-\sin^2\theta=0 \Rightarrow \cos 2\theta=0 \Rightarrow 2\theta=\frac{\pi}{2} \qquad \theta=\frac{\pi}{4}$$

Using this values of $\theta=\frac{\pi}{4}$ we get

$$S^T AS=\begin{bmatrix} 1.1 & 0 \\ 0 & 0.9 \end{bmatrix}=D$$

$$S^T A^{10} S=D^{10}$$

$$A^{10}=SD^{10}S^T$$

Since $\theta=\frac{\pi}{4}$, $S=\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$

$$A^{10}=\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} (1.1)^{10} & 0 \\ 0 & (0.9)^{10} \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} \frac{(1.1)^{10}}{\sqrt{2}} & \frac{(1.1)^{10}}{\sqrt{2}} \\ -\frac{(0.9)^{10}}{\sqrt{2}} & \frac{(0.9)^{10}}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{(1.1)^{10}}{2}+\frac{(0.9)^{10}}{2} & \frac{(1.1)^{10}}{2}-\frac{(0.9)^{10}}{2} \\ \frac{(1.1)^{10}}{2}-\frac{(0.9)^{10}}{2} & \frac{(1.1)^{10}}{2}+\frac{(0.9)^{10}}{2} \end{bmatrix}$$

13. Find all the eigenvalues and the corresponding eigen vectors of the matrix $\begin{bmatrix} 1 & -2 & 4 \\ -2 & 5 & -2 \\ 4 & -2 & 1 \end{bmatrix}$ using Jacobi method.

Answer :

The largest off diagonal element in magnitude is a_{13}.

$$\tan 2\theta = \frac{2a_{13}}{a_{11} - a_{33}} = \frac{8}{0} = \infty \text{ i.e. } \theta = \frac{\pi}{4}$$

$$S_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

The first rotation gives,

$$B_1 = S_1^{-1} A S_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & -2 & 4 \\ -2 & 5 & -2 \\ 4 & -2 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{2}} & -2 & \frac{3}{\sqrt{2}} \\ -\frac{4}{\sqrt{2}} & 5 & 0 \\ \frac{5}{\sqrt{2}} & -2 & -\frac{3}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} 5 & -\frac{4}{\sqrt{2}} & 0 \\ -\frac{4}{\sqrt{2}} & 5 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

The largest off diagonal element in B_1 is $(1,2)$

$$\tan 2\theta = \frac{(1,2)}{(1,1)-(2,2)} = \frac{-4/\sqrt{2}}{5-5} = \infty \text{ i.e. } \theta = \frac{\pi}{4}$$

$$S_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The second rotation gives,

$$B_2 = S_2^{-1} B_1 S_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & -\frac{4}{\sqrt{2}} & 0 \\ -\frac{4}{\sqrt{2}} & 5 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{2}} - 2 & -\frac{5}{\sqrt{2}} - 2 & 0 \\ -2 + \frac{5}{\sqrt{2}} & 2 + \frac{5}{\sqrt{2}} & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

$$= \begin{bmatrix} 5 - 2\sqrt{2} & 0 & 0 \\ 0 & 5 + 2\sqrt{2} & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

Thus eigenvalues are $\lambda_1 = 5 - 2\sqrt{2}$, $\lambda_2 = 5 + 2\sqrt{2}$, $\lambda_2 = -3$.

We have the matrix of eigenvectors as,

$$S = S_1 S_2$$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

The eigenvectors are,

$$V_1 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{\sqrt{2}} \\ \frac{1}{2} \end{bmatrix}, \quad V_2 = \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{2} \end{bmatrix}, \quad V_3 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

14. Using the Householder's transformation reduce the matrix $A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ into a tridiagonal matrix.

ANswer : We have $w = [0, x_2, x_3]^T$, $S_1 = \sqrt{{a_{12}}^2 + {a_{13}}^2} = \sqrt{2}$

$${x_2}^2 = \frac{1}{2}\left[1 + \frac{a_{12}\ \text{sign}\ (a_{12})}{S_1}\right] = \frac{1}{2}\left[1 + \frac{1}{\sqrt{2}}\right] = \frac{\sqrt{2}+1}{2\sqrt{2}}$$

$$1 - 2{x_2}^2 = -\frac{1}{\sqrt{2}}$$

$$x_3 = \frac{a_{13}\left(\text{sign } a_{13}\right)}{2S_1 x_2} = \frac{1}{2S_1 x_2};\ 2x_2 x_3 = \frac{1}{\sqrt{2}}$$

$$1 - 2{x_3}^2 = \frac{1}{\sqrt{2}}$$

Thus, $$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-2{x_2}^2 & -2x_2x_3 \\ 0 & -2x_2x_3 & 1-2{x_3}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$B = P_1 A P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} 2 & -\sqrt{2} & 0 \\ -\sqrt{2} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

15. Redice the matrix $\begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix}$ into tridiagonal form using the Householder method.

Answer : We have $w = [0, x_2, x_3]^T$, $S_1 = \sqrt{{a_{12}}^2 + {a_{13}}^2} = \sqrt{5}$

$${x_2}^2 = \frac{1}{2}\left[1 + \frac{a_{12}\ \text{sign}\left(a_{12}\right)}{\text{S}_1}\right] = \frac{1}{2}\left[1 + \frac{2}{\sqrt{5}}\right] = \frac{\sqrt{5}+2}{2\sqrt{5}}$$

$$1 - 2{x_2}^2 = -\frac{\sqrt{5}+2}{\sqrt{5}} = -\frac{2}{\sqrt{5}}$$

$$x_3 = \frac{a_{13}\text{sign}\left(a_{13}\right)}{2S_1 x_2} = \frac{-1}{2S_1 x_2};\ \ 2x_2 x_3 = -\frac{2x_2}{2S_1 x_2} = -\frac{1}{S_1} = -\frac{1}{\sqrt{5}}$$

$$x_3^2 = 1 - x_2^2 = 1 - \frac{\sqrt{5}+2}{2\sqrt{5}} = \frac{\sqrt{5}-2}{2\sqrt{5}}$$

$$\therefore 1 - 2x_3^2 = 1 - \frac{2(\sqrt{5}-2)}{2\sqrt{5}} = \frac{2}{\sqrt{5}}$$

Thus, $P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-2x_2^2 & -2x_2x_3 \\ 0 & -2x_2x_3 & 1-2x_3^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ 0 & -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix}$

$$B = P_1 A P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ 0 & -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 1 & -2 & -1 \\ 2 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ 0 & -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ 0 & -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 1 & -\sqrt{5} & 0 \\ 2 & -\frac{4}{\sqrt{5}} & \sqrt{5} \\ -1 & -\frac{3}{\sqrt{5}} & \frac{4}{\sqrt{5}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -\frac{5}{\sqrt{5}} & 0 \\ -\frac{5}{\sqrt{5}} & -\frac{3}{5} & -\frac{6}{5} \\ 0 & -\frac{6}{5} & \frac{3}{5} \end{bmatrix} = \begin{bmatrix} 1 & -\sqrt{5} & 0 \\ -\sqrt{5} & -\frac{3}{5} & -\frac{6}{5} \\ 0 & -\frac{6}{5} & \frac{3}{5} \end{bmatrix}$$

16. Find the largest eigenvalues in modulus and the corresponding eigenvector of the matrix

$\begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix}$ using power method.

Answer : Let $\bar{v}_0$ be a non-zero arbitrary initial vector.

Define $\bar{y}_{k+1} = A\bar{v}_k$

$\bar{v}_{k+1} = \dfrac{\bar{y}_{k+1}}{m_{k+1}}$, m_{k+1} is the largest (in magnitude) element in $\bar{y}_{k+1}$.

$\lambda_1 = \lim_{k\to\infty}\left(\dfrac{\bar{y}_{k+1}}{\bar{v}_k}\right)_r$, r = 1, 2, 3, ..., n.

$\bar{v}_{k+1}$ is the required eigenvector,

I. $\bar{v}_0 = [1,1,1]^T$

$$\bar{y}_1 = \begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -8 \\ 4 \\ 18 \end{bmatrix};\ \bar{v}_1 = \begin{bmatrix} -4/9 \\ 2/9 \\ 1 \end{bmatrix}$$

$$\left(\frac{\bar{y}_1}{\bar{v}_0}\right)_{1,2,3} = \begin{pmatrix} -8 \\ 4 \\ 18 \end{pmatrix}$$ not compatible.

II. $$\bar{y}_2 = \begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix}\begin{bmatrix} -4/9 \\ 2/9 \\ 1 \end{bmatrix} = \begin{bmatrix} 10.55 \\ -1.11 \\ -7.77 \end{bmatrix};\ \bar{v}_2 = \begin{bmatrix} 1 \\ -0.105 \\ -0.736 \end{bmatrix}$$

$$\left(\frac{\bar{y}_2}{\bar{v}_1}\right)_{1,2,3} = \begin{pmatrix} -23.73 \\ -4.99 \\ -7.77 \end{pmatrix}$$ not compatible.

III. $$\bar{y}_3 = \begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -0.105 \\ -0.736 \end{bmatrix} = \begin{bmatrix} -17.628 \\ 6.844 \\ 18.948 \end{bmatrix}; \bar{v}_3 = \begin{bmatrix} -0.9303 \\ 0.3612 \\ 1 \end{bmatrix}$$

$$\left(\frac{\bar{y}_3}{\bar{v}_2}\right)_{(1,2,3)} = \begin{pmatrix} -17.628 \\ -65.18 \\ -24.45 \end{pmatrix} \text{ not compatible.}$$

IV. $$\bar{y}_4 = \begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix} \begin{bmatrix} -0.9303 \\ 0.3612 \\ 1 \end{bmatrix} = \begin{bmatrix} 18.39 \\ -7.6374 \\ -18.0508 \end{bmatrix}; \bar{v}_4 = \begin{bmatrix} 1 \\ -0.415 \\ -0.98 \end{bmatrix}$$

$$\left(\frac{\bar{y}_4}{\bar{v}_3}\right)_{(1,2,3)} = \begin{pmatrix} -19.76 \\ -21.14 \\ -18.05 \end{pmatrix} \text{ not compatible.}$$

V. $$\bar{y}_5 = \begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -0.415 \\ -0.98 \end{bmatrix} = \begin{bmatrix} -19.6 \\ 9.1 \\ 19.7 \end{bmatrix}; \bar{v}_5 = \begin{bmatrix} -0.99 \\ 0.46 \\ 1 \end{bmatrix}$$

$$\left(\frac{\bar{y}_5}{\bar{v}_4}\right)_{(1,2,3)} = \begin{pmatrix} -19.6 \\ -21.92 \\ -20.10 \end{pmatrix} \text{ not compatible.}$$

VI. $$\bar{y}_6 = \begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix} \begin{bmatrix} -0.99 \\ 0.46 \\ 1 \end{bmatrix} = \begin{bmatrix} 19.69 \\ -9.42 \\ -19.64 \end{bmatrix}; \bar{v}_6 = \begin{bmatrix} 1 \\ -0.48 \\ -0.99 \end{bmatrix}$$

$$\left(\frac{\bar{y}_6}{\bar{v}_5}\right)_{(1,2,3)} = \begin{pmatrix} -19.88 \\ -20.47 \\ -19.64 \end{pmatrix} \text{ not compatible.}$$

VII. $\bar{y}_7 = \begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -0.48 \\ -0.99 \end{bmatrix} = \begin{bmatrix} -19.89 \\ 9.82 \\ 19.94 \end{bmatrix}; \bar{v}_7 = \begin{bmatrix} -0.997 \\ 0.492 \\ 1 \end{bmatrix}$

$$\left(\frac{\bar{y}_7}{\bar{v}_6}\right)_{(1,2,3)} = \begin{pmatrix} -19.89 \\ -20.45 \\ -20.14 \end{pmatrix}$$ not compaitble.

VIII. $\bar{y}_8 = \begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix} \begin{bmatrix} -0.997 \\ 0.492 \\ 1 \end{bmatrix} = \begin{bmatrix} 19.923 \\ -9.874 \\ -19.908 \end{bmatrix}; \bar{v}_8 = \begin{bmatrix} 1 \\ -0.495 \\ -0.999 \end{bmatrix}$

$$\left(\frac{\bar{y}_8}{\bar{v}_7}\right)_{(1,2,3)} = \begin{pmatrix} -19.98 \\ -20.06 \\ -19.90 \end{pmatrix}$$

IX. $\bar{y}_9 = \begin{bmatrix} -15 & 4 & 3 \\ 10 & -12 & 6 \\ 20 & -4 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -0.495 \\ -0.999 \end{bmatrix} = \begin{bmatrix} -19.977 \\ 9.946 \\ 19.982 \end{bmatrix}$

$$\left(\frac{\bar{y}_9}{\bar{v}_8}\right)_{(1,2,3)} = \begin{pmatrix} -19.98 \\ -20.08 \\ -20.00 \end{pmatrix}$$ compatible.

The approximation to the largest eigenvalue in modulus is $|\lambda| = 20$. (i.e. $\lambda = -20$) and the eigenvector (approximate) is $[1 - 0.495 - 0.999]^T$.

17. Find the largest eigenvalue of the matrix $A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}$ using power method. Correct upto two decimal places.

Answer : $\bar{v}_0 = [1,1,1,1]^T$

I. $$\overline{y}_1 = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \\ 3 \\ 4 \end{bmatrix}; \quad \overline{v}_1 = \begin{bmatrix} 1 \\ 0.75 \\ 0.75 \\ 1 \end{bmatrix}$$

II. $$\overline{y}_2 = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0.75 \\ 0.75 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 2.75 \\ 2.75 \\ 3.5 \end{bmatrix}; \quad \overline{v}_2 = \begin{bmatrix} 1 \\ 0.786 \\ 0.786 \\ 1 \end{bmatrix}$$

$$\left(\frac{\overline{y}_2}{\overline{v}_1}\right)_{(1,2,3)} = \begin{pmatrix} 3.5 \\ 3.66 \\ 3.66 \\ 3.5 \end{pmatrix}$$ not compatible upto 2 decimal places.

III. $$\overline{y}_3 = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0.786 \\ 0.786 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.572 \\ 2.786 \\ 2.786 \\ 3.572 \end{bmatrix}; \quad \overline{v}_3 = \begin{bmatrix} 1 \\ 0.7799 \\ 0.7799 \\ 1 \end{bmatrix}$$

$$\left(\frac{\overline{y}_3}{\overline{v}_2}\right)_{(1,2,3)} = \begin{pmatrix} 3.572 \\ 3.544 \\ 3.544 \\ 3.572 \end{pmatrix}$$ not correct upto two decimal places.

IV. $$\overline{y}_4 = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0.7799 \\ 0.7799 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.5598 \\ 2.7799 \\ 2.7799 \\ 3.5598 \end{bmatrix}; \quad \overline{v}_4 = \begin{bmatrix} 1 \\ 0.7809 \\ 0.7809 \\ 1 \end{bmatrix}$$

$$\left(\frac{\overline{y}_4}{\overline{v}_3}\right)_{(1,2,3)} = \begin{pmatrix} 3.5596 \\ 3.564 \\ 3.564 \\ 3.5596 \end{pmatrix}$$ not correct upto two decimal places.

V. $\bar{y}_5 = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0.7809 \\ 0.7809 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.5618 \\ 2.7809 \\ 2.7809 \\ 3.5618 \end{bmatrix}$;

$\left(\frac{\bar{y}_5}{\bar{v}_4}\right)_{(1,2,3)} = \begin{pmatrix} 3.5688 \\ 3.5611 \\ 3.5611 \\ 3.5618 \end{pmatrix}$ correct upto two decimal place.

The approximate largest root is 3.561 and the corresponding eigenvector is $\begin{bmatrix} 1 \\ 0.7809 \\ 0.7809 \\ 1 \end{bmatrix}$.

EXERCISE

1. Determine the LU decomposition of the matrix $\begin{bmatrix} 2 & -6 & 10 \\ 1 & 5 & 1 \\ -1 & 15 & -1 \end{bmatrix}$ assuming $\ell_{ii} = 1$, i = 1, 2, 3.

2. Solve the system of equations.

$4x_1 + x_2 + x_3 = 4$

$x_1 + 4x_2 - 2x_3 = 4$

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

by Doolittle method.

3. Solve the following system of equation by Crout method.

(i) $x_1 + x_2 - x_3 = 2$

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

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

(ii) $4x_1 - x_2 = 1$

$-x_1 + 4x_2 - x_3 = 0$

$-x_2 + 4x_3 = 0$

4. Given the matrix A = I + L + U where

$$A = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix}$$

L and U are strictly lower and upper triangular matrices respectively, decide whether (a) Jacobi, (b) Gauss Seidel methods converge to the solution $A\bar{x} = \bar{b}$.

5. Show that the Gauss Seidel method for solving the system of equation.

(i) $\begin{bmatrix} 1 & 1 & -1 \\ 2 & 3 & 5 \\ 3 & 2 & -3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -1 \\ -6 \\ 4 \end{bmatrix}$ and

(ii) $\begin{bmatrix} 1 & 2 & 4 \\ 2 & 1 & 2 \\ 4 & 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -1 \\ 5 \\ 3 \end{bmatrix}$ diverges.

6. Setup the Jacobi iteration scheme in matrix form for the system

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

(i) Locate the eigenvalues of the iteration matrix H.

(ii) Determine the largest (in magnitude) eigenvalue using the Newton Rapnson method.

(iii) Find the rate of convergence of the iteration scheme.

7. The matrix $A = \begin{bmatrix} 1 & -2 & 3 \\ 6 & -13 & 18 \\ 4 & -10 & 18 \end{bmatrix}$ is transformed to diagonal for by the matrix $T = \begin{bmatrix} 1 & 0 & 1 \\ 3 & 3 & 4 \\ 2 & 2 & 3 \end{bmatrix}$

i.e. $T^{-1}AT$. Calculate the eigenvalues and the corresponding eigenvectors of A.

8. Find the intervals which contain all the eigenvalues of the following matrix

(i) $\begin{bmatrix} 1 & 2 & -3 \\ 2 & 1 & -1 \\ -3 & -1 & 2 \end{bmatrix}$ (ii) $\begin{bmatrix} 2 & 3 & 1 \\ 3 & 2 & 2 \\ 1 & 2 & 1 \end{bmatrix}$ (iii) $\begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix}$

9. Find eigenvalues and the corresponding eigenvalues of the following matrice using Jacobi method.

(i) $\begin{bmatrix} 3 & 2 & 2 \\ 2 & 5 & 2 \\ 2 & 2 & 3 \end{bmatrix}$ (ii) $\begin{bmatrix} 1 & -2 & 4 \\ -2 & 5 & -2 \\ 4 & -2 & 1 \end{bmatrix}$ (iii) $\begin{bmatrix} 2 & \sqrt{2} & 4 \\ \sqrt{2} & 6 & \sqrt{2} \\ 4 & \sqrt{2} & 2 \end{bmatrix}$

10. Determine the largest eigenvalues and the corresponding eigenvector of the matrix $\begin{bmatrix} 4 & 1 & 0 \\ 1 & 20 & 1 \\ 0 & 1 & 4 \end{bmatrix}$ to 3 correct decimal placed using power method.

11. Use Householder method and convert the following matrix in tridiagonal form $\begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix}$.

12. Find the largest eigenvalues of the matrix $A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}$ using power method.

❖ ❖ ❖

UNIT - III

INTERPOLATION, DIFFERENTIATION AND INTEGRATION

Interpolation is used to approximate given function by a polynomial or it is used to fit a polynomial when the data is given in tabular form. There are two main uses of interpolation. The first use is the reconstruction of function when it is not given explicity and second use is to replace a function $f(x)$ by an interpolating polynomial $P(x)$ so that many common operations such as determination of roots, differentiation, integration etc. which are required to perform on $f(x)$ may be performed using $P(x)$.

In this unit we first discuss the methods of constructing the interpolating polynomial $P(x)$ to a given function. We determine the deviation of the given function $f(x)$ from the approximating polynomial $P(x)$ by estimating truncation error bounds. We discuss numerical methods of differentiation and integration of a given function $f(x)$.

3.1 Interpolation

Definition 3.1.1 : Interpolating Polynomial

A polynomial $P(x)$ is called an interpolating polynomial if the value of $P(x)$ and/or its certain order derivatives coincide with those of $f(x)$ and/or its derivatives at one or more tabular points.

3.1.1 Lagrange Interpolation

Linear Interpolation :

We want to determine a polynomial of degree one denoted by

$$P_1(x) = a_1 x + a_0$$

where a_0 and a_1 are arbitrary constants.

Satisfying $p_1(x_0) = f(x_0)$ and $p_1(x_1) = f(x_1)$.

i.e. we want to interpolate a function by a polynomial of degree one.

Since $p_1(x_0) = f(x_0)$, $f(x_0) = a_1 x_0 + a_0$

and $p_1(x_1) = f(x_1) \Rightarrow f(x_1) = a_1 x_1 + a_0$

Above equations are two linear equations in two unknowns a_0 and a_1. Simultaneous evaluation of these two equations gives

$$a_1 = \frac{f(x_0) - f(x_1)}{x_0 - x_1} \text{ and } a_0 = f(x_0) - x_0\left[\frac{f(x_0) - f(x_1)}{x_0 - x_1}\right]$$

Thus $$p_1 = f(x_0) - x_0\left[\frac{f(x_0) - f(x_1)}{x_0 - x_1}\right] + \left[\frac{f(x_0) - f(x_1)}{x_0 - x_1}\right]x$$

$$\therefore p_1(x) = \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1) \quad \text{.... (3.1)}$$

In other words we want to determine a polynomial $p_1(x) = a_1 x + a_0$ which satisfies

$$f(x_0) = a_1 x_0 + a_0$$

$$f(x_1) = a_1 x_1 + a_0$$

Above equations are three equations in two unknowns and therefore they have to be linearly dependent.

$$\begin{vmatrix} p_1(x) & x & 1 \\ f(x_0) & x_0 & 1 \\ f(x_1) & x_1 & 1 \end{vmatrix} = 0$$

$$\therefore p_1(x)[x_0 - x_1] - x[f(x_0) - f(x_1)] + 1[x_1 f(x_0) - x_0 f(x_1)] = 0$$

$$\therefore p_1(x) = \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1)$$

$$= \ell_0(x) f(x_0) + \ell_1(x) f(x_1) \quad \text{..... (3.1.1.1)}$$

where $\ell_0(x) = \frac{x - x_1}{x_0 - x_1}$ and $\ell_1(x) = \frac{x - x_0}{x_1 - x_0}$

The functions $\ell_0(x)$ and $\ell_1(x)$ are called Lagrange fundamental polynomials. These polynomials satisfy

$$\ell_0(x) + \ell_1(x) = 1,\ \ell_i(x_j) = \delta_{ij},\ i, j = 0, 1$$

In general, if we have (n + 1) distinct points $a \le x_0 < x_1 < x_2 ..... < x_n \le b$ of [a, b] and a value of the function $f(x)$ is known at these points, we can determine the polynomial

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$$

which satisfies $p_n(x_i) = f(x_i)$, i = 0, 1, 2, ..., n.

In other words we have,

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_n x^n \quad .... (1)$$

$$f(x_0) = p_n(x_0) = a_0 + a_1 x_0 + a_2 {x_0}^2 + a_3 {x_0}^3 + ... + a_n {x_0}^n \quad .... (2)$$

$$f(x_1) = p_n(x_1) = a_0 + a_1 x_1 + a_2 {x_1}^2 + a_3 {x_1}^3 + ... + a_n {x_1}^n \quad .... (3)$$

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

$$f(x_n) = p_n(x_n) = a_0 + a_1 x_n + a_2 {x_n}^2 + a_3 {x_n}^3 + ... + a_n {x_n}^n \quad ... (n+2)$$

Equation (2) to (n + 1) are n + 1 equations in (n + 1) unknowns a_0 , a_1 a_2 , ..., a_n . Equations (2) to (n + 2) has unique solution if the corresponding coefficient matrix is non-singular i.e.

$$\begin{bmatrix} 1 & x_0 & {x_0}^2 & \cdots & {x_0}^n \\ 1 & x_1 & {x_1}^2 & \cdots & {x_1}^n \\ 1 & x_2 & {x_2}^2 & \cdots & {x_2}^n \\ \vdots & & & \cdots & \\ 1 & x_n & {x_n}^2 & \cdots & {x_n}^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \quad ..... (3.2.1.2)$$

has a unique solution if the Vandermonde's determinant

$$v(x_0, x_1, x_2, ..., x_n) = \begin{bmatrix} 1 & x_0 & {x_0}^2 & {x_0}^3 & \cdots & {x_0}^n \\ 1 & x_1 & {x_1}^2 & {x_1}^3 & \cdots & {x_1}^n \\ 1 & x_2 & {x_2}^2 & {x_2}^3 & \cdots & {x_2}^n \\ \vdots & & & & & \\ 1 & x_n & {x_n}^2 & {x_n}^3 & \cdots & {x_n}^n \end{bmatrix} \neq 0$$

If we determine a_0 , a_1 a_2 , ..., a_n from system (3.2.1.2) then the system (1), (2), ... (n + 2) will be linearly dependent and therefore its determinant should be zero.

i.e.
$$\begin{vmatrix} p_n(x) & 1 & x & x^2 & \cdots & x^n \\ f(x_0) & 1 & x_0 & {x_0}^2 & \cdots & {x_0}^n \\ f(x_1) & 1 & x_1 & {x_1}^2 & \cdots & {x_1}^n \\ f(x_2) & 1 & x_2 & {x_2}^2 & \cdots & {x_2}^n \\ \vdots & & & & & \\ f(x_n) & 1 & x_n & {x_n}^2 & \cdots & {x_n}^n \end{vmatrix} = 0 \qquad \text{..... (3.2.13)}$$

To evaluate the l.h.s. of equation (3.2.1.3) we use the following property of the determinants.

Let
$$v(x_0, x_1, x_2, ..., x_{n-1}x) = \begin{bmatrix} 1 & x_0 & {x_0}^2 & {x_0}^3 & \cdots & {x_0}^n \\ 1 & x_1 & {x_1}^2 & {x_1}^3 & \cdots & {x_1}^n \\ 1 & x_2 & {x_2}^2 & {x_2}^3 & \cdots & {x_2}^n \\ \vdots & & & & & \\ 1 & x_{n-1} & {x_{n-1}}^2 & {x_{n-1}}^3 & \cdots & {x_{n-1}}^n \\ 1 & x & x^2 & x^3 & \cdots & x^n \end{bmatrix}$$

$$= (x-x_0)(x-x_1)(x-x_2)....(x-x_{n-1}) \begin{bmatrix} 1 & x_0 & {x_0}^2 & \cdots & {x_0}^{n-1} \\ 1 & x_1 & {x_1}^2 & \cdots & {x_1}^{n-1} \\ 1 & x_2 & {x_2}^2 & \cdots & {x_2}^{n-1} \\ \vdots & & & \cdots & \\ 1 & x_{n-1} & {x_{n-1}}^2 & \cdots & {x_{n-1}}^{n-1} \end{bmatrix}$$

$$= (x-x_0)(x-x_1)(x-x_2)....(x-x_n) \vee (x_0, x_1, x_2, ..., x_{n-1})$$

To evaluate determinant in equation (3.2.1.3) we expand the determinant with respect to first column. The evaluation of determinant in equation (3.2.1.3) gives

$$p_n(x) \begin{bmatrix} 1 & x_0 & {x_0}^2 & \cdots & {x_0}^n \\ 1 & x_1 & {x_1}^2 & \cdots & {x_1}^n \\ 1 & x_2 & {x_2}^2 & \cdots & {x_2}^n \\ \vdots & & & \cdots & \\ 1 & x_n & {x_n}^2 & \cdots & {x_n}^n \end{bmatrix} - f(x_0) \begin{bmatrix} 1 & x & x^2 & \cdots & x^n \\ 1 & x_1 & {x_1}^2 & \cdots & {x_1}^n \\ 1 & x_2 & {x_2}^2 & \cdots & {x_2}^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & {x_n}^2 & \cdots & {x_n}^n \end{bmatrix} + ...$$

$$...+(-1)^{n+1}\begin{bmatrix} 1 & x & x^2 & \cdots & x^n \\ 1 & x_0 & {x_0}^2 & \cdots & {x_0}^n \\ 1 & x_1 & {x_1}^2 & \cdots & {x_1}^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n-1} & {x_{n-1}}^2 & \cdots & {x_{n-1}}^n \end{bmatrix} = 0$$

$$\therefore p_n(x)v(x_0,x_1,x_2,....,x_n) - f(x_0)v(x,x_1,x_2,x_3,...,x_n) + f(x_1)v(x,x_0,x_2,x_3,...,x_n)$$

$$....+(-1)^{n+1}v(x,x_0,x_1,x_2,....,x_{n-1}) = 0$$

$$\therefore p_n(x) = \frac{v(x,x_1,x_2,....,x_n)}{v(x_0,x_1,x_2,....,x_n)}f(x_0) - \frac{v(x,x_0,x_2,x_3,....,x_n)}{v(x_0,x_1,x_2,....,x_n)}f(x_1)$$

$$+\frac{v(x,x_0,x_1,x_3,x_4....,x_n)}{v(x_0,x_1,x_2,....,x_n)}f(x_2)+.....+(-1)^n\frac{v(x,x_0,x_1,....,x_{n-1})}{v(x_0,x_1,x_2,....,x_n)}f(x_n)$$

$$=\frac{v(x,x_1,x_2,x_3,....,x_n)}{v(x_0,x_1,x_2,....,x_n)}f(x_0)+\frac{v(x,x_0,x_2,x_3,....,x_n)}{v(x_1,x_0,x_2,....,x_n)}f(x_1)$$

$$+\frac{v(x,x_0,x_1,x_3,....,x_n)}{v(x_2,x_0,x_1,x_3,....,x_n)}f(x_2)+....+\frac{v(x,x_0,x_1,....,x_{n-1})}{v(x_n,x_0,x_1,x_2,....,x_{n-1})}f(x_n) \quad ... (3.2.1.3)$$

[If we interchange any row of determinant, the value of determinant changes its sign]

From the elimentary properties of determinant observe that

$$v(x,x_1,x_2,x_3,....,x_n) = (x-x_1)(x-x_2).....(x-x_n)v(x_1,x_2,...,x_n)$$

and $$v(x_0,x_1,x_2,x_3,....,x_n) = (x_0-x_1)(x_0-x_2).....(x_0-x_n)v(x_1,x_2,...,x_n)$$

Thus $$\frac{v(x,x_1,x_3,....,x_n)}{v(x_0,x_1,x_2,....,x_n)} = \frac{(x-x_1)(x-x_2)(x-x_3)....(x-x_n)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)....(x_0-x_n)} = \ell_0(x)$$ (say)

Similarly, $$\frac{v(x,x_0,x_2,x_3,....,x_n)}{v(x_1,x_0,x_2,x_3,....,x_n)} = \frac{(x-x_0)(x-x_2)(x-x_3)....(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)....(x_1-x_n)} = \ell_1(x)$$ (say)

In general,

$$\frac{v(x,x_0,x_1,....x_{i-1},x_{i+1},....,x_n)}{v(x_i,x_0,x_1,x_2,....,x_{i-1},x_{i+1},....,x_n)} = \frac{(x-x_0)(x-x_1)....(x-x_{i-1})(x-x_{i+1})....(x-x_n)}{(x_i-x_0)(x_i-x_1).....(x_i-x_{i-1})(x_i-x_{i+1}).....(x_i-x_n)}$$

$$= \ell_i(x) \quad \text{(say)}$$

Now equation (3.2.1.3) becomes

$$p_n(x) = \ell_0(x) f(x_0) + \ell_i(x) f(x_i) + ... + \ell_n(x) f(x_n)$$

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

$p_n(x) = \sum_{i=0}^{n} \ell_i(x) f(x_i)$ where

$\ell_i(x) = \dfrac{\prod_{k \neq i} (x - x_k)}{\prod_{k \neq i} (x_i - x_k)}$ is called Lagrarge interpolating polynomial.

3.1.2 Newtons Interpolating Polynomial

In the last section we have seen that if we have (n + 1) distinct points $a \leq x_0 < x_1 < x_2 ..... < x_n \leq b$ and the value of a function $f(x)$ at these points then we can fit a polynomial $p_n(x)$ of degree n such that

$$p_n(x_i) = f(x_i), \text{ i = 0, 1, 2, ..., n.}$$

Now represent this polynomial $p_n(x)$ in form

$$p_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + ... + a_n(x - x_0)(x - x_1)....(x - x_{n-1})$$

We want to calculate a_0, a_1, a_2, ..., a_n in such a way that

$$p_n(x_i) = f(x_i), \text{ i = 0, 1, 2, ..., n.}$$

$$f(x_0) = p_n(x_0) = a_0$$

$$f(x_1) = p_n(x_1) = a_0 + a_1(x_1 - x_0)$$

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

$$\vdots$$

$$f(x_n) = p_n(x_n) = a_0 + a_1(x_n - x_0) + a_2(x_n - x_0)(x_n - x_1) + ...$$

$$+ a_n(x_n - x_0)(x_n - x_1)....(x_n - x_{n-1})$$

By forward substitution we get, a_0, a_1, a_2, ..., a_n as follows,

$$a_0 = f(x_0)$$

$$a_1 = \frac{f(x_1) - a_0}{x_1 - x_0} = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f[x_0, x_1] \quad \text{(say)}$$

$$a_2 = \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)} - \frac{f(x_0)}{(x_2 - x_0)(x_2 - x_1)} - \frac{(x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)}\left[\frac{f(x_1)}{x_1 - x_0} - \frac{f(x_0)}{x_1 - x_0}\right]$$

$$\therefore a_2 = \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)} - \frac{f(x_1)}{(x_1 - x_0)(x_2 - x_1)} + f(x_0)\left[\frac{-1}{(x_2 - x_0)(x_2 - x_1)} + \frac{1}{(x_2 - x_1)(x_1 - x_0)}\right]$$

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

$$= f[x_0, x_1, x_2] \quad \text{(say)}$$

In general by induction we can prove that

$$a_i = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)....(x_0 - x_i)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)....(x_1 - x_i)} + ...$$

$$+ \frac{f(x_i)}{(x_i - x_0)(x_i - x_1)....(x_i - x_{i-1})}$$

$$= f[x_0, x_1, ..., x_i] \quad \text{(say)}$$

Thus we have

$$p_n(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + ...$$

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

The polynomial $p_n(x)$ is called Newton's divided difference interpolating polynomial.

The coefficients $a_1 = f[x_0, x_1]$, $a_2 = f[x_0, x_1, x_2]$, ..., $a_n = f[x_0, x_1, x_2, ..., x_n]$ are called Newton's divided differences.

3.1.2.1 Properties of Newton's Divided Differences

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

$$f[x_0, x_1] = f[x_1, x_0] \quad \text{(symmetry)}$$

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

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

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

(iii) $$f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

$$= \frac{1}{(x_2 - x_0)}\left[f(x) + \frac{(x_2 - x_0) f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \right]$$

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

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

$$= \frac{1}{(x_2 - x_0)}\left[-f[x_0, x_1] + f[x_1, x_2] \right]$$

$$= \frac{f[x_1, x_2] - f[x_0, x_1]}{(x_2 - x_0)}$$

In general,

$$f[x_0, x_1, x_2, ..., x_{k-1}, x_k] = \frac{f[x_1, x_2, ..., x_k] - f[x_0, x_1, x_2, ..., x_{k-1}]}{x_k - x_0}, \; k = 3, 4, ..., n$$

Thus Newton's divided differences are calculated as follows,

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

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0} \text{.....}$$

The Newton's divided differences may be calculaed with the help of following table.

$$\begin{array}{lll}
x_0 & f(x_0) & \\
 & \Big\}\Rightarrow \quad f[x_0, x_1] & \\
x_1 & f(x_1) & \Big\}\Rightarrow f[x_0, x_1, x_2] \\
 & \Big\}\Rightarrow \quad f[x_1, x_2] & \Big\}\Rightarrow f[x_0, x_1, x_2, x_3] \\
x_2 & f(x_2) & \Big\}\Rightarrow f[x_1, x_2, x_3] \qquad \Big\}\Rightarrow f[x_0, x_1, x_2, x_3, x_4] \\
 & \Big\}\Rightarrow \quad f[x_2, x_3] & \Big\}\Rightarrow f[x_1, x_2, x_3, x_4] \\
x_3 & f(x_3) & \Big\}\Rightarrow f[x_2, x_3, x_4] \\
 & \Big\}\Rightarrow \quad f[x_3, x_4] & \\
x_4 & f(x_4) &
\end{array}$$

(iv) Newtons divided differences are symmjetries in all the variables

$$f[x_0, x_1] = f[x_1, x_0]$$

$$f[x_0, x_1, x_2] = f[x_1, x_0, x_2] = f[x_2, x_1, x_0] \quad \text{etc.}$$

$$f[x_0, x_1, x_2, x_3] = f[x_1, x_0, x_2, x_3] = f[x_0, x_1, x_3, x_2] \quad \text{etc.}$$

3.1.3 Uniqueness of Interpolating Polynomials

In section 3.1.1 we have seen how to calculate Lagrange's interpolating polynomial $p_n(x)$ if the values of function $f(x)$ are known at (n + 1) nodes x_0, x_1, x_2, ..., x_n. In section 3.1.2 we have studied the evaluation of Newton's divided interpolating polynomial if the values of function $f(x)$ are known at (n + 1) points x_0, x_1, x_2, ..., x_n. We have seen that both the polynomials are polynomials of degree n. In this section we prove that the polynomials obtained by two different methods are same.

Suppose $p(x)$ is an interpolating polynomial of the given function $f(x)$ satisfying $f(x_i) = p(x_i)$, $i = 0, 1, 2, ..., n$ and $p^*(x)$ is another interpolating polynomial of same degree n satisfying $f(x_i) = p^*(x_i)$, i = 0, 1, 2, ..., n. We show that $p(x) = p^*(x)$.

Define $Q(x) = p(x) - p^*(x)$

Since $p(x)$ and $p^*(x)$ are polynomials of degree n. $Q(x)$ is a polynomial of degree at most n.

$$Q(x_i) = p(x_i) - p^*(x_i) = f(x_i) - f(x_i) = 0 \text{ , i = 0, 1, 2, ..., n}$$

Thus $Q(x)$ is a polynomial of degree less than or equal to n has (n + 1) distinct roots x_0, x_1, x_2, ..., x_n. But a polynomial of degree n has exactly n roots (real or complex). Therefore $Q(x)$ has at the most n roots. But $Q(x)$ has (n + 1) distinct roots x_0, x_1, x_2, ..., x_n. This implies that $Q(x) \equiv 0$.

Therefore $Q(x) = p(x) - p^*(x) \equiv 0$ i.e. $p(x) = p^*(x)$

Thus, the interpolating polynomials obtained in two different ways may be different in form, but are identical.

3.1.4 Truncation Error Bounds

In the last section we have seen that interpolating polynomial of degree n is unique Linear interpolation gives a polynomial $p_1(x)$ of degree one. The polynomial $p_1(x)$ coincides with the function $f(x)$ at x_0 and x_1. It deviates from $f(x)$ at all other points in the interval (x_0, x_1). This deviation is called the truncation error and is written as

$$E_1(f; x) = f(x) - p_1(x)$$

The expression for $E_1(f; x)$ is derived by using Rolle's theorem.

Rolle's Theorem : If $g(x)$ is a continuous function on some interval [a, b] and differentiable on (a, b) and if $g(a) = g(b) = 0$, then there is at least one point ξ inside (a, b) for which $g'(\xi) = 0$.

Error Bound for Linear Interpolation

Suppose $p_1(x)$ is a linear interpolating polynomial for the function $f(x)$. The polynomial $p_1(x)$ coincides with $f(x)$ at x_0 and x_1. Define a function $g(t)$ as

$$g(t) = f(t) - p_1(t) - \left[f(x) - p_1(x)\right]\frac{(t-x_0)(t-x_1)}{(x-x_0)(x-x_1)}$$

where $x \in (x_0, x_1)$ is a fixed point. Now function $g(t)$ is continuous and differentiable.

$$g(x_0) = f(x_0) - p_1(x_0) = 0 \text{ and } g(x_1) = f(x_1) - p_1(x_1) = 0$$

at $t = x$, $g(x) = f(x) - p_1(x_1) - \left[f(x) - p_1(x)\right] = 0$

Thus in the interval (x_0, x), $g(x_0) = g(x) = 0$ and by Rolle's theorem $\exists$ at least one point ξ, inside (x_0, x) for which $g'(\xi_1) = 0$.

Similarly in the interval (x, x_1), $g(x) = g(x_1) = 0$ and by Rolle's theorem $\exists$ at least one point ξ_2 inside (x, x_1) for which $g'(\xi_2) = 0$.

Thus $x_0 < \xi_1 < x < \xi_2 < x_1$ and $g'(\xi_1) = g'(\xi_2) = 0$.

Now on applying Rolle's theorem for $g'(t)$ on the interval (ξ_1, ξ_2) we get $\xi \in (\xi_1, \xi_2)$ such that $g''(\xi) = 0$. Since $x_0 < \xi_1 < x < \xi_2 < x_1$, $\xi \in (x_0, x_1)$.

Now differentiating $g(t)$ twice with respect to t we obtain,

$$g'(t) = f'(t) - p_1'(t) - \left[f(x) - p_1(x)\right]\frac{(t-x_1)+(t-x_0)}{(x-x_0)(x-x_1)}$$

$$g''(t) = f''(t) - p_1''(t) - \left[f(x) - p_1(x)\right]\frac{2}{(x-x_0)(x-x_1)}$$

$$= f''(t) - \frac{2\left[f(x) - p_1(x)\right]}{(x-x_0)(x-x_1)}$$

$[\because p_1(t)$ is polynomial of degree 1 $p_1''(t) = 0]$

$$g''(\xi) = 0 \Rightarrow f''(\xi) = \frac{2\left[f(x) - p_1(x)\right]}{(x-x_0)(x-x_1)}$$

Thus $f(x)-p_1(x)=(x-x_0)(x-x_1)\dfrac{f''(\xi)}{2}$

Therefore, the truncation error in linear interpolation is given by,

$$E_1(f;x)=f(x)-p_1(x)=\frac{1}{2}(x-x_0)(x-x_1)f''(\xi)$$

If $|f''(x)|\le M_2 \quad \forall x\in[x_0,x_1]$ then,

$$|f(x)-p_1(x)|=\frac{1}{2}|(x-x_0)(x-x_1)f''(\xi)|$$

$$\le\frac{1}{2}\underset{x_0\le x\le x_1}{Max}|(x-x_0)(x-x_1)|M_2$$

Let $w(x)=(x-x_0)(x-x_1)$ then $w'(x)=(x-x_0)+(x-x_1)$

and $w'(x)=0\Rightarrow x=\dfrac{x_0+x_1}{2}$. Hence maximum value of $|(x-x_0)(x-x_1)|$ is attained at $x=\dfrac{x_0+x_1}{2}$.

$$\underset{x}{Max}|(x-x_0)(x-x_1)|=\left|\left(\frac{x_0+x_1}{2}-x_0\right)\left(\frac{x_0+x_1}{2}-x_1\right)\right|=\frac{(x_1-x_0)^2}{4}$$

Thus $|f(x)-p_1(x)|\le\dfrac{1}{2}\dfrac{(x_1-x_0)^2}{4}M_2$

Thus bound for truncation error $E_1(f;x)$ is $\dfrac{1}{2}\dfrac{(x_1-x_0)^2}{4}M_2$.

Thus the truncation error

$$|E_1(f;x)|\le\frac{1}{2}\frac{(x_1-x_0)^2}{4}\underset{x\in(x_0,x_1)}{Max}|f''(x)| \qquad \text{..... (3.1.4.1)}$$

Truncation Error for higher order interpolating polynomial

Interpolating polynomial $p_n(x)$ of degree n coincides with (n + 1) times differentiable function $f(x)$ at $\{x_0, x_1, x_2, ..., x_n\}$ (n + 1) points. The truncation error for this polynomial approximation is

$$E_n(f;x) = f(x) - p_n(x)$$

For $x \in (a,b)$ and $x \neq x_i$, i = 0, 1, 2, ...m define

$$g(t) = f(t) - p_n(t) - \left[f(x) - p_n(x)\right]\frac{(t-x_0)(t-x_1).....(t-x_n)}{(x-x_0)(x-x_1).....(x-x_n)} \quad .... (3.1.4.2)$$

Observe that $g(x_i) = 0$, $\forall$ i = 0, 1, 2, ..., n.

In the interval (x_0, x_1), $g(x_0) = g(x_i) = 0$. Function *g* is continuous and differentiable. Therefore by Rolle's theorem $\exists \xi_1 \in (x_0, x_1)$ such that $g'(\xi_1) = 0$.

Similarly in the interval $(x_1, x_2) \exists \xi_2$ such that $g'(\xi_2) = 0$. In general $\exists \xi_i \in (x_{i-1}, x_i)$ such that $g'(\xi_i) = 0$.

Now $g'(\xi_1) = g'(\xi_2) = 0$. g' is continuous and differentiable therefore by Rolle's theorem

$\exists \eta_1 \in (\xi_1, \xi_2)$ such that $g''(\eta_1) = 0$.

$\exists \eta_2 \in (\xi_2, \xi_3)$ such that $g''(\eta_2) = 0$.

In general $\exists \eta_i (\xi_i, \xi_{i+1})$ such that $g''(\eta_i) = 0$, i = 1, 2, ..., (n – 1).

Thus repeated application of Rolle's theorem for $g(t)$, $g'(t)$, $g''(t)$, ..., $g^{(n)}(t)$ gives $\xi \in (a,b)$ such that $g^{(n+1)}(\xi) = 0$.

Differentiating function *g* defined in equation (3.1.4.2), (n + 1) times with respect to *t*, we get,

$$g^{(n+1)}(t) = f^{(n+1)}(t) - p_n{}^{(n+1)}(t) - \frac{(n+1)!\left[f(x) - p_n(x)\right]}{(x-x_0)(x-x_1).....(x-x_n)}$$

Since p_n is a polynomial of degree (n) $p_n{}^{(n+1)}(t) = 0$.

$$g^{(n+1)}(\xi) = 0 \Rightarrow f^{(n+1)}(\xi) - \frac{(n+1)!\left[f(x) - p_n(x)\right]}{(x-x_0)(x-x_1).....(x-x_n)} = 0$$

and $\left[f(x)-p_n(x)\right]=\dfrac{(x-x_0)(x-x_1).....(x-x_n)}{(n+1)!}f^{(n+1)}(\xi)$

Thus the truncation error in intempolating polynomial is given by

$$E_n(f;x)=f(x)-p_n(x)=\frac{(x-x_0)(x-x_1).....(x-x_n)}{(n+1)!}f^{(n+1)}(\xi)$$

3.2 Finite Difference Operators

Let the tabular points $x_0, x_1, x_2, ..., x_n$ be equally spaced i.e. $x_1-x_0=h$, $x_2-x_1=h$,...., $x_n-x_{n-1}=h$.

In general $x_i=x_0+ih$, i = 1, 2, 3, ..., n.

Define the following operators.

(i) Forward differene operator $\Delta f(x_i)=f(x_{i+1})-f(x_i)$.

(ii) Backward difference operator $\nabla f(x_i)=f(x_i)-f(x_{i-1})$.

(iii) Central difference operator $\delta f(x_i)=f\left(x_i+\dfrac{h}{2}\right)-f\left(x_i-\dfrac{h}{2}\right)$.

(iv) Average operator $\mu f(x_i)=\dfrac{1}{2}\left[f\left(x_i-\dfrac{h}{2}\right)+f\left(x_i+\dfrac{h}{2}\right)\right]$.

(v) Shift operator $Ef(x_i)=f(x_i+h)$.

3.2.1 Relations between finite difference operators

(i) $\Delta f(x_i)=f(x_{i+1})-f(x_i)=\nabla f(x_{i+1})\Rightarrow \Delta f_i=\nabla f_{i+1}$

where $f(x_i)$ is denoted by f_i.

(ii) $\Delta f(x_i)=f(x_i+h)-f(x_i)=f\left(x_i+\dfrac{h}{2}+\dfrac{h}{2}\right)-f\left(x_i+\dfrac{h}{2}-\dfrac{h}{2}\right)$

$$=\delta f\left(x_i+\frac{h}{2}\right)$$

$$\therefore \Delta f(x_0+ih)=\delta f\left(x_0+\left(i+\frac{1}{2}\right)h\right)\Rightarrow \Delta f_i=\delta f_{i+\frac{1}{2}}$$

(iii) $\Delta f(x_i) = f(x_{i+1}) - f(x_i) \Rightarrow f(x_i + h) - f(x_i) = Ef(x_i) - f(x_i)$

$= (E-1) f(x_i)$

Thus $\Delta f(x_i) = (E-1) f(x_i) \Rightarrow \Delta \equiv E - 1.$

(iv) $\nabla f(x_i) = f(x_i) - f(x_{i-1})$

Since $Ef(x_i) = f(x_{i+1}) \Rightarrow E^{-1} f(x_{i+1}) = f(x_i)$

Thus $\nabla f(x_i) = f(x_i) - E^{-1} f(x_i)$

$= (1 - E^{-1}) f(x_i)$

Therefore $\nabla \equiv 1 - E^{-1}$

(v) $E^1 f(x_i) = f(x_i + h)$

$\therefore E^{1/2} f(x_i) = f\left(x_i + \frac{1}{2}h\right)$ and $E^{-1/2} f(x_i) = f\left(x_i - \frac{1}{2}h\right)$

But $f\left(x_i + \frac{h}{2}\right) - f\left(x_i - \frac{h}{2}\right) = \delta f(x_i)$

$\therefore E^{1/2} f(x_i) - E^{-1/2} f(x_i) = \delta f(x_i) \Rightarrow E^{1/2} - E^{-1/2} = \delta$

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

$= \frac{1}{2}\left[E^{1/2} f(x_i) + E^{-1/2} f(x_i)\right]$

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

(vii) $E\left[Ef(x_i)\right] = Ef(x_i + h) = f(x_i + h + h) = f(x_i + 2h)$

In general $E^n f(x_i) = f(x_i + nh)$

$\Delta\left[\Delta f(x_i)\right] = \Delta\left[f(x_i + h) - f(x_i)\right]$

$= \Delta f(x_i + h) - \Delta f(x_i)$

$= f(x_i + h + h) - f(x_i + h) - \left[f(x_i + h) - f(x_i)\right]$

$$= f(x_i + 2h) - 2f(x_i + h) + f(x_i)$$

$$\Delta f(x_i) = f(x_i + h) - f(x_i)$$

$$= (E-1) f(x_i)$$

$$\Rightarrow \Delta^n f(x_i) = (E-1)^n f(x_i) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} f(x_{i+n-k})$$

$$\Delta^n f(x_i) = (1 - E^{-1})^n f(x_i) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} f(x_{i-k})$$

Similarly $\delta^n f(x_i) = \left(E^{1/2} - E^{-1/2}\right)^n f(x_i)$

$$= \sum (-1)^k \binom{n}{k} \left(E^{1/2}\right)^{n-k} \left(E^{-1/2}\right)^k f(x_i)$$

$$= \sum (-1)^k \binom{n}{k} E^{\frac{n-k}{2}} E^{-\frac{k}{2}} f(x_i)$$

$$= \sum (-1)^k \binom{n}{k} E^{\frac{n-k}{2}} f\left(x_i - \frac{k}{2}h\right)$$

$$= \sum (-1)^k \binom{n}{k} f\left(x_i - \frac{k}{2}h + \left(\frac{n-k}{2}\right)h\right)$$

$$= \sum_{k=0}^{n} (-1)^k \binom{n}{k} f\left(x_i + \left(\frac{n}{2} - k\right)h\right) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} f_{i+\frac{n}{2}-k}$$

3.2.2 Relations Between Differences and Derivative

In section 3.2.1 we have derived relations between different finite difference operators. If the relation between derivative and any finite difference operator is obtained then the relation between derivative and any finite difference operator is known.

Suppose f is smooth function

$$\Delta f(x_i) = f(x_i + h) - f(x_i)$$

$$= f(x_i) + hf'(x_i) + \frac{h^2}{2!} f''(x_i + \theta h) - f(x_i)$$

$$= hf'(x_i) + \frac{h^2}{2!} f''(x_i + \theta h)$$

Therefore, $$f'(x_i) = \frac{1}{h}\Delta f(x_i) - \frac{h}{2} f''(x_i + \theta h)$$

$$= \frac{1}{h}\Delta f(x_i) + 0(h)$$

Thus $$Df(x_i) = \frac{\Delta}{h} f(x_i) + 0(h)$$

Since finite difference operators are defined for discrete equidistant points and derivative is defined for containuous functions, exact relation between derivative and differences cannot be obtained and we get an error.

3.3 Numerical Differenciation

In this section we discuss methods for approximating the derivatives of a given function. Numerical differentiation methods are developed by using one of the following techniques.

(i) Methods based on interpolation

In this method the function values or table values are used to approximate a function by a polynomial (by using Lagranges interpolation or Newton's divided difference formula) and this polynomial is differentiated to get the derivative of a function.

(ii) Methods Based on Finite Difference Operators

In this method derivative is obtained by considerng Newton's forward difference operator or Newton's backward difference operator or using shift operator. Since there is relation between every pair of finite difference operators, derivatives are obtained by using any finite difference operator.

(ii) Methods Based on Undetermined Coefficients

In this method function is written in the form of linear combination of values of a function at some points and the coefficients of this linear combination are obtained by using Taylor series expansion. Now we discuss each of this methods in detail.

Methods Based on Interpolation :

Given the values of function $f(x)$ at a set of points $x_0, x_1, x_2, ..., x_n$, obtain an interpolation polynomial $p_n(x)$ by using any interpolation technique (Lagrenge interpolation or Newton's divided difference formula). Thus

$$f(x) = p_n(x) + E_n(f;x)$$

where $E_n(f;x)$ is an error in the approximation.

In section 3.1.4 we have seen that

$$E_n(f;x) = \frac{(x-x_0)(x-x_1).....(x-x_n)}{(n+1)!} f^{(n+1)}(\xi)$$

$$f'(x) = p_n'(x) + E_n'(f;x)$$

In general the values of $p_n^{(r)}(x_k)$ gives the approximate values at the point x_k.

The quantity $E_n^{(r)}(f;x) = f^{(r)}(x) - p_n^{(r)}(x)$ is called the error of approximation in the rth order derivative at any point x.

Example : The following data for the function $f(x) = x^4$ is given.

x	0.2	0.3	0.4
$f(x)$	0.0016	0.0081	0.0256

Find $f'(0.4)$ and $f''(0.4)$ using quadratic interpolation. Compare the results with exact solution obtain the bound on the truncation error.

Answer : Using Lagrange interpolation we have

$$p_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2)$$

$$= \frac{(x-0.3)(x-0.4)}{(0.2-0.3)(0.2-0.4)}(0.0016) + \frac{(x-0.2)(x-0.4)}{(0.3-0.2)(0.3-0.4)}(0.008)$$

$$+ \frac{(x-0.2)(x-0.3)}{(0.4-0.2)(0.4-0.3)}(0.0256)$$

$$= (x^2 - 0.7x + 0.12)\frac{0.0016}{0.02} + (x^2 - 0.6x + 0.08)\frac{(0.0081)}{-0.01}$$

$$+ (x^2 - 0.5x + 0.06)\frac{(0.0256)}{0.02}$$

$$= (0.08)(x^2 - 0.7x + 0.12) - 0.81(x^2 - 0.6x + 0.08) + (1.28)(x^2 - 0.5x + 0.06)$$

$$= 0.55x^2 - 0.21x + 0.0216$$

$$f'(0.4) \cong p_2'(0.4) = (0.55)(2)(0.4) - 0.21$$

$$= 0.440 - 0.21 = 0.230$$

$$f''(0.4) \cong p_2''(0.4) = (0.55)(2) = 1.10$$

Thus the approximate solutions are $f'(0.4) \cong 0.23$, $f''(0.4) = 1.1$.

The exact solutions are

$$f'(0.4) = 4(0.4)^3 = 0.256$$

$$f''(0.4) = 4(3)x^2 = 1.92$$

$$f'(x) = 4x^3,\ f''(x) = 12x^2,\ f'''(x) = 24x$$

$$\therefore M_3 = \max_{0.2 \le x \le 0.4} |f'''(x)| = 24(0.4) = 9.6$$

$$|E_2'(0.4)| \le \frac{h^2}{3} M_3 = \frac{(0.1)^2}{3} 9.6 = (0.01)(3.2) = 0.032$$

$$|E_2''(0.4)| \le hM_3 = (0.1)(9.6) = 0.96$$

Methods Based on Finite Difference Operators

In this section we derive relation between derivative and finite difference operators.

$$Ef(x) = f(x + h)$$

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

$$= f(x) + hDf(x) + \frac{h^2}{2!} D^2 f(x) + ... \qquad \left[D = \frac{d}{dx} \right]$$

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

$$= e^{hD} f(x)$$

Thus $E = e^{hD}$

$$\therefore hD = \log E$$

$$= \log(1+\Delta) = \Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{2} - ....$$

$$= -\log(1-\nabla) = \Delta + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + ...$$

$$h^2D^2 = \left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} + ...\right)^2$$

$$= \left(\nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + ...\right)^2$$

Keeping only first term in each of the above series we have

$$f'(x_k) \cong \frac{f(x_{k+1}) - f(x_k)}{h}$$ Forward differences

$$\cong \frac{f(x_k) - f(x_{k-1})}{h}$$ Backward differenes

$$\cong \frac{f(x_{k+1}) - f(x_{k-1})}{2h}$$ Central differences

Similarly,

$$f''(x_k) \cong \frac{f(x_{k+2}) - 2f(x_{k+1}) + f(x)}{h^2}$$ Forward differences

$$\cong \frac{f(x_k) - 2f(x_{k-1}) + f(x_{k-2})}{h^2}$$ Backward differences

$$\cong \frac{f(x_{k+1}) - 2f(x_k) + f(x_{k-1})}{h^2}$$ Central differences

First two expressions for both the representations are of first order whereas the third representation is of second order.

Methods Based on Undetermined Coeficients

In this method we write $f^{(r)}(x)$ as a linear combination of the value of $f(x)$ at an arbitrary chosen set of tabular points. Determine the coefficients in linear combination by using Taylor series expansions of function at some point and by equating the equal powers of derivatives.

For example, assume that the tabular points are equispaced with step length h. Write

$$h^r f^{(r)}(x_k) = \sum_{p=-v}^{v} a_p f(x_{k+p})$$

The local truncation error is defined as

$$E^{(r)}(x_k) = \frac{1}{h^r}\left[h^r f^{(r)}(x_k) - \sum_{p=-v}^{v} a_p f(x_{k+p}) \right]$$

The coefficients a_p are determined by requiring the method to be of particular order.

Example : A differentiation rule of the form

$$f'(x_0) = \alpha_0 f(x_0) + \alpha_1 f(x_1) + \alpha_2 f(x_2)$$

where $x_k = x_0 + kh$, is given. Find the values of α_0, α_1, α_2 so that the rule is exact for $f \in p_2$. Find the error term.

Answer :

$$\begin{aligned} f'(x_0) &= \alpha_0 f(x_0) + \alpha_1 f(x_1) + \alpha_2 f(x_2) \\ &= \alpha_0 f(x_0) + \alpha_1 f(x_0 + h) + \alpha_2 f(x_0 + 2h) \\ &= \alpha_0 f(x_0) + \alpha_1 \left\{ f(x_0) + hf'(x_0) + \frac{h^2}{2!} f''(x_0) + \frac{h^3}{3!} f'''(x_0) + ... \right\} \\ &\quad + \alpha_2 \left\{ f(x_0) + (2h) f'(x_0) + \frac{(2h)^2}{2!} f''(x_0) + \frac{(2h)^3}{3!} f'''(x_0) + ... \right\} \\ &= (\alpha_0 + \alpha_1 + \alpha_2) f(x_0) + \left[\alpha_1 h + \alpha_2 (2h) \right] f'(x_0) \\ &\quad + \frac{h^2}{2!} \left[\alpha_1 + 4\alpha_2 \right] f''(x_0) + \frac{h^3}{3!} \left[\alpha_1 + 2^3 \alpha_2 \right] f'''(x_0) + ... \end{aligned}$$

On equating the coefficients of equal powers of derivatives we have,

$$f(x_0): \alpha_0 + \alpha_1 + \alpha_2 = 0$$

$$f'(x_0): (\alpha_1 + 2\alpha_2)h = 1$$

$$f''(x_0): \alpha_1 + 4\alpha_2 = 0$$

$$f'''(x_0): \alpha_1 + 8\alpha_2 = 0....$$

Since above system of equations contain three arbitrary values α_0, α_1, α_2 we can consider only first three equations to obtain the values of α_0, α_1, α_2. From first three equations we have the following

$$\alpha_1 + 4\alpha_2 = 0 \Rightarrow \alpha_1 = -4\alpha_2$$

$$(\alpha_1 + 2\alpha_2)h = 1 \Rightarrow (-4\alpha_2 + 2\alpha_2)h = 1 \Rightarrow \alpha_2 = \frac{-1}{2h}$$

Thus $\alpha_1 = \frac{2}{h}$, $\alpha_2 = \frac{-1}{2h}$ and $\alpha_0 + \alpha_1 + \alpha_2 = 0$ gives

$$\alpha_0 = -\alpha_1 - \alpha_2 = -\frac{2}{h} + \frac{1}{2h} = \frac{-3}{2h}$$

The leading term in the error expression is

$$(\alpha_1 + 8\alpha_2)\frac{h^3}{3!}f'''(\xi) = \left(\frac{2}{h} - \frac{8}{2h}\right)f'''(\xi)\frac{h^3}{3!}$$

$$= \left(\frac{-4}{2h}\right)\frac{h^3}{6}f'''(\xi)$$

$$= -\frac{h^2}{3}f'''(\xi)$$

Thus the error term $E = -\frac{h^2}{3}f'''(\xi)$

Since the error term contains third derivative, the method is exact for the functions whose third derivative is zero. i.e. method is exact for $f \in p_2$.

3.4 Numerical Integration

The general problem of numerical integration is to find the numerical value of the integral

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

We assume that $w(x)$ and $f(x)$ are Riemann integrable functions on [a, b]. $w(x) > 0$ defined on [a, b] is called weight function. The integral I is written as a finite linear combination of values of $f(x)$ in the form

$$I = \int_a^b w(x) f(x) dx \cong \sum_{k=0}^{n} \lambda_k f(x_k) \qquad \text{..... (3.4.1)}$$

$x_k \in [a, b]$ are called nodes and are distributed on the interval [a, b] with $x_{k-1} < x_k$, k = 1, 2, 3, ...n. The coefficients λ_k, k = 1, 2, 3, ...n are called weighs of integration rule or quadrature formula (3.4.1).

The error is given by

$$R_n = \int_a^b w(x) f(x) dx - \sum_{k=0}^{n} \lambda_k f(x_k)$$

Definition : An integration method (3.4.1) is said to be of order p, if it produces exact results ($R_n = 0$) for all polynomials of degree less than or equal to p.

Methods Based on Interpolation

Given the (n + 1) nodal values and the corresponding values of $f(x_k)$, the Lagrange interpolating polynomial is given by

$$f(x) = \sum_{k=0}^{n} \ell_k(x) f(x_k) + \pi(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}; \quad x_0 < \xi < x_n$$

where $\ell_k(x) = \dfrac{\pi(x)}{(x - x_k)\pi'(x_k)}$ and $\pi(x) = (x - x_0)(x - x_1).....(x - x_n)$

From equation (3.4.1) we get,

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

$$= \int_a^b w(x)\left[\sum_{k=0}^{n} \ell_k(x) f(x_k) + \frac{\pi(x) f^{(n+1)}(\xi)}{(n+1)!}\right] dx$$

$$= \sum_{k=0}^{n}\left[\int_a^b w(x)\,\ell_k(x)\,dx\right] f(x_k) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\int_a^b \pi(x)\,w(x)\,dx$$

$$= \sum_{k=0}^{n} \lambda_k f(x_k) + R_n \qquad \text{..... (3.4.2)}$$

where $\lambda_k = \int_a^b w(x)\,\ell_k(x)\,dx$ and error $R_n = \frac{1}{(n+1)!}\int_a^b \pi(x)\,w(x)\,f^{(n+1)}(\xi)\,dx$

The error $R_n = \frac{1}{(n+1)!}\int_a^b \pi(x)\,w(x)\,f^{(n+1)}(\xi)\,dx$

$$= \frac{f^{(n+1)}(\eta)}{(n+1)!}\int_a^b w(x)|\pi(x)|\,dx \quad \text{for some } \eta \in (a,b).$$

$$|R_n| \le \frac{|f^{(n+1)}(\eta)|}{(n+1)!}\int_a^b w(x)|\pi(x)|\,dx$$

The error term can also be determined by

$$R_n = \frac{C}{(n+1)!} f^{(n+1)}(\eta)$$

where $C = \int_a^b w(x)\,x^{n+1}dx - \sum_{k=0}^{n} \lambda_k x_k^{\,n+1}$

C is called error constant. If C is zero for $f(x) = x^{n+1}$ then we take the next term $f(x) = x^{n+2}$ and naturally the error will be zero for all polynomials of degree (n + 1).

Thus our aim is to determine the weights λ_k and nodal pionts x_k such that the error term R_n is minimum. For simplicity we assume that all the nodal points are equispaced and end points are nodal points.

The approximate value of the integral is written as

$$I = \int_a^b w(x)\,f(x)\,dx = \sum_{k=0}^{n} \lambda_k f(x_k)$$

Newton Cotes Methods

In this method we assume that $w(x) \equiv 1$ and the nodes are equispaced with $x_0 = a$, $x_n = b$ and $h = \dfrac{b-a}{n}$. The weights λ_k are called cotes numbers. We calculate the weights λ_k by using Lagrange interpolation. From equation (3.4.2) we know that

$$I = \int_a^b w(x) f(x)\, dx = \sum_{k=0}^{n} \lambda_k f(x_k) + R_n$$

$$\lambda_k = \int_a^b w(x) \ell_k(x)\, dx \text{ and } R_n = \frac{f^{(n+1)}(\eta)}{(n+1)!} \int_a^b \pi(x) w(x)\, dx .$$

$$\lambda_k = \int_a^b w(x) \ell_k(x)\, dx$$

Now $$\ell_k(x) = \frac{(x-x_0)(x-x_1).....(x-x_{k-1})(x-x_{k+1}).....(x-x_n)}{(x_k-x_0)(x_k-x_1)....(x_k-x_{k-1})(x_k-x_{k+1}).....(x_k-x_n)}$$

Since all nodes are equispaced, $x_i = x_0 + ih$ and

$$\ell_k(x) = \frac{(x-x_0)(x-x_1).....(x-x_{k-1})(x-x_{k+1}).....(x-x_n)}{(kh)(k-1)h(k-2)h....h(-h)(-2h).....(-(n-k)h)}$$

$$= \frac{(x-x_0)(x-x_1).....(x-x_{k-1})(x-x_{k+1}).....(x-x_n)}{k!h^k(-1)^{n-k}(n-k)!h^{n-k}}$$

Substitute $x = x_0 + sh$ then $x - x_i = x_0 + sh - (x_0 + ih) = (s-i)h$.

$$\ell_k(x) = \frac{sh(s-1)h(s-2)h....(s-k+1)h(s-k-1)h....(s-n)h}{k!(n-k)!(-1)^{n-k}h^n}$$

$$= \frac{s(s-1)(s-2)....(s-k+1)(s-k-1)....(s-n)h^n}{k!(n-k)!(-1)^{n-k}h^n}$$

Since $w(x) = 1$ and $x = x_0 + sh$, $dx = hds$.

$x = a = x_0 \Rightarrow s = 0$ and $x = b = x_n = x_0 + nh = x_0 + sh \Rightarrow s = n$.

$$\therefore \lambda_k = \int_a^b w(x) \ell_k(x)\, dx = \int_{x_0}^{x_n} \ell_k(x)\, dx$$

$$= \int_0^n \frac{s(s-1)(s-2)....(s-k+1)(s-k-1)....(s-n)}{k!(n-k)!(-1)^{n-k}} hds \qquad \text{..... (3.4.3)}$$

$$\pi(x) = (x-x_0)(x-x_1).....(x-x_n)$$

$$= sh(s-1)h(s-2)h.....(s-n)h$$

$$\pi(x) = h^{n+1}s(s-1)(s-2)(s-3).....(s-n) \text{ where } x = x_0 + sh$$

In equation (3.4.2) we have

$$R_n = \frac{f^{(n+1)}(\eta)}{(n+1)!}\int_a^b \pi(x)\, w(x)\, dx$$

$$= \frac{f^{(n+1)}(\eta)}{(n+1)!}\int_0^n h^{n+1}s(s-1)(s-2)(s-3).....(s-n)\cdot hds$$

$$= \frac{h^{n+2}f^{(n+1)}(\eta)}{(n+1)!}\int_0^n s(s-1)(s-2).....(s-n)\cdot ds \qquad \text{... (3.4.4)}$$

Thus from equation (3.4.3), (3.4.4) and (3.4.2) we have

$$I = \int_a^b f(x)\, dx = \sum_{k=0}^{n} \lambda_k f(x_k) + R_n$$

where $\lambda_k = \int_0^n \frac{s(s-1)(s-2)....(s-k+1)(s-k-1)....(s-n)hds}{k!(n-k)!(-1)^{n-k}}$

and $R_n = \frac{h^{n+2}f^{(n+1)}(\eta)}{(n+1)!}\int_0^n s(s-1)(s-2).....(s-n)ds$ ..... (3.4.5)

From equation (3.4.5) for different values of n we get different numerical methods of integration.

Case 1 : n = 1 : Trapezoidal Rule

$$\int_a^b f(x)\, dx = \lambda_0 f(x_0) + \lambda_1 f(x_1)$$

$$= \lambda_0 f(a) + \lambda_1 f(b)$$

$$\lambda_0 = \frac{(-1)^{1-0}}{0!(1-0)!} h \int_0^1 (s-1)ds = -h \frac{(s-1)^2}{2}\Big|_0^1 = \frac{h}{2}$$

$$\lambda_1 = \frac{(-1)^{1-1}}{1!(1-1)!} h \int_0^1 s ds = h \cdot \frac{s^2}{2}\Big|_0^1 = \frac{h}{2}$$

Since n = 1, $h = \frac{b-a}{n} = (b-a)$ and we get

$$\int_a^b f(x)\,dx = \frac{(b-a)}{2} f(a) + \frac{(b-a)}{2} f(b) = \frac{(b-a)}{2}\left[f(a) + f(b)\right]$$

$$\int_a^b f(x)\,dx = \frac{(b-a)}{2}\left[f(a) + f(b)\right] \qquad \text{..... (3.4.6)}$$

equation (3.4.6) is called Trapezoidal rule.

From equation (3.4.5) we get error in Trapezoide rule as

$$R_1 = \frac{h^3}{2!} \int_0^1 s(s-1)dx \cdot f''(\xi)$$

$$= \frac{h^3}{2}\left[\frac{s^3}{3} - \frac{s^2}{2}\right]_0^1 f''(\xi)$$

$$= \frac{h^3}{2}\left(\frac{1}{3} - \frac{1}{2}\right) f''(\xi)$$

$$= -\frac{h^3}{12} f''(\xi) \qquad \text{where } \xi \in (a,b)$$

Alternatively $R_1 = \frac{C}{2!} f''(\eta)$ since trapezoidal rule is exact for a polynomial of degree one we calculate C for $f(x) = x^2$.

$$C = \int_a^b x^2 dx - \frac{(b-a)}{2}\left[f(a) + f(b)\right]$$

$$= \frac{x^3}{3}\Big|_a^b - \frac{(b-a)}{2}\left[a^2 + b^2\right]$$

$$= \frac{b^3 - a^3}{3} - \frac{(b-a)(a^2+b^2)}{2}$$

$$= (b-a)\left[\frac{b^2+ab+a^2}{3} - \frac{a^2+b^2}{2}\right]$$

$$= \frac{(b-a)}{6}[2b^2 + 2ab + 2a^2 - 3a^2 - 3b^2]$$

$$= -\frac{(b-a)^3}{6}$$

The error

$$R_1 = \frac{C}{2!} f''(\eta) = -\frac{(b-a)^3}{12} f''(\eta) \qquad \text{where } \eta \in (a,b)$$

Thus the trapezoidal rule for numerical integration is

$$\int_a^b f(x)\,dx = \frac{(b-a)}{2}[f(a) + f(b)]$$

and the error in the formula is

$$R_1 = -\frac{(b-a)^3}{12} f''(\eta)$$

To determine the error bound we calculate maximum absolute value of R_1 by evaluating values of $f''(\eta)$.

Case 2 : n = 2 : Simpson's Rule

Here $h = \frac{b-a}{2}$. $x_0 = a$, $x_1 = a + \frac{b-a}{2} = \frac{a+b}{2}$, $x_2 = b$.

From equation (3.4.5) we have

$$\int_a^b f(x)\,dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) + \lambda_2 f(x_2)$$

$$\lambda_0 = \frac{(-1)^{2-0}}{0!(2-0)!} h \int_0^2 (s-1)(s-2)\,ds$$

$$= \frac{h}{2}\left[(s-1)\frac{(s-2)^2}{2} - \frac{(s-2)^3}{6} \right]_0^2$$

$$= -\frac{h}{2}\left[(-1)\frac{4}{2} - \frac{(-2)^3}{6} \right]$$

$$= -\frac{h}{2}\left[-2 + \frac{4}{3} \right]$$

$$= \frac{h}{3}$$

$$\lambda_1 = \frac{(-1)^{2-1}}{1!(2-1)!} h \int_0^2 s(s-2)ds$$

$$= -h\left[\frac{s^3}{3} - \frac{2s^2}{2} \right]_0^2$$

$$= -h\left[\frac{8}{3} - 4 \right]$$

$$= \frac{4h}{3}$$

$$\lambda_2 = \frac{(-1)^{2-2}}{2!(2-2)!} h \int_0^2 s(s-1)ds$$

$$= \frac{h}{2}\left[\frac{s^3}{3} - \frac{s^2}{2} \right]_0^2$$

$$= \frac{h}{2}\left[\frac{8}{3} - \frac{4}{2} \right]$$

$$= \frac{h}{3}$$

Thus we have

$$\int_a^b f(x)\,dx = \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2} \right) + f(b) \right]$$

Which is called the Simpson's rule or Simpson's $\frac{1}{3}$ rd rule. Since three observations are used to derive the formula, the formula is exact for all polynomials upto order two. The error may occur for $f(x)=x^3$. Let us calculate C for $f(x)=x^3$.

$$C=\int_a^b x^3dx-\frac{(b-a)}{6}\left[a^3+4\left(\frac{a+b}{2}\right)^3+b^3\right]$$

$$=\frac{1}{4}(b^4-a^4)-\frac{(b-a)}{12}\left[2a^3+(a+b)^3+2b^3\right]$$

$$=\frac{b^4-a^4}{4}-\frac{(b-a)}{12}\left[2a^3+a^3+2a^2b+3ab^2+b^3+2b^3\right]$$

$$=\frac{b^4-a^4}{4}-\frac{(b-a)}{4}\left[a^3+a^2b+ab^2+b^3\right]$$

$$=\frac{b^4-a^4}{4}-\frac{(b-a)}{4}\left[(a+b)(a^2+b^2)\right]$$

$$=0$$

This shows that method is exact for polynomials of degree upto three. Let use caulculate C for $f(x)=x^4$.

$$C=\int_a^b x^4dx-\frac{(b-a)}{6}\left[a^4+4\left(\frac{a+b}{2}\right)^4+b^4\right]$$

$$=\frac{1}{5}(b^5-a^5)-\frac{(b-a)}{6}\left[a^4+\frac{1}{4}(a+b)^4+b^4\right]$$

$$C=\frac{1}{5}(b^5-a^5)-\frac{(b-a)}{24}\left[4a^4+a^4+4a^3b+6a^2b^2+4ab^3+b^4+4b^4\right]$$

$$=\frac{(b-a)}{120}\left[24(b^4+ab^3+a^2b^2+a^3b+a^4)-5(5a^4+4a^3b+6a^2b^2+4ab^3+5b^4)\right]$$

$$=\frac{(b-a)}{120}\left[-a^4-b^4+4ab^3-6a^2b^2+4a^3b\right]$$

$$=-\frac{(b-a)(b-a)^4}{120}$$

$$= -\frac{(b-a)^5}{120}$$

Then error in Simpson's $\frac{1}{3}$ rd rule is

$$R = -\frac{(b-a)^5}{4!(120)} f^{(iv)}(\eta)$$

Since $b-a = 2h$ the error term

$$R = -\frac{(2h)^5}{24(120)} f^{(iv)}(\eta)$$

$$= -\frac{32h^5}{(24)(120)} f^{(iv)}(\eta)$$

$$= -\frac{h^5}{(3)(30)} f^{(iv)}(\eta)$$

$$= -\frac{h^5}{90} f^{(iv)}(\eta) \qquad \text{where } \eta \in (a,b)$$

From above two cases we observe that for large value of *n* we get better approximation. However for large n $(n \geq 8, n \neq 9)$ some of λ_k 's become negative and therefore higher order Newton Cotes formulas are not commonly used.

Conclusions :

In this unit we have seen how to approximate a function or a given tabular values of function by a polynomial. If (n + 1) observations or values of function are known we can fit a unique polynomial of degree *n*. Lagrange interpolation method and Newtons divided difference formula generate the same polynomial.

Finite difference operators and polynomial interpolation is used to calculate the derivative of a given function. Method of undetermined coefficients can also be used to find out the derivatives of a given function.

Methods of numerical integration are developed by approximating a function by a polynomial and on integrating this approximating polynomials. It is shown that trapezoidal method is first order method whereas Simpson's one third rule is of order three.

ILLUSTRATIVE EXAMPLES

1. Given $f(2)=4$, $f(2.5)=5.5$, find the linear interpolation polynomial using (i) Lagrage interpolation (ii) Newton's divided difference interpolation.

Answer : Here $x_0=2$, $x_1=2.5$, $f(x_0)=4$, $f(x_1)=5.5$.

(i) Lagrage fundamental polynomials are given by

$$\ell_0(x)=\frac{x-x_1}{x_0-x_1}=\frac{x-2.5}{(-0.5)},\ \ell_1(x)=\frac{x-x_0}{x_1-x_0}=\frac{x-2}{(0.5)}$$

Interpolating polynomial

$$P_1(x)=\ell_0(x)f(x_0)+\ell_1(x)f(x_1)$$

$$=\frac{x-2.5}{-(0.5)}(4)+\frac{x-2}{0.5}(5.5)$$

$$=-8(x-2.5)+11(x-2)$$

$$=3\times-2$$

Thus $f(x)\cong P_1(x)=3\times-2$

ii) By Newton's divided difference interpolation formula we have

$$P_1(x)=f(x_0)+f[x_0,x_1](x-x_0)$$

$$f[x_0,x_1]=\frac{f(x_1)-f(x_0)}{x_1-x_0}=\frac{5.5-4}{2.5-2}=\frac{1.5}{0.5}=3$$

$$P_1(x)=4+3(x-2)=3x-2$$

2. Using the data sin (0.1) = 0.09983 and sin (0.2) = 0.19867 find the Lagrange interpolation polynomial. Obtain a bound on the truncation error.

Answer : Here $x_0=0.1$, $x_1=0.2$, $f(x_0)=0.09983$, $f(x_1)=0.19867$.

Lagrange fundamental polynomials are given by

$$\ell_0(x)=\frac{(x-x_1)}{(x_0-x_1)}=\frac{x-0.2}{(-0.1)},\ \ell_1(x)=\frac{(x-x_0)}{(x_1-x_0)}=\frac{x-0.1}{0.1}$$

Interpolating polynomial

$$P_1(x) = \ell_0(x) f(x_0) + \ell_1(x) f(x_1)$$

$$= \frac{x-0.2}{-(0.1)}(0.09983) + \frac{x-0.1}{(0.1)}(0.19867)$$

$$= 0.9884x - 0.00099$$

The truncation error

$$E_1(f;x) = \frac{(x-x_0)(x-x_1)}{2!} f''(\xi)$$

$$= \frac{(x-0.1)(x-0.2)}{2}(-\sin\xi)$$

The maximum value of $|-\sin\xi|, \xi \in [0.1, 0.2]$ is $\sin(0.2) = 0.19867$.

Thus $E_1(f;x) \le \left|\frac{(x-x_0)(x-x_1)}{2}\right| 0.19867$ where $x \in [0.1, 0.2]$

3. In the following problems, find the maximum value of the step size *h* that can be used to tabulate $f(x)$ on [*a, b*] using linear interpolation such that $|\text{Error}| \le \varepsilon$.

i) $f(x) = (1+x)^6$, $[a,b] = [0,1]$, $\varepsilon = 5 \times 10^{-5}$.

ii) $f(x) = 2^x$, $[a,b] = [0,1]$, $\varepsilon = 1 \times 10^{-5}$.

iii) $f(x) = xe^x$, $[a,b] = [1,2]$, $\varepsilon = 1 \times 10^{-5}$.

Answer : The truncation error in linear interpolation is given by

$$E_1(f;x) = \frac{1}{2}(x-x_0)(x-x_1) f''(\xi)$$

$$\max_{x_0 \le x \le x_1} = \left|(x-x_0)(x-x_1)\right|$$

$$w(x) = (x-x_0)(x-x_1)$$

$$w'(x) = (x-x_1) + (x-x_0) = 0 \Rightarrow x = \frac{x_0 + x_1}{2}$$

$$\therefore \max_{x_0 \le x \le x_1} = \left|(x-x_0)(x-x_1)\right| \text{ occurs at } x = \frac{x_0 + x_1}{2}.$$

$$\max_{x_0 \le x \le x_1} = \left|(x-x_0)(x-x_1)\right| = \left|\left(\frac{x_0+x_1}{2}-x_0\right)\left(\frac{x_0+x_1}{2}-x_1\right)\right|$$

$$= \left(\frac{x_1-x_0}{2}\right)\left(\frac{x_0-x_1}{2}\right) = \frac{(x_1-x_0)^2}{4} = \frac{h^2}{4}$$

$$\therefore \left|E_1(f;x)\right| < \varepsilon \Rightarrow \left|\frac{(x_1-x_0)^2}{4} f''(\xi)\right| < \varepsilon$$

$$\therefore \left|E_1(f;x)\right| \le \frac{h^2}{8} M \qquad \text{where } M = \max_{x_0 \le \xi \le x_1} \left|f''(\xi)\right|$$

i) For $f(x) = (1+x)^6$, $f''(x) = 30(1+x)^4$

$\max_{0 \le x \le 1} \left|f''(x)\right| = 30(2)^4 = 480$ Thus M = 480

$$\therefore \left|E_1(f;x)\right| \le \frac{h^2}{8} \cdot 480 = 60h^2 < 5 \times 10^{-5}$$

$$\therefore h < 0.0009128$$

ii) $f(x) = 2^x$

$f'(x) = 2^x \ln 2$, $\quad f''(x) = 2^x (\ln 2)^2$

$\max_{0 \le x \le 1} f''(x) = 2(\ln 2)^2 = 0.960906027$

Thus M = 0.960906027

$$\therefore \frac{h^2}{8} \cdot M < 1 \times 10^{-5}$$

$$\therefore h^2 < \frac{1 \times 10^{-5}}{0.120113253} \Rightarrow h < 0.009124$$

iii) $f(x) = xe^x$, $f'(x) = e^x + xe^x = (1+x)e^x$

$f''(x) = e^x + (1+x)e^x = (2+x)e^x$

$\therefore \max_{1 \le x \le 2} \left|f''(x)\right| = (2+2)e^2 = 29.5562244$

$$\frac{h^2}{8}M < 1\times10^{-5} \Rightarrow h^2 < \frac{1\times10^{-5}\times8}{29.5563244}$$

$$\therefore h < 0.001645$$

4. Find the interpolaring polynomial that fits the following data. Find an approximation to $f(x)$ at 3.0.

x	0	1	2	4	5	6
$f(x)$	1	14	15	5	6	19

Answer :

By Langrange interpolation polynomial,

$$f(x) \cong P_5(x) = \frac{(x-1)(x-2)(x-4)(x-5)(x-6)}{(-1)(-2)(-4)(-5)(-6)}(1) + \frac{x(x-2)(x-4)(x-5)(x-6)}{(1)(-1)(-3)(-4)(-5)}(14)$$

$$+\frac{(x-0)(x-1)(x-4)(x-5)(x-6)}{(2)(1)(-2)(-3)(-4)}(15) + \frac{x(x-1)(x-2)(x-5)(x-6)}{4(3)(2)(-1)(-2)}(5)$$

$$+\frac{(x-0)(x-1)(x-2)(x-4)(x-6)}{5(4)(3)(1)(-1)}(6) + \frac{x(x-1)(x-2)(x-4)(x-5)}{6(5)(4)(2)(1)}(19)$$

$$P_3(5) = \frac{(2)(1)(-1)(-2)(-3)(1)}{-240}(1) + \frac{3(1)(-1)(-2)(-3)}{60}(14)$$

$$+\frac{3(2)(-1)(-2)(-3)}{-48}(15) + \frac{3(2)(1)(-2)(-3)}{48}(5)$$

$$+\frac{3(2)(1)(-1)(-3)}{-60}(6) + \frac{3(2)(1)(-1)(-2)}{240}(19)$$

$$= \frac{12}{240} - \frac{18\times14}{60} + \frac{36\times5}{48} + \frac{36\times5}{48} - \frac{18\times6}{60} + \frac{12\times19}{240} = 10$$

5. Given the following values of $f(x) = \ln x$, find the approximate value of $f'(2.0)$ using linear and quadratic interpolation and $f''(2.0)$ using quadratic interpolation. Also obtain an upper bound on the error.

i	0	1	2
x_i	2.0	2.2	2.6
f_i	0.69315	0.78846	0.95551

Answer :

With linear interpolation we have

$$f'(2.0) = \frac{f(2.2) - f(2.0)}{2.2 - 2.0}$$

i.e. $$f'(2.0) = \frac{0.78846 - 0.69315}{0.2} = 0.47655$$

By Lagrange interpolation we have

$$f(x) \cong P_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2)$$

$$= \frac{(x-2.2)(x-2.6)}{(-0.2)(-0.6)}(0.69315) + \frac{(x-2)(x-2.6)}{(2)(-0.4)}(0.78846)$$

$$+ \frac{(x-2)(x-2.2)}{(0.6)(0.4)}(0.95551)$$

Thus,

$$f'(x) \cong P_2'(x) = \frac{0.69315}{0.12}[(x-2.2)+(x-2.6)] - \frac{0.78846}{0.8}[(x-2)+(x-2.6)]$$

$$+ \frac{0.95551}{0.24}[(x-2)+(x-2.2)]$$

$$\therefore f'(2.0) \cong P_2'(2.0) = \frac{0.69315(-0.8)}{0.12} - \frac{0.78846(-0.6)}{0.8} + \frac{0.95551(-0.2)}{0.24} = 0.49619$$

$$f''(x) \cong P_2''(x) = \frac{0.69315}{0.12}(2) - \frac{0.78846}{0.8}(2) + \frac{0.95551}{0.24}(2) = -0.19642$$

The errors associated with methods are given by,

$$E_1'(x_0) = \frac{x_0 - x_1}{2} f''(\xi) \qquad x_0 < \xi < x_1$$

$$E_2'(x_0) = \frac{1}{6}(x_0 - x_1)(x_0 - x_2) f'''(\xi) \qquad x_0 < \xi < x_2$$

$$E_2''(x_0) = \frac{1}{3}(2x_0 - x_1 - x_2) f'''(\xi) + \frac{1}{24}(x_0 - x_1)(x_0 - x_2)\left[f^{(iv)}(\eta_1) + f^{(iv)}(\eta_2) \right]$$

$$x_0 < \xi, \eta_1, \eta_2 < x_2$$

For $f(x) = \log_n x$ we have

$$M_1 = \max_{x_0 \le \xi \le x_1} |f'(x)| = \max_{2 \le x \le 2.2} \left| \frac{1}{x} \right| = 0.5$$

$$M_2 = \max_{x_0 \le \xi \le x_2} |f''(x)| = \max_{2 \le x \le 2.6} \left| -\frac{1}{x^2} \right| = 0.25$$

$$M_3 = \max_{x_0 \le \xi \le x_2} |f'''(x)| = \max_{2 \le \xi \le 2.6} \left| +\frac{2}{x^3} \right| = 0.25$$

$$M_4 = \max_{x_0 \le \xi \le x_2} |f^{(iv)}(x)| = \max_{2 \le \xi \le 2.6} \left| \frac{-6}{x^4} \right| = \frac{6}{16} = 0.375$$

Thus $|E_1'(2.0)| \le \left| \frac{2 - 2.2}{2} \right| (0.25) = 0.025$

$$|E_2'(2.0)| \le \frac{1}{6} |(2 - 2.2)(2 - 2.6)| (0.25) = 0.005$$

$$|E_2''(2.0)| \le \frac{1}{3} |2(2) - 2.2 - 2.6| \cdot (0.25)$$

$$+ \frac{1}{24}(2 - 2.2)(2 - 2.6)[0.375 + 0.375]$$

$$\le \frac{(0.8)(0.25)}{3} + \frac{(0.12)(0.750)}{24} = 0.0704$$

7. A differentiation rule of the form

$$f'(x_0) = \alpha_0 f_0 + \alpha_1 f_1 + \alpha_2 f_2 \qquad (x_k = x_0 + kh,\ f_k = f(x_k))$$

is given. Find the values of α_0, α_1, α_2 so that the rule is exact for $f \in P_2$. Find the error term.

Answer :

$$f'(x_0) = \alpha_0 f(x_0) + \alpha_1 f(x_1) + \alpha_2 f(x_2)$$

$$= \alpha_0 f(x_0) + \alpha_1 f(x_0 + h) + \alpha_2 f(x_0 + 2h)$$

$$= \alpha_0 f(x_0) + \alpha_1 \left[f(x_0) + hf'(x_0) + \frac{h^2}{2!} f''(x_0) + \frac{h^3}{3!} f'''(x_0) + ... \right]$$

$$+ \alpha_2 \left[f(x_0) + (2h) f'(x_0) + \frac{(2h)^2}{2!} f''(x_0) + \frac{(2h)^3}{3!} f'''(x_0) + ... \right]$$

$$= (\alpha_0 + \alpha_1 + \alpha_2) f(x_0) + h(\alpha_1 + 2\alpha_2) f'(x_0) + \frac{h^2}{2!} f''(x_0)[\alpha_1 + 4\alpha_2]$$

$$+ \frac{h^3}{3!} f'''(x_0)[\alpha_1 + 8\alpha_2] + ...$$

On comparing the coefficients of $f(x_0)$, $f'(x_0)$, $f''(x_0)$ .... we get,

$\alpha_0 + \alpha_1 + \alpha_2 = 0$, $h(\alpha_1 + 2\alpha_2) = 1$, $\alpha_1 + 4\alpha_2 = 0$.

Since the formula is exact for a polynomial of degree 2, the error is in the form $f'''(\xi)$.

$$\alpha_1 + 4\alpha_2 = 0 \Rightarrow \alpha_1 = -4\alpha_2$$

$$h(\alpha_1 + 2\alpha_2) = 1 \Rightarrow h(-2\alpha_2) = 1 \Rightarrow \alpha_2 = -\frac{1}{2h}$$

Therefore, $\alpha_1 = -4\alpha_2 = \frac{2}{h}$

and $\alpha_0 + \alpha_1 + \alpha_2 = 0 \Rightarrow \alpha_0 + \frac{2}{h} - \frac{1}{2h} = 0 \Rightarrow \alpha_0 = -\frac{3}{2h}$.

The error will be

$$\frac{h^3}{3!} f'''(\xi)(\alpha_1 + 8\alpha_2)$$

$$= \frac{h^3}{3!} f'''(\xi)\left[\frac{2}{h} - \frac{8}{2h}\right]$$

$$= \frac{h^2}{3!} f'''(\xi)(-2)$$

$\therefore$ Error bound $|E_2| \le \frac{2h^2}{6} f'''(\xi) = \frac{h^2}{3} f'''(\xi)$.

8. Show that

(i) $\Delta \equiv E\nabla$ (ii) $\nabla \equiv E^{-1}\Delta$ (iii) $E \equiv 1 + \Delta$ (iv) $E^{-1} \equiv 1 - \Delta$

(i) $\Delta f(x_i) = f(x_{i+1}) - f(x_i)$ .... (i)

and $E\nabla f(x_i) = E\left[\nabla f(x_i)\right] = E\left[f(x_i) - f(x_{i-1})\right] = Ef(x_i) - Ef(x_{i-1})$

$\therefore E\nabla f(x_i) = (x_{i+1}) - f(x_i)$ ..... (ii)

From (i) and (ii) we have $\Delta \equiv E\nabla$.

(ii) $\nabla f(x_i) = f(x_i) - f(x_{i-1})$ ..... (i)

$(E^{-1}\Delta) f(x_i) = E^{-1}\left(\Delta f(x_i)\right) = E^{-1}\left(f(x_{i+1}) - f(x_i)\right) = E^{-1} f(x_{i+1}) = E^{-1} f(x_i)$

$\therefore (E^{-1}\Delta) f(x_i) = f(x_i) - f(x_{i-1})$ .... (ii)

From (i) and (ii) we have $\nabla \equiv E^{-1}\Delta$.

(iii) $(1 + \Delta) f(x_i) = f(x_i) + \Delta f(x_i) = f(x_i) + f(x_{i+1}) - f(x_i) = f(x_{i+1}) = Ef(x_i)$

Thus $E \equiv 1 + \Delta$.

(iv) $(1 - \nabla) f(x_i) = f(x_i) - \nabla f(x_i) = f(x_i) - \left[f(x_i) - f(x_{i-1})\right] = f(x_{i-1}) = E^{-1} f(x_i)$

Thus $E^{-1} \equiv 1 - \Delta$.

9. Using the following data find $f'(6.0)$, error = 0 (h) and $f''(6.3)$ error = 0 (h^2).

x	6.0	6.1	6.2	6.3	6.4
$f(x)$	0.1750	– 0.1998	– 0.2223	– 0.2422	– 0.2596

Answer :

With linear interpolation error is 0 (h)

$$f'(6.0) = \frac{f(6.1) - f(6.0)}{6.1 - 6.0} = \frac{-0.1998 - 0.1750}{0.1} = -3.748$$

$$f''(x_k) = \frac{f(x_{k+1}) - 2f(x_k) + f(x_{k-1})}{h^2} + 0(h^2)$$

Here $h = 0.1$ and $x_{k+1} = x_k + h$.

$$\therefore f(6.3) = \frac{f(6.4) - 2f(6.3) + f(6.2)}{(0.1)^2}$$

$$= \frac{-0.2596 - 2(-0.2422) + (-0.2223)}{(0.01)} = 0.25$$

10. Calculate the nth divided difference of $\frac{1}{x}$ based on the points $x_0, x_1, x_2, ..., x_n$.

Answer : We have $f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{\frac{1}{x_1} - \frac{1}{x_0}}{x_1 - x_0} = -\frac{1}{x_0 x_1}$

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{(x_2 - x_0)} = \frac{-\frac{1}{x_1 x_2} + \frac{1}{x_0 x_1}}{(x_2 - x_0)} = -\frac{1}{x_0 x_1 x_2}$$

Let $f[x_0, x_1, x_2, ..., x_k] = \frac{(-1)^k}{x_0 x_1 x_2 .... x_k}$.

Then $f[x_0, x_1, x_2, ..., x_{k+1}] = \frac{f[x_1, x_2, ..., x_{k+1}] - f[x_0, x_1, x_2, ..., x_k]}{x_{k+1} - x_0}$.

$$= \frac{\frac{(-1)^k}{x_1 x_2 x_3 .... x_{k+1}} - \frac{(-1)^k}{x_0 x_1 x_2 .... x_k}}{x_{k+1} .... x_0}$$

$$= \frac{(-1)^k (x_0 - x_{k+1})}{(x_0 x_1 x_2 .... x_{k+1})(x_{k+1} - x_0)} = \frac{(-1)^{k+1}}{x_0 x_1 x_2 .... x_{k+1}}$$

Hence by induction we have,

$$f[x_0, x_1, x_2, ..., x_n] = \frac{(-1)^n}{x_0 x_1 x_2 .... x_n}$$

11. If $f(x) = \frac{1}{x^2}$, find the divided difference $f[x_1, x_2, x_3, x_4]$.

Answer : $f[x_1, x_2] = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{\frac{1}{x_2^2} - \frac{1}{x_1^2}}{x_2 - x_1} = \frac{(x_1 + x_2)(x_1 - x_2)}{x_1^2 x_2^2 (x_2 - x_1)}$

$$= -\frac{(x_1 + x_2)}{x_1^2 x_2^2}$$

$$f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1} = \frac{-\frac{(x_2 + x_3)}{x_2^2 x_3^2} + \frac{x_1 + x_2}{x_1^2 \cdot x_2^2}}{x_3 - x_1}$$

$$= \frac{-x_1^2 (x_2 + x_3) + x_3^2 (x_1 + x_2)}{x_1^2 x_2^2 x_3^2 (x_3 - x_1)}$$

$$= \frac{-x_1^2 (x_2 + x_3) + x_3^2 (x_1 + x_2) + x_1 x_2 x_3 - x_1 x_2 x_3}{x_1^2 x_2^2 x_3^2 (x_3 - x_1)}$$

$$= \frac{(x_3 - x_1)(x_1 x_2 + x_1 x_3 + x_2 x_3)}{x_1^2 x_2^2 x_3^2 (x_3 - x_1)}$$

$$= \frac{x_1 x_2 + x_1 x_3 + x_2 x_3}{x_1^2 x_2^2 x_3^2}$$

$$f[x_1, x_2, x_3, x_4] = \frac{f[x_2, x_3, x_4] - f[x_1, x_2, x_3]}{x_4 - x_1}$$

$$= \frac{\dfrac{x_2x_3 + x_2x_4 + x_3x_4}{x_2^{\,2}x_3^{\,2}x_4^{\,2}} - \dfrac{x_1x_2 + x_1x_3 + x_2x_3}{x_1^{\,2}x_2^{\,2}x_3^{\,2}}}{x_4 - x_1}$$

$$= \frac{x_1^{\,2}\left(x_2x_3 + x_2x_4 + x_3x_4\right) - x_4^{\,2}\left(x_1x_2 + x_1x_3 + x_2x_3\right)}{x_1^{\,2}x_2^{\,2}x_3^{\,2}x_4^{\,2}\left(x_4 - x_1\right)}$$

$$= -\frac{\left(x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4\right)}{x_1^{\,2}x_2^{\,2}x_3^{\,2}x_4^{\,2}}$$

12. If $f(x) = e^{ax}$ show that $\Delta^n f(x) = \left(e^{ah} - 1\right)^n e^{ax}$.

Answer :

$$\Delta f(x) = f(x+h) - f(x)$$

$$= e^{a(x+h)} - e^{ax}$$

$$= e^{ax}\left(e^{ah} - 1\right)$$

$$\Delta^2 f(x) = \Delta\left[\Delta f(x)\right] = \Delta\left[e^{ax}\left(e^{ah} - 1\right)\right]$$

$$= \left(e^{ah} - 1\right)\Delta e^{ax}$$

$$= \left(e^{ah} - 1\right)\left(e^{ah} - 1\right)e^{ax}$$

$$= \left(e^{ah} - 1\right)^2 e^{ax}$$

Let $\Delta^k f(x) = \left(e^{ah} - 1\right)^k e^{ax}$

Then $\Delta^{k+1} f(x) = \Delta\left[\left(e^{ah} - 1\right)^k e^{ax}\right]$

$$= \left(e^{ah} - 1\right)^k \Delta e^{ax}$$

$$= \left(e^{ah} - 1\right)^k \left(e^{ah} - 1\right)e^{ax}$$

$$= \left(e^{ah} - 1\right)^{k+1} e^{ax}$$

Hence by induction $\Delta^n f(x) = \left(e^{ax} - 1\right)^n e^{ax}$.

13. The following table of values represents a polynomial of degree ≤ 3. Locate any error in the table of values

x	0	0.1	0.2	0.3	0.4
$f(x)$	2.0	2.11	2.28	2.39	2.56

Answer : Observe that $\Delta f_0 = 0.11$, $\Delta f_1 = 0.17$, $\Delta f_2 = 0.11$, $\Delta f_3 = 0.17$. Since there is sudden change in value at Δf_2. The error is expected at $x = 0.3$. Let $f(0.3) = 2.39 + \varepsilon$. Since the data represents a polynomial of degree ≤ 3, $\Delta^4 f = 0$.

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
0	2.0	0.11	0.06	$-0.12 + \varepsilon$	$0.24 - 4\varepsilon$
0.1	2.11	0.17	$-0.06 + \varepsilon$		
0.2	2.28	$0.11 + \varepsilon$		$+0.12 - 3\varepsilon$	
0.3	$2.39 + \varepsilon$		$0.06 - 2\varepsilon$		
0.4	2.56	$0.17 - \varepsilon$			

$$\Delta^4 f = 0.24 - 4\varepsilon = 0 \Rightarrow \varepsilon = \frac{0.24}{4} = 0.06$$

Thus $f(0.3) = 2.39 + 0.06 = 2.45$

14. Determine the step size h that can be used in the tabulation of a function $f(x)$, $a \leq x \leq b$, at equally spaced nodal points so that the truncation error of the quadratic interpolation is less than ε.

Answer : Let x_{i-1}, x_i, x_{i+1} denote three consecutive equipaced points with step size h. The truncation error of the quadratic Lagrange interpolation is bounded by

$$\left|E_2(f:x)\right| \leq \frac{M_3}{6} \max \left|(x - x_{i+1})(x - x_i)(x - x_{i-1})\right|$$

where $x_{i-1} \leq x \leq x_{i+1}$ and $M_3 = \max_{a \leq x \leq b} \left|f'''(x)\right|$

Put $t = \dfrac{x - x_i}{h}$

Then $x - x_{i-1} = x_i + th - x_{i-1} = (1+t)\,h$

$x - x_i = x_i + th - x_i = th$

$x - x_{i+1} = x_i + th - x_{i+1} = (t-1)\,h$

Since $x \in (x_{i-1}, x_{i+1}), t \in (-1,1)$

and $(x - x_{i+1})(x - x_i)(x - x_{i-1}) = (t-1)\,t\,(t+1)\,h^3$

Define $g(t) = (t-1)\,t\,(t+1) = t(t^2 - 1)$

Then $g'(t) = 3t^2 - 1 = 0 \Rightarrow t^2 = \frac{1}{3}$ and $t = \pm\sqrt{\frac{1}{3}}$

$$\max_{x_{i-1} \le x \le x_{i+1}} \left|(x - x_{i-1})(x - x_i)(x - x_{i+1})\right| = h^2 \max_{-1 \le t \le 1} \left|(t-1)\,t\,(t+1)\right|$$

$$= h \max_{-1 \le t \le t} \left|t(t^2 - 1)\right|$$

$$= h^3 \cdot \left|\sqrt{\frac{1}{3}}\left(\frac{1}{3} - 1\right)\right|$$

$$= h^3 \cdot \frac{2}{3\sqrt{3}}$$

Hence the truncation error in quadratic interpolation is bounded by

$$\left|E_2(f:x)\right| \le \frac{2h^3}{6(3\sqrt{3})} M_3$$

Now choose h such that $\frac{2h^3}{6(3\sqrt{3})} M_3 < \varepsilon$.

i.e. $$h < \left[\frac{9\sqrt{3}\varepsilon}{M_3}\right]^{\frac{1}{3}}$$

15. Find the maximum value of the uniform mesh size *h* that can be used to tabulate $f(x)$ on [*a, b*] using quadratic interpolation. Where $f(x) = x^2 e^x$, $[a,b] = [0,1]$ such that $|\text{error}| < 5 \times 10^{-6}$.

Answer : From example 14 we see that,

$$h < \left[\frac{9\sqrt{3}\varepsilon}{M_3} \right]^{\frac{1}{3}}$$

where $M_3 = \max_{0 \le x \le 1} x^2 e^x = e$

$$\therefore h < \left[\frac{9\sqrt{3} \times 5 \times 10^{-6}}{e} \right]^{\frac{1}{3}}$$

16. Determine the step size *h* that can be used in the tabulation of a function $f(x)$, $a \le x \le b$, at equally spaced nodal points so that the truncation error of the cubic interpolation is less than ε.

Answer : Let x_0, x_1, x_2, x_3 denote four consecutive equispaced points with step size *h*. The truncation error of the cubic interpolation is bounded by

$$\left| E_3(f;x) \right| \le \frac{M_4}{4!} \max_{x_0 \le x \le x_3} \left| (x - x_0)(x - x_1)(x - x_2)(x - x_3) \right|$$

and $M_4 = \max_{a \le x \le b} \left| f^{(4)}(x) \right|$

Put $t = \frac{1}{h}\left[x - \frac{x_1 + x_2}{2} \right]$ i.e. $x = \frac{x_1 + x_2}{2} + th$

Since $x \in (x_0, x_3)$, $t \in \left(-\frac{3}{2}, \frac{3}{2} \right)$

$$(x - x_0) = th + \frac{x_1 + x_2}{2} - x_0 = \left(t + \frac{3}{2} \right) h$$

$$(x - x_1) = th + \frac{x_1 + x_2}{2} - x_1 = \left(t + \frac{1}{2} \right) h$$

$$(x - x_2) = th + \frac{x_1 + x_2}{2} - x_2 = \left(t - \frac{1}{2} \right) h$$

$$(x - x_3) = th + \frac{x_1 + x_2}{2} - x_3 = \left(t - \frac{3}{2} \right) h$$

Thus $(x-x_0)(x-x_1)(x-x_2)(x-x_3)=\left(t+\frac{3}{2}\right)\left(t+\frac{1}{2}\right)\left(t-\frac{1}{2}\right)\left(t-\frac{3}{2}\right)h^2$

Define $g(t)=\left(t+\frac{3}{2}\right)\left(t+\frac{1}{2}\right)\left(t-\frac{1}{2}\right)\left(t-\frac{3}{2}\right)$

To determine optimum value of $g(t)$, consider $g'(t)=0$.

$$g(t)=\left(t^2-\frac{9}{4}\right)\left(t^2-\frac{1}{4}\right)$$

$$g'(t)=2t\left(t^2-\frac{1}{4}\right)+2t\left(t^2-\frac{9}{4}\right)=2t\left(t^2-\frac{1}{4}+t^2-\frac{9}{4}\right)$$

$$=2t\left(2t^2-\frac{5}{2}\right)$$

$$g'(t)=0\Rightarrow t=0,\ t=\pm\frac{\sqrt{5}}{2}$$

$$g(0)=\frac{9}{16}\text{ and }\left|g\left(\pm\frac{\sqrt{5}}{2}\right)\right|=\left(\frac{5}{4}-\frac{9}{4}\right)\left(\frac{5}{4}-\frac{1}{4}\right)=1$$

$\therefore$ Maximum absolute value of g is obtained for $t^2=\frac{5}{4}$.

Hence $\left|E_3(f;x)\right|\le\frac{M_4}{4!}\cdot h^2(1)$

Now choose h such that

$$\frac{h^4}{24}M_4<\varepsilon\quad\text{or}\quad h<\left(\frac{24\varepsilon}{M_4}\right)^{1/4}$$

17. Determine the step size that can be used in the tabulation of $f(x)=\cos 2x$ in the interval $\left[0,\frac{\pi}{4}\right]$ at equally spaced nodal points so that the truncation error of the cubic interpolation is less than 1×10^{-6}.

Answer : We have

$$\left|E_3(f;x)\right|\le\frac{h^4}{24}\cdot M_4$$

For $f(x) = \cos 2x$, $f'(x) = -2\sin 2x$, $f''(x) = -4\cos 2x$, $f'''(x) = 8\sin 2x$.

$f^{(iv)}(x) = 16\cos 2x$ and $M_4 = \max_{0 \le x \le \pi/4} \left| f^{(4)}(x) \right|$

$$= \max_{0 \le x \le \pi/4} |16\cos 2x| = 16$$

Hence, $\frac{h^2}{24} \cdot 16 < 1 \times 10^{-6}$

$$h < \left(\frac{24 \times 10^{-6}}{16} \right)^{\frac{1}{4}}$$

18. A table of values of $f(x) = e^{3x}$ in [0, 1] is constructed with step size 0.05. Find the maximum total error if cubic interpolation is to be used to interpolate in this interval.

Answer : $\left| E_3(f;x) \right| \le \frac{h^4}{24} \cdot M_4$

For $f(x) = e^{3x}$, $f'(x) = 3e^{3x}$, $f''(x) = 9e^{3x}$, $f'''(x) = 27e^{3x}$, $f^{(iv)}(x) = 81e^{3x}$.

and $M_4 = \max_{0 \le x \le 1} \left| f^{(4)}(x) \right|$

$$= \max_{0 \le x \le 1} \left| 81e^{3x} \right| = 81e$$

Therefore, maximum total error

$$\left| E_3(f;x) \right| \le \frac{(0.05)^4}{24} \cdot 81e$$

19. Evaluate the integral

$$I = \int_0^1 \frac{dx}{1+x}$$

Using (i) Composite trapezoidal rule.

(ii) Composite Simpson's rule with 2, 4 and 8 equal subintervals.

Answer : Let I_T and I_S denote the values obtained by using Trapezoidal and Simpson's rule respectively.

i) For N = 2, $h = \frac{1}{2}$ and 0, $\frac{1}{2}$, 1 are three nodes. We have two subintervals for trapezoidal rule and one interval for Simpson's 1/3 rd rule.

$$I_T = \frac{h}{2}\left[f(x_0) + 2f(x_1) + f(x_2)\right]$$

$$= \frac{1}{4}\left[f(0) + 2f\left(1/2\right) + f(1)\right]$$

$$= \frac{1}{4}\left[1 + \frac{2}{1+\frac{1}{2}} + \frac{1}{2}\right] = \frac{17}{24} = 0.708333$$

$$I_S = \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right]$$

$$= \frac{1}{6}\left[f(0) + 4f\left(1/2\right) + f(1)\right]$$

$$= \frac{1}{6}\left[1 + \frac{8}{3} + \frac{1}{2}\right] = \frac{25}{36} = 0.694444$$

ii) For N = 4, $h = \frac{1}{4}$ and 0, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, 1 are five nodes. We have 4 subintervals for trapezoidal rule and two subintervals for Simpson's rule. $x_0 = 0$, $x_1 = \frac{1}{4}$, $x_2 = \frac{1}{2}$, $x_3 = \frac{3}{4}$, $x_4 = 1$.

$$I_T = \frac{h}{2}\left[f(x_0) + 2f(x_1) + 2f(x_2) + 2f(x_3) + f(x_4)\right]$$

$$= \frac{1}{8}\left[f(0) + 2f\left(1/4\right) + 2f\left(1/2\right) + 2f\left(3/4\right) + f(1)\right]$$

$$= 0.697024$$

$$I_S = \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + f(x_4)\right]$$

$$= \frac{1}{12}\left[f(0) + 4f\left(1/4\right) + 2f\left(1/2\right) + 4f\left(3/4\right) + f(1)\right]$$

$$= 0.693254$$

For N = 8, $h = \frac{1}{8}$, $x_0 = 0$, $x_1 = \frac{1}{8}$, $x_2 = \frac{1}{4}$, $x_3 = \frac{3}{8}$, $x_4 = \frac{1}{2}$, $x_5 = \frac{5}{8}$, $x_6 = \frac{6}{8}$, $x_7 = \frac{7}{8}$, $x_0 = 1$.

$$I_T = \frac{h}{2}\left[f(x_0) + 2\{f(x_1) + f(x_2) + f(x_3) + f(x_4) + f(x_5) + f(x_6) + f(x_7)\} + f(x_8)\right]$$

$$= \frac{1}{16}\left[f(0) + 2\left\{f\left(\tfrac{1}{8}\right) + f\left(\tfrac{1}{4}\right) + f\left(\tfrac{3}{8}\right) + f\left(\tfrac{1}{2}\right) + f\left(\tfrac{5}{8}\right) + f\left(\tfrac{6}{8}\right) + f\left(\tfrac{7}{8}\right)\right\} + f(1)\right]$$

$= 0.694122$

$$I_S = \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + 4f(x_5) + 2f(x_6) + 4f(x_7) + f(x_8)\right]$$

$$= \frac{1}{24}\left[f(0) + 4f\left(\tfrac{1}{8}\right) + 2f\left(\tfrac{1}{4}\right) + 4f\left(\tfrac{3}{8}\right) + 2f\left(\tfrac{1}{2}\right) + 4f\left(\tfrac{5}{8}\right) + 2f\left(\tfrac{6}{8}\right) + 4f\left(\tfrac{7}{8}\right) + f(1)\right]$$

$= 0.693155.$

20. Evaluate $\int_0^2 e^x dx$ using the Simpson's rule with $h = 1$ and $h = \frac{1}{2}$. Find the bound on the error in each case. Compare with the exact solution.

Answer : For $h = 1$, $x_0 = 0$, $x_1 = 1$, $x_2 = 2$.

$$I = \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right]$$

$$= \frac{1}{3}\left[f(0) + 4f(1) + f(2)\right] = \frac{1}{3}\left[e^0 + 4e^1 + e^2\right]$$

Since for $h = 1$ we have one interval for Simpson's rule, the error in the integration $R_2 = -\frac{h^5}{90}\left[f^{(iv)}(\xi)\right]$, $0 \le \xi \le 2$, $f(x) = e^x$ and therefore $f^{(iv)}(x) = e^x$, $\max_{0\le 2\le 2}\left|f^{(iv)}(x)\right| = e^2$.

$$\therefore |R_2| < \frac{1}{90} \cdot e^2$$

ii) For $h = \frac{1}{2}$, $x_0 = 0$, $x_1 = \frac{1}{2}$, $x_2 = 1$, $x_3 = \frac{3}{4}$, $x_4 = 1$. We have two subintervals for Simpson's rule and

$$I = \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_0) + f(x_4)\right]$$

$$= \frac{1}{6}\left[f(0) + 4f\left(\tfrac{1}{2}\right) + 2f(1) + 4f\left(\tfrac{5}{2}\right) + f(2) \right]$$

$$= \frac{1}{6}\left[1 + 4\sqrt{e} + 2e + 4e\sqrt{e} + e^2\right]$$

Since we have two intervals for Simpson's rule

$$R_2 = -\frac{h^5}{90}\left[f^{(iv)}(\xi_1) + f^{(iv)}(\xi_2) \right]$$

where $x_0 \le \xi_1 \le x_2$ and $x_2 \le \xi_2 \le x_4$

i.e. $0 \le \xi_1 \le 1$ and $1 \le \xi_2 \le 2$.

$f(x) = e^x$

$\therefore \max\limits_{0 \le \xi_1 \le 1} f^{(iv)}(\xi_1) = e^1$

and $\max\limits_{0 \le \xi_2 \le 2} f^{(iv)}(\xi_2) = e^2$ and we have

$$|R_2| \le \frac{h^5}{90}\left[e + e^2\right] = \frac{1}{90(2^5)}\left[e + e^2\right]$$

EXERCISE

1. In the following problems, the values of a function $f(x)$ are given. Find the interpolating polynomial that fits the data.

(i)

x	–2	–1	0	1	3	4
$f(x)$	9	16	17	18	44	51

Calculate $f(0.5)$ and $f(3.1)$.

(ii)

x	1	3	4	5	7	10
$f(x)$	3	31	69	131	351	1011

Calculate $f(3.5)$.

(iii)

x	0	1	2	4	5	6
$f(x)$	1	14	15	5	6	19

Calculate $f(5.5)$.

(iv)

x	–1	1	4	7
$f(x)$	–2	0	63	342

Calculate $f(5.0)$.

(v)

x	–1	2	4	5
$f(x)$	–5	13	255	625

Calculate $f(3.0)$.

2. In the following problems, find the maximum value of stepsize h that can be used to tabulate $f(x)$ on $[a, b]$ using linear interpolation such that $|\text{Error}| < \varepsilon$.

(i) $f(x) = 1 + x^6$, $[a,b] = [0,1]$, $\varepsilon = 5 \times 10^{-5}$.

(ii) $f(x) = \dfrac{1}{1+x^2}$, $[a,b] = [1,2]$, $\varepsilon = 1 \times 10^{-4}$.

3. Prove the following relations.

(i) $\displaystyle\sum_{k=0}^{n-1} \Delta^2 f_k = \Delta f_n - \Delta f_0$

(ii) $\Delta\left(f_i \cdot g_i\right) = f_i \Delta g_i + g_{i+1} \Delta f_i$

(iii) $\Delta f_i^2 = \left(f_i + f_{i+1}\right) \Delta f_i$

(iv) $\Delta\left(\dfrac{f_i}{g_i}\right) = \left(g_i \Delta f_i - f_i \Delta g_i\right) / g_i g_{i+1}$

(v) $\Delta - \nabla = -\Delta\nabla$

4. The following data represents the function $f(x) = e^x$.

x	1	1.5	2.0	2.5
$f(x)$	2.7183	4.4817	7.3891	12.1825

Evaluate the value of $f(2.25)$ using Newton's divided difference interpolation.

5. In the following problems find the maximum value of the uniform mesh size *h* that can be using to tabulate $f(x)$ on [*a, b*] using quadratic interpolation such that $|\text{Error}| < \varepsilon$.

(i) $f(x) = (2+x)^4$, $[a,b] = [1,2]$, $\varepsilon = 1\times10^{-4}$.

(ii) $f(x) = e^{x+1}$, $[a,b] = [0,1]$, $\varepsilon = 1\times10^{-4}$.

(iii) $f(x) = x^2e^x$, $[a,b] = [0,1]$, $\varepsilon = 5\times10^{-6}$.

(iv) $f(x) = x^2 \ln x$, $[a,b] = [5,10]$, $\varepsilon = 1\times10^{-5}$.

6. In the following problems find the maximum value of uniform mesh size *h* that can be used to tabulate $f(x)$ on [*a, b*] using cubic interpolation such that $|\text{Error}| < \varepsilon$.

(i) $f(x) = e^x$, $[a,b] = [1,2.5]$, $\varepsilon = 1\times10^{-4}$.

(ii) $f(x) = \cos 2x$, $[a,b] = \left[0, \pi/4\right]$, $\varepsilon = 1\times10^{-6}$.

(iii) $f(x) = xe^x$, $[a,b] = [1,2]$, $\varepsilon = 5\times10^{-5}$.

7. Determine $\alpha, \beta, \gamma, \delta$ such that the relation

$$y'\left(\frac{a+b}{2}\right) = \alpha y(a) + \alpha y(b) + y''(a) + \delta y''(b)$$

is exact for polynomial of as high degree as possible.

8. Evaluate $\int_0^2 e^x dx$ using the Simpson's rule with $h = 1$ and $\frac{1}{2}$. Find a bound on the error in each case.

9. Compute $I_P = \int_0^1 \frac{x^P}{x^3 + 0} dx$ for p = 0, 1 using trapezoidal and Simpson's rules with the number of points 3, 5 and 9.

10. Compute $\int_{\pi/4}^{\pi/2} \frac{\cos \ln(\sin x)}{\sin^2 x + 1} dx$ correct to 3 decimal places, using trapezoidal rule and Simpson's rule.

11. Evaluate $\int_0^1 \left(1 + \frac{\sin x}{x}\right) dx$ using trapezoidal and Simpson's rule.

NUMERICAL SOLUTION OF DIFFERENTIAL EQUATION

Euler's Method :

Consider the interval [a, b] and the initial value problem $x' = f(t, x)$ with $x(a) = x_0$ (initial point).

Divide interval [a, b] into equal sub-interval and select mesh point

$$t_k = a + hk \text{ , k = 0, 1, ... N and } h = \frac{b-a}{N}$$

Suppose x, x', x'' are all continuous. The Euler Method iteration formula is $\xi_0 = x_0$.

$$\xi_{i+1} = \xi_i + hf\left[t_i, \xi_i\right]$$

Example 1 : Solve the initial value problem $x' = \frac{t-x}{2}$, on the interval $[0, 3]$, $x(0) = 1$.

Solution : Let $h = 0.5$, $\xi_0 = x(0) = 1$.

$$\xi_1 = \xi_0 + hf\left[t_0, \xi_0\right]$$

$$= 1 + (0.5)\left(\frac{0-1}{2}\right) \qquad \text{This for } t_0 = 0,\ x_1 = 0.5$$

$$= 0.75$$

$$\xi_2 = \xi_1 + hf\left[t_1, \xi_1\right] = 0.75 + 0.5\left(\frac{0.5-0.75}{2}\right) = 0.6875 \qquad \text{for } t_1 = 0.5, x_2 = 1$$

$$\xi_3 = 0.6875 + 0.5\left(\frac{1-0.6875}{2}\right) = 0.765625 \qquad \text{for } t_2 = 1, x_3 = 1.5$$

$$\xi_4 = 0.765625 + 0.5\left(\frac{1.5-0.765625}{2}\right) = 0.9492187 \qquad \text{for } t_3 = 1.5, x_4 = 2$$

$$\xi_5 = 0.9492187 + 0.5\left(\frac{2-0.9492187}{2}\right) = 1.2119141 \qquad \text{for } t_4 = 2, x_5 = 2.5$$

$$\xi_6 = 1.2119141 + 0.5\left(\frac{2.5 - 1.2119141}{2}\right) = 1.5339355 \qquad \text{for } t_5 = 2.5,\ x_6 = 3$$

Now, $x' = \frac{t}{2} - \frac{x}{2} \Rightarrow x' + \frac{x}{2} = \frac{t}{2}$ multiply by $e^{t/2}$ we have,

$$e^{t/2}\left(x' + \frac{x}{2}\right) = e^{t/2}\frac{t}{2} \Rightarrow \left(e^{t/2}x\right)^1 = \frac{t}{2}e^{t/2} \qquad \text{integrate}$$

$$\int \left(e^{t/2}x\right)^1 dt = \int \frac{t}{2}e^{t/2}dt$$

$$\Rightarrow e^{t/2}x = \frac{t}{2}\frac{e^{t/2}}{1/2} - \int \frac{1}{2}\frac{e^{t/2}}{1/2}dt + C$$

$$= te^{t/2} - 2e^{t/2} + C = (t-2)e^{t/2} + C$$

$\therefore x = t - 2 + Ce^{-t/2}$ by initial condition 1 =0.2 + C $\Rightarrow$ C = 3

$$x = t - 2 + 3e^{-t/2}$$

Example 2 : Solve $x' = -2 + x^2$ with $x(0) = 1$, with h = 0.2, 0.1 and 0.05 on [0, 1].

Solution : Let h = 0.2

$$X_{i+1} = X_i - 2ht_iX_i^2 \qquad \text{i = 0, 1, 2, 3, 4, } X_0 = 1$$

For i = 0, $t_0 = 0$, $X_0 = 1$.

$$X(0.2) \approx X_1 = X_0 - 2ht_0x_0^2 = 1$$

For i = 1, $t_1 = 0.2$, $X_1 = 1$.

$$\Rightarrow X_2 = 1 - 2(0.2)(0.2)(1)^2 = 0.92$$

For i = 2, $t_2 = 0.4$, $X_2 = 0.92$.

$$\Rightarrow X_3 = 0.92 - 2(0.2)(0.4)(0.92)^2 = 0.78458$$

For i = 3, $t_3 = 0.6$, $X_3 = 0.78458$.

$$\Rightarrow X_4 = 0.63684$$

$$X_5 = 0.50706$$

Now for,

h = 0.1

$u_0 = 1,\ u_1 = 1,\ u_2 = 0.98,\ u_3 = 0.94158,\ u_4 = 0.88839$

$u_5 = 0.82525,\ u_6 = 0.75715,\ u_7 = 0.68835,\ u_8 = 0.62202$

$u_9 = 0.56011,\ u_{10} = 0.50364.$

Order of Euler's Method

Lemma :

Let $a > 0$, $b \geq 0$ and let x_n, n = 0, 1, ... be a sequence of non-negative numbers satisfying the inequality

$$x_{n+1} \leq (1+a)x_n + b \qquad \text{..... (1)}$$

Then $$x_n \leq e^{na}x_0 + b\frac{e^{na}-1}{a} \qquad \text{..... (2)}$$

Proof : We have $x_1 \leq (1+a)x_0 + b$

$$x_2 \leq (1+a)x_1 + b \leq (a+1)\left[(1+a)x_0 + b\right] + b = (1+a)^2 x_0 + (1+a)b + b$$

$$x_3 \leq (1+a)x_2 + b \leq (a+1)\left[(1+a)^2 x_0 + (1+a)b + b\right] + b = (1+a)^3 x_0 + (1+a)^2 b + (1+a)b + b$$

Continuing in this way we see that,

$$x_n \leq (1+a)^n x_0 + b\sum_{j=0}^{n-1}(1+a)^j = (1+a)^n x_0 + b\frac{(1+a)^n - 1}{a} \qquad \text{..... (3)}$$

From the finite geometric series formula.

From the Maclaurin expansion of e^a, we have $1 + a < e^a$,

For $a > 0$ and thus that $(1+a)^n < e^{na}$. Substituting this into (3) gives

$$x_n \leq e^{na}x_0 + b\frac{e^{na}-1}{a}$$

Hence the result.

Theorem : Let X be the solution to

$$X' = f(t,x) \qquad \text{..... (1)}$$

and suppose that $C = X(t) \le d$ for $t \in (t_0, t_k)$.

Let ξ_i, i = 0,...., N be the Euler Method approximation to $x(t_i)$.

Where $t_i = t_0 + ih$ with $h = \dfrac{t_k - t_0}{N}$.

If $|X''(t)| \le M$ for some constant M and all $t \in (t_0, t_k)$ and the function f in (1) is Lipschitz continuous with constant L in the reactangle

$$R = \{(t,x) \mid t_0 \le t \le t_k ; c \le x \le d\}$$

Then $$|x(t_i) - \xi_i| \le \frac{Mh}{2L}\left(e^{L(t_i - t_0)} - 1\right) \qquad \text{..... (2)}$$

Proof : Let $E_i = X(t_i) - \xi_i$ be the error made by Euler's Method at the ith step. From Taylor's theorem applied to $X(t)$ and (1) we have

$$\begin{aligned} E_{i+1} &= X(t_{i+1}) - \xi_{i+1} \\ &= X(t_i) + hX'(t_i) + \frac{1}{2}h^2 X''(E_i) - \left[\xi_i + hf(t_i, \xi_i)\right] \\ &= E_i + h\left[X'(t_i) - f(t_i, \xi_i)\right] + \frac{1}{2}h^2 X''(\tau_i) \\ &= E_i + h\left[f(t_i)x(t_i) - f(t_i, \xi_i)\right] + \frac{1}{2}h^2 X''(\tau_i) \qquad \text{..... (3)} \end{aligned}$$

For some $\tau_i \in (t_i, t_{i+1})$ since X'' is bounded by M and f is Linchitz continuous.

$$|E_{i+1}| \le |E_i| + hL|X(t_i) - \xi_i| + \frac{Mh^2}{2} = |E_i|(1 + hL) + \frac{Mh^2}{2} \qquad \text{..... (4)}$$

Thus from above lemma

$$|E_i| \le \frac{Mh^2}{2}\frac{e^{ihL} - 1}{hL} = \frac{Mh}{2L}\left(e^{ihL} - 1\right)$$

Since $E_0 = 0$. The result now following by putting $ih = t_i - t_0$.

Theorem : Let $\delta = \max_{0 \le i \le N} |\delta_i|$.

Then under the same hypothesis as above Theorem

$$|X(t_i) - \xi_i| \le \left(\frac{hM}{2L} + \frac{\delta}{hL} \right) \left(e^{L(t_i - t_0)} - 1 \right) + |\delta_0| e^{L(t_i - t_0)}$$

Let $\xi_0 = x_0 + \delta_0$, $\xi_{i+1} = \xi_i + hf(t_i, \xi_i) + \delta_{i+1}$.

Runge-Kutta Method

One Step Method

$$y_0 = y(x_0)$$

$$y_{i+1} = y_i + h\phi(x_i y_i, h)$$

The function ϕ predicts the direction that the solution will take for the point (x_i, y_i) if ϕ satisfying the condition

$$\lim_{h \to 0} \phi(x_i, y_i; h) = f(x, y)$$

Then we get the improved solution and the method is said to be consistent.

Modified Euler Method

Instead of using Y_{Euler} as the approximation to $Y(x_i)$ we use it to locate another point near the trajections of $y(x_i)$ take the Estimated slope ϕ to be the average of slope of (x_i, y_i) and $(x_{i+1}, y_{\text{Euler}})$.

$$y_0 = x_0$$

$$M_1 = f(x_i, y_i)$$

$$M_2 = f(x_i + h, y_i + hM_1)$$

Then $y_{i+1} = y_i + h \dfrac{M_1 + M_2}{2}$

This is modified Euler method.

Example : Solve $y' = \frac{x-y}{2}$, $[0, 3]$, $y_0 = y(0) = 1$, h = 0.5.

Solution : Let $x = 0$, $M_1 = -0.5$, $M_2 = f\left(x_{i+h}, y_i + hM_1\right) = -0.125$

$$y_1 = y_0 + h\left(\frac{M_1 + M_2}{2}\right) = 0.84375$$

$x = 0.5$, $M_1 = -0.171875$, $M_2 = 0.1210937$.

$y_2 = 0.8310546$

if $x = 1$, $M_1 = 0.0844727$, $M_2 = 0.3133845$, then $y_3 = 0.9305114$

if $x = 1.5$, $M_1 = 0.2847443$, $M_2 = 0.4635582$, then $y_4 = 1.117887$

if $x = 2$, $M_1 = 0.4412062$, $M_2 = 0.5809048$, then $y_5 = 1.3731148$

if $x = 2.5$, $M_1 = 0.5634426$, $M_2 = 0.6725819$, then $y_6 = 1.6821209$

Mid Point Method

Using Euler's method approximate the solution at the mid point of x_i, x_{i+1} and take estimated slope ϕ to be f at that point

$$y_0 = x_0 \qquad M_1 = f\left(x_i, y_i\right) \qquad M_2 = f\left(x_i + \frac{h}{2}, y_i + h\frac{M_1}{2}\right)$$

$$\therefore y_{i+1} = y_i + hM_2$$

Note that Local truncation error for one step method defined as

$$E_i(h) = y\left(x_{i+1}\right) - y\left(x_i\right) - h\phi\left(x_i, y_i, h\right)$$

Question : Find the order of local truncation error for the modified Euler Method.

Answer : Assume that f has continuous third order partial derivatives in a rectangle containing the solution $y(x)$.

From Taylor's theorem.

$$y\left(x_{i+1}\right) = y\left(x_i + h\right) = y\left(x_i\right) + hy'\left(x_i\right) + \frac{h^2}{2!}y''\left(x_i\right) + \frac{h^3}{3!}y'''\left(x_i\right) + 0\left(h^4\right)$$

$$y(x_{i+1}) - y(x_i) = h\left[y'(x_i) + \frac{h}{2} y''(x_i) + \frac{h^2}{6} y'''(x_i) \right] \quad \text{..... (1)}$$

Differentiating the differential equation $y' = f(x, y)$ we obtain

$$y'' = \frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y} y' = \frac{\partial f(x,y)}{\partial x} + f(x,y) \frac{\partial f(x,y)}{\partial y}$$

$$y''' = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial x \partial y} y' + f(x,y) \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y} f(x,y) \right) + \frac{\partial}{\partial y} f(x,y) \cdot \frac{\partial}{\partial x} f(x,y)$$

$$+ \frac{\partial}{\partial y} f(x,y) \left(\frac{\partial}{\partial y} f(x,y) \right) f(x,y) + f(x,y) \frac{\partial^2}{\partial y^2} f(x,y) \cdot f(x,y)$$

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

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

$$y'(x_i) = f(x_i, y_i) = f_i$$

$$\therefore (1) \Rightarrow y_{i+1} - y_i = h f_i + \frac{h^2}{2!}\left(f_{i,x} + f_{i,y} f_i \right) + \frac{h^3}{3!}\left[f_{i,xx} + 2 f_{i,xy} f_i + f_{i,yy} f_i^2 \right.$$

$$\left. + f_{i,y} f_{i,x} + f_{i,y}{}^2 f_i \right] + 0(h^4) \quad \text{..... (2)}$$

$$\phi(x_i, y_i; h) = \frac{M_1 + M_2}{2} = \frac{1}{2}\left[f(x_i, y_i) + f(x_i + h, y_i + h M_1) \right]$$

$$= \frac{1}{2}\left[f(x_i, y_i) + f\left(x_i + h, y_i + h f(x_i, y_i) \right) \right]$$

Taylor's Theorem for two variables

$$f(x, y, h) = f(x, y) + h \frac{\partial}{\partial x} f(x, y) + k \frac{\partial}{\partial y} f(x, y)$$

$$+ \frac{1}{2}\left[h^2 \frac{\partial^2}{\partial x^2} f(x,y) + 2hk \frac{\partial^2}{\partial x \partial y} f(x,y) + k^2 \frac{\partial^2}{\partial y^2} f(x,y) \right]$$

Apply to the slope estimate ϕ for modified Euler method we obtained.

$$\phi(x_i, y_i; h) = \frac{1}{2}\left[f_i + f_i + hf_{i,x} + hf_i f_{i,y} + \frac{1}{2}\left(h^2 f_{i,xx} + 2L^2 f_{i,xy} f_i + h^2 f_{i,yy} f_i^2\right) + 0(h^3)\right]$$

$$= f_i + \frac{1}{2} f_{i,x} + \frac{h}{2} f_{i,y} f_i + \frac{h^2}{4} f_{i,xx} + \frac{1}{2} h^2 f_{i,xy} f_i + \frac{h^2}{4} f_{i,yy} f_i^2 \qquad \text{..... (3)}$$

$$h\phi(x_i, y_i, h) = hf_i + \frac{h^2}{2} f_{i,x} + \frac{h^2}{2} f_{i,y} f_i + \frac{h^3}{4} f_{i,xx} + \frac{h^3}{2} f_{i,xy} f_i + \frac{h^3}{4} f_{i,yy} f_i^2 \qquad \text{..... (4)}$$

Subtract equation (4) from (2) we get,

$$y_{i+1} - y_i - h\phi(x_i, y_i; h) = h^3\left[-\frac{1}{12} f_{i,xx} - \frac{1}{6} f_{i,xy} f_i - \frac{1}{10} f_{i,yy} f_i^2 \right.$$

$$\left. + \frac{1}{6} f_{i,y} f_{i,x} + \frac{1}{6} f_{i,y}{}^2 f_i \right] + 0(h^4) - 0(h^3)$$

$|\in_i (h)| \le kh^3$ k is constant.

Provided 2nd partial derivative of f is continuous.

Therefore, the local direction error of Euler modified method of order 3 in h.

Theorem : Consider the one step method

$$y_{i+1} = y_i + h\phi(x_i, y_i h)$$

Let ϕ be Lipschitz continuous with consistent L in the variable x in the reactangle R.

$R = \{(x, y) : x_0 \le x \le x_k ; c \le y \le d\}$ where $c \le y(x_i)$ and $\xi_i \le d$ for the space point $x_i = 0, 1, ..., N$ then

$$|y(x_i) - \xi_i| \le \frac{\in(L)}{hL}\left(e^{L(x_i - x_0)} - 1\right)$$

Proof : ξ_i is approximation to y_i.

$$|y_{i+1} - \xi_{i+1}| = |y_{i+1} - [\xi_i + h\phi(x_i, \xi_i; h)]|$$

Add and subtract $h\phi(x_i, y_i; h) + y_i$

$$\therefore |y_{i+1} - \xi_{i+1}| = |y_{i+1} - [h\phi(x_i, y_i; h) + y_i] + [h\phi(x_i, y_i; h) + y_i] - [\xi_i + h\phi(x_i, \xi_i; h)]|$$

$$= \left| \in_i (h) + y_i - \xi_i + h\left[\phi(x_i, y_i; h) - \phi(x_i, \xi_i; h) \right] \right|$$

$$\leq \left| \in_i (h) \right| + \left| y_i - \xi_i \right| + h \left| \left[\phi(x_i, y_i; h) - \phi(x_i, \xi_i; h) \right] \right|$$

$$\leq \left| \in_i (h) \right| + \left| y_i - \xi_i \right| + hL \left| y_i - \xi_i \right|$$

$$= \left| \in (h) \right| + (1 + hL) \left| y_i - \xi_i \right|$$

From the assumption of Lipschitz continuity for ϕ and by the lemma above

$$\left[x_{n+1} \leq (1+a) x_n + b \Rightarrow x_n \leq e^{na} x_0 + b \frac{e^{na} - 1}{a} \right]$$

We have, $\left| y(x_i) - \xi_i \right| \leq \frac{\in (h)}{hL} \left(e^{L(x_i - x_0)} - 1 \right)$ $\quad \because y_0 = \xi_0$

Question : Show that the global truncation error made by the modified Euler Method is $O(h^2)$.

Answer : We seen that the local truncation error is order three.

This follows from the above theorem, once we have established that

$$\phi(x_i, y_i; h) = \frac{1}{2}\left[f(x, y) + f\left(x + h, x + hf(x, y) \right) \right]$$

is Lipschitz continuous. Assume that f itself is Lipschitz continuous with constant L_f. we have

$$\left| \phi(x, y_1; h) - \phi(x, y_2; h) \right| = \left[\frac{1}{2} f(x, y_1) + f\left(x + h, y_1 + hf(x, y) \right) \right]$$

$$- \frac{1}{2}\left[f(x, y_2) + f\left(x + h, y_2 + hf(x, y_2) \right) \right]$$

$$\leq \frac{1}{2} \left| f(x, y_1) - f(x, y_2) \right| + \frac{1}{2} \left| f\left(x + h, y_1 + hf(x, y_1) \right) - f\left(x + h, y_2 + hf(x, y_2) \right) \right|$$

$$\leq \frac{1}{2} L_f \left| y_1 - y_2 \right| + \frac{1}{2} L_f \left| \left[y_1 + hf(x, y_1) \right] - \left[y_2 + hf(x, y_2) \right] \right|$$

$$\leq \frac{1}{2} L_f \left| y_1 - y_2 \right| + \frac{1}{2} \left(L_f \left| y_1 - y_2 \right| \right) + \frac{1}{2} L_f h \left| f(x, y_1) - f(x, y_2) \right|$$

$$\leq L_f \left| y_1 - y_2 \right| + \frac{1}{2} L^2{}_f h \left| y_1 - y_2 \right|$$

$$\leq \left(L_f + \frac{1}{2} h L^2{}_f \right) (y_1 - y_2)$$

Thus ϕ is Lipschitz continuous with constant $L_\phi = L_f + \frac{1}{2}hL^2{}_f$.

Note that, similarly we can show that the midpoint method can have a global truncation error of order two.

Example : $y' = -ty$, [0, 0.15], h = 0.05 with $y(0) = 1$

Solve by Euler modified method.

Solution : Euler Method : $\xi_{i+1} = \xi_i + hf(t, \xi_i)$ $\xi_0 = 1 \to 0$

$\xi_1 = 1 + 0.05(0 \times 1) = 1$ for $\to 0.05$

$\xi_2 = 1 + 0.05(-0.05 \times 1) = 0.9975$ for $\to 0.1$

$\xi_3 = 0.9975 + 0.05(-0.1 \times 0.9975) = 0.9925125$ for $\to 0.15$

Euler Modified $y_{i+1} = y_i + h\frac{M_1 + M_2}{2}$ (0, 0.05, 0.1, 0.15)

$M_1 = f(x_i, y_i)$ $M_2 = f(x_i + h, y_i + hM_1)$

$M_1 = 0$ $M_2 = -0.05$

$y_1 = 1 + 0.05\frac{(0 + (-0.05))}{2} = 0.99375$ $y_2 = 0.9923205$

Mid point $y_{i+1} = y_i + hM_2$

$M_2 = f\left(x_i + \frac{h}{2}, y_i + \frac{hM_1}{2}\right)$ $y_1 = 1 + 0.15(-0.075) = 0.98875$

Example : Solve $y' = -ty$.

Solution : $\frac{y'}{y} = -t \Rightarrow (\log y)' = -t$

intigrating both side we have $\int \log(y)'dy = -\int tdt$

$\Rightarrow \log y = -\frac{t^2}{2} + c$ $\Rightarrow y = e^{-\frac{t^2}{2} + c}$ $1 = e^c \Rightarrow c = 0$

$\Rightarrow y = e^{-\frac{t^2}{2}}$ $\therefore y = 0.988813$

Exercise :

1) Solve y' = – $2xy^2$, if y (0) = 1, h = 0.2, 0.1, 0.05 on [0, 1]

2) Solve y' = $x^2 + y^2$, if y (0) = 0, h = 0.5 on [0, 2].

Runge Kutta Methods

The most general Runge-Kutta Method involving two slope calculations is

$$y_0 = x_0 \qquad M_1 = f(x_i, y_i) \qquad M_2 = f(x_i + \alpha h, y_i + h\beta M_1)$$

$$y_{i+1} = y_i + h(w_1 M_1 + w_2 M_2) \text{ where } \alpha \in [0,1].$$

Note : 1. This gives the modified Euler Method when $\alpha = \beta = 1$ and $w_1 = w_2 = \frac{1}{2}$ and the midpoint method when $\alpha = \beta = \frac{1}{2}$ and $w_1 = 0$, $w_2 = 1$.

2. Not every choice of α and β will lead to a method that has order three local truncation error, however indeed with $y_i = y(x_i)$, $f_i = f(x_i, y_i)$, $f_{i,x} = \frac{\partial f}{\partial x}$ and so forth we have from Taylor's Theorem by sequence of computation (Similar to that used above).

$$\in_i (h) = y_{i+1} - y_i - h\left[w_1 f_i + w_2 f(x_i + \alpha h, y_i + \beta h f_i)\right]$$

$$= y_i' h + \frac{1}{2} y_i'' h^2 - \left[w_1 h f_i + w_2 h\left(f_i + f_{i,x}\alpha h + f_{i,y}\beta h f_i\right)\right] + 0(h^3)$$

$$= (1 - w_1 - w_2) h f_i + \left(\frac{1}{2} - \alpha w_2\right) f_{i,x} h^2 + \left(\frac{1}{2} - \beta w_2\right) f_{i,y} f_i h^2 + 0(h^3)$$

Thus since $f(x, y)$ is arbitrary for $\in_i$ to be $0(h^3)$ we must have

$$w_1 + w_2 = 1 \qquad \alpha w_2 = \frac{1}{2} \qquad \beta w_2 = \frac{1}{2}$$

$$\Rightarrow \beta = \alpha \qquad w_2 = \frac{1}{2\alpha} \qquad w_1 = 1 - \frac{1}{2\alpha}$$

It can be shown that no choice of α can lead to an order of local truncation error greater than three. Put

$$y_{i+1} = y_i + h\left[\left(1 - \frac{1}{2\alpha}\right) M_1 + \frac{1}{2\alpha} M_2\right] = y_i + \frac{h}{2\alpha}\left[(2\alpha - 1) M_1 + M_2\right]$$

Note : Every Runge-Kutta Method should reduce to a quadrature formula when $f(x, y)$ is independent of y with w's as weights and α's as abscissas.

If $\alpha = \frac{1}{2}$ we get

$$y_{i+1} = y_i + hM_2 \qquad M_1 = f(x_i, y_i) \text{ and } M_2 = f\left(x_i + \frac{h}{2}, y_i + \frac{hM_1}{2}\right)$$

Which is the Euler method with spacing $\frac{h}{2}$, i.e. is midpoint quadrature rule when $f(x, y)$ is independent of y.

For $\alpha = 1$ we get

$$y_{i+1} = y_i + \frac{h}{2}[M_1 + M_2] \qquad M_1 = f(x_i, y_i),\ M_2 = f(x_i + h, y_i + hM_1)$$

Which is Eular modified method.

Which reduces to the trapezoidal rule when $f(x, y)$ independent of y.

Note that the general form of a Runge-Rutta Method involving n slope calculations is

$$y_0 = x_0,\ M_1 = f(x_i, y_i),\ M_2 = f(x_i + \alpha_2 h, y_i + \beta_2 hM_1)$$

$$\ldots\ M_n = f\left(x_i + \alpha_n h, y_i + h\sum_{j=1}^{n-1} \beta_{nj} M_j\right)$$

$$y_{i+1} = y_i + h\sum_{j=1}^{n} w_j M_j$$

2. Thus a three slope Runge-Kutta Method has the form

$$y_0 = x_0,\ M_1 = f(x_i, y_i),\ M_2 = f(x_i + \alpha_2 h, y_i + \beta_{21} M_1 h)$$

$$M_3 = f(x_i + \alpha_3 h, y_i + \beta_{31} M_1 h + \beta_{32} M_2 h)$$

$$y_{i+1} = y_i + h(w_1 m_1 + w_2 m_2 + w_3 m_3)$$

By expanding the local truncation error in a manner similar to that used in above

$$w_1 + w_2 + w_3 = 1$$

$$\beta_{21} w_2 + (\beta_{31} + \beta_{32}) w_3 = \frac{1}{2}$$

$$\alpha_2 \beta_{21} w_2 + \alpha_3 (\beta_{31} + \beta_{32}) w_3 = \frac{1}{3},\ \alpha_2 \beta_{31} w_3 = \frac{1}{6},\ \alpha_2 w_2 + \alpha_3 w_3 = \frac{1}{2}$$

$$\frac{1}{2}{\alpha_2}^2 w_2 + \frac{1}{2}{\alpha_3}^2 w_3 = \frac{1}{6},\ \frac{1}{2}{\beta_{21}}^2 w_2 + \frac{1}{2}\left(\beta_{31}+\beta_{32}\right)^2 w_3 = \frac{1}{6},\ \beta_{21}\beta_{31}w_3 = \frac{1}{6}$$

$$\Rightarrow \beta_{21} = \alpha_2,\ \ \beta_{31}+\beta_{32} = \alpha_3$$

$$\Rightarrow w_1 + w_2 + w_3 = 1$$

$$\beta_{21}w_2 + \left(\beta_{31}+\beta_{32}\right)w_3 = \frac{1}{2}$$

$$\beta_{21}{}^2 w_2 + \left(\beta_{31}+\beta_{32}\right)^2 w_3 = \frac{1}{2},\ \ \beta_{21}\beta_{31}w_3 = \frac{1}{6}$$

$$\in_i (h) = y_{i+1} - y_i - h\left\{\left(w_1 f_i + w_2 f\left(x_i\right) + \alpha_2 h, y_i + \beta_{21} h f_i\right)\right.$$

$$\left. + w_3 f\left(x_i + \alpha_3 h, y_i + \beta_{31} h f_i + \beta_{32} M_2 h\right)\right\}$$

$$= h y_i{}' + \frac{h^2}{2} y_i{}'' - \left\{\left[h w_1 f_i + w_2 h\left(f_i + f_{i,x}\alpha_2 h + f_{i,y}\beta_{21} h f_i\right)\right] + \right.$$

$$\left. + w_3 h\left(f_i + f_{i,x}\alpha_3 h + f_{i,y}\left(\beta_{31} h f_i + \beta_{32} h f\left(x_i + \alpha_2 h, y_i + \beta_{21} h f_i\right)\right)\right)\right\}$$

$$= h f_i + \frac{h^2}{2}\left(f_{i,x} + f_{i,y} f_i\right) - h w_1 f_i - w_2 h f_i - w_2 h^2 f_{i,x}\alpha_2 - w_2 h^2 \beta_{21} f_i f_{i,y} -$$

$$- w_3 h f_i - w_3 h^2 \alpha_3 f_{i,x} - f_{i,y}\left[\beta_{31} h f_i + \beta_{32} h\left(f_i + f_{i,x}\alpha_2 h + f_{i,y}\beta_{21} h f_i\right)\right]$$

$$= \left(1 - w_1 - w_2 - w_3\right) h f_i + \left(\frac{1}{2} - w_2\alpha_2 - w_3\alpha_3\right) h^2 f_{i,x} + h^2\left(\frac{1}{2} - w_2\beta_{21} - w_3\left(\beta_{31}+\beta_{32}\right)\right) f_{i,y} f_i +$$

$$+ \frac{h^3}{2}\left(\frac{1}{3} - \alpha_2\beta_{21}w_2 - \alpha_3\left(\beta_{31}+\beta_{32}\right)w_3\right) f_{i,xy} f_i + \frac{h^3}{2}\left(\frac{1}{3} - {\alpha_2}^2 w_2 - {\alpha_3}^2 w_3\right) f_{i,xx} +$$

$$+ \frac{h^3}{2}\left(\frac{1}{3} - {\beta_{21}}^2 w_2 - \left(\beta_{31}+\beta_{32}\right)^2 w_3\right) f_{i,yy} f_i^2 + \frac{h^3}{2}\left(\frac{1}{6} - \alpha_2\beta_{31}w_3\right) f_{i,x} f_{i,y} +$$

$$+ h^3\left(\frac{1}{6} - \beta_{21}\beta_{31}w_3\right) f_{i,y}{}^2 f_i$$

α_2	β_{21}		
α_3	β_{31}	β_{32}	
	w_1	w_2	w_3

$\frac{1}{2}$	$\frac{1}{2}$		
1	-1	2	
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

... (1)

$$\alpha_2 = \frac{1}{2} = \beta_{21},\ \alpha_3 = \frac{3}{4},\ \beta_{31} = 0,\ \beta_{32} = \frac{3}{4},\ w_1 = \frac{2}{9},\ w_2 = \frac{3}{9},\ w_3 = \frac{4}{9} \qquad \text{... (2)}$$

$$\alpha_2 = \frac{1}{3} = \beta_{21},\ \alpha_3 = \frac{2}{3} = \beta_{32},\ \beta_{31} = 0,\ w_1 = \frac{1}{4},\ w_2 = 0,\ w_3 = \frac{3}{4} \qquad \text{... (3)}$$

Runge Kutta Method for Four Slopes

Order four method

$$y_0 = x_0, \qquad M_1 = f(x_i, y_i), \qquad M_2 = f\left(x_i + \frac{h}{2}, y_i + \frac{hM_1}{2}\right)$$

$$M_3 = f\left(x_i + \frac{h}{2} + y_i + \frac{hM_2}{2}\right) \qquad M_4 = f(x_i + h, y_i + hM_3)$$

$$y_{i+1} = y_i + \frac{h}{6}(M_1 + 2M_2 + 2M_3 + M_4)$$

Define

$$M_1 = f(x_i, y_i)$$

$$M_2 = f(x_i + \alpha_2 h, y_i + \beta_{21} hM_1)$$

$$M_3 = f(x_i + \alpha_3 h, y_i + \beta_{31} hM_1 + \beta_{32} hM_2) \qquad \text{... (1)}$$

$$M_4 = f(x_i + \alpha_4 h, y_i + \beta_{41} hM_1 + \beta_{42} hM_2 + \beta_{43} hM_3)$$

$$y_{i+1} = y_i + h(w_1 M_1 + w_2 M_2 + w_3 M_3 + w_4 M_4) \qquad \text{... (2)}$$

where the ----- α_2, α_3, α_4, β_{21}, β_{31}, β_{32}, β_{41}, β_{42}, β_{43}, and w_1, w_2, w_3, w_4 are chosen to make y_{i+1} closer to $y(x_{i+1})$.

Expanding as before (and matches coefficients of powers of h) we obtain the following system of equation

$$\alpha_2 = \beta_{21},\ \alpha_3 = \beta_{31} + \beta_{32},\ \alpha_4 = \beta_{41} + \beta_{42} + \beta_{43}$$

$$w_1 + w_2 + w_3 + w_4 = 1,\quad w_2\alpha_2 + w_3\alpha_3 + w_4\alpha_4 = \frac{1}{2}$$

$$w_2{\alpha_2}^2 + w_3{\alpha_3}^2 + w_4{\alpha_4}^2 = \frac{1}{3},\quad w_3\alpha_2\beta_{32} + w_4\left(\alpha_4\beta_{42} + \alpha_3\beta_{43}\right) = \frac{1}{6}$$

$$w_2{\alpha_2}^3 + w_3{\alpha_3}^3 + w_4{\alpha_4}^3 = \frac{1}{4},\ w_3{\alpha_2}^2\beta_{32} + w_4\left({\alpha_2}^2\beta_{42} + {\alpha_3}^2\beta_{43}\right) = \frac{1}{12}$$

$$w_3\alpha_2\alpha_3\beta_{32} + w_4\left(\alpha_2\beta_{42} + \alpha_3\beta_{43}\right)\alpha_4 = \frac{1}{8},\ w_4\alpha_2\beta_{32}\beta_{43} = \frac{1}{24}$$

The equations of the above form occur in all Runge-Kutta Methods. We have 11 equations in 13 unknowns.

The method (2) will correspond to Simpson's rule of integration.

If $\alpha_2 = \alpha_3$ and $w_2 = w_3$ the solution of equations above is given by

$$\alpha_2 = \alpha_3 = \frac{1}{2} \qquad \alpha_4 = 1 \qquad w_2 = w_3 = \frac{1}{3} \qquad w_1 = w_4 = \frac{1}{6}$$

$$\beta_{41} = 0 = \beta_{42}, \beta_{43} = 1$$

Thus the equation in (1) and (2) gives above.

To solve by ---- choice $\alpha_2 = \frac{1}{2},\ \beta_{31} = 0$.

α_2	β_{21}			
α_3	β_{31}	β_{32}		
α_4	β_{41}	β_{42}	β_{43}	
	w_1	w_2	w_3	w_4

$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

$$\alpha_2 = \frac{1}{3} = \beta_{21},\ \alpha_3 = \frac{2}{3},\ \beta_{31} = -\frac{1}{3},\ \beta_{32} = 1,\ \alpha_4 = 1,\ \beta_{41} = 1,\ \beta_{42} = -1,\ \beta_{43} = 1,$$

$$w_1 = \frac{1}{8} = w_4,\ w_2 = \frac{3}{8} = w_3$$

Example : Solve

$$y' = \frac{x - y}{2} \qquad [0, 3] \qquad h = 0.5 \qquad y_0 = 1$$

Answer :

Case 1 $h = 0.5$ $M_1 = 0.5$ $M_2 = -0.3125$ $M_3 = -0.3359375$ $M_4 = -0.1660156$

if $x = 0$ $y_1 = 1 + 0.08333(-0.5 - 2 \times 0.3125 - 2 \times 0.3359 - 0.1660156)$

$= 0.8364258$

if $x = 0.5$ $M_1 = 0.1682119$ $M_2 = -0.0221862$ $M_3 = -0.0404396$ $M_4 = 0.091897$

$y_2 = 0.8196285$

if $x = 1$ $M_1 = 0.0901857$ $M_2 = 0.2039125$ $M_3 = 0.1896966$ $M_4 = 0.2927615$

$y_3 = 0.9171423$

if $x = 1.5$ $M_1 = 0.2914288$ $M_2 = 0.3800002$ $M_3 = 0.3689288$ $M_4 = 0.44919$

$y_4 = 1.1036826$

if $x = 2$ $M_1 = 0.4481587$ $M_2 = 0.5171388$ $M_3 = 0.508516$ $M_4 = 0.571029$

$y_5 = 1.3595575$

if $x = 2.5$ $M_1 = 0.5702212$ $M_2 = 0.623943$ $M_3 = 0.6172283$ $M_4 = 0.6659141$

$y_6 = 1.6694308$

Case 2 : $h = 1$ $y_0 = 1$

if $x = 0$ $M_1 = -0.5$ $M_2 = -0.125$ $M_3 = -0.21875$ $M_4 = 0.10975$

$y_1 = 0.8203125$

if $x = 1$ $M_1 = 0.0898439$ $M_2 = 0.3173828$ $M_3 = 0.260498$ $M_4 = 0.4595947$

$y_2 = 1.1045125$

if $x = 2$ $M_1 = 0.4477437$ $M_2 = 0.5858078$ $M_3 = 0.5512918$ $M_4 = 0.61209$

$y_3 = 1.670186$

Case3 : $h = 1.5$ $y_0 = 1$

if $x = 0$ $M_1 = -0.5$ $M_2 = -0.0625$ $M_3 = -0.1015625$ $M_4 = 0.3261718$

$y_1 = 0.8745117$

if $x = 1.5$ $M_1 = 0.3127441$ $M_2 = 0.5704651$ $M_3 = 0.4738197$ $M_4 = 0.7073793$

$y_2 = 1.621685$

Case 4 : $h = 3$ $y_0 = 1$

if $x = 0$ $M_1 = -0.5$ $M_2 = 0.625$ $M_3 = -0.21875$ $M_4 = 1.328125$

$y_1 = 1.8203125$

Exercise :

1) $y' = 5(x-1)y$, y (0) = 5, [0, 2], h = 0.5, 0.2, 0.1

2) $y' = -2xy^2$, y (0) = 1, [0, 1], h = 0.2, 0.1, 0.05

3) $y' = x^2 + y^2$, y (0) = 0, [0, 2], h = 0.5

Example :

$h = 0.25$, $y_0 = 1$

$y_1 = 0.8974915$ $y_2 = 0.8364037$ $y_3 = 0.8128696$ $y_4 = 0.8195840$

(y_4') $y_5 = 0.9121021$ (y_5') $y_6 = 1.1036408$

(y_6') $y_7 = 1.3595168$ (y_7') $y_8 = 1.6693928$

$x = 0$ $M_1 = -0.5$ $M_2 = -0.40625$ $M_3 = -0.4121084$ $M_4 = -0.3134863$

$x = 0.25$ $M_1 = -0.375$ $M_2 = -0.2378082$ $M_3 = -0.2463827$ $M_4 = -0.1679479$

Find, $x = 0.5$

$x = 0.75$

$x = 1$

$x = 1.25$

$x = 1.5$

$x = 1.75$

$x = 2$

$x = 2.25$, and

$x = 2.5$

Systems of Differential Equations

$$\frac{dx}{dt} = f(t, x, y) \qquad x(t_0) = x_0$$

$$\frac{dy}{dt} = g(t, x, y) \qquad y(t_0) = y_0 \qquad \text{... (1)}$$

This can be written as,

$$x'(t) = f(t, x(t), y(t)) \qquad x(t_0) = x_0$$

$$y'(t) = g(t, x(t), y(t)) \qquad y(t_0) = y_0$$

Example :

$$\frac{dx}{dt} = x + 2y \qquad x(0) = 6$$

$$\frac{dy}{dt} = 3x + 2y \qquad y(0) = 4$$

Solution to the I.V.P is

$$x(t) = 4e^{4t} + 2e^{-t}$$

$$y(t) = 6e^{4t} - 2e^{-t}$$

The Runge-Kutta formulas of order 4 are

$$x_{k+1} = x_k + \frac{4}{6}(f_1 + 2f_2 + 2f_3 + f_4)$$

$$y_{k+1} = y_k + \frac{4}{6}(g_1 + 2g_2 + 2g_3 + g_4)$$

where

$$f_1 = f(t_k, x_k, y_k) \qquad g_1 = g(t_k, x_k, y_k)$$

$$f_2 = f\left(t_k + \frac{h}{2}, x_k + \frac{h}{2} f_1, y_k + \frac{h}{2} g_1\right) \qquad g_2 = g\left(t_k + \frac{h}{2}, x_k + \frac{h}{2} f_1, y_k + \frac{h}{2} g_1\right)$$

$$f_3 = f\left(t_k + \frac{h}{2}, x_k + \frac{h}{2} f_2, y_k + \frac{h}{2} g_2\right) \qquad g_3 = g\left(t_k + \frac{h}{2}, x_k + \frac{h}{2} f_2, y_k + \frac{h}{2} g_2\right)$$

$$f_4 = f(t_k + h, x_k + hf_3, y_k + hg_3) \qquad g_4 = g(t_k + h, x_k + hf_3, y_k + hg_3)$$

Example : Solve

$x' = x + 2y \qquad x(0) = 6 \qquad [0, 0.2] \qquad \text{h} = 0.02$

$y' = 3x + 2y \qquad y(0) = 4$

Solution :

$$f_1 = f(0,6,4) = 6 + 2 \times 4 = 14 \qquad g_1 = g(0,8,4) = 26$$

$$x_0 + \frac{h}{2} f_1 = 6.14 \qquad y_0 + \frac{h}{2} g_1 = 4.26$$

$$f_2 = f(0.01, 6.14, 4.26) = 14.66 \qquad g_2 = g(0.01, 6.14, 4.26) = 26.94$$

$$x_0 + \frac{h}{2} f_2 = 6.1466 \qquad y_0 + \frac{h}{2} f_2 = 4.2694$$

$$f_3 = f(0.01, 6.1466, 4.2694) = 14.6854 \qquad g_3 = g(0.01, 6.1466, 4.2694) = 26.9786$$

$$x_0 + hf_3 = 6.293708 \qquad y_0 + hg_3 = 4.539572$$

$$f_4 = f(0.02, 6.293708, 4.539572) = 15.372852$$

$$g_4 = g(0.02, 6.293708, 4.539572) = 27.96028$$

$$x_1 = 6 + \frac{0.02}{6}(14 + 2 \times 14.66 + 2 \times 14.6854 + 15.372852) = 6.29354551$$

$$y_1 = 4 + \frac{0.02}{6}(26 + 2 \times 26.94 + 2 \times 26.9786 + 27.960268) = 4.5393249$$

k	t_k	f_1	f_2	f_3	f_4	g_1	g_2	g_3	g_4	x_k	y_k
0	0	--	--	--	--	--	--	--	--	6	4
1	0.02	14	14.66	14.6854	15.372852	26	26.94	26.9786	27.960268	6.2935	4.8393
2	0.04	15.3721				27.9591				6.6150	5.11948
3	0.06									6.968525	5.74396
4	0.08									7.35474	6.41653
5	0.1									7.7769	7.1412
6	0.12									8.2381	7.9226
7	0.14									8.7414	8.7653
8	0.16									9.290209	9.6745
9	0.18									9.888271	10.6560
10	0.2									10.53962	11.715780

Exercise :

1) Solve the system $x' = 2x + 3y$, $y' = 2x + y$ with initial condition $x(0) = -2.7$, $y(0) = 2.8$ over the interval [0, 1] use h = 0.05.

2) Solve the system $x' = 3x - y$, $y' = 4x - y$ with initial condition $x(0) = 0.2$, $y(0) = 0.5$ over the interval [0, 2] use h = 0.05.

3) Solve the system $x' = x - 4y$, $y' = x + y$ with initial condition $x(0) = 2$, $y(0) = 3$ over the interval [0, 2] use h = 0.05.

4) Solve the system $x' = y - 4x$, $y' = x + y$ with initial condition $x(0) = 1$, $y(0) = 1$ over the interval [0, 1.2] use h = 0.05.