Basics in Computing

1.1 INTRODUCTION

We begin this chapter with some of the basic concepts of representation of numbers on computers and errors introduced during computation. Problem-solving using computers and the steps involved are also discussed in brief.

Many of the available digital computing systems fall mainly under four categories: personal computers, workstations, mainframe computers and super computers, based on their speed, cost and facilities. Mainframe and super computers being costly and are used only for research and development purposes. These computers involve large-scale programmes with huge data. In the new millennium, computers with storage capacities of several hundred billion words and capable of making 15 billion calculations a second are made available at select installations over the globe.

With the availability of such powerful digital computers and vastly improved numerical methods, scientists and engineers will be able to develop models that can be used for numerous purposes: weather prediction, effect of solar storms on Earth, the performance of an aircraft as a whole, some aspects of space flight simulations and many more practical problems meant for the welfare of the mankind.

Personal computers with local area networking (LAN) and pentium processors can meet most of the demands of teaching, project evaluations, computer-aided design and almost all business applications. In fact, many computer installations provide the user with the necessary software of routine nature in the name of utility sub-programmes.

1.2 REPRESENTATION OF NUMBERS

It is known that in our daily life, we use numbers based on the decimal system. In this system, we use ten symbols 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 and the number 10 is called the base of the system. Thus, when a base N is given, we need N different symbols 0, 1, 2, . . . , $(N-1)$ to represent an arbitrary number. The number systems commonly used in computers are shown as in Table 1.1.

Scanned with CamScanner

Table 1.1 Number Systems

Base, N	Number
2	Binary
8	Octal
10	Decimal
16	Hexadecimal

Thus, if a number system has only two symbols 0 and 1, then, its base is 2 and so on. In general, an arbitrary real number, a can be written as

$$a = a_m N^m + a_{m-1} N^{m-1} + \cdots + a_1 N^1 + a_0 + a_{-1} N^{-1} + \cdots + a_{-m} N^{-m}$$

In binary system, it has the form,

$$a = a_m 2^m + a_{m-1} 2^{m-1} + \cdots + a_1 2^1 + a_0 + a_{-1} 2^{-1} + \cdots + a_{-m} 2^{-m}$$

In hexadecimal system, we write it as

$$a = a_m 16^m + a_{m-1} 16^{m-1} + \cdots + a_1 16^1 + a_0 + a_{-1} 16^{-1} + \cdots + a^{-m} 16^{-m}$$

For example, the value of the decimal number 1729 is represented and calculated as

$$(1729)_{10} = 1 \times 10^3 + 7 \times 10^2 + 2 \times 10^1 + 9 \times 10^0$$

While the decimal equivalent of binary number 1.0011001 is

$$1 \times 2^0 + 0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} + 0 \times 2^{-5} + 0 \times 2^{-6} + 1 \times 2^{-7}$$

$$= 1 + \frac{1}{8} + \frac{1}{16} + \frac{1}{128} = (1.1953125)_{10}$$

Electronic computers use binary system whose base is 2. The two symbols used in this system are 0 and 1, which are called *binary digits* or simply *bits*. The internal representation of any data within a computer is in binary form. However, we prefer data input and output of numerical results in decimal system. Within the computer, the arithmetic is carried out in binary form. Infact, there is a built-in circuit design in every computer, which converts decimal input to binary and binary result to decimal output, and carry-out binary addition, subtraction, multiplication and division. For example, the method of converting decimal to binary equivalent can be seen as follows: we divide the given decimal number by 2 and the resulting successive quotients and continue to do so till the quotient becomes zero. Then, the binary equivalent of the decimal number is obtained as a string of remainders and the process can be seen through examples.

Similarly, for converting a decimal fraction into its binary equivalent, we multiply it by 2, the resultant integer part gives the most significant bit of the binary fraction. We keep multiplying by 2 and extract the next significant digit, and the process is continued until the fractional part becomes zero.

Scanned with CamScanner

Example 1.1 Convert the decimal number 47 into its binary equivalent.

Solution

		Remainder
2	47	↓
2	23	1
2	11	1
2	5	1
2	2	1
2	1	0
	0	1 ← Most significant bit

Thus,

$$(47)_{10} = (101111)_2$$

Example 1.2 Find the binary equivalent of the decimal fraction 0.7625.

Solution

	Product	Integer part	
0.7625 × 2	1.5250	1	← Most significant digit
0.5250 × 2	1.0500	1	
0.05 × 2	0.1	0	
0.1 × 2	0.2	0	
0.2 × 2	0.4	0	
0.4 × 2	0.8	0	
0.8 × 2	1.6	1	
0.6 × 2	1.2	1	
0.2 × 2	0.4	0	Repeated hereafter

Therefore,

$$(0.7625)_{10} = (0.11000011(0011))_2$$

Suppose, we consider 8 bits only, that is, $(0.11000011)_2$; its decimal value is equal to $(0.7617187)_{10}$. While if we take 12 bits, its decimal equivalent is $(0.76245)_{10}$.

Example 1.3 Convert $(59)_{10}$ into binary and then into octal.

Solution

		Remainder
2	59	
2	29	1
2	14	1
2	7	0
2	3	1
2	1	1
	0	1 ← Most significant bit

Scanned with CamScanner

Thus,
$$(59)_{10} = (111011)_2$$

Now, if we group the binary digits such that each group contains three bits each, we can easily go from binary to octal. Thus,

$$(111011)_2 = 111\ 011 = (73)_8$$

1.2.1 Floating-point Representation

In general, two types of arithmetic operations are carried out in computers: integer arithmetic and floating point arithmetic. However, most scientific and engineering calculations are essentially carried out in floating point arithmetic. For example, an n digit floating point number in base b can be represented as

$$a = \pm(.d_1 d_2, \ldots, d_n)_b b^e$$

where, $(.d_1 d_2, \ldots, d_n)_b$ is called *mantissa* and e is the *exponent*. Let us consider a binary number as

$$.10110101 \times 2^{11} = .10110101E01011$$

Here, .10110101 is called the mantissa and 1011 is the exponent. The precision of floating point numbers on any computer is determined by the number of digits used in the mantissa, which in general, varies.

Computers having 48 bits to represent single precision real number, in general, allocate 1 bit for sign, 7 bits for exponent and 40 bits for mantissa. Thus, the range is limited from 2.939E−39 to 1.701E+38. The numerical precision in this case is 11 decimal digits. Similarly, those computers having 32 bits to represent a single precision real number, in general, allocate 1 bit for sign, 7 bits for exponent and 24 bits for mantissa. In this case, the limit ranges from 2.939E−39 to 1.701E+38. The numerical precision in this case is only six decimal digits.

When 64-bit double precision arithmetic is used, the computer output may be accurate even up to 16 decimal digits.

1.3 ERRORS IN COMPUTATIONS

Numerically, computed solutions are subject to certain errors. It may be fruitful to identify the error sources and their growth while classifying the errors in numerical computation. There are essentially three error sources: inherent errors, local round-off errors and local truncation errors.

1.3.1 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.

Scanned with CamScanner

1.3.2 Local Round-off Errors

Every computer has a finite word length, and therefore, it is possible to store only a fixed number of digits of a given input number. Since computers store information in binary form, storing an exact decimal number in its binary form into the computer memory gives an error. This error is computer-dependent. Also, at the end of computation of a particular problem, the final results in the computer, which is obviously in binary form, should be converted into decimal form — a form understandable to the user — before their print out. Therefore, an additional error is committed at this stage too. This error is called *local round-off error*.

For example, in Section 1.2, we have noted that

$$(0.7625)_{10} = (0.11000011\ (0011))_2$$

If a particular computer system has a word length of 12 bits only, then the decimal number 0.7625 is stored in the computer memory in binary form as 0.110000110011. However, it is equivalent to 0.76245. Thus, in storing the number 0.7625, we have committed an error equal to 0.00005, which is the round-off error; inherent with the computer system considered. Thus, we define the *error* as

$$\text{Error} = \text{True value} - \text{Computed value}$$

Now, in order to determine the accuracy of an approximate solution, errors are measured in different ways. *Absolute error*, denoted by |Error|, while, the *relative error* is defined as

$$\text{Relative error} = \frac{|\text{Error}|}{|\text{True value}|}$$

For example, consider the value of $\sqrt{2}$ = (1.414213 . . .) up to four decimal places, then

$$\sqrt{2} = 1.4142 + \text{Error}$$

Hence, we get

$$\text{Absolute error} = |\text{Error}| = 0.00001$$

$$\text{Relative error} = \frac{0.00001}{1.4142}$$

We are aware of the fact that $\sqrt{2}$ is irrational. However, widely used value up to four decimal digits is taken as the true value for the computation of relative error. These error measures are generally used in numerical analysis for measuring the accuracy of the results

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

Scanned with CamScanner

we say that the number N has t significant digits. Here, d_1 is called the *most significant digit*. For example, 0.3 agrees with 1/3 to one significant digit, while 0.3333 agrees with 1/3 to four significant digits.

1.3.3 Local Truncation Error

It is generally easier to expand a function into a power series using Taylor series expansion and evaluate it by retaining the first few terms. For example, we may approximate the function $f(x) = \cos x$ by the series

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + \cdots \tag{1.1}$$

In fact, it is an infinite series expansion. If we use only the first three terms to compute $\cos x$ for a given x, we get an approximate answer. Here, the error is due to truncating the series. Suppose, we retain the first n terms, the *truncation error* (TE) is given by

$$\text{TE} \leq \frac{x^{2n+2}}{(2n+2)!} \tag{1.2}$$

It may be noted that the TE is independent of the computer used.

If we wish to compute $\cos x$ for $|x| < \pi/2$ accurate with five significant digits, the question is, how many terms in the expansion (1.1) are to be included? In this situation

$$\frac{x^{2n+2}}{(2n+2)!} < .5 \times 10^{-5} = 5 \times 10^{-6}$$

Taking logarithm on both sides, we get

$$(2n+2)\log x - \log[(2n+2)!] < \log_{10} 5 - 6 \log_{10} 10 = 0.699 - 6 = -5.3$$

or

$$\log[(2n+2)!] - (2n+2)\log x > 5.3$$

We can observe that, for x in the interval $[-\pi/2, \pi/2]$, the above inequality is satisfied for $n = 7$. Hence, seven terms in the expansion (1.1) are required to get the value of $\cos x$, with the prescribed accuracy, in the interval $[-\pi/2, \pi/2]$. In this example, the truncation error is given by

$$\text{TE} \leq \frac{x^{16}}{16!}$$

1.4 PROBLEM-SOLVING USING COMPUTERS

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

(i) Choosing an appropriate numerical method,

Scanned with CamScanner

(ii) Designing an algorithm,

(iii) Programming and debugging, and

(iv) Computer execution.

These steps are briefly explained as follows.

We define the numerical method as a mathematical formula for finding the solution to a given problem. There may be many methods available to solve the same problem. For example, in Chapter 2, we shall present various computer-based numerical methods, such as, bisection method, regula-falsi method, method of iteration, Newton–Raphson method, Muller's method, Graeffe's root squaring method, etc., for solving an algebraic or transcendental equation. One should choose an appropriate method, which suits best in the given situation, as a first step.

Once we chose a particular method for solving a problem, we should write down the sequence of steps to be followed in order, precisely and unambiguously, to obtain the solution. This is called *designing an algorithm.*

Now, the flow chart for the algorithm is drawn and then translated into a programming language, which we call *computer programme.* This programme should be debugged and free from coding errors. The choice of the languages is for the user to decide. It can be FORTRAN, BASIC, COBOL, Pascal, C, etc.

As a last step, the computer programme is fed to a personal computer or to a mainframe computer, with the necessary data through an input unit. Then, the central processing unit (CPU) of the computing system interprets the programme steps and executes them if the programme is free from coding errors. When it encounters output statement, the numerical answers to the problem are sent to the output unit chosen by the user; it may be a printer. This completes the problem-solving task using a computer.

Scanned with CamScanner

Chapter 2

Solution of Algebraic and Transcendental Equations

2.1 INTRODUCTION

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

$$x^3 + 5x^2 - 6x + 3 = 0$$

is an algebraic equation, whereas

$$M = E - e \sin E \quad \text{and} \quad ax^2 + \log (x - 3) + e^x \sin x = 0$$

are transcendental equations.

To find the solution of an equation $f(x) = 0$, we find those values of x for which $f(x) = 0$ is satisfied. Such values of x are called the *roots* of $f(x) = 0$. Thus a is a root of an equation $f(x) = 0$, if and only if, $f(a) = 0$.

Before, we develop various numerical methods, we shall list below some of the basic properties of an algebraic equation:

(i) Every algebraic equation of nth degree, where n is a positive integer, has n and only n roots.

(ii) Complex roots occur in pairs. That is, if $(a + ib)$ is a root of $f(x) = 0$, then $(a - ib)$ is also a root of this equation.

(iii) If $x = a$ is a root of $f(x) = 0$, a polynomial of degree n, then $(x - a)$ is a factor of $f(x)$. On dividing $f(x)$ by $(x - a)$ we obtain a polynomial of degree $(n - 1)$.

(iv) Descartes rule of signs: The number of positive roots of an algebraic 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 an equation

$$x^3 - 3x^2 + 4x - 5 = 0$$

Scanned with CamScanner

As there are three changes in sign, also, the degree of the equation is three, and hence the given equation will have all the three positive roots.

(v) Intermediate value property: If $f(x)$ is a real valued continuous function in the closed interval $a \le x \le b$. If $f(a)$ and $f(b)$ have opposite signs, then the graph of the function $y = f(x)$ crosses the x-axis at least once; that is $f(x) = 0$ has at least one root ξ such that $a < \xi < b$.

Broadly speaking, all the known numerical methods for solving either a transcendental equation or an algebraic equation can be classified into two groups: *direct methods* and *iterative methods*. Direct methods require no knowledge of the initial approximation of a root of the equation $f(x) = 0$, while iterative methods do require first approximation to initiate iteration. How to get the first approximation? We can find the approximate value of the root of $f(x) = 0$ either by a *graphical method* or by an *analytical method* as explained below:

Graphical method

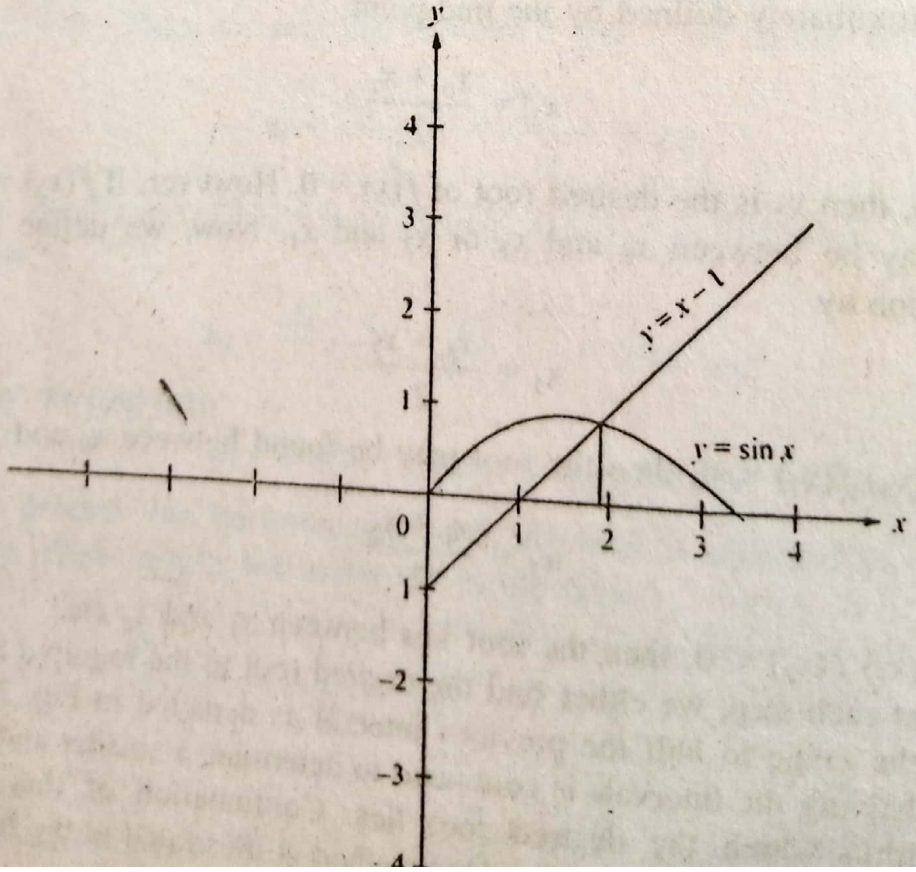
Often, the equation $f(x) = 0$ can be rewritten as $f_1(x) = f_2(x)$ and the first approximation to a root of $f(x) = 0$ can be taken as the abscissa of the point of intersection of the graphs of $y = f_1(x)$ and $y = f_2(x)$. For example, consider,

$$f(x) = x - \sin x - 1 = 0$$

It can be written as $x - 1 = \sin x$. Now, we shall draw the graphs of

$$y = x - 1 \quad \text{and} \quad y = \sin x$$

as shown in Fig. 2.1. The approximate value of the root is found to be 1.9.



Scanned with CamScanner

Analytical method

This method is based on 'intermediate value property'. We shall illustrate it through an example. Let,

$$f(x) = 3x - \sqrt{1 + \sin x} = 0$$

We can easily verify

$$f(0) = -1$$

$$f(1) = 3 - \sqrt{1 + \sin\left(1 \times \frac{180}{\pi}\right)} = 3 - \sqrt{1 + 0.84147} = 1.64299$$

We observe that $f(0)$ and $f(1)$ are of opposite signs. Therefore, using intermediate value property we infer that there is at least one root between $x = 0$ and $x = 1$. This method is often used to find the first approximation to a root of either transcendental equation or algebraic equation. Hence, in analytical method, we must always start with an initial interval (a, b), so that $f(a)$ and $f(b)$ have opposite signs.

2.2 BISECTION METHOD

This method is due to Bolzano. Suppose, we wish to locate the root of an equation $f(x) = 0$ in an interval, say (x_0, x_1). Let $f(x_0)$ and $f(x_1)$ are of opposite signs, such that $f(x_0)\, f(x_1) < 0$.

Then the graph of the function crosses the x-axis between x_0 and x_1, which guarantees the existence of at least one root in the interval (x_0, x_1). The desired root is approximately defined by the mid-point

$$x_2 = \frac{x_0 + x_1}{2}$$

If $f(x_2) = 0$, then x_2 is the desired root of $f(x) = 0$. However, if $f(x_2) \neq 0$, then the root may be between x_0 and x_2 or x_2 and x_1. Now, we define the next approximation by

$$x_3 = \frac{x_0 + x_2}{2}$$

provided $f(x_0)\, f(x_2) < 0$, then the root may be found between x_0 and x_2 or by

$$x_3 = \frac{x_1 + x_2}{2}$$

provided $f(x_1)\, f(x_2) < 0$, then the root lies between x_1 and x_2 etc.

Thus, at each step, we either find the desired root to the required accuracy or narrow the range to half the previous interval as depicted in Fig. 2.2. This process of halving the intervals is continued to determine a smaller and smaller interval within which the desired root lies. Continuation of this process eventually gives us the desired root. This method is illustrated in the following example.

Scanned with CamScanner

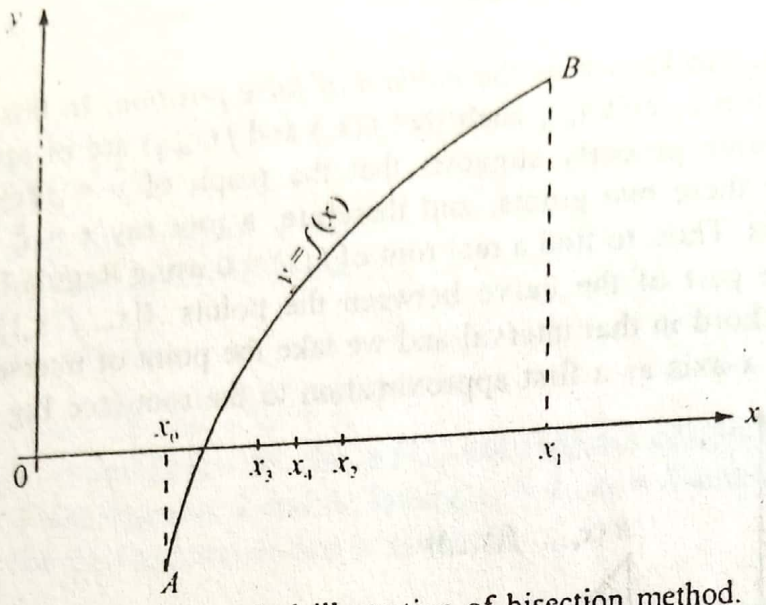


Fig. 2.2 Geometrical illustration of bisection method.

Example 2.1 Solve $x^3 - 9x + 1 = 0$ for the root between $x = 2$ and $x = 4$ by the bisection method.

Solution Given $f(x) = x^3 - 9x + 1$. We can verify $f(2) = -9, f(4) = 29$. Therefore, $f(2)\, f(4) < 0$ and hence the root lies between 2 and 4. Let $x_0 = 2$, $x_1 = 4$. Now, we define

$$x_2 = \frac{x_0 + x_1}{2} = \frac{2 + 4}{2} = 3$$

as a first approximation to a root of $f(x) = 0$ and note that $f(3) = 1$, so that $f(2)\, f(3) < 0$. Thus, the root lies between 2 and 3. We further define,

$$x_3 = \frac{x_0 + x_2}{2} = \frac{2 + 3}{2} = 2.5$$

and note that $f(x_3) = f(2.5) < 0$, so that $f(2.5)\, f(3) < 0$. Therefore, we define the mid-point,

$$x_4 = \frac{x_3 + x_2}{2} = \frac{2.5 + 3}{2} = 2.75, \text{ etc.}$$

Similarly, we find that

$$x_5 = 2.875 \quad \text{and} \quad x_6 = 2.9375$$

and the process can be continued until the root is obtained to the desired accuracy. These results are presented in the table.

n	x_n	$f(x_n)$
2	3	1.0
3	2.5	–5.875
4	2.75	–2.9531
5	2.875	–1.1113
6	2.9375	–0.0901

Scanned with CamScanner

2.3 REGULA–FALSI METHOD

This method is also known as the *method of false position*. In this method, we choose two points x_n and x_{n-1} such that $f(x_n)$ and $f(x_{n-1})$ are of opposite signs. Intermediate value property suggests that the graph of $y = f(x)$ crosses the x-axis between these two points, and therefore, a root say $x = \xi$ lies between these two points. Thus, to find a real root of $f(x) = 0$ using Regula-Falsi method, we replace the part of the curve between the points $A[x_n, f(x_n)]$ and $B[x_{n-1}, f(x_{n-1})]$ by a chord in that interval and we take the point of intersection of this chord with the x-axis as a first approximation to the root (see Fig. 2.3).

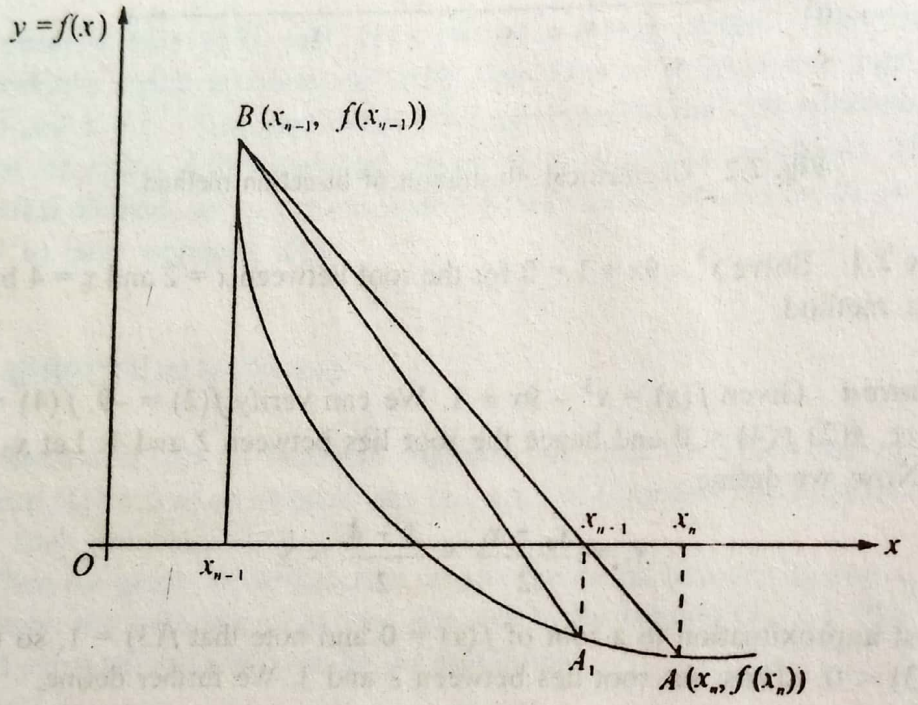


Fig. 2.3 Geometrical illustration of Regula-Falsi method.

Now, the equation of the chord joining the points A and B is

$$\frac{y - f(x_n)}{f(x_{n-1}) - f(x_n)} = \frac{x - x_n}{x_{n-1} - x_n} \tag{2.1}$$

Setting $y = 0$ in Eq. (2.1), we get

$$x = x_n - \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)} f(x_n)$$

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

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) \tag{2.2}$$

From Fig. 2.3, we observe that $f(x_{n-1})$ and $f(x_{n+1})$ are of opposite sign. Thus, it is possible to apply the above procedure, to determine the line through B and A_1 and so on. Hence, the successive approximations to the root of $f(x) = 0$ is

Scanned with CamScanner

given by Eq. (2.2). This method can best be understood through the following examples.

Example 2.2 Use the Regula-Falsi method to compute a real root of the equation $x^3 - 9x + 1 = 0$,

(i) if the root lies between 2 and 4
(ii) if the root lies between 2 and 3.

Comment on the results.

Solution Let $f(x) = x^3 - 9x + 1$.

(i) $f(2) = -9$ and $f(4) = 29$. Since $f(2)$ and $f(4)$ are of opposite signs, the root of $f(x) = 0$ lies between 2 and 4. Taking $x_1 = 2$, $x_2 = 4$ and using Regula-Falsi method, the first approximation is given by

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} f(x_2) = 4 - \frac{2 \times 29}{38} = 2.47368$$

and $f(x_3) = -6.12644$. Since $f(x_2)$ and $f(x_3)$ are of opposite signs, the root lies between x_2 and x_3. The second approximation to the root is given as

$$x_4 = x_3 - \frac{x_3 - x_2}{f(x_3) - f(x_2)} f(x_3) = 2.73989$$

and $f(x_4) = -3.090707$. Now, since $f(x_2)$ and $f(x_4)$ are of opposite signs, the third approximation is obtained from

$$x_5 = x_4 - \frac{x_4 - x_2}{f(x_4) - f(x_2)} f(x_4) = 2.86125$$

and $f(x_5) = -1.32686$. This procedure can be continued till we get the desired result. The first three iterations are shown as in the table.

n	x_{n+1}	$f(x_{n+1})$
2	2.47368	–6.12644
3	2.73989	–3.090707
4	2.86125	–1.32686

(ii) $f(2) = -9$ and $f(3) = 1$. Since $f(2)$ and $f(3)$ are of opposite signs, the root of $f(x) = 0$ lies between 2 and 3. Taking $x_1 = 2$, $x_2 = 3$ and using Regula-Falsi method, the first approximation is given by

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} f(x_2) = 3 - \frac{1}{10} = 2.9$$

and $f(x_3) = -0.711$. Since $f(x_2)$ and $f(x_3)$ are of opposite signs, the root lies between x_2 and x_3. The second approximation to the root is given as

$$x_4 = x_3 - \frac{x_3 - x_2}{f(x_3) - f(x_2)} f(x_3) = 2.94156$$

Scanned with CamScanner

and $f(x_4) = -0.0207$. Now, we observe that $f(x_2)$ and $f(x_4)$ are of opposite signs, the third approximation is obtained from

$$x_5 = x_4 - \frac{x_4 - x_2}{f(x_4) - f(x_2)} f(x_4) = 2.94275$$

and $f(x_5) = -0.0011896$. This procedure can be continued till we get the desired result. The first three iterations are shown as in the table.

n	x_{n+1}	$f(x_{n+1})$
2	2.9	−0.711
3	2.94156	−0.0207
4	2.94275	−0.0011896

From the above computations, we observe that the value of the root as a third approximation is evidently different in both the cases, while the value of x_5, when the interval considered is (2, 3), is closer to the root. Hence, an important observation in this method is that the interval (x_1, x_2) chosen initially in which the root of the equation lies must be sufficiently small.

Example 2.3 Use Regula-Falsi method to find a real root of the equation

$$\log x - \cos x = 0$$

accurate to four decimal places after three successive approximations.

Solution Given $f(x) = \log x - \cos x$. We observe that

$$f(1) = 0 - 0.5403 = -0.5403$$

and

$$f(2) = 0.69315 + 0.41615 = 1.1093$$

Since $f(1)$ and $f(2)$ are of opposite signs, the root lies between $x_1 = 1$, $x_2 = 2$. The first approximation is obtained from

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} f(x_2) = 2 - \frac{1.1093}{1.6496} = 1.3275$$

and

$$f(x_3) = 0.2833 - 0.2409 = 0.0424$$

Now, since $f(x_1)$ and $f(x_3)$ are of opposite signs, the second approximation is obtained as

$$x_4 = 1.3275 - \frac{(.3275)(.0424)}{0.0424 + 0.5403} = 1.3037$$

and

$$f(x_4) = 1.24816 \times 10^{-3}$$

Similarly, we observe that $f(x_1)$ and $f(x_4)$ are of opposite signs, so, the third approximation is given by

$$x_5 = 1.3037 - \frac{(0.3037)(0.001248)}{0.001248 + 0.5403} = 1.3030$$

Scanned with CamScanner

and

$$f(x_5) = 0.62045 \times 10^{-4}$$

Hence, the required real root is 1.3030.

Example 2.4 Using Regula-Falsi method, find the real root of the following equation correct to three decimal places:

$$x \log_{10} x = 1.2$$

Solution Let $f(x) = x \log_{10} x - 1.2$. We observe that $f(2) = -0.5979$, $f(3) = 0.2314$. Since $f(2)$ and $f(3)$ are of opposite signs, the real root lies between $x_1 = 2, x_2 = 3$. The first approximation is obtained from

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} f(x_2) = 3 - \frac{0.2314}{0.8293} = 2.72097$$

and $f(x_3) = -0.01713$. Since $f(x_2)$ and $f(x_3)$ are of opposite signs, the root of $f(x) = 0$ lies between x_2 and x_3. Now, the second approximation is given by

$$x_4 = x_3 - \frac{x_3 - x_2}{f(x_3) - f(x_2)} f(x_3) = 2.7402$$

and $f(x_4) = -3.8905 \times 10^{-4}$. Thus, the root of the given equation correct to three decimal places is 2.740.

2.4 METHOD OF ITERATION

The method of iteration can be applied to find a real root of the equation $f(x) = 0$ by rewriting the same in the form,

$$x = \phi(x) \tag{2.3}$$

For example, $f(x) = \cos x - 2x + 3 = 0$. It can be rewritten as

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

Let $x = \xi$ is the desired root of Eq. (2.3). Suppose x_0 is its initial approximation. The first and successive approximations to the root can be obtained as

$$\left.\begin{aligned} x_1 &= \phi(x_0) \\ x_2 &= \phi(x_1) \\ &\vdots \\ x_{n+1} &= \phi(x_n) \end{aligned}\right\} \tag{2.4}$$

Definition 2.1 Let $\{x_i\}$ be the sequence obtained by a given method and let $x = \xi$ denotes the root of the equation $f(x) = 0$. Then, the method is said to be *convergent*, if and only if

$$\lim_{n \to \infty} |x_n - \xi| = 0$$

The convergence of the above sequence to the root is stated as in Theorem 2.1.

Scanned with CamScanner

Theorem 2.1 Support $x = \xi$ be a root of the equation $f(x) = 0$, which can be rewritten as $x = \phi(x)$, contained in an interval I. Also, let $\phi(x)$ and $\phi'(x)$ be continuous in I. Then, if $|\phi'(x)| < 1$ for all x in I, the iterative process defined by $x_{n+1} = \phi(x_n)$ converges to the root $x = \xi$, if and only if, the initially chosen approximation $x_0 \in I$.

This method is illustrated through the following examples.

Example 2.5 Use the method of iteration to determine the real root of the equation $e^{-x} = 10x$ correct to four decimal places.

Solution Let $f(x) = e^{-x} - 10x = 0$, we observe that $f(0) = 1$ and $f(1) = -9.6321$. Since $f(0) < f(1)$ numerically, the root is near to $x = 0$. Now, we shall rewrite the given equation in the form

$$x = \frac{1}{10}e^{-x} = \phi(x)$$

Therefore,

$$\phi'(x) = -\frac{1}{10}e^{-x}$$

and

$$|\phi'(x)| = \frac{1}{10}e^{-x} = \frac{1}{10\,e^{x}} < 1$$

for all x in (0, 1). Hence, the method of iteration can be applied. Thus, we start with the initial value $x_0 = 0$, then

$$x_1 = \phi(x_0) = \frac{1}{10} = 0.1, \qquad f(x_1) = -0.09516$$

Similarly, the successive approximations are

$$x_2 = \phi(x_1) = \frac{1}{10}e^{-0.1} = \frac{0.904837}{10} = 0.09048, \quad f(x_2) = 0.00869$$

$$x_3 = \phi(x_2) = 0.091349, \qquad f(x_3) = -7.90877 \times 10^{-4}$$

$$x_4 = \phi(x_3) = 0.091274, \qquad f(x_4) = 2.75784 \times 10^{-5}.$$

Hence, the required root is 0.0913.

Example 2.6 Find a real root of the equation

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

by the method of iteration.

Solution We observe that $f(0) = -1, f(1) = 1$ which shows that there is a real root between $x = 0$ and $x = 1$. To find the real root, we rewrite the equation in the form

$$x^2(x + 1) = 1 \qquad \text{or} \qquad x = \frac{1}{\sqrt{x + 1}} = \phi(x)$$

Scanned with CamScanner

Therefore,

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

We note that $|\phi'(x)| < 1$, for all x in $(0, 1)$. Hence, the method of iteration is applicable here.

Taking the initial value $x_0 = 1$, we successively obtain the following values:

$$\begin{aligned}
x_1 &= \phi(x_0) = 1/\sqrt{2} = 0.70711, & f(x_1) &= -0.14644\\
x_2 &= \phi(x_1) = 0.76537, & f(x_2) &= 0.03414\\
x_3 &= \phi(x_2) = 0.75263, & f(x_3) &= 7.2213 \times 10^{-3}\\
x_4 &= \phi(x_3) = 0.75536, & f(x_4) &= 1.55658 \times 10^{-3}\\
x_5 &= \phi(x_4) = 0.75477, & f(x_5) &= -3.44323 \times 10^{-4}\\
x_6 &= \phi(x_5) = 0.7549, & f(x_6) &= 7.38295 \times 10^{-5}
\end{aligned}$$

Hence, the required root is 0.7549.

Note: The given equation can be rewritten in many ways. Suppose, we rewrite

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

Then

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

if we take $x = 1$, in the interval $(0, 1)$, $|\phi'(x)| = \infty$, then the condition $|\phi'(x)| < 1$ is violated.

2.5 NEWTON–RAPHSON METHOD

This is a very powerful method for finding the real root of an equation in the form, $f(x) = 0$. Suppose, x_0 is an approximate root of $f(x) = 0$. Let $x_1 = x_0 + h$, where h is small, be the exact root of $f(x) = 0$, then $f(x_1) = 0$. Now, expanding $f(x_0 + h)$ by Taylor's theorem, we get

$$f(x_0 + h) = f(x_0) + h f'(x_0) + \frac{h^2}{2} f''(x_0) + \cdots = 0 \tag{2.5}$$

Since h is small, we neglect terms containing h^2 and its higher powers, then

$$f(x_0) + h f'(x_0) = 0 \qquad \text{or} \qquad h = \frac{-f(x_0)}{f'(x_0)}$$

Therefore, a better approximation to the root is given by

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

Still better and successive approximations $x_2, x_3, \ldots, x_n$ to the root can obviously be obtained from the iteration formula,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.6}$$

Scanned with CamScanner

This is known as Newton–Raphson iteration formula, which has the following geometrical interpretation:

Suppose, the graph of the function $y = f(x)$ crosses the x-axis at α (*see* Fig. 2.4), then $x = \alpha$ is the root of the equation $f(x) = 0$.

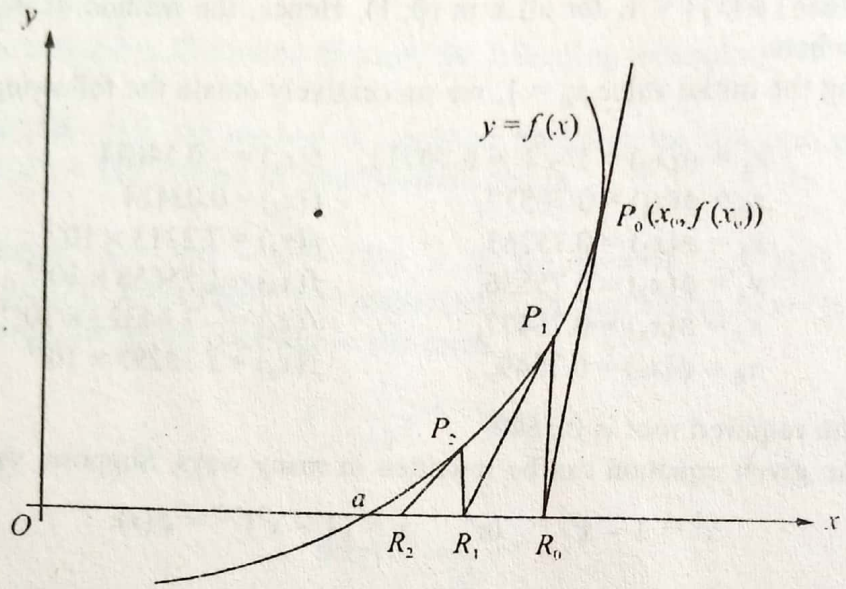


Fig. 2.4 Geometrical interpretation of Newton–Raphs

Let x_0 be a point closer to the root α, then the equation of the tangent at $P_0(x_0, f(x_0))$ is

$$y - f(x_0) = f'(x_0)\,(x - x_0) \tag{2.7}$$

This tangent cuts the x-axis at $R_0\,(x_1, 0)$. Therefore,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{2.8}$$

which is a first approximation to the root α. If P_1 is a point on the curve corresponding to x_1, then the tangent at P_1 cuts the x-axis at $R_1(x_2, 0)$, which is still closer to α, than x_1. Therefore, x_2 is a second approximation to the root. Continuing this process, we arrive at the root α, very rapidly, which is evident from Fig. 2.4. Thus, in this method, we have replaced the part of the curve between the point P_0 and x-axis by a tangent to the curve at P_0 and so on. In order to illustrate this method, we shall consider the following examples.

Example 2.7 Find the real root of the equation $xe^x - 2 = 0$ correct to two decimal places, using Newton–Raphson method.

Solution Given $f(x) = xe^x - 2$, we have

$$f'(x) = xe^x + e^x \text{ and } f''(x) = xe^x + 2e^x$$

clearly, we have

$$f(0) = -2 \quad \text{and} \quad f(1) = e - 2 = 0.71828$$

Hence, the required root lies in the interval (0, 1) and is nearer to 1.

Also, $f'(x)$ and $f''(x)$ do not vanish in (0, 1) and $f(x)$ and $f''(x)$ will have the

Scanned with CamScanner

same sign at $x = 1$. Therefore, we take the first approximation $x_0 = 1$, and using Newton–Raphson method, we get

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = \frac{e+2}{2e} = 0.867879$$

and

$$f(x_1) = 6.71607 \times 10^{-2}$$

The second approximation is

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 0.867879 - \frac{0.06716}{4.44902} = 0.85278$$

and

$$f(x_2) = 7.655 \times 10^{-4}$$

Thus, the required root is 0.853.

Example 2.8 Find a real root of the equation $x^3 - x - 1 = 0$ using Newton–Raphson method, correct to four decimal places.

Solution Let $f(x) = x^3 - x - 1$, then we observe that $f(1) = -1, f(2) = 5$. Therefore, the root lies in the interval (1, 2). We also observe

$$f'(x) = 3x^2 - 1, \qquad f''(x) = 6x$$

and

$$f(1) = -1, \qquad f''(1) = 6, \qquad f(2) = 5, \qquad f''(2) = 12$$

Since $f(2)$ and $f''(2)$ are of the same sign, we choose $x_0 = 2$ as the first approximation to the root. The second approximation is computed using Newton–Raphson method as

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 2 - \frac{5}{11} = 1.54545 \quad \text{and} \quad f(x_1) = 1.14573$$

The successive approximations are

$$x_2 = 1.54545 - \frac{1.14573}{6.16525} = 1.35961, \qquad f(x_2) = 0.15369$$

$$x_3 = 1.35961 - \frac{0.15369}{4.54562} = 1.32579, \qquad f(x_3) = 4.60959 \times 10^{-3}$$

$$x_4 = 1.32579 - \frac{4.60959 \times 10^{-3}}{4.27316} = 1.32471, \qquad f(x_4) = -3.39345 \times 10^{-5}$$

$$x_5 = 1.32471 + \frac{3.39345 \times 10^{-5}}{4.26457} = 1.324718, \qquad f(x_5) = 1.823 \times 10^{-7}$$

Hence, the required root is 1.3247.

Scanned with CamScanner

Convergence of Newton–Raphson method

To examine the convergence of Newton–Raphson formula (2.6), that is,

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

We compare it with the general iteration formula $x_{n+1} = \phi(x_n)$, and thus obtain

$$\phi(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}$$

In general, we write it as

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

We have already noted in Theorem 2.1 that the iteration method converges if $|\phi'(x)| < 1$. Therefore, Newton–Raphson formula (2.6) converges, provided

$$|f(x)f''(x)| < |f'(x)|^2 \tag{2.9}$$

in the interval considered. Newton–Raphson formula therefore converges, provided the initial approximation x_0 is chosen sufficiently close to the root and $f(x)$, $f'(x)$ and $f''(x)$ are continuous and bounded in any small interval containing the root.

Definition 2.2 Let

$$x_n = \alpha + \varepsilon_n, \qquad x_{n+1} = \alpha + \varepsilon_{n+1}$$

where α is a root of $f(x) = 0$. If we can prove that $\varepsilon_{n+1} = K\varepsilon_n^p$, where K is a constant and ε_n is the error involved at the n th step, while finding the root by an iterative method, then the rate of convergence of the method is p.

We can now establish that Newton–Raphson method converges quadratically. Let

$$x_n = \alpha + \varepsilon_n, \qquad x_{n+1} = \alpha + \varepsilon_{n+1}$$

where α is a root of $f(x) = 0$ and ε_n is the error involved at the nth step, while finding the root by Newton–Raphson formula (2.6). Then, Eq. (2.6) gives,

$$\alpha + \varepsilon_{n+1} = \alpha + \varepsilon_n - \frac{f(\alpha + \varepsilon_n)}{f'(\alpha + \varepsilon_n)}$$

i.e.

$$\varepsilon_{n+1} = \varepsilon_n - \frac{f(\alpha + \varepsilon_n)}{f'(\alpha + \varepsilon_n)} = \frac{\varepsilon_n f'(\alpha + \varepsilon_n) - f(\alpha + \varepsilon_n)}{f'(\alpha + \varepsilon_n)}$$

Using Taylor's expansion, we get

$$\varepsilon_{n+1} = \frac{1}{f'(\alpha) + \varepsilon_n f''(\alpha) + \cdots} \left\{ \varepsilon_n [f'(\alpha) + \varepsilon_n f''(\alpha) + \cdots] \right.$$

$$\left. - \left[f(\alpha) + \varepsilon_n f'(\alpha) + \frac{\varepsilon_n^2}{2} f''(\alpha) + \cdots \right] \right\}$$

Scanned with CamScanner

Since α is a root, $f(\alpha) = 0$. Therefore, the above expression simplifies to

$$\varepsilon_{n+1} = \frac{\varepsilon_n^2}{2} f''(\alpha) \frac{1}{f'(\alpha) + \varepsilon_n f''(\alpha)}$$

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

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

or

$$\varepsilon_{n+1} = \frac{\varepsilon_n^2}{2} \frac{f''(\alpha)}{f'(\alpha)} + O\left(\varepsilon_n^3\right)$$

On neglecting terms of order ε_n^3 and higher powers, we obtain

$$\varepsilon_{n+1} = K\varepsilon_n^2 \tag{2.10}$$

where

$$K = \frac{f''(\alpha)}{2f'(\alpha)} \tag{2.11}$$

It shows that Newton–Raphson method has second order convergence or converges quadratically.

Example 2.9 Set up Newton's scheme of iteration for finding the square root of a positive number N.

Solution The square root of N can be carried out as a root of the equation $x^2 - N = 0$. Let $f(x) = x^2 - N$. By Newton's method, we have

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

In this problem, $f(x) = x^2 - N$, $f'(x) = 2x$. Therefore,

$$x_{n+1} = x_n - \frac{x_n^2 - N}{2x_n} = \frac{1}{2}\left(x_n + \frac{N}{x_n}\right) \tag{2.12}$$

Example 2.10 Evaluate $\sqrt{12}$, by Newton's formula.

Solution Since $\sqrt{9} = 3$, $\sqrt{16} = 4$, we take $x_0 = (3 + 4)/2 = 3.5$. Using Eq. (2.12), we have

$$x_1 = \frac{1}{2}\left(x_0 + \frac{N}{x_0}\right) = \frac{1}{2}\left(3.5 + \frac{12}{3.5}\right) = 3.4643$$

$$x_2 = \frac{1}{2}\left(3.4643 + \frac{12}{3.4643}\right) = 3.4641$$

Scanned with CamScanner

$$x_3 = \frac{1}{2}\left(3.4641 + \frac{12}{3.4641}\right) = 3.4641$$

Hence, $\sqrt{12} = 3.4641$.

Example 2.11 Obtain the Newton–Raphson extended formula

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} - \frac{1}{2}\frac{[f(x_0)]^2}{[f'(x_0)]^3} f''(x_0)$$

for finding the root of the equation $f(x) = 0$.

Solution Expanding $f(x)$ by Taylor's series, in the neighbourhood of x_0, we obtain after retaining the first order term only

$$0 = f(x) = f(x_0) + (x - x_0) f'(x_0) + \cdots$$

Which gives

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

This is the first approximation to the root. Therefore,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{2.13}$$

Again, expanding $f(x)$ by Taylor's series and retaining up to second order term, we have

$$0 = f(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{(x - x_0)^2}{2} f''(x_0)$$

Therefore,

$$f(x_1) = f(x_0) + (x_1 - x_0) f'(x_0) + \frac{(x_1 - x_0)^2}{2} f''(x_0) = 0$$

Using Eq. (2.13), the above equation reduces to the form

$$f(x_0) + (x_1 - x_0) f'(x_0) + \frac{1}{2}\frac{[f(x_0)]^2}{[f'(x_0)]^2} f''(x_0) = 0$$

Thus, the Newton–Raphson extended formula is given by

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} - \frac{1}{2}\frac{[f(x_0)]^2}{[f'(x_0)]^3} f''(x_0) \tag{2.14}$$

This is also known as Chebyshev's formula of third order.

2.6 MULLER'S METHOD

In Muller's method, $f(x) = 0$ is approximated by a second degree polynomial; that is by a quadratic equation that fits through three points in the vicinity of a root.

Scanned with CamScanner

and

$$g(x_2, y_2) = -0.01055.$$

Further continuation gives

$$f_x|_{(x_2,y_2)} = 1.2617, \quad f_y|_{(x_2,y_2)} = -3.1829$$

$$g_x|_{(x_2,y_2)} = 3.1829, \quad g_y|_{(x_2,y_2)} = 1.2617$$

and

$$J = \begin{vmatrix} f_x & f_y \\ g_x & g_y \end{vmatrix} = 11.7227 \neq 0$$

which yield

$$h = 0.006379, \qquad k = -0.0773$$

Thus, the third approximation is

$$x_3 = x_2 + h = 0.8901, \qquad y_3 = y_2 + k = 0.5926$$

and

$$f(x_3, y_3) = -0.0127, \quad g(x_3, y_3) = 0.0152$$

The three iterations are tabulated as

i	x_i	y_i	$f(x_i, y_i)$	$g(x_i, y_i)$
1	0.9	0.7	–0.394	–0.042
2	0.8837	0.6003	–0.0327	–0.0106
3	0.8901	0.5926	–0.0127	0.0152

EXERCISES

2.1 Find the real root of the equation, $x^3 - 3x - 5 = 0$ by the bisection method.

2.2 Find the real root of the equation, $x^3 + x - 3 = 0$, using Regula-Falsi method, correct to four places of decimal.

2.3 Find the real root of the equation $x^6 - x^4 - x^3 - 1 = 0$, which lies between 1.4 and 1.5, correct to four places of decimal, by the method of false position, obtained after three successive approximations.

2.4 Use Regula-Falsi method to find the real roots of the equation $x^3 - \sin x + 1 = 0$ correct to four decimal places after three successive approximations between $(-2, -1)$.

2.5 Explain the method of false position for finding a real root of the equation $f(x) = 0$, and hence derive the general formula.

Scanned with CamScanner

2.6 Use Regula-Falsi method to compute the root of the equation $\cos x - xe^x = 0$.

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

2.8 Find a root of the equation $x \log_{10} x = 4.77$ by Newton–Raphsor method, correct to two decimal places.

2.9 Explain the Newton–Raphson method to find a root of the equation $f(x) = 0$, and hence derive its iteration formula.

2.10 Geometrically explain Newton–Raphson method to find a root of the equation $f(x) = 0$ and hence derive the general formula.

2.11 Obtain the real root of the equation $x^3 - 3x - 5 = 0$ using Newton–Raphson method, after third iteration.

2.12 Find a real root of the equation, $x^4 - x - 10 = 0$ using Newton–Raphson method correct to four decimal places.

2.13 Apply Newton–Raphson method to determine a root of the equation $\cos x = x e^x$ correct to three decimal places, using the initial approximation, $x_0 = 1$.

2.14 Set up the Newton's scheme of iteration for finding the p-th root of a positive number N. Hint: Ex 2.9, 2.10

2.15 Obtain the cube root of 12 using Newton–Raphson iteration.

2.16 Find the first approximation of the root of the equation $x^3 - 3x - 5 = 0$ using Muller's method, which lies between 2 and 3.

2.17 Find the first approximation to the root of the equation

$$f(x) = \sin x - \frac{x}{2} = 0$$

near $x = 2.0$, using Muller's method.

2.18 Using Graeffe's root squaring method, find the roots of the equation $x^3 - 4x^2 + 3x + 1 = 0$ with the help of a calculator.

2.19 Using the method of false position, find the root of $x \sin x - 1 = 0$ which lies in the interval $(0, 2)$.

2.20 Find the quadratic factors of

$$x^4 - 5.7x^3 + 26.7x^2 - 42.21x + 69$$

Using Bairstows method with $(x^2 - 1.5x + 4.3)$ as a starting factor.

2.21 Using Bairstows method, find the quadratic factors of the polynomial

$$2x^4 + 7x^3 - 4x^2 + 29x + 14$$

with $(x^2 + 5x + 2)$ as a starting factor.

Scanned with CamScanner

2.22 Find the solution of

$$f(x, y) = x^3 - 3xy^2 + 1 = 0$$
$$g(x, y) = 3x^2y - y^3 = 0$$

taking (1, 1) as the initial approximation using Newtons method.

2.23 Using Newtons method, find the solution of

$$f(x, y) = 4x^2 + y^2 + 2xy - y - 2 = 0$$
$$g(x, y) = 2x^2 + 3xy + y^2 - 3 = 0$$

taking (0.4, 0.9) as the initial approximation.

Scanned with CamScanner

Chapter 3

Solution of Linear System of Equations and Matrix Inversion

3.1 INTRODUCTION

Many real-life problems in engineering give rise to a system of linear equations. For example, such systems occur in certain applications of statistical analysis and in finding the numerical solution of partial differential equations and so on. It is therefore, natural to seek efficient methods for solving these equations numerically.

The general form of a system of m linear equations in n unknowns x_1, x_2, x_3, ..., x_n can be represented in matrix form as under:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \tag{3.1}$$

Using matrix notation, the above system can be written in compact form as

$$[A]\,(X) = (B) \tag{3.2}$$

The solution of the system of equations (3.2) gives n unknown values x_1, x_2, ..., x_n, which satisfy the system simultaneously. If $m > n$, we may not be able to find a solution, in principle, which satisfy all the equations. If $m < n$, the system usually will have an infinite number of solutions. However, in this chapter, we shall restrict to the case $m = n$. In this case, if $|A| \neq 0$, then the system will have a unique solution, while, if $|A| = 0$, then there exists no solution.

Various numerical methods are available for finding the solution of the system of equations (3.2), and they are classified as *direct* and *iterative methods*. In direct methods, we get the solution of the system after performing all the steps involved in the procedure. The direct methods consist of *elimination methods* and *decomposition methods*. In this chapter, under elimination methods, we consider, *Gaussian elimination* and *Gauss–Jordan elimination methods*. *Crout's reduction* also known as *Cholesky's reduction* is

Scanned with CamScanner

considered under decomposition methods. Under iterative methods, the initial approximate solution is assumed to be known and is improved towards the exact solution in an iterative way. We consider *Jacobi, Gauss–Seidel* and *relaxation methods* under iterative methods. All these methods are easily adoptable to computers and can be used to solve even hundred or more simultaneous linear equations.

3.2 GAUSSIAN ELIMINATION METHOD

In the Gaussian elimination method, the solution to the system of Eqs. (3.2) is obtained in two stages. In the first stage, the given system of equations is reduced to an equivalent upper triangular form using elementary transformations. In the second stage, the upper triangular system is solved using back substitution procedure by which we obtain the solution in the order x_n, x_{n-1}, x_{n-2}, ..., x_2, x_1.

This method is explained by considering a system of n equations in n unknowns in the form as follows

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \quad\quad \vdots \quad\quad\quad\quad \vdots \quad\quad \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}\right\} \tag{3.3}$$

Stage I: We divide the first equation by a_{11} and then subtract this equation multiplied by a_{21}, a_{31}, ..., a_{n1} from the 2nd, 3rd, ..., n th equation. Then the system (3.3) reduces to the following form:

$$\left.\begin{aligned} x_1 + a'_{12}x_2 + \cdots + a'_{1n}x_n &= b'_1 \\ a'_{22}x_2 + \cdots + a'_{2n}x_n &= b'_2 \\ \vdots \quad\quad\quad\quad \vdots \quad\quad \vdots & \\ a'_{n2}x_2 + \cdots + a'_{nn}x_n &= b'_n \end{aligned}\right\} \tag{3.4}$$

Here, we can observe that the last $(n - 1)$ equations are independent of x_1, that is, x_1 is eliminated from the last $(n - 1)$ equations.

This procedure is repeated with the second equation of (3.4), that is, we divide the second equation by a'_{22} and then x_2 is eliminated from 3rd, 4th, ..., nth equations of (3.4). The same procedure is repeated again and again till the given system assumes the following upper triangular form:

$$\left.\begin{aligned} c_{11}x_1 + c_{12}x_2 + \cdots + c_{1n}x_n &= d_1 \\ c_{22}x_2 + \cdots + c_{2n}x_n &= d_2 \\ \vdots \quad\quad\quad\quad \vdots & \\ c_{nn}x_n &= d_n \end{aligned}\right\} \tag{3.5}$$

Scanned with CamScanner

Stage II: Now, the values of the unknowns are determined by back substitution procedure, in which we obtain x_n from the last equation of (3.5) and then substituting this value of x_n in the preceding equation, we get the value of x_{n-1}. Continuing this way, we can find the values of all other unknowns in the order $x_n, x_{n-1}, \ldots, x_2, x_1$.

In this method, we observe that the determinant of the coefficient matrix is obtained as a by-product, that is,

$$|A| = c_{11}c_{22}\ldots c_{nn} \tag{3.6}$$

To familiarize with the method, we consider the following example:

Example. 3.1 Solve the following system of equations using Gaussian elimination method

$$2x + 3y - z = 5$$
$$4x + 4y - 3z = 3$$
$$-2x + 3y - z = 1$$

Solution The given system of equations is solved in two stages.

Stage I (Reduction to upper-triangular form): We divide the first equation by 2 and then subtract the resulting equation (multiplied by 4 and –2) from the second and third equations respectively. Thus, we eliminate x from the 2nd and 3rd equations. The resulting new system is given by

$$\left.\begin{aligned} x + \frac{3}{2}y - \frac{z}{2} &= \frac{5}{2} \\ -2y - z &= -7 \\ 6y - 2z &= 6 \end{aligned}\right\} \tag{1}$$

Now, we divide the second equation of (1) by –2 and eliminate y from the last equation and the modified system is given by

$$\left.\begin{aligned} x + \frac{3}{2}y - \frac{z}{2} &= \frac{5}{2} \\ y + \frac{z}{2} &= \frac{7}{2} \\ -5z &= -15 \end{aligned}\right\} \tag{2}$$

Stage II (Back substitution): From the last equation of (2), we immediately get

$$z = 3 \tag{3}$$

using this value of z, the second equation of (2) gives

$$y = \frac{7}{2} - \frac{3}{2} = 2 \tag{4}$$

Scanned with CamScanner

Using these values of y and z in the first equation of (2), we get

$$x = \frac{5}{2} + \frac{3}{2} - 3 = 1 \tag{5}$$

Thus, the solution of the given system is given by Eqs. (3) – (5).

Partial and full pivoting

The Gaussian elimination method fails if any one of the pivot elements becomes zero. In such a situation, we rewrite the equations in a different order to avoid zero pivots. Changing the order of equations is called *pivoting*.

We now introduce the concept of partial pivoting. In this technique, if the pivot a_{ii} happens to be zero, then the ith column elements are searched for the numerically largest element. Let the jth row ($j > i$) contains this element, then we interchange the ith equation with the jth equation and proceed for elimination. This process is continued whenever pivots become zero during elimination. For example, let us examine the solution of the following simple system

$$10^{-5}x_1 + x_2 = 1$$
$$x_1 + x_2 = 2$$

Using Gaussian elimination method with and without partial pivoting, assuming that we require the solution accurate to only four decimal places. The solution by Gaussian elimination gives $x_1 = 0$, $x_2 = 1$. If we use partial pivoting, the system takes the form

$$x_1 + x_2 = 2$$
$$10^{-5}x_1 + x_2 = 1$$

Using Gaussian elimination method, the solution is found to be $x_1 = 1$, $x_2 = 1$, which is a meaningful and perfect result.

In full pivoting which is also known as *complete pivoting*, we interchange rows as well as columns, such that the largest element in the matrix of the system becomes the pivot element. In this process, the position of the unknown variables also get changed. Full pivoting, in fact, is more complicated than the partial pivoting. Partial pivoting is preferred for hand computation.

Example 3.2 Solve the system of equations

$$x + y + z = 7$$
$$3x + 3y + 4z = 24$$
$$2x + y + 3z = 16$$

by Gaussian elimination method with partial pivoting.

Solution In matrix notation, the given system can be written as

$$\begin{bmatrix} 1 & 1 & 1 \\ 3 & 3 & 4 \\ 2 & 1 & 3 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 7 \\ 24 \\ 16 \end{pmatrix} \tag{1}$$

Scanned with CamScanner

To start with, we observe that the pivot element $a_{11} = 1$ ($\neq 0$). However, a glance at the first column reveals that the numerically largest element is 3 which is in the second row. Hence, we interchange the first row with the second row and then proceed for elimination. Thus, Eq. (1) takes the form

$$\begin{bmatrix} 3 & 3 & 4 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 24 \\ 7 \\ 16 \end{pmatrix} \tag{2}$$

after partial pivoting.

Stage I (Reduction to upper triangular form): By dividing the first row of the system (2) by 3 and then subtracting the resulting row, multiplied by 1 and 2 from the second and third rows of the system (2), we get

$$\begin{bmatrix} 1 & 1 & \frac{4}{3} \\ & & -\frac{1}{3} \\ & -1 & \frac{1}{3} \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ -1 \\ 0 \end{pmatrix} \tag{3}$$

The second row in Eq. (3) cannot be used as the pivot row, as $a_{22} = 0$. Interchanging the second and third rows, we obtain

$$\begin{bmatrix} 1 & 1 & \frac{4}{3} \\ & -1 & \frac{1}{3} \\ & & -\frac{1}{3} \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ 0 \\ -1 \end{pmatrix} \tag{4}$$

which is in the upper triangular form.

Stage II (Back substitution): From the last row of Eq. (4), we at once get

$$z = 3 \tag{5}$$

The second row of Eq. (4) with this value of z gives

$$-y + 1 = 0 \quad \text{or} \quad y = 1 \tag{6}$$

Using these values of y and z, the first row of Eq. (4) gives

$$x + 1 + 4 = 8 \quad \text{or} \quad x = 3 \tag{7}$$

Thus, Eqs. (5)–(7) constitute the solution to the given system of equations.

Example 3.3 Solve by Gaussian elimination method with partial pivoting, the following system of equations:

Scanned with CamScanner

$$0x_1 + 4x_2 + 2x_3 + 8x_4 = 24$$
$$4x_1 + 10x_2 + 5x_3 + 4x_4 = 32$$
$$4x_1 + 5x_2 + 6.5x_3 + 2x_4 = 26$$
$$9x_1 + 4x_2 + 4x_3 + 0x_4 = 21$$

Solution In matrix notation, the given system can be written as

$$\begin{bmatrix} 0 & 4 & 2 & 8 \\ 4 & 10 & 5 & 4 \\ 4 & 5 & 6.5 & 2 \\ 9 & 4 & 4 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 24 \\ 32 \\ 26 \\ 21 \end{pmatrix} \tag{1}$$

To start with, we observe that the pivot row, that is, the first row has a zero pivot element ($a_{11} = 0$). This row should be interchanged with any row following it, which on becoming a pivot row should not have a zero pivot element. While interchanging rows it is better to interchange with a row having largest pivotal element. Thus, we interchange the first and fourth rows, which is called partial pivoting and get,

$$\begin{bmatrix} 9 & 4 & 4 & 0 \\ 4 & 10 & 5 & 4 \\ 4 & 5 & 6.5 & 2 \\ 0 & 4 & 2 & 8 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 21 \\ 32 \\ 26 \\ 24 \end{pmatrix} \tag{2}$$

We observe that, in partial pivoting, the unknown vector remains unaltered, while the right-hand side vector gets changed.

Now, we shall carry out Gaussian elimination process in two stages.

Stage I (Reduction to upper-triangular form): In this stage, by dividing the first row of the system (2) by 9 and then subtracting this resulting row, multiplied by 4 and 4 from the second and third rows of Eq. (2), we get

$$\begin{bmatrix} 1 & \frac{4}{9} & \frac{4}{9} & 0 \\ 0 & 8.2222 & 3.2222 & 4 \\ 0 & 3.2222 & 4.7222 & 2 \\ 0 & 4 & 2 & 8 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2.3333 \\ 22.6666 \\ 16.6666 \\ 24 \end{pmatrix} \tag{3}$$

Now, we divide the second pivot row by 8.2222 and subtract the resultant row multiplied by 3.2222 and 4 from the third and fourth rows of Eq. (3) to get

$$\begin{bmatrix} 1 & \frac{4}{9} & \frac{4}{9} & 0 \\ 0 & 1 & 0.3919 & 0.4865 \\ 0 & 0 & 3.4594 & 0.4324 \\ 0 & 0 & 0.4324 & 6.0540 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2.3333 \\ 2.7568 \\ 7.7836 \\ 12.9728 \end{pmatrix} \tag{4}$$

Scanned with CamScanner

Finally, we divide the third pivot row by 3.4594 and subtract the resultant row multiplied by 0.4324 from fourth row of Eq. (4), thus getting the upper triangular form

$$\begin{bmatrix} 1 & \frac{4}{9} & \frac{4}{9} & 0 \\ 0 & 1 & 0.3919 & 0.4865 \\ 0 & 0 & 1 & 0.1250 \\ 0 & 0 & 0 & 5.9999 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2.3333 \\ 2.7568 \\ 2.2500 \\ 11.9999 \end{pmatrix} \tag{5}$$

Stage II (Back substitution): From the last row of Eq. (5), we immediately get $x_4 = 2.0000$. Using this value of x_4 into the third row of Eq. (5), we obtain

$$x_3 + 0.25 = 2.25 \quad \text{or} \quad x_3 = 2.0000 \tag{6}$$

Similarly, we get

$$x_2 = 1.0000, \quad x_1 = 1.0000$$

Thus, the solution of the given system is given by

$$x_1 = 1.0, \quad x_2 = 1.0, \quad x_3 = 2.0, \quad x_4 = 2.0$$

3.3 GAUSS–JORDAN ELIMINATION METHOD

This method is a variation of Gaussian elimination method. In this method, the elements above and below the diagonal are simultaneously made zero and thereby the given system is reduced to an equivalent diagonal form using elementary transformations. Then the solution of the resulting diagonal system can be readily obtained.

Sometimes, we normalize the pivot row with respect to the pivot element, before elimination. Partial pivoting is also used whenever the pivot element becomes zero.

This method is illustrated through the following examples.

Example 3.4 Solve the system of equations

$$\left.\begin{aligned} x + 2y + z &= 8 \\ 2x + 3y + 4z &= 20 \\ 4x + 3y + 2z &= 16 \end{aligned}\right\} \tag{1}$$

using Gauss–Jordan elimination method.

Solution In matrix notation, the given system (1) can be written as

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 4 \\ 4 & 3 & 2 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ 20 \\ 16 \end{pmatrix} \tag{2}$$

Scanned with CamScanner

We subtract the first row multiplied by 2 and 4 from the second and third rows respectively of Eq. (2), and eliminate x

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & -1 & 2 \\ 0 & -5 & -2 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ 4 \\ -16 \end{pmatrix} \tag{3}$$

Now, we eliminate y from the first and third rows using the second row. Thus, we get

$$\begin{bmatrix} 1 & 0 & 5 \\ 0 & -1 & 2 \\ 0 & 0 & -12 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 16 \\ 4 \\ -36 \end{pmatrix} \tag{4}$$

Before, eliminating z from the first and second row, normalizing the third row with respect to the pivot element, we get

$$\begin{bmatrix} 1 & 0 & 5 \\ 0 & -1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 16 \\ 4 \\ 3 \end{pmatrix} \tag{5}$$

Using the third row of Eq. (5), eliminating z from the first and second rows of Eq. (5), we obtain

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} \tag{6}$$

From Eq. (16), we get the solution directly as $x = 1$, $y = 2$, $z = 3$.

3.4 CROUT'S REDUCTION METHOD

This method is based on the fact that the coefficient matrix $[A]$ of the system of equations (3.3) can be decomposed into the product of two matrices $[L]$ and $[U]$, where $[L]$ is a lower-triangular matrix and $[U]$ is an upper-triangular matrix with 1's on its main diagonal. The rules for getting $[L]$ and $[U]$ can be obtained from the fact

$$[L]\,[U] = [A] \tag{3.7}$$

For the purpose of illustration, let us consider a (3×3) general matrix in the form

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{3.8}$$

Scanned with CamScanner

The sequence of steps for getting $[L]$ and $[U]$ are given below:

Step I: Multiplying all the rows of $[L]$ by the first column of $[U]$, we get

$$l_{11} = a_{11}, \qquad l_{21} = a_{21}, \qquad l_{31} = a_{31} \tag{3.9}$$

Thus, we observe that the first column of $[L]$ is same as the first column of $[A]$.

Step II: Now, multiplying the first row of $[L]$ by the second and third columns of $[U]$, we obtain

$$l_{11}u_{12} = a_{12}, \qquad l_{11}u_{13} = a_{13}$$

or

$$u_{12} = \frac{a_{12}}{l_{11}}, \qquad u_{13} = \frac{a_{13}}{l_{11}} \tag{3.10}$$

Thus, the first row of $[U]$ is obtained. Now, we continue this process, thus getting alternately the column of $[L]$ and a row of $[U]$.

Step III: Multiply the second and third rows of $[L]$ by the second column of $[U]$ to get

$$l_{21}u_{12} + l_{22} = a_{22}, \qquad l_{31}u_{12} + l_{32} = a_{32}$$

which gives

$$l_{22} = a_{22} - l_{21}u_{12}, \qquad l_{32} = a_{32} - l_{31}u_{12} \tag{3.11}$$

Thus, the second column of $[L]$ is obtained.

Step IV: Now, multiply the second row of $[L]$ by the third column of $[U]$ which yields

$$l_{21}u_{13} + l_{22}u_{23} = a_{23} \qquad \text{or} \qquad u_{23} = \frac{a_{23} - l_{21}u_{13}}{l_{22}} \tag{3.12}$$

Step V: Lastly, we multiply the third row of $[L]$ by the third column of $[U]$ and get

$$l_{31}u_{13} + l_{32}u_{23} + l_{33} = a_{33}$$

which gives

$$l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23} \tag{3.13}$$

Thus, the above five steps determine $[L]$ and $[U]$. This algorithm can be generalized to any linear system of order n.

Now, to obtain the solution of the linear system

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}\right\} \tag{3.14}$$

in matrix notation as $[A]\,(X) = (B)$. Let $[A] = [L]\,[U]$, then we get,

$$[L]\,[U]\,(X) = (B) \tag{3.15}$$

Substituting $[U]\,(X) = (Z)$ in Eq. (3.15), we obtain

$$[L]\,(Z) = (B) \tag{3.16}$$

Scanned with CamScanner

Now, Eq. (3.16) is equivalent to

$$\left.\begin{aligned} l_{11}z_1 &= b_1 \\ l_{21}z_1 + l_{22}z_2 &= b_2 \\ l_{31}z_1 + l_{32}z_2 + l_{33}z_3 &= b_3 \end{aligned}\right\} \tag{3.17}$$

The first of these equations gives z_1. Knowing z_1, the second equation of (3.17) gives z_2; then the third equation of (3.17) can be solved and z_3 is obtained. Having computed z_1, z_2 and z_3, we can compute x_1, x_2 and x_3 from equation $[U]\,(X) = (Z)$ or from

$$\begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

This method is also known as *Cholesky reduction method*. This technique is widely used in the numerical solutions of partial differential equations.

This method is very popular from computer programming point of view, since the storage space reserved for matrix $[A]$ can be used to store the elements of $[L]$ and $[U]$ at the end of computation.

It may be noted that this method fails if any $a_{ii} = 0$. In that case, the system is singular. In order to familiarize with the Crout's reduction method, we consider the following examples.

Example 3.5 Solve the following system of equations

$$5x_1 - 2x_2 + x_3 = 4$$
$$7x_1 + x_2 - 5x_3 = 8$$
$$3x_1 + 7x_2 + 4x_3 = 10$$

by Crout's reduction method using hand computation.

Solution Let the coefficient matrix $[A]$ be written as $[L]\,[U]$. Thus,

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & -2 & 1 \\ 7 & 1 & -5 \\ 3 & 7 & 4 \end{bmatrix} \tag{1}$$

Step I: Multiply all the rows of $[L]$ by the first column of $[U]$, we get

$$l_{11} = 5, \quad l_{21} = 7, \quad l_{31} = 3 \tag{2}$$

Step II: Multiply the first row of $[L]$ by the second and third columns of $[U]$, we have

$$l_{11}u_{12} = -2, \quad l_{11}u_{13} = 1$$

Using Eq. (2), we get

$$u_{12} = -\frac{2}{5}, \quad u_{13} = \frac{1}{5} \tag{3}$$

Scanned with CamScanner

Step III: Multiply the second and third rows of $[L]$ by the second column of $[U]$. Using Eqs. (2) and (3), we get

$$\left.\begin{aligned} l_{21}u_{12} + l_{22} &= 1 \quad \text{or} \quad l_{22} = 1 + \frac{14}{5} = \frac{19}{5} \\ l_{31}u_{12} + l_{32} &= 7 \quad \text{or} \quad l_{32} = 7 + \frac{6}{5} = \frac{41}{5} \end{aligned}\right\} \tag{4}$$

Step IV: Multiply the second row of $[L]$ by the third column of $[U]$ which yields $l_{21}u_{13} + l_{22}u_{23} = -5$. Using Eqs. (2)–(4), we get

$$\frac{19}{5}u_{23} = -5 - \frac{7}{5}$$

Therefore,

$$u_{23} = -\frac{32}{19} \tag{5}$$

Step V: Finally, multiply the third row of $[L]$ with the third column of $[U]$, we obtain

$$l_{31}u_{13} + l_{32}u_{23} + l_{33} = 4$$

Using Eqs. (2)–(5), we get

$$l_{33} = \frac{327}{19} \tag{6}$$

Thus, the given system of equations takes the form $[L]\,[U]\,[X] = (B)$. That is,

$$\begin{bmatrix} 5 & 0 & 0 \\ 7 & \frac{19}{5} & 0 \\ 3 & \frac{41}{5} & \frac{327}{19} \end{bmatrix} \begin{bmatrix} 1 & -\frac{2}{5} & \frac{1}{5} \\ 0 & 1 & -\frac{32}{19} \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 8 \\ 10 \end{pmatrix} \tag{7}$$

Let $[U](X) = (Z)$, then

$$[L](Z) = (4\ 8\ 10)^{\mathrm{T}}$$

or

$$\begin{bmatrix} 5 & 0 & 0 \\ 7 & \frac{19}{5} & 0 \\ 3 & \frac{41}{5} & \frac{327}{19} \end{bmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 8 \\ 10 \end{pmatrix} \tag{8}$$

which gives

$$z_1 = \frac{4}{5}, \qquad z_2 = \frac{12}{19}, \qquad z_3 = \frac{46}{327} \tag{9}$$

Scanned with CamScanner

utilizing these values of z, Eq. (7) becomes

$$\begin{bmatrix} 1 & -\frac{2}{5} & \frac{1}{5} \\ 0 & 1 & -\frac{32}{19} \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \frac{4}{5} \\ \frac{12}{19} \\ \frac{46}{327} \end{pmatrix} \tag{10}$$

By back substitution method, we obtain

$$x_3 = \frac{46}{327}, \qquad x_2 = \frac{284}{327}, \qquad x_1 = \frac{366}{327}$$

This is the required solution.

3.5 JACOBI'S METHOD

Jacobi's method is an iterative method, where initial approximate solution to a given system of equations is assumed and is improved towards the exact solution in an iterative way. In general, when the coefficient matrix of the system of equations is a sparse matrix (many elements are zero), iterative methods have definite advantage over direct methods in respect of economy in computer memory. Such spare matrices arise in computing the numerical solution of partial differential equations.

To illustrate Jacobi's method, let us consider a linear system given by

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\quad &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}\right\} \tag{3.18}$$

In this method, we assume that the coefficient matrix $[A]$ is strictly diagonally dominant, that is, in each row of $[A]$ the modulus of the diagonal element exceeds the sum of the off-diagonal elements. We also assume that the diagonal element a_{ii} do not vanish. If any diagonal element vanishes, the equations can always be rearranged to satisfy this condition. Now the system (3.18) can be written as

$$\left.\begin{aligned} x_1 &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2 - \cdots - \frac{a_{1n}}{a_{11}}x_n \\ x_2 &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1 - \cdots - \frac{a_{2n}}{a_{22}}x_n \\ \vdots \quad &\vdots \qquad\qquad \vdots \\ x_n &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1 - \cdots - \frac{a_{n(n-1)}}{a_{nn}}x_{n-1} \end{aligned}\right\} \tag{3.19}$$

Scanned with CamScanner

We shall take this solution vector $(x_1, x_2, \ldots, x_n)^T$ as a first approximation to the exact solution of system (3.18). For convenience, let us denote the first approximation vector by $(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)})$ got after taking $(0, 0, \ldots, 0)$ as an initial starting vector. Substituting this first approximation in the right-hand size of system (3.19), we obtain the second approximation to the given system in the form

$$\left.\begin{aligned} x_1^{(2)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2^{(1)} - \cdots - \frac{a_{1n}}{a_{11}}x_n^{(1)} \\ x_2^{(2)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1^{(1)} - \cdots - \frac{a_{2n}}{a_{22}}x_n^{(1)} \\ &\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad\quad \vdots \\ x_n^{(2)} &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1^{(1)} - \cdots - \frac{a_{n(n-1)}}{a_{nn}}x_{n-1}^{(1)} \end{aligned}\right\} \tag{3.20}$$

This second approximation is substituted into the right-hand side of Eqs. (3.20) and obtain the third approximation and so on. This process is repeated and $(r + 1)$th approximation is calculated from

$$\left.\begin{aligned} x_1^{(r+1)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2^{(r)} - \cdots - \frac{a_{1n}}{a_{11}}x_n^{(r)} \\ x_2^{(r+1)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1^{(r)} - \cdots - \frac{a_{2n}}{a_{22}}x_n^{(r)} \\ &\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad\quad \vdots \\ x_n^{(r+1)} &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1^{(r)} - \cdots - \frac{a_{n(n-1)}}{a_{nn}}x_{n-1}^{(r)} \end{aligned}\right\} \tag{3.21}$$

Briefly, we can rewrite Eqs. (3.21) as

$$x_i^{(r+1)} = \frac{b_i}{a_{ii}} - \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(r)}, \qquad r = 1, 2, \cdots, \quad i = 1, 2, \cdots, n \tag{3.22}$$

This method is due to Jacobi and is called *Jacobi's iterative method*. It is also known as method of *simultaneous displacements*, since no element of $x_i^{(r+1)}$ is used in this iteration until every element is computed.

A sufficient condition for convergence of the iterative solution to the exact solution is

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \qquad i = 1, 2, \cdots, n \tag{3.23}$$

When this condition (diagonal dominance) is true, Jacobi's method converges.

Example 3.6 Find the solution to the following system of equations

$$\begin{aligned} 83x + 11y - 4z &= 95 \\ 7x + 52y + 13z &= 104 \\ 3x + 8y + 29z &= 71 \end{aligned}$$

using Jacobi's iterative method for the first five iterations.

Scanned with CamScanner

Solution At first, we rewrite the given system in the form

$$\left.\begin{aligned} x &= \frac{95}{83} - \frac{11}{83}y + \frac{4}{83}z \\ y &= \frac{104}{52} - \frac{7}{52}x - \frac{13}{52}z \\ z &= \frac{71}{29} - \frac{3}{29}x - \frac{8}{29}y \end{aligned}\right\} \quad (1)$$

Taking the initial starting of solution vector as $(0, 0, 0)^T$, from Eq. (1), we have the first approximation as

$$\begin{pmatrix} x^{(1)} \\ y^{(1)} \\ z^{(1)} \end{pmatrix} = \begin{pmatrix} 1.1446 \\ 2.0000 \\ 2.4483 \end{pmatrix} \quad (2)$$

Now, using Eq. (1), the second approximation is computed from the equations

$$\left.\begin{aligned} x^{(2)} &= 1.1446 - 0.1325y^{(1)} + 0.0482z^{(1)} \\ y^{(2)} &= 2.0 - 0.1346x^{(1)} - 0.25z^{(1)} \\ z^{(2)} &= 2.4483 - 0.1035x^{(1)} - 0.2759y^{(1)} \end{aligned}\right\} \quad (3)$$

Substituting Eq. (2) into Eq. (3), we get the second approximation as

$$\begin{pmatrix} x^{(2)} \\ y^{(2)} \\ z^{(2)} \end{pmatrix} = \begin{pmatrix} 0.9976 \\ 1.2339 \\ 1.7424 \end{pmatrix} \quad (4)$$

Similar procedure yields the third, fourth and fifth approximations to the required solution and they are tabulated as below:

Iteration number, r	Variables		
	x	y	z
1	1.1446	2.0000	2.4483
2	0.9976	1.2339	1.7424
3	1.0651	1.4301	2.0046
4	1.0517	1.3555	1.9435
5	1.0587	1.3726	1.9655

3.6 GAUSS–SEIDEL ITERATION METHOD

It is another well-known iterative method for solving a system of linear equations of the form of system (3.18). In Jacobi method, the $(r + 1)$th approximation to the system (3.18) is given by Eqs. (3.21), from which we can

Scanned with CamScanner

observe that no element of $x_i^{(r+1)}$ replaces $x_i^{(r)}$ entirely for the next cycle of computation.

However, in Gauss–Seidel method, the corresponding elements of $x_i^{(r+1)}$ replaces those of $x_i^{(r)}$ as soon as they become available. Hence, it is called the method of *successive displacements*. For illustration, consider the system (3.18). In Gauss–Seidel iteration, the $(r + 1)$th approximation or iteration is computed from

$$\left.\begin{aligned}
x_1^{(r+1)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2^{(r)} - \cdots - \frac{a_{1n}}{a_{11}}x_n^{(r)} \\
x_2^{(r+1)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1^{(r+1)} - \cdots - \frac{a_{2n}}{a_{22}}x_n^{(r)} \\
&\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad\quad\quad \vdots \\
x_n^{(r+1)} &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1^{(r+1)} - \cdots - \frac{a_{n,(n-1)}}{a_{nn}}x_{n-1}^{(r+1)}
\end{aligned}\right\} \tag{3.24}$$

Thus, the general procedure can be written in the following compact form

$$x_i^{(r+1)} = \frac{b_i}{a_{ii}} - \sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^{(r+1)} - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^{(r)} \tag{3.25}$$

for all $i = 1, 2, \ldots, n$ and $r = 1, 2, \ldots$

To describe system (3.24); in the first equation, we substitute the rth approximation into the right-hand side and denote the result by $x_1^{(r+1)}$. In the second equation, we substitute $(x_1^{(r+1)}, x_3^{(r)}, \ldots, x_n^{(r)})$ and denote the result by $x_2^{(r+1)}$. In the third equation, we substitute $(x_1^{(r+1)}, x_2^{(r+1)}, x_4^{(r)}, \ldots, x_n^{(r)})$ and denote the result by $x_3^{(r+1)}$, and so on. This process is continued till we arrive at the desired result. For illustration, we consider the following example.

Example 3.7 Find the solution of the following system of equations

$$x_1 - \frac{1}{4}x_2 - \frac{1}{4}x_3 = \frac{1}{2}$$

$$-\frac{1}{4}x_1 + x_2 - \frac{1}{4}x_4 = \frac{1}{2}$$

$$-\frac{1}{4}x_1 + x_3 - \frac{1}{4}x_4 = \frac{1}{4}$$

$$-\frac{1}{4}x_2 - \frac{1}{4}x_3 + x_4 = \frac{1}{4}$$

using Gauss–Seidel method and perform the first-five iterations.

Scanned with CamScanner

Solution The given system of equations can be rewritten as

$$\left.\begin{aligned} x_1 &= 0.5 + 0.25x_2 + 0.25x_3 \\ x_2 &= 0.5 + 0.25x_1 + 0.25x_4 \\ x_3 &= 0.25 + 0.25x_1 + 0.25x_4 \\ x_4 &= 0.25 + 0.25x_2 + 0.25x_3 \end{aligned}\right\} \quad (1)$$

Taking $x_2 = x_3 = x_4 = 0$ on the right-hand side of the first equation of system (1), we get $x_1^{(1)} = 0.5$. Taking $x_3 = x_4 = 0$ and the current value of x_1, we get

$$x_2^{(1)} = 0.5 + (0.25)\,(0.5) + 0 = 0.625$$

from the second equation of system (1). Further, we take $x_4 = 0$ and the current value of x_1, we obtain

$$x_3^{(1)} = 0.25 + (0.25)\,(0.5) + 0 = 0.375$$

from the third equation of system (1). Now, using the current values of x_2 and x_3, the fourth equation of system (1) gives

$$x_4^{(1)} = 0.25 + (0.25)\,(0.625) + (0.25)\,(0.375) = 0.5$$

The Gauss–Seidel iterations for the given set of equations can be written as

$$\begin{aligned} x_1^{(r+1)} &= 0.5 + 0.25x_2^{(r)} + 0.25x_3^{(r)} \\ x_2^{(r+1)} &= 0.5 + 0.25x_1^{(r+1)} + 0.25x_4^{(r)} \\ x_3^{(r+1)} &= 0.25 + 0.25x_1^{(r+1)} + 0.25x_4^{(r)} \\ x_4^{(r+1)} &= 0.25 + 0.25x_2^{(r+1)} + 0.25x_3^{(r+1)} \end{aligned}$$

Now, by Gauss–Seidel procedure, the second and subsequent approximations can be obtained and the sequence of the first-five approximations are tabulated as below:

Iteration number r	Variables			
	x_1	x_2	x_3	x_4
1	0.5	0.625	0.375	0.5
2	0.75	0.8125	0.5625	0.59375
3	0.84375	0.85938	0.60938	0.61719
4	0.86719	0.87110	0.62110	0.62305
5	0.87305	0.87402	0.62402	0.62451

3.7 THE RELAXATION METHOD

This method is an iterative method and is due to Southwell. To explain the details, consider again the system of equations (3.18). Let

$$X^{(p)} = (x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})^T$$

be the solution vector obtained iteratively after pth iteration. If $R_i^{(p)}$ denotes the residual of the ith equation of system (3.18), that is of

Scanned with CamScanner

Iteration number	Residuals R_1	R_2	R_3	Maximum R_i	Difference dx_i	Variables x_1	x_2	x_3
0	11	10	−15	−15	15/8 = 1.875	0	0	0
1	9.125	8.125	0	9.125	−9.125/(−6) = 1.5288	0	0	1.875
2	0.0478	6.5962	−3.0576	6.5962	−6.5962/7 = −0.9423	1.5288	0	1.875
3	−2.8747	0.0001	−2.1153	−2.8747	2.8747/(−6) = −0.4791	1.5288	−0.9423	1.875
4	−0.0031	0.4792	−1.1571	−1.1571	1.1571/8 = 0.1446	1.0497	−0.9423	1.875
5	−0.1447	0 3346	0.0003	.3346	−.3346/7 = −0.0478	1.0497	−0.9423	2.0196
6	0.2881	0.0000	0.0475	.2881	−.2881/(−6) = 0.0480	1.0497	−0.9901	2 0196
7	−0.0001	0.048	0.1435	0.1435	−0.1435/8 = −0.0179	1.0017	−0.9901	2.0196
8	0.0178	0.0659	0.0003	–	–	1.0017	−0.9901	2.0017

At this stage, we observe that all the residuals R_1, R_2 and R_3 are small enough, and therefore, we may take the corresponding values of x_i at this iteration as the solution. Hence, the numerical solution to the given system is given by

$$x_1 = 1.0017, \qquad x_2 = -0.9901, \qquad x_3 = 2.0017,$$

The exact solution is found to be

$$x_1 = 1.0, \qquad x_2 = -1.0, \qquad x_3 = 2.0$$

3.8 MATRIX INVERSION

Consider a system of equations in the form

$$[A]\,(X) = (B) \tag{3.27}$$

One way of writing its solution is in the form

$$(X) = [A]^{-1}\,(B) \tag{3.28}$$

Thus, the solution to the system (3.27) can also be obtained if the inverse of the coefficient matrix $[A]$ is known. Alternatively, if the product of two square matrices is an identity matrix, that is, if

$$[A]\,[B] = [I] \tag{3.29}$$

then,

$$[B] = [A]^{-1} \quad \text{and} \quad [A] = [B]^{-1}$$

Every square non-singular matrix will have an inverse. Gauss elimination and Gauss–Jordan methods are popular among many methods available for finding the inverse of a matrix.

Scanned with CamScanner

3.8.1 Gaussian Elimination Method

In this method, if A is a given matrix, for which we have to find the inverse; at first, we place an identity matrix, whose order is same as that of A, adjacent to A which we call an *augmented matrix*. Then the inverse of A is computed in two stages. In the first stage, A is converted into an upper triangular form, using Gaussian elimination method as discussed in Section 3.2. In the second stage, the above upper triangular matrix is reduced to an identity matrix by row transformations. All these operations are also performed on the adjacently placed identity matrix. Finally, when A is transformed into an identity matrix, the adjacent matrix gives the inverse of A. In order to increase the accuracy of the result, it is essential to employ partial pivoting. To understand the sequence of the steps involved, we consider an example.

Example 3.9 Use the Gaussian elimination method to find the inverse of the matrix

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

Solution At first, we place an identity matrix of the same order adjacent to the given matrix. Thus, the augmented matrix can be written as

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

Stage I (Reduction to upper triangular form): Let R_1, R_2 and R_3 denote the first, second and third rows of a matrix. In the first column of Eq. (1), 4 is the largest element, thus interchanging R_1 and R_2 to bring the pivot element 4 to the place of a_{11}, we have the augmented matrix in the form

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

Divide R_1 by 4 to get

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

Scanned with CamScanner

Perform $R_2 - R_1 \longrightarrow R_2$, which gives

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

Perform $R_3 - 3R_1 \longrightarrow R_3$ in Eq. (4), which yields

$$\left[\begin{array}{ccc|ccc} 1 & \frac{3}{4} & -\frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{1}{4} & \frac{5}{4} & 1 & -\frac{1}{4} & 0 \\ 0 & \frac{11}{4} & \frac{15}{4} & 0 & -\frac{3}{4} & 1 \end{array}\right] \tag{5}$$

Now, looking at the second column for the pivot, the max (1/4, 11/4) is 11/4. Therefore, we interchange R_2 and R_3 in Eq. (5) and get

$$\left[\begin{array}{ccc|ccc} 1 & \frac{3}{4} & -\frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{11}{4} & \frac{15}{4} & 0 & -\frac{3}{4} & 1 \\ 0 & \frac{1}{4} & \frac{5}{4} & 1 & -\frac{1}{4} & 0 \end{array}\right] \tag{6}$$

Now, divide R_2 by the pivot $a_{22} = 11/4$, and obtain

$$\left[\begin{array}{ccc|ccc} 1 & \frac{3}{4} & -\frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 1 & \frac{15}{11} & 0 & -\frac{3}{11} & \frac{4}{11} \\ 0 & \frac{1}{4} & \frac{5}{4} & 1 & -\frac{1}{4} & 0 \end{array}\right] \tag{7}$$

Performing $R_3 - (1/4)R_2 \longrightarrow R_3$ in (7) yields

$$\left[\begin{array}{ccc|ccc} 1 & \frac{3}{4} & -\frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 1 & \frac{15}{11} & 0 & -\frac{3}{11} & \frac{4}{11} \\ 0 & 0 & \frac{10}{11} & 1 & -\frac{2}{11} & -\frac{1}{11} \end{array}\right] \tag{8}$$

Scanned with CamScanner

Finally, we divide R_3 by (10/11), thus getting an upper triangular form

$$\left[\begin{array}{ccc|ccc} 1 & \frac{3}{4} & -\frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 1 & \frac{15}{11} & 0 & -\frac{3}{11} & \frac{4}{11} \\ 0 & 0 & 1 & \frac{11}{10} & -\frac{1}{5} & -\frac{1}{10} \end{array}\right] \tag{9}$$

Stage II (Reduction to an identity matrix): Multiply R_3 by $-1/4$ and $15/11$ respectively and subtract it from R_1 and R_2 of Eq. (9), we get

$$\left[\begin{array}{ccc|ccc} 1 & \frac{3}{4} & 0 & \frac{11}{40} & \frac{1}{5} & -\frac{1}{40} \\ 0 & 1 & 0 & -\frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & \frac{11}{10} & -\frac{1}{5} & -\frac{1}{10} \end{array}\right] \tag{10}$$

Finally, performing $R_1 - (3/4)\, R_2 \longrightarrow R_1$ in Eq. (10), we obtain

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

Thus, we have

$$A^{-1} = \begin{bmatrix} \frac{7}{5} & \frac{1}{5} & -\frac{2}{5} \\ -\frac{3}{2} & 0 & \frac{1}{2} \\ \frac{11}{10} & -\frac{1}{5} & -\frac{1}{10} \end{bmatrix} \tag{11}$$

We can easily cheque $[A]\,[A^{-1}] = [I]$.

3.8.2 Gauss–Jordan Method

This method is similar to Gaussian elimination method, with the essential difference that the stage *I* of reducing the given matrix to an upper triangular form is not needed. However, the given matrix can be directly reduced to an identity matrix using elementary row transformations. This technique is illustrated in the following example.

Scanned with CamScanner

Example 3.10 Find the inverse of

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

by Gauss–Jordan method.

Solution Let R_1, R_2 and R_3 denote the first, second and third rows of a matrix. We place an identity matrix adjacent to the given matrix as a first step and the resulting augmented matrix is given by

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

Performing $R_2 - 4R_1 \longrightarrow R_2$, we get

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

Now, performing $R_3 - 3R_1 \longrightarrow R_3$, we obtain

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

Carrying out further operations $R_2 + R_1 \longrightarrow R_1$ and $R_3 + 2R_2 \longrightarrow R_3$, we arrive at

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -4 & -3 & 1 & 0 \\ 0 & -1 & -5 & -4 & 1 & 0 \\ 0 & 0 & -10 & -11 & 2 & 1 \end{array}\right] \tag{4}$$

Now, dividing the third row by -10, we get

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -4 & -3 & 1 & 0 \\ 0 & -1 & -5 & -4 & 1 & 0 \\ 0 & 0 & 1 & \frac{11}{10} & -\frac{1}{5} & -\frac{1}{10} \end{array}\right] \tag{5}$$

Scanned with CamScanner

Further, we perform $R_1 + 4R_3 \longrightarrow R_1$, and $R_2 + 5R_3 \longrightarrow$ get

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{7}{5} & \frac{1}{5} & -\frac{2}{5} \\ 0 & -1 & 0 & \frac{3}{2} & 0 & -\frac{1}{2} \\ 0 & 0 & 1 & \frac{11}{10} & -\frac{1}{5} & -\frac{1}{10} \end{array}\right] \quad (6)$$

Finally, multiplying R_2 by -1, we obtain

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{7}{5} & \frac{1}{5} & -\frac{2}{5} \\ 0 & 1 & 0 & -\frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & \frac{11}{10} & -\frac{1}{5} & -\frac{1}{10} \end{array}\right] \quad (7)$$

Hence, we have

$$A^{-1} = \begin{bmatrix} \frac{7}{5} & \frac{1}{5} & -\frac{2}{5} \\ -\frac{3}{2} & 0 & \frac{1}{2} \\ \frac{11}{10} & -\frac{1}{5} & -\frac{1}{10} \end{bmatrix} \quad (8)$$

It can be easily verified that $[A]\,[A^{-1}] = [I]$.

EXERCISES

3.1 Solve the following systems of equations

(i) $x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 = 1$

$\frac{1}{2}x_1 + \frac{1}{3}x_2 + \frac{1}{4}x_3 = 0$

$\frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 = 0$

(ii) $4x_1 + x_2 + x_3 = 4$

$x_1 + 4x_2 - 2x_3 = 4$

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

(iii) $10x - 7y + 3z + 5w = 6$

$-6x + 8y - z - 4w = 5$

$3x + y + 4z + 11w = 2$

$5x - 9y - 2z + 4w = 7$

by Gaussi...

Scanned with CamScanner

3.2 Using Gaussian elimination method with partial pivoting, solve the following systems of equations

(i) $x_1 + x_2 - 2x_3 = 3$
$4x_1 - 2x_2 + x_3 = 5$
$3x_1 - x_2 + 3x_3 = 8$

(ii) $2.5x - 3y + 4.6z = -1.05$
$-3.5x + 2.6y + 1.5z = -14.46$
$-6.5x - 3.5y + 7.3z = -17.735$

3.3 Using Gauss–Jordan elimination method, solve the systems of equations

(i) $10x_1 + x_2 + x_3 = 12$
$x_1 + 10x_2 + x_3 = 12$
$x_1 + x_2 + 10x_3 = 12$

(ii) $x + y + z = 7$
$3x + 3y + 4z = 24$
$2x + y + 3z = 16$

3.4 Using Crout's reduction method, solve the following systems of equations

(i) $x + y + z = 3$
$2x - y + 3z = 16$
$3x + y - z = -3$

(ii) $6x_1 - x_2 = 3$
$-x_1 + 6x_2 - x_3 = 4$
$-x_2 + 6x_3 = 3$

3.5 Solve the following system of equations

$$4x_1 - 3x_2 + 2x_3 = 11$$
$$2x_1 + x_2 + 7x_3 = 2$$
$$3x_1 - x_2 + 5x_3 = 8$$

using Cholesky's reduction method.

3.6 Using Crout's reduction, decompose the matrix

$$[A] = \begin{bmatrix} 5 & -2 & 1 \\ 7 & 1 & -5 \\ 3 & 7 & 4 \end{bmatrix}$$

into $[L]$ $[U]$ form and hence solve the system of equations

$$5x - 2y + z = 4$$
$$7x + y - 5z = 8$$
$$3x + 7y + 4z = 10$$

3.7 Explain how Jacobi's method is used to obtain numerical solution of a system of linear equations. What is the condition of convergence for any choice for the first approximation.

3.8 Find the solution of the following system of equations

$$x_1 - \frac{1}{4}x_2 - \frac{1}{4}x_3 = \frac{1}{2}$$
$$-\frac{1}{4}x_1 + x_2 - \frac{1}{4}x_4 = \frac{1}{2}$$
$$-\frac{1}{4}x_1 + x_3 - \frac{1}{4}x_4 = \frac{1}{4}$$
$$-\frac{1}{4}x_2 - \frac{1}{4}x_3 + x_4 = \frac{1}{4}$$

sing Jacobi method of iteration. Carry computation up to the seventh iteration.

Scanned with CamScanner

3.9 Solve the following system of equations

$$\begin{aligned} 2x_1 - x_2 &= 7 \\ -x_1 + 2x_2 - x_3 &= 1 \\ -x_2 + 2x_3 &= 1 \end{aligned}$$

using Gauss–Seidel method of iteration and perform the first-five iterations. The exact solution is $(6\ 5\ 3)^T$.

3.10 Solve the system of equations

$$\begin{aligned} 20x + y - 2z &= 17 \\ 3x + 20y - z &= -18 \\ 2x - 3y + 20z &= 25 \end{aligned}$$

by Gauss–Seidel iterative method and perform the first-three iterations.

3.11 Solve the following system of equations

$$\begin{aligned} 5x - 2y + z &= 13 \\ 3x + 7y - 11z &= 2 \\ x + 20y - 2z &= 8 \end{aligned}$$

by relaxation method.

3.12 Using relaxation method, find the solution to the system of equations

$$\begin{aligned} 8x_1 + x_2 - x_3 &= 8 \\ 2x_1 + x_2 + 9x_3 &= 12 \\ x_1 - 7x_2 + 2x_3 &= -4 \end{aligned}$$

taking the initial solution vector (0, 0, 0).

3.13 Using Gaussian elimination method, find the inverse of the matrix

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

3.14 Find the inverse of the matrix

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 3 & 2 & 5 \\ 1 & -1 & 0 \end{bmatrix}$$

3.15 Find the inverse of the following matrices

(i) $A = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 3 & -3 \\ -2 & -4 & -4 \end{bmatrix}$ (ii) $A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 2 & 4 & 7 \end{bmatrix}$

by Gauss–Jordan method.

Scanned with CamScanner

Chapter 4

Eigenvalue Problems

4.1 INTRODUCTION

Computation of eigenvalues and the corresponding eigenvectors of a matrix is of practical importance. For example, in solid mechanics, where we consider an element in a continuum, subjected to normal and shear stresses, usually one will be interested in finding the principal stresses, which are the maximum and minimum stresses in an element.

Consider the wedge of unit thickness, subjected to normal and shear stresses. If it has to be in equilibrium, a system of equations written in matrix notation as

$$\begin{bmatrix} \sigma_x - \sigma_\theta & \sigma_{xy} \\ \sigma_{xy} & \sigma_y - \sigma_\theta \end{bmatrix} \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{4.1}$$

has to be satisfied. For non-trivial solution to exist, the determinant of the matrix must be zero. Its characteristic equation is

$$\sigma_\theta^2 - \sigma_\theta(\sigma_x + \sigma_y) + (\sigma_x\sigma_y - \sigma_{xy}^2) = 0 \tag{4.2}$$

This is a quadratic equation whose roots give two eigenvalues corresponding to principal stresses.

In general, let $[A]$ be an $n \times n$ square matrix. Suppose, there exists a scalar λ and a vector $X = (x_1\ x_2\ \ldots\ x_n)^T$ such that

$$[A]\,(X) = \lambda\,(X) \tag{4.3}$$

then λ is the eigenvalue and X is the corresponding eigenvector of the matrix $[A]$. Equation (4.3) can also be written as

$$[A - \lambda I]\,(X) = (O) \tag{4.4}$$

This represents a set of n homogeneous equations possessing non-trivial solution, provided

$$|A - \lambda I| = 0 \tag{4.5}$$

This determinant, on expansion, gives an nth degree polynomial in λ, which is called *characteristic* polynomial of $[A]$, which has n roots. Corresponding to each root, we can solve Eq. (4.4) in principle, and determine a vector called *eigenvector*. However, finding the roots of the characteristic equation is laborious. Hence, we look for better methods suitable from the point of view of

Scanned with CamScanner

computation. Depending upon the type of matrix $[A]$ and on what one is looking for, various numerical methods are available. Power method, Jacobi's method, Given's method, Householder, Lanczos method, Ruthishauser and Francis method are well known in the literature.

In this chapter, we shall consider only real and real-symmetric matrices and discuss power method and Jacobi's method in detail. For further study, one can consult Wilkinson (1965).

4.2 POWER METHOD

To compute the largest eigenvalue and the corresponding eigenvector of the system

$$[A](X) = \lambda(X)$$

where $[A]$ is a real, symmetric or unsymmetric matrix, the power method is widely used in practice. It is an iterative technique. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the distinct eigenvalues of an $(n \times n)$ matrix $[A]$, such that

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| \tag{4.6}$$

and suppose $v_1, v_2, \ldots, v_n$ are the corresponding eigenvectors. Power method is applicable if the above eigenvalues are real and distinct, and hence, the corresponding eigenvectors are linearly independent. Then, any eigenvector v in the space spanned by the eigenvectors $v_1, v_2, \ldots, v_n$ can be written as their linear combination. Therefore,

$$v = c_1 v_1 + c_2 v_2 + \cdots + c_n v_n \tag{4.7}$$

Pre-multiplying Eq. (4.7) by A and substituting

$$Av_1 = \lambda_1 v_1, \qquad Av_2 = \lambda_2 v_2, \ \ldots, \qquad Av_n = \lambda_n v_n$$

We get

$$Av = \lambda_1 \left(c_1 v_1 + c_2 \frac{\lambda_2}{\lambda_1} v_2 + \cdots + c_n \frac{\lambda_n}{\lambda_1} v_n \right) \tag{4.8}$$

Again, pre-multiplying by A and simplifying, we obtain

$$A^2 v = \lambda_1^2 \left[c_1 v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1} \right)^2 v_2 + \cdots + c_n \left(\frac{\lambda_n}{\lambda_1} \right)^2 v_n \right]$$

Similarly, we have

$$A^r v = \lambda_1^r \left[c_1 v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1} \right)^r v_2 + \cdots + c_n \left(\frac{\lambda_n}{\lambda_1} \right)^r v_n \right] \tag{4.9}$$

and

$$A^{r+1} v = (\lambda_1)^{r+1} \left[c_1 v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1} \right)^{r+1} v_2 + \cdots + c_n \left(\frac{\lambda_n}{\lambda_1} \right)^{r+1} v_n \right] \tag{4.10}$$

Scanned with CamScanner

Since $\lambda_i/\lambda_1 < 1$ for $i = 2, 3, \ldots, n$ and as $r \to \infty$, the right-hand sides of Eqs. (4.9) and (4.10) tend to $\lambda_1^r c_1 v_1$ and $\lambda_1^{r+1} c_1 v_1$ respectively. Now, the eigenvalue λ_1 can be computed as the limit of the ratio of the corresponding components of $A^r v$ and $A^{r+1} v$. That is,

$$\lambda_1 = \frac{\lambda_1^{r+1}}{\lambda_1^r} = \underset{r\to\infty}{\text{Lt}} \frac{(A^{r+1}v)_p}{(A^r v)_p}, \qquad p = 1, 2, \ldots, n \tag{4.11}$$

Here, the index p stands for the pth component in the corresponding vector.

Sometimes, we may be interested in finding the least eigenvalue and the corresponding eigenvector. In that case, we proceed as follows. We note that $[A]\,(X) = \lambda\,(X)$. Pre-multiplying by $[A^{-1}]$, we get

$$[A^{-1}]\,[A]\,(X) = [A^{-1}]\,\lambda\,(X) = \lambda\,[A^{-1}]\,(X)$$

or

$$(X) = \lambda\,[A^{-1}]\,(X)$$

which can be rewritten as

$$[A^{-1}]\,(X) = \frac{1}{\lambda}(X) \tag{4.12}$$

which shows that the inverse matrix has a set of eigenvalues which are the reciprocals of the eigenvalues of $[A]$. Thus, for finding the eigenvalue of the least magnitude of the matrix $[A]$, we have to apply power method to the inverse of $[A]$. In order to see that the power method converges fast, the following numerical algorithm is adopted, particularly when working with numerical examples.

Step 1: Choose the initial vector such that the largest element is unity.

Step 2: This normalized vector $v^{(0)}$ is pre-multiplied by the given matrix $[A]$.

Step 3: The resultant vector is again normalized.

Step 4: This process of iteration is continued and the new normalized vector is repeatedly pre-multiplied by the matrix $[A]$ until the required accuracy is obtained. At this point, the result looks like

$$u^{(k)} = [A]v^{(k-1)} = q_k v^{(k)}$$

Here, q_k is the desired largest eigenvalue and $v^{(k)}$ is the corresponding eigenvector.

Example 4.1 Find the eigenvalue of largest modulus, and the associated eigenvector of the matrix

$$[A] = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 3 & 5 \\ 3 & 2 & 9 \end{bmatrix}$$

by power method.

Solution We choose an initial vector $v^{(0)}$ as $(1, 1, 1)^T$. Then, compute the first iteration

Scanned with CamScanner

$$u^{(1)} = [A]v^{(0)} = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 3 & 5 \\ 3 & 2 & 9 \end{bmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 7 \\ 12 \\ 14 \end{pmatrix}$$

and normalize the resultant vector to get

$$u^{(1)} = 14 \begin{pmatrix} \frac{1}{2} \\ \frac{6}{7} \\ 1 \end{pmatrix} = q_1 v^{(1)}$$

The second iteration gives,

$$u^{(2)} = [A]v^{(1)} = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 3 & 5 \\ 3 & 2 & 9 \end{bmatrix} \begin{pmatrix} \frac{1}{2} \\ \frac{6}{7} \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{39}{7} \\ \frac{67}{7} \\ \frac{171}{14} \end{pmatrix} = 12.2143 \begin{pmatrix} 0.456140 \\ 0.783626 \\ 1.0 \end{pmatrix} = q_2 v^{(2)}$$

Similarly, continuing this procedure, the third and subsequent iterations are given as

$$u^{(3)} = [A]v^{(2)} = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 3 & 5 \\ 3 & 2 & 9 \end{bmatrix} \begin{pmatrix} 0.456140 \\ 0.783626 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 5.263158 \\ 9.175438 \\ 11.935672 \end{pmatrix}$$

$$= 11.935672 \begin{pmatrix} 0.44096 \\ 0.76874 \\ 1.0 \end{pmatrix} = q_3 v^{(3)}$$

$$u^{(4)} = [A]v^{(3)} = \begin{pmatrix} 5.18814 \\ 9.07006 \\ 11.86036 \end{pmatrix} = 11.86036 \begin{pmatrix} 0.437435 \\ 0.764737 \\ 1.0 \end{pmatrix} = q_4 v^{(4)}$$

$$u^{(5)} = [A]v^{(4)} = \begin{pmatrix} 5.16908 \\ 9.04395 \\ 11.84178 \end{pmatrix} = 11.84178 \begin{pmatrix} 0.436512 \\ 0.763732 \\ 1.0 \end{pmatrix} = q_5 v^{(5)}$$

Scanned with CamScanner

After rounding-off, we take the largest eigenvalue as $\lambda = 11.84$ and the corresponding eigenvector as

$$(X) = \begin{pmatrix} 0.44 \\ 0.76 \\ 1.00 \end{pmatrix}$$

accurate to two decimals.

4.3 JACOBI'S METHOD

Definition 4.1 An $(n \times n)$ matrix $[A]$ is said to be *orthogonal* if

$$[A]^T [A] = [I], \qquad \text{i.e. } [A]^T = [A]^{-1}$$

In order to compute all the eigenvalues and the corresponding eigenvectors of a real symmetric matrix, Jacobi's method is highly recommended. It is based on an important property from matrix theory, which states that, if $[A]$ is an $(n \times n)$ real symmetric matrix, its eigenvalues are real, and there exists an orthogonal matrix $[S]$ such that $[S^{-1}]\,[A]\,[S]$ is a diag [D]. This diagonalization can be carried out by applying a series of orthogonal transformations $S_1, S_2, \ldots, S_n$, as explained below.

Let A be an $(n \times n)$ real symmetric matrix. Suppose $|a_{ij}|$ be numerically the largest element amongst the off-diagonal elements of A. We construct an orthogonal matrix S_1 defined as

$$s_{ij} = -\sin\theta, \quad s_{ji} = \sin\theta, \quad s_{ii} = \cos\theta, \quad s_{jj} = \cos\theta \tag{4.13}$$

while each of the remaining off-diagonal elements are zero, the remaining diagonal elements are assumed to be unity. Thus, we construct S_1 as under

$$S_1 = \begin{array}{c} \begin{matrix} & & & \overset{i\text{th column}}{\downarrow} & & \overset{j\text{th column}}{\downarrow} & \\ \end{matrix} \\ \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \cos\theta & \cdots & -\sin\theta & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \sin\theta & \cdots & \cos\theta & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \end{array} \begin{matrix} \\ \\ \\ \leftarrow i\text{th row} \\ \\ \leftarrow j\text{th row} \\ \\ \end{matrix} \tag{4.14}$$

where $\cos\theta$, $-\sin\theta$, $\sin\theta$ and $\cos\theta$ are inserted in (i, i), (i, j), (j, i), (j, j)th positions respectively, and elsewhere it is identical with a unit matrix. Now, we compute

$$D_1 = S_1^{-1} A S_1 = S_1^T A S_1$$

since S_1 is an orthogonal matrix, such that $S_1^{-1} = S_1^T$. After the transformation,

Scanned with CamScanner

the elements at the positions (i, j), (j, i) get annihilated, that is, d_{ij} and d_{ji} reduce to zero, which is seen as follows:

$$\begin{bmatrix} d_{ii} & d_{ij} \\ d_{ji} & d_{jj} \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$\begin{bmatrix} a_{ii}\cos^2\theta + 2a_{ij}\sin\theta\cos\theta + a_{jj}\sin^2\theta & (a_{jj} - a_{ii})\sin\theta\cos\theta + a_{ij}\cos 2\theta \\ (a_{jj} - a_{ii})\sin\theta\cos\theta + a_{ij}\cos 2\theta & a_{ii}\sin^2\theta + a_{jj}\cos^2\theta - 2a_{ij}\sin\theta\cos\theta \end{bmatrix}$$

Therefore, $d_{ij} = 0$, only if,

$$a_{ij}\cos 2\theta + \frac{a_{jj} - a_{ii}}{2}\sin 2\theta = 0$$

That is, if

$$\tan 2\theta = \frac{2a_{ij}}{a_{ii} - a_{jj}} \tag{4.15}$$

Thus, we choose θ such that, Eq. (4.15) is satisfied, thereby, the pair of off-diagonal elements d_{ij} and d_{ji} reduces to zero.

However, though it creates a new pair of zeros, it also introduces non-zero contributions at formerly zero positions. Also, Eq. (4.15) gives four values of θ, but to get the least possible rotation, we choose $-\pi/4 \le \theta \le \pi/4$.

As a next step, the numerically largest off-diagonal element in the newly obtained rotated matrix D_1 is identified and the above procedure is repeated using another orthogonal matrix S_2 to get D_2. That is, we obtain

$$D_2 = S_2^{-1}D_1S_2 = S_2^{\mathrm{T}}(S_1^{\mathrm{T}}AS_1)S_2$$

Similarly, we perform a series of such two-dimensional rotations or orthogonal transformations. After making r transformations, we obtain

$$\begin{aligned} D_r &= S_r^{-1}S_{r-1}^{-1} \ \ldots\ S_2^{-1}S_1^{-1}AS_1S_2 \ \ldots\ S_{r-1}S_r \\ &= (S_1S_2 \ \ldots\ S_{r-1}S_r)^{-1}A(S_1S_2 \ \ldots\ S_{r-1}S_r) \\ &= S^{-1}AS \end{aligned} \tag{4.16}$$

where $S = S_1S_2 \ \ldots\ S_{r-1}\ S_r$. Now, as $r \to \infty$, D_r approaches to a diagonal matrix, with the eigenvalues on the main diagonal. The corresponding eigenvectors are the columns of S.

It is estimated that the minimum number of rotations required to transform the given $(n \times n)$ real symmetric matrix $[A]$ into a diagonal form is $n(n-1)/2$.

Example 4.2 Find all the eigenvalues and the corresponding eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & \sqrt{2} & 2 \\ \sqrt{2} & 3 & \sqrt{2} \\ 2 & \sqrt{2} & 1 \end{bmatrix}$$

by Jacobi's method.

Scanned with CamScanner

Solution The given matrix is real and symmetric. The largest off-diagonal element is found to be $a_{13} = a_{31} = 2$. Now, we compute

$$\tan 2\theta = \frac{2a_{ij}}{a_{ii} - a_{jj}} = \frac{2a_{13}}{a_{11} - a_{33}} = \frac{4}{0} = \infty$$

which gives, $\theta = \pi/4$. Thus, we construct an orthogonal matrix S_1 as

$$S_1 = \begin{bmatrix} \cos\frac{\pi}{4} & 0 & -\sin\frac{\pi}{4} \\ 0 & 1 & 0 \\ \sin\frac{\pi}{4} & 0 & \cos\frac{\pi}{4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

The first rotation gives,

$$D_1 = S_1^{-1} A S_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & \sqrt{2} & 2 \\ \sqrt{2} & 3 & \sqrt{2} \\ 2 & \sqrt{2} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

we may observe that the elements d_{13} and d_{31} got annihilated. To make sure that our calculations are correct up to this step, we may also observe that the sum of the diagonal elements of D_1 is same as the sum of the diagonal elements of the original matrix A.

As a second step, we choose the largest off-diagonal element of D_1 and is found to be $d_{12} = d_{21} = 2$, and compute

$$\tan 2\theta = \frac{2d_{12}}{d_{11} - d_{22}} = \frac{4}{0} = \infty$$

which again gives $\theta = \pi/4$. Thus, we construct the second rotation matrix as

$$S_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Scanned with CamScanner

At the end of second rotation, we get

$$D_2 = S_2^{-1} D_1 S_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{1}$$

which turned out to be a diagonal matrix, and therefore, we stop the computation. From (1) we notice that the eigenvalues of the given matrix are 5, 1 and –1. The eigenvectors are the column vectors of $S = S_1 S_2$. Therefore,

$$S = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Example 4.31 Find all the eigenvalues of the matrix

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

by Jacobi's method.

Solution In this example, we find that all the off-diagonal elements are of the same order of magnitude. Therefore, we can choose any one of them. Suppose, we choose a_{12} as the largest element and compute

$$\tan 2\theta = \frac{-1}{0} = \infty$$

which gives, $\theta = \pi/4$. Then $\cos\theta = \sin\theta = 1/\sqrt{2}$ and we construct an orthogonal matrix S_1 such that

$$S_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Scanned with CamScanner

The first rotation gives

$$D_1 = S_1^{-1} A S_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 3 & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 2 \end{bmatrix}$$

Now, we choose $d_{13} = -1/\sqrt{2}$ as the largest element of D_1 and compute

$$\tan 2\theta = \frac{2d_{13}}{d_{11} - d_{33}} = \frac{-\sqrt{2}}{1-2}$$

which gives, $\theta = 27^\circ\ 22'\ 41''$.

Now we construct another orthogonal matrix S_2, such that

$$S_2 = \begin{bmatrix} 0.888 & 0 & -0.459 \\ 0 & 1 & 0 \\ 0.459 & 0 & 0.888 \end{bmatrix}$$

At the end of second rotation, we obtain

$$D_2 = S_2^{-1} D_1 S_2 = \begin{bmatrix} 0.634 & -0.325 & 0 \\ 0.325 & 3 & -0.628 \\ 0 & -0.628 & 2.365 \end{bmatrix}$$

Now, the numerically largest off-diagonal element of D_2 is found to be $d_{23} = -0.628$ and compute

$$\tan 2\theta = \frac{-2 \times 0.628}{3 - 2.365}$$

we get, $\theta = -31^\circ\ 35'\ 24''$. Thus, the orthogonal matrix S_3 is seen to be

$$S_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.852 & 0.524 \\ 0 & -0.524 & 0.852 \end{bmatrix}$$

Scanned with CamScanner

At the end of third rotation, we get

$$D_3 = S_3^{-1} D_2 S_3 = \begin{bmatrix} 0.634 & -0.277 & 0 \\ 0.277 & 3.386 & 0 \\ 0 & 0 & 1.979 \end{bmatrix}$$

To reduce D_3 to a diagonal form, some more rotations are required. However, we may take 0.634, 3.386 and 1.979 as eigenvalues of the given matrix.

Example 4.4 Find all the eigenvalues and eigenvectors of the matrix:

$$A = \begin{bmatrix} 5 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 5 \end{bmatrix}$$

by Jacobi's method.

Solution The given matrix is

$$A = \begin{bmatrix} 5 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 5 \end{bmatrix}$$

In this example, the largest off-diagonal element is found to be $a_{13} = a_{31} = 1$. Now, we compute

$$\tan 2\theta = \frac{2a_{13}}{a_{11} - a_{33}} = \frac{2}{5-5} = \frac{2}{0} = \infty$$

which gives $\theta = \pi/4$. Following Jacobi's method, we construct an orthogonal matrix S_1 as

$$S_1 = \begin{bmatrix} \cos(\pi/4) & 0 & -\sin(\pi/4) \\ 0 & 1 & 0 \\ \sin(\pi/4) & 0 & \cos(\pi/4) \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}$$

The first rotation gives

$$D_1 = S_1^{-1} A S_1 = \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 5 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 5 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}$$

$$= \begin{bmatrix} 6 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 4 \end{bmatrix} \qquad (1)$$

Scanned with CamScanner

Which is a diagonal matrix and hence we stop further computation. From (1), we observe that 6, –2 and 4 are the eigenvalues of the given matrix and the corresponding eigenvectors are respectively the column vectors of

$$S_1 = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}$$

4.4 GERSCHGORIN'S THEOREM

This is one of the useful theorems on the bounds for eigenvalues of a square matrix.

Let λ_i be an eigenvalue of the $n \times n$ matrix [A] and let x_i be the corresponding eigenvector. Suppose R_s be the sum of the moduli of the terms along sth row, excluding the diagonal element a_{ss}. Then, every eigenvalue of $[A]$ lies inside or on the boundary of atleast one of the circles $|\lambda - a_{ss}| = R_s$

Proof: Given that

$$[A]x_i = \lambda_i x_i \tag{4.17}$$

Let $v_1, v_2, \ldots, v_n$ are the components of x_i, then the above equation can be expanded as

$$\left.\begin{aligned} a_{11}v_1 + a_{12}v_2 + \cdots + a_{1n}v_n &= \lambda_i v_1 \\ a_{21}v_1 + a_{22}v_2 + \cdots + a_{2n}v_n &= \lambda_i v_2 \\ \vdots \qquad\qquad\qquad\qquad &\quad \vdots \\ a_{s1}v_1 + a_{s2}v_2 + \cdots + a_{sn}v_n &= \lambda_i v_s \\ \vdots \qquad\qquad\qquad\qquad &\quad \vdots \\ a_{n1}v_1 + a_{n2}v_2 + \cdots + a_{nn}v_n &= \lambda_i v_n \end{aligned}\right\} \tag{4.18}$$

Suppose v_s be the largest in modulus of $v_1, v_2, \ldots, v_n$. Now, let us divide the sth equation by v_s and get

$$\lambda_i = a_{s1}\left(\frac{v_1}{v_s}\right) + a_{s2}\left(\frac{v_2}{v_s}\right) + \cdots + a_{ss} + \cdots + a_{sn}\left(\frac{v_n}{v_s}\right) \tag{4.19}$$

Since $\left|\frac{v_i}{v_s}\right| \le 1$, $i = 1, 2, \ldots, n$, it follows that

$$|\lambda_i| \le |a_{s1}| + |a_{s2}| + \cdots + |a_{ss}| + \cdots + |a_{sn}| \tag{4.20}$$

Eq. (4.19) can also be written as

$$|\lambda_i - a_{ss}| \le |a_{s1}| + |a_{s2}| + \cdots + |a_{sn}|$$

or

$$|\lambda_i - a_{ss}| \le \sum_{\substack{j=1 \\ j \ne s}}^{n} |a_{sj}| = R_s \tag{4.21}$$

Hence the proof. This theorem also holds for any column.

Scanned with CamScanner

An immediate consequence of Gerschgorin's theorem, when applied to identity matrix or permutation matrix is that, its eigenvalues lie within a circle having center at 1 and radius 0. Here follows an example.

Example 4.4 Apply Gerschgorin's theorem to the matrix

$$[A] = \begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & -1 & -1 \\ -1 & -1 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}$$

Solution This matrix has a dominant diagonal.

In this example $a_{ss} = 4$, Max. $R_s = 3$. Thus, Gerschgorin's theorem states that all the eigenvalues of the given matrix lie inside the circle with center at 4 and radius 3. In view of its symmetry, the eigenvalues are also real.

Definition 4.2 (Spectral Norm). Let λ_i be the largest eigenvalue of AA^* or A^*A, where A^* is the conjugate transpose of A, then the spectral norm of the matrix A, denoted by $\sigma(A)$ is defined as

$$\sigma(A) = (\lambda_i)^{1/2} \tag{4.22}$$

Definition 4.3 (Determinant). The determinant of an $n \times n$ matrix A is the product of its eigenvalues.

Definition 4.4 (Trace of a Matrix). The sum of the diagonal elements of an $n \times n$ matrix A is called the trace of the matrix A. Trace of the matrix A is also defined as the sum of its eignvalues.

EXERCISES

4.1 Find the largest eigenvalue of the matrix

$$\begin{bmatrix} 1 & & 2 \\ 4 & & -1 \\ 6 & & 5 \end{bmatrix}$$

and the corresponding eigenvector, by power method after sixth iteration

4.2 Find the largest eigenvalue of the matrix

$$\begin{bmatrix} 4 & 1 & 0 \\ 1 & 20 & 1 \\ 0 & 0 & 4 \end{bmatrix}$$

and the corresponding eigenvector, by power method after fourth iteration starting with the initial vector $v^{(0)} = (0, 0, 1)^T$.

Scanned with CamScanner

4.3 Find the dominant eigenvalue and the corresponding eigevector of the matrix

$$\begin{bmatrix} 8 & 1 & 2 \\ 0 & 10 & -1 \\ 6 & 2 & 15 \end{bmatrix}$$

by power method with unit vector as the initial vector.

4.4 Find the largest eigenvalue and the corresponding eigenvector of the matrix

$$\begin{bmatrix} 3 & 1 & 4 \\ 0 & 2 & 6 \\ 0 & 0 & 5 \end{bmatrix}$$

by power method at the end of sixth iteration, taking unit vector as the initial vector.

4.5 Using Jacobi's method, find all the eigenvalues and eigenvectors of the Hilbert matrix

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

Give result after two rotations.

4.6 Use Jacobi's method to find all the eigenvalues and the corresponding eigenvectors of the matrix

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

4.7 Find all the eigenvalues and the corresponding eigenvectors of the matrix

(i) $A = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix}$ (ii) $A = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}$

by Jacobi's method. Give results at the end of third rotation.

4.8 Find the dominant eigenvalue of

(i) $A = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}$ (ii) $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

and the corresponding eigenvector.

Scanned with CamScanner