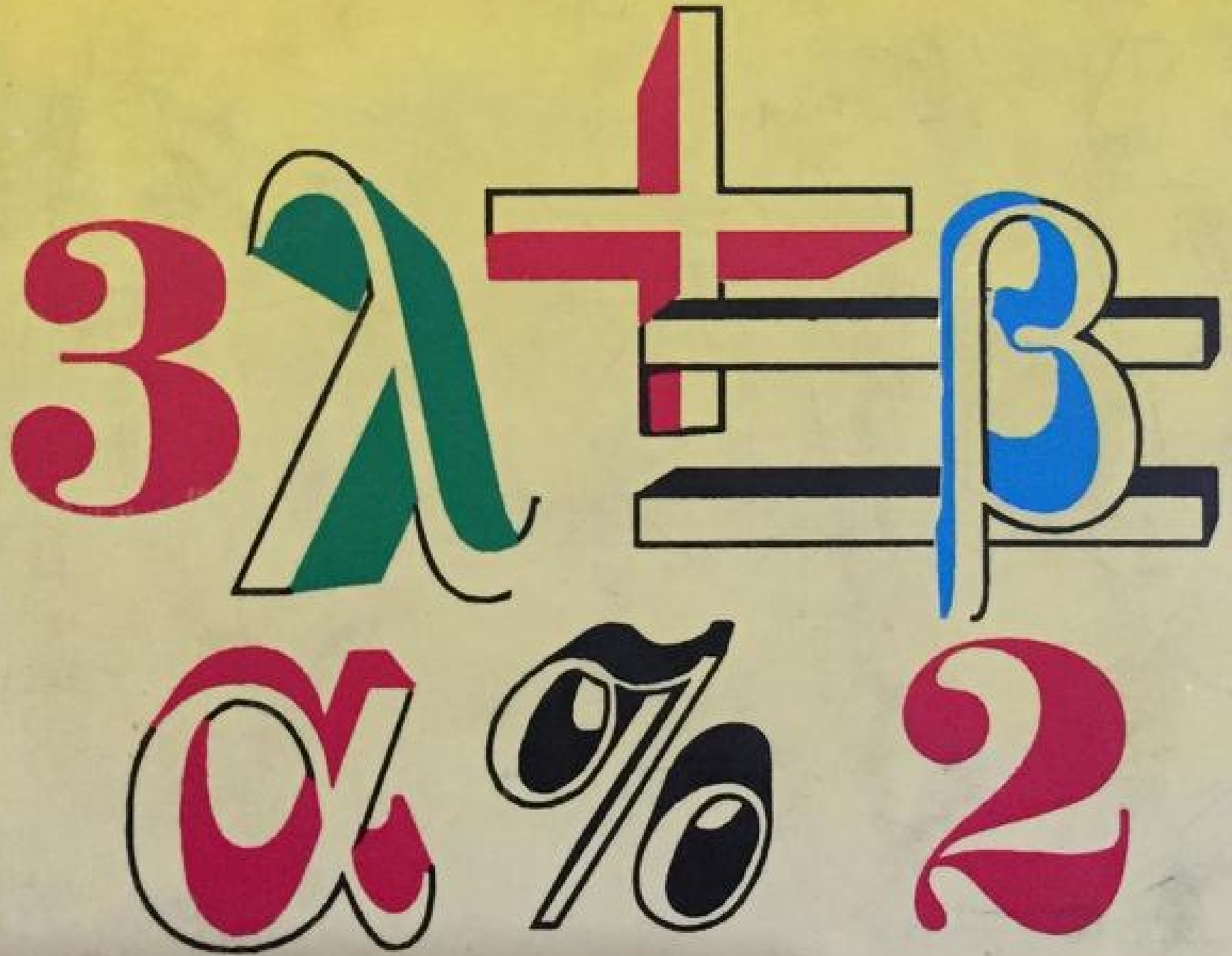
ELEMENTS OF NUMERICAL ANALYSIS

$$(-) \times (-) = +$$
$$(+) \times (-) = -$$



Dr. Faiz Ahmad
Muhammad Afzal Rana

CONTENTS

CHAPTER 4

INTERPOLATION WITH UNEQUALLY SPACED DATA

CHAPTER 5

INTERPOLATION WITH EQUALLY SPACED DATA

CHAPTER 6

THE SOLUTION OF NONLINEAR EQUATIONS

CHAPTER 7

DIFFERENCE EQUATIONS

CHAPTER 4

INTERPOLATION WITH UNEQUALLY SPACED DATA

CHAPTER 5

INTERPOLATION WITH EQUALLY SPACED DATA

CHAPTER 6

THE SOLUTION OF NONLINEAR EQUATIONS

CHAPTER 7

DIFFERENCE EQUATIONS

CHAPTER 11

NUMERICAL INTEGRATION

CHAPTER 12

NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

CHAPTER 13

SYSTEMS OF DIFFERENTIAL EQUATIONS AND HIGHER ORDER EQUATIONS

CHAPTER 1

ERRORS IN COMPUTATION

1.1 INTRODUCTION

Almost all calculations with numbers, whether peformed by hand or by computer, are subject to numerical errors. This form of error is unavoidable in numerical computation and should not be confused with 'gross errors' arising from programming mistakes or from incorrect design or misuse of algorithms. For example, if we wish to add 1 to 7 we can obtain the exact answer '8', but if we wish to divide 1 by 7 the exact answer is a decimal fraction with an infinite number of figures. We have to be content with an approximate answer, say 0.143, which is in error by about 0.0001428.... We have incurred this error because, in practice, we can only cope with a finite number of significant figures: in this case three. All calculating devices, from pencil-and-paper to the most sophisticated computer, have this restriction. Typically, hand-held calculators are limited to eight or ten significant figures; large, high speed computers can handle a few more.

Whenever calculations are performed there are many possible sources of error, such as inaccurate data and inaccurate formulae, and errors will obviously affect the accuracy of the solution of a particular problem. It is therefore evident that the error in a computed result may be due to one or both of two sources: errors in the data and errors of calculation. Errors of the first type cannot be remedied, but those of the second type can usually be made as small as we please. Thus, when such a number as e is replaced by its approximate value in a computation, we can decrease the error due to the approximation by taking e to as many figures as desired.

Nearly all numerical calculations are in some way approximate, and the aim of the computer should be to obtain results consistent with the data with a minimum of labour and with better approximation. The object of the present chapter is to identify the sources of numerical errors, assess the effects on our computations and to set forth some basic ideas and methods relating to

approximate calculations and to give methods for estimating the accuracy of the results obtained.

1.2 ERRORS AND THEIR SOURCES

The errors involved in mathematical problems may be divided into the following five groups.

1. Errors involved in the statement of the problem. More generally, there are almost always idealizations made in setting up a mathematical model before any computation can begin. The discrepancy between the model and the physical system is a source of error which can be termed as error of the problem. (Mathematical models and their relation to the real world are investigated in applied mathematics, physics, engineering, etc.)

   It sometimes happens that it is either difficult or even impossible to solve a given problem when formulated precisely. If that is the case, it is replaced by an approximate problem yielding almost the same results. This is the source of an error regarded as the error of the method.

2. Errors occur due to the presence of infinite processes in mathematical analysis. Many numerical methods express the desired result as the limit of an infinite sequence or the sum of an infinite series. Since, generally speaking, an infinite process cannot be completed in a finite number of steps, we are forced to stop at some term of the sequence and consider it to be an approximation to the required solution. Naturally, such a termination of the process gives rise to an error known as truncation error.

3. Errors associated with the system of numeration. In numerical computation there are fundamental limitations imposed by the finite nature of the computing medium at our disposal. Every possible means of computation is subject to the restrictions of finite space and finite time. Finite space gives rise to rounding error and finite time to truncation error. For example, if we wish to divide 2 by 3 the exact answer is a decimal fraction with an infinite number of figures $2/3=0.666\ldots$. We have to be content with an approximate number, say 0.67, which is in error by about $-0.003333\ldots$. We have incurred this error because, in practice, we can only use a finite number of digits in our computations.

4. Errors due to operations involving approximate numbers (errors of operation). Experimental data involving real numbers (measurements of length, time etc.) are always subject to uncertainty. When performing computations with approximate numbers, we naturally carry (to some extent) the errors

the original data into the final results. In this way, errors of operation are inherent.

5. Errors due to numerical parameters (in formulae) whose values can only be determined approximately. Such, for instance, are all physical constants and mathematical constants such as π and e. Although such constants are defined precisely in mathematical terms, their representation requires an infinite sequence of digits, for example,

$$\pi = 3.1415926 \ldots, \quad e = 2.7182818 \ldots.$$

To use these in numerical computation we must approximate the values to a finite number of digits, for instance,

$\pi \approx 3.142$ (rounded to 3 decimal places or 4 significant figures),
$e \approx 2.71828$ (rounded to 5 decimal places or 6 significant figures).

The consequence is a data error which may be called the initial error.

In a specific problem, quite naturally, some errors are absent and others exert a negligible effect. But, generally, a complete analysis must deal with all types of errors.

1.3 APPROXIMATE NUMBERS AND SIGNIFICANT FIGURES

In the discussion of approximate computation, it is convenient to make a distinction between numbers which are exact and those which express approximate values.

DEFINITION 1.1 (Approximate Numbers): An approximate value of a number, called approximate number, is defined as a number which is used as an approximation to an exact number and differs only slightly from the exact number for which it stands.

The numbers such as 5, 1/3, 200 etc. are exact numbers because there is no approximation or uncertainty associated with them. Although the numbers such as e, π, $\sqrt{2}$, etc. are exact numbers but they cannot be expressed exactly by a finite number of digits. When expressed in decimal form, they must be written as 2.7183, 3.1416, 1.4142, etc. Such numbers are therefore only approximations to the true values and are called approximate numbers.

The difference between the exact number x and its approximate value $\bar{x}$ is called **error**, that is,

$$\text{Error} = \text{Exact number} - \text{Approximate number} = x - \bar{x}.$$

DEFINITION 1.2 (Significant Figures): Significant figures (o
digits) of an approximate number is any one of the digi
1,2,..., 9, in its decimal representation, or any zero exce
when it is used to fix the decimal point or to fill the places
unknown or discarded digits.

For example, in the number 0.00475 the significant figures a
4,7,5; the first three zeros are used merely to fix the positi
of the decimal point and indicate the place values of the oth
digits and are therefore not significant. In the number 830
however, all the digits, including zero, are significant figure
In the number 0.0006070 the first four zeros are not significa
digits. All the remaining digits, including the other two zer
are significant figures. If the last digit of 0.0006070 is n
significant, then the number must be written as 0.000607. Fr
this point of view, the numbers 0.0006070 and 0.000607 are n
the same, because the former has four significant figures and t
later only three.

When writing large numbers, the zeros on the right can ser
both to indicate the significant figures and to fix the pla
values of the other digits. This can lead to misunderstandi
when the numbers are written in the ordinary way. Consider,
example, the number 675,000. It is not clear how many signific
figures there are, although we can say we have at least thr
This ambiguity can be removed by using powers-of-ten notat
(scientific notation) as 6.75×10^5 if the number has three sig
ficant figures, or as 6.750×10^5 if it has four significant fig
es, or as 6.7500×10^5 if the number has five significant figu
etc. Speaking generally, this notation is convenient for numb
containing a large number of non significant zeros, such
$0.00000230 = 2.30\times10^{-6}$, and the like.

1.4 ROUNDING OF NUMBERS AND ROUND-OFF ERRORS

If we divide 13 by 7, we obtain

$$13/7 = 1.8571428571\ldots,$$

a quotient which never terminates. In order to use such a nu
in a practical computation, we must cut it down to a manage
form, such as 1.86, or 1.857, or 1.85714, or 1.857143, etc.
process of cutting off superfluous digits and retaining as
as desired is called rounding off.

ROUNDING-OFF R
significant figure
the nth place, or
as place holder. I

(a) if the firs
the remain
(b) if the firs
retained di
(c) if the fir
non-zero d
retained dig
(d) if the firs
all other
digit is lef
it is even o

EXAMPLE 1.1: Roun
four, and three sig

SOLUTION: Roundi
obtain respectively
3.142, 3.14.

EXAMPLE 1.2: Roun
ures:

7.4727, 7

SOLUTION: We obta

7.473, 7

Let us now introd

DEFINITION 1.3 (Ro
requires more signif
the resulting error
must be rounded of
In other words, the
epresented to its fu
If we attempt to
raction with an inf
ent with an appro
ound-off error of
case, we do not kno
ome calculating devi
e that the result

ROUNDING-OFF RULE. To round off or simply round a number to n significant figures or digits, drop all digits to the right of the nth place, or replace them by zeros if the zeros are needed as place holder. In this rounding-off rule, note the following:

(a) if the first of the discarded digits is less than 5, leave the remaining digits unchanged;
(b) if the first discarded digit exceeds 5, add 1 to the last retained digit;
(c) if the first discarded digit is exactly 5 and there are non-zero digits among those discarded, add 1 to the last retained digit;
(d) if the first discarded digit, however, is exactly 5 and all other discarded digits are zeros, the last retained digit is left unchanged or is increased by 1 according as it is even or odd (the even-digit rule).

EXAMPLE 1.1: Round-off the number $\pi = 3.14159265...$ to six, five, four, and three significant figures.

SOLUTION: Rounding the number to given significant figures we obtain respectively the approximate numbers 3.14159, 3.1416, 3.142, 3.14.

EXAMPLE 1.2: Round the following number to four significant figures:

7.4727, 76346, 15.235 and 15.245

SOLUTION: We obtain respectively the approximate numbers

7.473, 76350, 15.24, 15.24.

Let us now introduce the concept of the round-off error.

DEFINITION 1.3 (Round-off error): If the result of a calculation requires more significant figures than the number available, then the resulting error is called a round-off error, the exact value must be rounded off to a certain number of significant figures. In other words, the round-off error arises when a number is not represented to its full precision.

If we attempt to divide 1 by 7 the exact answer is a decimal fraction with an infinite number of figures. We have to be content with an approximate answer, say 0.143, which contains a round-off error of about 0.00014. But when, as is usually the case, we do not know the exact value of an answer obtained using some calculating device, we assume the worst. For example, suppose that the result of some calculation worked to three signifi-

cant figures is 0.527. In fact, the answer may be anything be ween 0.52650... and 0.52749.... In other words, 0.527 may be i error due to round-off by ± 0.0005, i.e. by five units in the po sition after the last significant figure which is retained.

EXAMPLE 1.3: Find the round-off errors introduced in Example 1.2

SOLUTION: The round-off error introduced in each case of th example is as follows:

$$7.4727 - 7.473 = -0.0003$$

$$76346 - 76350 = -4$$

$$15.235 - 15.24 = -0.005$$

$$15.245 - 15.24 = 0.005$$

1.5 TRUNCATION ERROR

Many numerical methods express the desired results as the limi of an infinite sequence or the sum of an infinite series. I practice only a finite number of terms can be computed; the con sequence of neglecting the remaining terms is the error known a the truncation error. A truncation error, therefore, leads to a approximate formula being used instead of an exact one. Conside a convergent infinite series

$$S = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots \frac{1}{2^n} + \ldots.$$

Since we cannot include all of the infinite number of terms the series, and since we see that successive terms are contribu ing progressively smaller and smaller amounts to S, we conclu that we can truncate the series after some finite number of ter (n) and so obtain an approximation to S, which we might denote S_n. The error $S - S_n$ is called the truncation error in S_n.

It is important to note that it should be possible to redu the truncation error by taking more terms in the series. Tab 1.1 shows the values of S_n and the corresponding error for se ral values of n. The true value of S is 1. We can see that error decreases as n increases, and that S_n is converging to limit.

Table 1.1 Truncation Error

n	S_n	Error in S_n
3	0.875	0.125
5	0.96875	0.03125
10	0.99902	0.00098
15	0.99997	0.00003

1.6 ABSOLUTE, RELATIVE, AND PERCENTAGE ERRORS

DEFINITION 1.4 (Absolute Error): If $\bar{x}$ is an approximation to a number x, then the absolute error E_a of $\bar{x}$ is

$$E_a = |x - \bar{x}|. \tag{1.1}$$

DEFINITION 1.5: An approximation $\bar{x}$ to x is said to be correct to n decimal places or positions (or agree to n decimal places) if n is the largest non-negative integer for which

$$|x - \bar{x}| < 0.5\times10^{-n} \tag{1.2}$$

For $\pi = 3.14159...$ and $\bar{\pi} = 3.142$, the absolute error is

$$|\pi-\bar{\pi}| = 0.41\times10^{-3} < 0.5\times10^{-3},$$

and we see that $\bar{\pi}$ is indeed correct to 3 decimal places.

DEFINITION 1.6 (Limiting Absolute Error): The limiting absolute error of an approximate number is any number not less than the absolute error of that number

Thus, if E_{la} is the limiting absolute error of an approximate number $\bar{x}$ which takes the place of the exact number x, then

$$E_a = |x - \bar{x}| \le E_{la}. \tag{1.3}$$

From this it follows that the exact number x lies within the range

$$\bar{x} - E_{la} \le x \le \bar{x} + E_{la} \tag{1.4}$$

For brevity, we can then write $x = \bar{x} \pm E_{la}$.

EXAMPLE 1.4: Determine the limiting absolute error of the number $\bar{x} = 3.14$ which is used instead of the number π.

SOLUTION: It follows from the inequality

$$3.14 < \pi < 3.15$$

that $$|x - \bar{x}| = |3.15 - 3.14| = 0.01,$$

we can take $E_{la} = 0.01$.

We have a better estimate from the inequality

$$3.14 < \pi < 3.142$$

that

$$E_{la} = 0.002.$$

The absolute error (or the limiting absolute error) is not sufficient to describe the accuracy of a computation. Suppose that in measuring the lengths of two rods we get $l_1 = 105.8\text{cm} \pm 0.1\text{cm}$ and $l_2 = 200\text{cm} \pm 0.1\text{cm}$. Despite the fact that the limiting absolute errors coincide, the second measurement is better than the first one. For this reason, an essential point in the accuracy of measurement is to define the relative error.

DEFINITION 1.7 (Relative Error): If $\bar{x}$ is an approximation to x, then the relative error E_r of $\bar{x}$ is given by

$$E_r = \frac{|x - \bar{x}|}{|x|}, \qquad (x \neq 0),$$

or

$$E_r = \frac{E_a}{|x|}. \qquad (1.5)$$

EXAMPLE 1.5: a) If $x = .3000 \times 10$ and $\bar{x} = .3100 \times 10$, the absolute error is .1 and the relative error is $.3333 \times 10^{-1}$.

b) If $x = .3000 \times 10^{-3}$ and $\bar{x} = .3100 \times 10^{-3}$ the absolute error is $.1 \times 10^{-4}$ and the relative error is $.3333 \times 10^{-1}$.

c) If $x = .3000 \times 10^{4}$ and $\bar{x} = .3100 \times 10^{4}$, the absolute error is $.1 \times 10^{}$

and the relative error is $.3333\times10^{-1}$.

This example shows that the same relative error, $.3333\times10^{-1}$, occurs for widely varying absolute errors. Consequently, as a measure of accuracy, the absolute error may be misleading and the relative error more meaningful.

DEFINITION 1.8: An approximation $\bar{x}$ to x is said to be correct to n significant figures (or digits) if n is the largest non- negative integer for which

$$|x - \bar{x}| \leq \frac{1}{2}10^{m-n+1} \tag{1.6}$$

where m, an integer, is the highest power of ten when $\bar{x}$ is represented in the form of decimal. For example, the number $\bar{x} = 36.00$ is an approximation to the exact number $x = 35.97$ correct to three figures, since

$$|x - \bar{x}| = 0.03 < \frac{1}{2}\times10^{1-3+1} = \frac{1}{2}\times0.1$$

DEFINITION 1.9: The limiting relative error E_{lr} of a given approximate number $\bar{x}$ is any number not less than the relative error of that number. Thus, by definition we have

$$E_r \leq E_{lr} \tag{1.7}$$

or
$$\frac{E_a}{|x|} \leq E_{lr}$$

and hence
$$E_a \leq |x|E_{lr}.$$

Thus, for the limiting absolute error of a number $\bar{x}$ we can take

$$E_{la} = |x|E_{lr} \tag{1.8}$$

Since, in practical situations, $x \approx \bar{x}$, in place of (1.8) one frequently uses

$$E_{la} = |\bar{x}|E_{lr} \tag{1.8'}$$

EXAMPLE 1.6: The weight of 1 dm^3 of water at 0^0C is given as $p = 999.847$ gf $\pm$ 0.001 gf (gf = gram(force)). Determine the limiting relative error of the result of weighing the water.

SOLUTION: We have

$$E_{la} = 0.001 \text{ gf} \quad \text{and} \quad \bar{x} = p \leq 999.847 \text{ gf}.$$

Thus
$$E_{lr} = \frac{E_{la}}{\bar{x}} = \frac{0.001}{999.847} \approx 10^{-4}\%$$

DEFINITION 1.10: The percentage error is 100 times the relative error. Thus

percentage error of an approximate quantity = $100E_r$.

It is to be noted that relative and percentage errors are independent of the unit of measurement, whereas absolute errors are expressed in terms of the unit used.

EXERCISE 1.1

1. Round-off the following number to four significant figures: 11.64498, 81.9772, 48.365, 67.495, 4.4995002, 39.63244.

2. Round-off the number $\pi = 3.1415926535...$ to three, four and five significant figures and also find the absolute error in each case.

3. What will 7.52, 6.345, 7.4727, 76340, 15.235, 15.245, and 122377 become when rounded to one digit?

4. What will 2.73235, 35.671 and 430050 become when rounded to two digits?

5. What will 73.6547, 180.273 and 541500499 become when rounded to three digits?

6. Find round-off errors in Ex. 3, 4 and 5.

7. Find the relative error when $c = 485165195.4...(c = e^{20})$ is rounded to 4 significant figures.

8. Round-off the following numbers:

a) 8632574 to 4 significant figures
b) 3.1415926 to 5 decimal places
c) 8.5250 to 2 decimal places
d) 1.6750 to 2 decimal places.

Find round-off error in each case.

9. A machinist measures a 0.5 inch bolt to the nearest thousandth of an inch, and a carpenter a 10 feet beam to the nearest eighth of an inch. Which measurement is more accurate?

1.7 THE ERROR OF A SUM

THEOREM 1.1: The absolute error of an algebraic sum of several approximate numbers does not exceed the sum of the absolute errors of the numbers.

PROOF: Let $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ be the given approximate numbers. Consider their algebraic sum

$$u = \pm \bar{x}_1 \pm \bar{x}_2 \pm \ldots \pm \bar{x}_n$$

Then $$\Delta u = \pm \Delta\bar{x}_1 \pm \Delta\bar{x}_2 \pm \ldots \pm \Delta\bar{x}_n$$

and, hence $$|\Delta u| = E_a \leq |\Delta\bar{x}_1| + |\Delta\bar{x}_2| + \ldots + |\Delta\bar{x}_n|. \quad (1.9)$$

COROLLARY 1.1: The limiting absolute error of an algebraic sum is equal to the sum of the limiting absolute errros of the terms. Thus

$$E_{la} = E_{1_{la}} + E_{2_{la}} + \ldots + E_{n_{la}} \quad (1.10)$$

Following is the practical rule for the addition of approximate numbers.

RULE FOR ADDITION:

i) Find the approximate numbers with the least number of decimal places and leave them unchanged.
ii) Round off the remaining approximate numbers, retaining one or two more decimal places than those with the small-

est number of decimals.

iii) Add the numbers, taking into account all retained decimals.

v) Round off the result, reducing it by one decimal.

EXAMPLE 1.7: Find the sum of the approximate numbers 0.348, 0.1834, 345.4, 235.2, 11.75, 9.27, 0.0849, 0.0214, 0.000354, each correct to the indicated significant figures.

SOLUTION: We find the least accurate numbers 345.4 and 235.2 whose absolute error may attain 0.1. Rounding the remaining numbers to two decimal places, we obtain

$$\begin{array}{r} 345.4 \\ 235.2 \\ 11.75 \\ 9.27 \\ 0.08 \\ 0.02 \\ 0.35 \\ 0.18 \\ 0.00 \\ \hline 602.25 \end{array}$$

Round the result to one decimal place by the even-digit rule. We get the approximate value of the sum; 602.2.

The total absolute error E_a of the result is made up of three terms:

i) the sum of limiting errors of the original data:

$$E_{1_a} = 10^{-3} + 10^{-4} + 10^{-1} + 10^{-1} + 10^{-2} + 10^{-2} + 10^{-4} + 10^{-4} + 10^{-6}$$

$$= 0.221301 < 0.222.$$

ii) the absolute value of the sum of the rounding errors of the terms:

$$E_{2_a} = |0.0049+0.0014-0.002+0.0034+0.000354| = 0.008054 < 0.009.$$

iii) the final rounding error of the result:

$$E_{3_a} = |602.25 - 602.2| = 0.05.$$

Hence $E_a = E_{1_a} + E_{2_a} + E_{3_a} = 0.222 + 0.009 + 0.050 = 0.281 < 0.3$

and thus the desired sum is 602.2 ± 0.3.

EXAMPLE 1.8: Find the sum of 36490, 994, 557.32, 295000, and 86939, assuming that the number 29500 is known to only three significant figures.

SOLUTION: Since the number 29500 is known only to the nearest hundred, we round off the others to the nearest ten, and add to give the sum as shown below:

$$\begin{array}{r} 29500 \\ 86940 \\ 36490 \\ 990 \\ 560 \\ \hline 154480 \end{array}$$

Round the sum to hundred, we get the approximate value of the sum; 154500 or 1.545×10^5.

1.8 THE ERROR OF A DIFFERENCE

We consider the difference $u = \bar{x}_1 - \bar{x}_2$ of two approximate numbers $\bar{x}_1$, $\bar{x}_2$.

From equation (1.10), the limiting absolute error E_{1a} of the difference is

$$E_{1a} = E_{1_{1a}} + E_{2_{1a}}.$$

The limiting relative error of the difference is

$$E = \frac{E_{1a}}{x} = \frac{E_{1_{1a}} + E_{2_{1a}}}{x}$$

where x is the exact value of the absolute magnitude of the difference between the numbers $\bar{x}_1$ and $\bar{x}_2$.

When one approximate number is to be subtracted from another, they must both be rounded off to the same place before subtraction.

EXAMPLE 1.9: Subtract 46.365 from 779.8, assuming that each number is approximate and correct only to its last figure.

SOLUTION: Since 779.8 is correct to one decimal place, therefore, we round off the number 46.365 to one decimal place to obtain 46.4. Thus

$$779.8 - 46.4 = 733.4.$$

1.9 THE ERROR OF A PRODUCT

THEOREM 1.2: The relative error of a product of several approximate nonzero numbers does not exceed the sum of the relative errors of the numbers.

PROOF: Let $u = \bar{x}_1 . \bar{x}_2 \ldots \bar{x}_n .$

If all the approximate numbers $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ are positive, then

$$\ln u = \ln \bar{x}_1 + \ln \bar{x}_2 + \ldots + \ln \bar{x}_n .$$

Using the approximate formula $\Delta \ln \bar{x} \approx d \ln \bar{x} = \dfrac{\Delta \bar{x}}{\bar{x}}$, we obtain

$$\frac{\Delta u}{u} = \frac{\Delta \bar{x}_1}{\bar{x}_1} + \frac{\Delta \bar{x}_2}{\bar{x}_2} + \ldots + \frac{\Delta \bar{x}_n}{\bar{x}_n} .$$

Thus

$$\left|\frac{\Delta u}{u}\right| \leq \left|\frac{\Delta \bar{x}_1}{\bar{x}_1}\right| + \left|\frac{\Delta \bar{x}_2}{\bar{x}_2}\right| + \ldots + \left|\frac{\Delta \bar{x}_n}{\bar{x}_n}\right| .$$

If $x_i (i=1,2,\ldots,n)$ are the exact values of the factors $\bar{x}_i$, we can write approximately

$$\left|\frac{\Delta \bar{x}_i}{\bar{x}_i}\right| \approx \left|\frac{\Delta \bar{x}_i}{x_i}\right| = E_{i_r} \quad \text{and} \quad \left|\frac{\Delta u}{u}\right| = E_r$$

where E_{i_r} are the relative errors of the factors $\bar{x}_i (i=1,2,\ldots$

and E_r is the relative error of the product.

Consequently $$E_r \leq E_{1_r} + E_{2_r} + \ldots + E_{n_r}. \qquad (1.11)$$

It is clear that (1.11) also holds true if the factors $\bar{x}_i (i = 1, 2, \ldots, n)$ have different signs.

COROLLARY 1.2: The limiting relative error of a product is equal to the sum of the limiting relative errors of the factors:

$$E_{lr} = E_{1_{lr}} + E_{2_{lr}} + \ldots + E_{n_{lr}} \qquad (1.12)$$

EXAMPLE 1.10: Find the product 349.1×863.4 and state how many figures of the result are trustworthy if the factors are correct to all written figures.

SOLUTION: We have

$$E_{1_{la}} = \frac{1}{2}10^{2-4+1} = 0.05 \text{ and } E_{2_{la}} = \frac{1}{2}.10^{2-4+1} = 0.05.$$

Hence
$$E_{lr} = \frac{0.05}{349.1} + \frac{0.05}{863.4} = .000143 + .0000579 = 0.00020.$$

Now, $u = 349.1 \times 863.4 = 301412.94 \approx 301413.$

The limiting absolute error of this product is

$$E_{la} = u\, E_{lr} = 301413 \times 0.00020 = 60.2826 \approx 60.$$

And so u is correct to only three figures and the result should be written as

$$u = 301413 \pm 60.$$

Thus the true result lies between 301473 and 301353. There is some uncertainty about the fourth figure.

It is interesting to note that when an approximate number is multiplied by an exact number k, the limiting relative error remains unchanged, while the limiting absolute error is increased by $|k|$ times.

REMARKS: When it is desired to find the product of two or more approximate numbers of different accuracies, it is sufficient to:

1. round them off so that each contains one or two significant figures more than the least accurate factor;
2. in the final result, retain as many significant figures as there are correct figures in the least accurate factor (or keep one extra digit).

EXAMPLE 1.11: Find the product of the approximate numbers $\bar{x}_1 = 2.5$ and $\bar{x}_2 = 72.397$ correct to the number of digits written.

SOLUTION: Rounding $\bar{x}_2$ to one decimal place, we have

$$\bar{x}_1 = 2.5, \quad \bar{x}_2 = 72.4. \text{ Thus}$$

$$\bar{x}_1 \bar{x}_2 = 2.5 \times 72.4 = 181 \approx 1.8 \times 10^2.$$

1.10 THE ERROR OF A QUOTIENT

Let a, b be two approximate numbers. If

$$u = \frac{a}{b}$$

then

$$\ln u = \ln a - \ln b$$

and

$$\frac{\Delta u}{u} = \frac{\Delta a}{a} - \frac{\Delta b}{b}$$

Hence

$$\left|\frac{\Delta u}{u}\right| \leq \left|\frac{\Delta a}{a}\right| + \left|\frac{\Delta b}{b}\right|. \tag{1.13}$$

Thus, we have established the following theorem:

THEOREM 1.3: The relative error of a quotient does not exceed the sum of the relative errors of the dividend and the divisor.

COROLLARY 1.3: If $u = \frac{a}{b}$, then

$$E_{u_{lr}} = E_{a_{lr}} + E_{b_{lr}} = \text{limiting relative error of dividend} + \text{limiting relative error of divisor}$$

EXAMPLE 1.12: Find the number of correct figures in the quotient $56.3/\sqrt{5}$, assuming that the numerator is correct to its last figure but no further.

SOLUTION: We take $\sqrt{5} = 2.236$. Thus

$$E_{lr} = \frac{0.05}{56.3} + \frac{0.0005}{2.236} = .000888 + .000224 = 0.001$$

Since

$$u = \frac{56.3}{2.236} = 25.18 = 25.2,$$

then

$$E_{la} = u \times E_{lr} = 25.18 \times .001 = 0.025.$$

As this error does not affect the third figure of the quotient, we take 25.2 as the correct result.

REMARKS: If one of the numbers (divisor or dividend) is more accurate than the other, the more accurate number should be rounded off so as to contain one more significant figure than the less accurate one. The result should be given to as many significant figures as the less accurate number(or one digit more).

1.11 THE ERROR OF A POWER

Here $u = a^m$ (m any natural number). Then $\ln u = m \ln a$, and

$$\left|\frac{\Delta u}{u}\right| = m\left|\frac{\Delta a}{a}\right|.$$

Thus, we have

$$E_{u_{lr}} = mE_{a_{lr}}. \tag{1.14}$$

Hence, the limiting relative error of the mth power of a number is m times the limiting relative error of the number.

EXAMPLE 1.13: Find the number of trustworthy figures in $(0.3862)^4$, assuming that the number in parentheses is correct to its last figure but no further.

SOLUTION: Let $a = 0.3862$. Then the limiting relative error of the number a is

$$E_{a_{lr}} = \frac{0.00005}{0.3862} \approx 0.00013.$$

The limiting relative error of the result is

$$E_{lr} = 4xE_{a_{lr}} = 4X0.00013 = 0.00052.$$

Also, $(0.3862)^4 = 0.022246.$

Therefore, $E_{la} = 0.022246x0.00052 = 0.0000116.$

Since this error affects the fourth significant figure of t result, the best we can do is to write.

$$(0.3862)^4 = 0.02225$$

1.12 THE ERROR OF A ROOT

Suppose $u = \sqrt[m]{a}$, then $\ln u = \frac{1}{m} \ln a$ and, hence

$$\left|\frac{\Delta u}{u}\right| = \frac{1}{m}\left|\frac{\Delta a}{a}\right|.$$

Thus $E_{u_{lr}} = \frac{1}{m} E_{a_{lr}}$. (1.15)

That is, the limiting relative error of the mth root of an appr ximate number is only (1/m)th of the limiting relative error the given number (i.e. radicand).

EXAMPLE 1.14: Find the number of trustworthy figures $(0.3862)^{1/4}$, assuming that the number in parentheses is corre to its last figure but no farther.

SOLUTION: Here

$$E_{lr} = \frac{1}{4}(0.00013) = 0.000032, \quad (0.3862)^{1/4} = 0.78832.$$

Thus $E_{la} = 0.78832x0.000032 = 0.000026.$

Hence the fourth root 0.7883 is correct to four figures.

ELEMENTS

1. Find
   3.94
   no f
2. Comp
   figur
3. Deter
   73.56
   facto
4. Evalu
   round
   which
   to as

   a) $\frac{2}$

. Find t
  68.243
  gits or
. Determ
  es in t
. The nu
  only t
  how ma
. Find th
  876.3/4
  true onl
. Find th
  assuming
  digit giv
). Evaluate
  first pu
  maximum
  efficients
  maximum

EXERCISE 1.2

1. Find the sum of the approximate numbers 561.32, 86.954, 3.9462, and 491.6, each being correct to its last figure but no farther.

2. Compute the difference $\sqrt{2.03} - \sqrt{2}$ correct to five significant figures.

3. Determine the product of the approximate numbers 12.2 and 73.56 and the number of correct digits in the product if the factors are correct to all written digits.

4. Evaluate the following expressions. Then, assuming maximum round-off error in each figure, estimate the limits between which your calculated value lies. Finally quote each result to as many significant figures as are reliable.

   a) $\dfrac{2.362 \times 1.76}{1.46 \times 0.785}$ b) $7.62 + \dfrac{54}{15.3}$ c) $\dfrac{1.32 - 0.463}{16.1 \times 2.17}$

5. Find the sum of the approximate numbers 136.421, 28.3, 321, 68.243, 17.482 if the numbers are correct to their last digits only.

6. Determine the relative error and the number of correct figures in the product of the numbers 93.87 and 9.236.

7. The number 6852.4 and 48.392 are both approximate and true only to their last digits. Find their difference and state how many figures in the result are trustworthy.

8. Find the number of trustworthy figures in the quotient of 876.3/494.2, assuming that both numbers are approximate and true only to the number of digits given.

9. Find the number of correct digits in the quotient u=25.7:3.6 assuming the dividend and divisor are correct to the last digit given.

10. Evaluate the following expressions for the given values of x, first putting each in an appropriate form and estimate the maximum error in each result, assuming that all integer coefficients are exact but that all decimals are subject to the maximum round-off error.

CHAPTER 2

LINEAR OPERATORS

2.1 INTRODUCTION

Let F be a linear real function space, that is, F is a set of functions f,g, ... with the property that $f \in F$ and $g \in F$ implies $\alpha f + \beta g \in F$. Here α and β are arbitrary constants. Let us define the following operators which map F into itself:

$$E\ f(x) = f(x + h) \tag{2.1}$$

$$\Delta\ f(x) = f(x + h) - f(x) \tag{2.2}$$

$$\nabla\ f(x) = f(x) - f(x - h) \tag{2.3}$$

$$\delta\ f(x) = f(x + h/2) - f(x - h/2) \tag{2.4}$$

$$\mu\ f(x) = (1/2)[f(x + h/2) + f(x - h/2)] \tag{2.5}$$

The operators E,Δ,∇,δ and μ are respectively called the **shifting operator**, the **forward difference operator**, the **backward difference operator**, the **central difference operator** and the **mean value operator.** All these operators are linear i.e. for any two functions f and g, arbitrary constants α and β and P any of the above operators, we have

$$P(\alpha f + \beta g) = \alpha Pf + \beta Pg.$$

When P denotes any of the above operators, sums, differences and products of these operators are defined in the usual manner. An operator Q is said to be inverse of P if

$$QPf = f \qquad \text{for all } f \in F.$$

The inverse of operator P is denoted by P^{-1}.

Thus $E^{-1}f(x) = f(x - h)$

We also define $E^{1/2}f(x) = f(x + h/2)$

$$E^{-1/2}f(x) = f(x - h/2)$$

Two operators P_1 and P_2 are said to be equal if

$$P_1 f = P_2 f \qquad \text{for all } f \in F$$

EXAMPLE 2.1: Find Δe^x.

SOLUTION:
$$\begin{aligned}\Delta e^x &= e^{x+h} - e^x \\ &= (e^h - 1)e^x.\end{aligned}$$

EXAMPLE 2.2: Find $\Delta E^{-1}(\sin x)$.

SOLUTION:
$$\begin{aligned}\Delta E^{-1}\sin x &= \Delta \sin (x - h) \\ &= \sin (x - h + h) - \sin (x - h) \\ &= \sin x - \sin (x - h) \\ &= 2 \sin \frac{(x - x + h)}{2} \cos \frac{(x + x - h)}{2} \\ &= 2 \sin \frac{h}{2} \cos \frac{(2x - h)}{2}\end{aligned}$$

EXAMPLE 2.3: Find $\mu^2 f(x)$.

SOLUTION:
$$\begin{aligned}\mu^2 f(x) &= \mu(\mu f(x)) = \mu(1/2)[f(x + h/2) + f(x - h/2)] \\ &= (1/2)[\mu[f(x + h/2) + f(x - h/2)]] \\ &= (1/2)[(1/2)\{f(x+h)+f(x)\}+(1/2)\{f(x)+f(x-h)\}] \\ &= (1/4)[f(x + h) + 2 f(x) + f(x - h)].\end{aligned}$$

2.2 FUNCTIONS OF OPERATORS

Suppose every $f \in F$ can be expanded in Taylor series:

$$f(x+h) = f(x) + hf'(x) + (h^2/2) f''(x) + \ldots \qquad (2.6)$$

Denoting the derivative operator by D, we can write (2.6) in the form

$$E f(x) = (1 + hD + (h^2/2)D^2 + \ldots)f(x)$$

Thus we have the operator identity

$$E = 1 + hD + (h^2/2) D^2 + \ldots \qquad (2.7)$$

or we write

$$E = e^{hD}$$

where e^{hD} is defined by means of the power series in (2.7). Similarly one can define, for any operator P,

$$\sin P = P - \frac{P^3}{3!} + \frac{P^5}{5!} - \cdots$$

$$\log (1+P) = P - \frac{P^2}{2} + \frac{P^3}{3} - \cdots$$

etc. An operator defined by means of an infinite series is meaningful only if the series obtained when it operates on any function $f \in F$ is convergent. For example, let Q be defined by

$$Q = 1 + D + \frac{D^2}{2} + \frac{D^3}{3} + \cdots$$

If F is the space of polynomials, the above definition is meaningful. Howeve if $e^x \in F$ then Qe^x does not have any meaning, since

$$Qe^x = e^x(1 + 1 + 1/2 + 1/3 + \ldots),$$

and the series on the right side does not converge.

2.3 OPERATOR IDENTITIES

Any of the above operators can be expressed in terms of any other operator. For example the relations

$$\Delta = E - 1 \tag{2.8}$$

$$\nabla = 1 - E^{-1} \tag{2.9}$$

$$\delta = E^{1/2} - E^{-1/2} \tag{2.10}$$

$$\mu = (1/2)[E^{1/2} + E^{-1/2}] \tag{2.11}$$

follow from the definition of these operators. From

$$\delta = E^{1/2} - E^{-1/2},$$

we have

$$\delta^2 = (E^{1/2} - E^{-1/2})(E^{1/2} - E^{-1/2})$$
$$= E + E^{-1} - 2 \tag{2.12}$$

and $$\mu = (1/2)[E^{1/2} + E^{-1/2}],$$

gives $$\mu^2 = (1/4)[E + E^{-1} + 2). \tag{2.13}$$

Thus $$\delta^2+2 = 4\mu^2 - 2.$$

This gives the useful relation

$$\mu^2 = 1 + \frac{\delta^2}{4} \tag{2.14}$$

EXAMPLE 2.4: Prove that $\delta^2 = \Delta^2 (1 + \Delta)^{-1}$.

SOLUTION: $$\delta = E^{1/2} - E^{-1/2}$$

Therefore $$\delta^2 = E + E^{-1} - 2$$

But $$E = \Delta + 1,$$

hence $$\delta^2 = \Delta + 1 + \frac{1}{\Delta+1} - 2$$

$$= \frac{(\Delta + 1)^2 + 1 - 2(\Delta + 1)}{\Delta + 1}$$

$$= \Delta^2/(\Delta + 1).$$

EXAMPLE 2.5: Show that $\mu^2 = (1 + \Delta/2)^2(1 + \Delta)^{-1}$.

SOLUTION: Since $\mu = (1/2)[E^{1/2} + E^{-1/2}),$

therefore $$\mu^2= (1/4)[E + E^{-1} + 2]$$

But $$E = 1 + \Delta,$$

therefore $$\mu^2= (1/4)[1 + \Delta + \frac{1}{1+\Delta} +2]$$

$$= (1/4)[\frac{(1 + \Delta)^2 + 1 + 2(1 + \Delta)}{1 + \Delta}]$$

$$=(1/4)[2 + \Delta]^2/(1 + \Delta)$$

$$= (1 + \Delta/2)^2(1 + \Delta)^{-1}.$$

EXAMPLE 2.6: Prove that $E = 1 + \frac{\delta^2}{2} + \delta \left[1 + \frac{\delta^2}{4} \right]^{1/2}$.

SOLUTION: We know $\delta = E^{1/2} - E^{-1/2}$.

and $$\mu = (1/2)[E^{1/2} + E^{-1/2}]$$

which give $$2E^{1/2} = \delta + 2\mu = \delta + 2 \left[1 + \frac{\delta^2}{4} \right]^{1/2}.$$

On squaring, we get the desired result.

2.4 DIFFERNCE OPERATORS AND THE DERIVATIVE OPERATORS

From the relation

$$E = e^{hD}$$

we have

$$\Delta = E - 1 = e^{hD} - 1, \tag{2.13}$$

$$\nabla = 1 - E^{-1} = 1 - e^{-hD} \tag{2.14}$$

$$\delta = E^{1/2} - E^{-1/2} = e^{hD/2} - e^{-hD/2},$$
$$= 2 \sinh (hD),$$

$$\mu = (1/2)[E^{1/2} + E^{-1/2}] = \cosh (hD).$$

Later on, in chapter 10, we shall derive the following formulae:

$$hD = \Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4} + \ldots \tag{2.15}$$

$$h^2D^2 = \Delta^2 - \Delta^3 + \frac{11}{12}\Delta^4 - \ldots \tag{2.16}$$

$$hD = \nabla + \frac{\nabla^2}{2} + \frac{\nabla^3}{3} + \frac{\nabla^4}{4} + \ldots \tag{2.17}$$

$$h^2D^2 = \nabla^2 + \nabla^3 + \frac{11}{12}\nabla^4 + \ldots \tag{2.18}$$

EXERCISE 2

1. Evaluate $\Delta(2^x)$, $\nabla(\cos x)$, $\mu\delta(x)$.

2. Prove the following identities:

(a) $\delta^2 = \nabla^2(1 - \nabla)^{-1} = \Delta^2(1 + \Delta)^{-1}$

(b) $\mu\delta = (1/2)[E - E^{-1}] = (1/2)[\Delta + \nabla]$

(c) $\mu^2 = [1 - \nabla/2]^2(1 - \nabla)^{-1}$

(d) $\Delta = (1/2)\delta^2 + \delta\left[1 + \frac{\delta^2}{4}\right]^{1/2}$

(e) $\nabla = -(1/2)\delta^2 + \delta\left[1 + \frac{\delta^2}{4}\right]^{1/2}$

(f) $\mu\delta = \Delta(1 + \Delta/2)(1 + \Delta)^{-1}$

(g) $1 + \mu^2\delta^2 = (1 + \delta^2/2)^2$

(h) $(1 + \Delta/2)(1 + \Delta)^{-1/2} = (1 - \nabla/2)(1 - \nabla)^{-1/2}$

CHAPTER 3

EIGENVALUES AND EIGENVECTORS OF MATRICES

3.1 INTRODUCTION

Eigenvalues play a very important role in obtaining deep understanding of many physical system. For example, the stability of aircraft in flight and the modes of vibration of a bridge are governed by eigenvalues. They are also important in the study of the growth of certain types of population. Suppose **X** denotes the number of females of a certain age in a given population. If the population is in the state of age stability, the vector **X** satisfies an equation of the form

$$AX = \lambda X$$

where A is a square matrix and λ is a positive real number which indicates whether the population increases or not. If $\lambda < 1$ the population decreases, If $\lambda > 1$ it increases, and if $\lambda = 1$ it remains the same.

In fact, for a given square matrix A there are only a limited number of values of λ which lead to a non-trivial solution. These are known as the **eigenvalues** of the matrix and the corresponding solution vectors are known as the **eigenvectors** of A.

3.2 EIGENVALUES AND EIGENVECTORS

Let A be a square matrix of order n. A number λ is called an eigenvalue or characteristic value of A if an $n \times 1$ non zero column vector X' can be found such that

$$AX = \lambda X. \qquad (3.1)$$

The above equation may also be written as

$$(A - \lambda I)X = 0$$

If $X^T = (x_1, x_2, \ldots, x_n)$, then we have a system of n equations in n unknowns:

$$\begin{aligned}
(a_{11} - \lambda)x_1 + a_{12}x_2 + a_{13}x_3 + \ldots a_{1n}x_n &= 0,\\
a_{21}x_1 + (a_{22} - \lambda)x_2 + \ldots a_{2n}x_n &= 0,\\
\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots &\\
\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots &\\
a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + (a_{nn} - \lambda)x_n &= 0.
\end{aligned}$$

It is well-known that a non-trivial solution exists if and only if the determinant of the system vanishes i.e.,

$$\begin{vmatrix}
a_{11}-\lambda & a_{12} & \ldots & a_{1n}\\
a_{21} & a_{22}-\lambda & \ldots & a_{2n}\\
\ldots & \ldots & \ldots & \ldots\\
\ldots & \ldots & \ldots & \ldots\\
\ldots & \ldots & \ldots & \ldots\\
a_{n1} & a_{n2} & \ldots & a_{nn}-\lambda
\end{vmatrix} = 0 \qquad (3.2)$$

This gives an equation of degree n in λ:

$$\lambda^n - (a_{11} + a_{22} + \ldots + a_{nn})\lambda^{n-1} + \ldots + (-1)^n k = 0, \qquad (3.3)$$

where k is yet to be found. It is clear that if we put $\lambda = 0$ in the left hand side of (3.2), we are left with det A. Hence k in (3.3) must be equal to det A. The equation

$$\lambda^n - (a_{11} + a_{22} + \ldots + a_{nn})\lambda^{n-1} + \ldots + (-1)^n \det A = 0, \qquad (3.4)$$

is called the **characteristic equation** of the matrix A and any root of this equation will be an eigenvalue of A. Thus a matrix of order n has n eigenvalues (although all of them may not be distint).

EXAMPLE 3.1: Find eigenvalues and corresponding eigenvectors of the matrix

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

SOLUTION: $|A - \lambda I| = \begin{vmatrix} 2-\lambda & 3 \\ 4 & 3-\lambda \end{vmatrix} = (2 - \lambda)(3 - \lambda) - 12$

$$= \lambda^2 - 5\lambda - 6.$$

Thus $\lambda^2-5\lambda-6 = 0$ is the characteristic equation of A and $\lambda = 6,-1$ are the two eigenvalues. To find eigenvcctors we have to solve

$$(2 - \lambda)x_1 + 3x_2 = 0,$$
$$4x_1 + (3 - \lambda)x_2 = 0.$$

First if $\lambda = 6$, then the system becomes

$$4x_1 - 3x_2 = 0,$$
$$4x_1 - 3x_2 = 0.$$

So x_2 may be chosen arbitrarily. Let $x_2 = 1$, then $x_1 = 3/4$. So

$$x_1 = \begin{bmatrix} 3/4 \\ 1 \end{bmatrix}$$

is an eigenvector corresponding to $\lambda_1 = 6$. Similarly we find that

$$x_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

is an eigenvector corresponding to $\lambda_2 = -1$.

EXAMPLE 3.2: Find the eigenvalues and the corresponding eigenvectors of

$$A = \begin{bmatrix} -2 & 6 & -24 \\ 0 & -3 & 10 \\ 1 & -4 & 13 \end{bmatrix}$$

SOLUTION: $|A - \lambda I| = \begin{vmatrix} -2-\lambda & 6 & -24 \\ 0 & -3-\lambda & 10 \\ 1 & -4 & 13-\lambda \end{vmatrix} = -\lambda^3 + 8\lambda^3 - 5\lambda - 14.$

The characteristic equation is

$$\lambda^3 - 8\lambda^2 + 5\lambda + 14 = 0,$$

with roots $\lambda = -1,2,7$. Hence A has eigenvalues -1,2 and 7. To find the eigenvectors we proceed as follows. Corresponding to -1 we have to solve

$$AX = -X$$

Let $X = (x_1, x_2, x_3)$. Thus we have to solve the system of equations

$$\begin{aligned} -2x_1 + 6x_2 - 24x_3 &= -x_1, \\ -3x_2 + 10x_3 &= -x_2, \\ x_1 - 4x_2 + 13x_3 &= -x_3, \end{aligned}$$

or

$$\begin{aligned} x_1 - 6x_2 + 24x_3 &= 0, \\ 2x_2 - 10x_3 &= 0, \\ x_1 - 4x_2 + 14x_3 &= 0. \end{aligned}$$

Since the equations are not linearly independent, we can choose x_3 arbitrarily. Let $x_3 = 1$, then $x_2 = 5$, $x_1 = 6$. Thus $\begin{bmatrix} 6 \\ 5 \\ 1 \end{bmatrix}$ is an eigenvector corresponding to the eigenvalue $\lambda_1 = -1$. Similarly we find that $\begin{bmatrix} -3 \\ 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$ respectively are the eigenvectors corresponding to the eigenvalues $\lambda_2 = 2$ and $\lambda_3 = 7$.

We now give some simple results about eigenvalues of a matrix. Suppose $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the (not necessarily distinct) eigenvalues of a square matrix A.

THEOREM 3.1: $\lambda_1 + \lambda_2 + \ldots + \lambda_n = a_{11} + a_{22} + \ldots + a_{nn}$

$$\lambda_1 \lambda_2 \ldots \lambda_n = \det A$$

This result follows from equation (3.4). The number $\lambda_1 + \lambda_2 + \ldots + \lambda_n$ is called the **trace** of A and is denoted by tr(A).

THEOREM 3.2: If **X** is an eigenvector of A corresponding to an eigenvalue λ then so is cX where c is any non zero constant.

PROOF:

$$AX = \lambda X \Rightarrow cAX = c\lambda X$$

or $$A(cX) = \lambda(cX).$$

DEFINITION 3.1: Two column vectors X_1 and X_2 are said to be orthogonal if $\overline{X}_1 X_2 = 0$.

Here $\overline{X}_1 = (X_1^T)^*$, for example if $X_1 = \begin{bmatrix} i \\ 2-i \\ 3 \end{bmatrix}$, then $\overline{X}_1 = (-i, 2+i, 3)$.

DEFINITION 3.2: A matrix is said to be hermitian if $\overline{A} = A$.

EXAMPLE 3.3: If $A = \begin{bmatrix} 1 & 1-i \\ 1+i & 3 \end{bmatrix}$, then $A^T = \begin{bmatrix} 1 & 1+i \\ 1-i & 3 \end{bmatrix}$

and $(A^T)^* = \overline{A} = \begin{bmatrix} 1 & 1-i \\ 1+i & 3 \end{bmatrix} = A.$

Thus A is a hermitian matrix.

THEOREM 3.3: A hermitian matrix has only real eigenvalues.

PROOF: Let A be hermitian and λ be an eigenvalue and X the corresponding eigenvector i.e.

$$AX = \lambda X.$$

By definition X is not the zero vector. Taking the transpose of both sides we have

$$(AX)^T = \lambda X^T,$$

or

$$X^T A^T = \lambda X^T.$$

Take complex conjugates

$$\overline{X}\ \overline{A} = \overline{\lambda}\ \overline{X},$$

or

$$\overline{X}\ A = \overline{\lambda}\ \overline{X},$$

because A is hermitian. Postmultiply with X

$$\overline{X}\ A\ X = \overline{\lambda}\ \overline{X}\ X,$$

or

$$\overline{X}\ \lambda X = \overline{\lambda}\ \overline{X}\ X,$$

or

$$(\lambda - \overline{\lambda})\ \overline{X}X = 0.$$

But $\overline{X}X \neq 0$, since X is assumed to be a non zero vector. Hence $\lambda = \overline{\lambda}$ which shows that λ is real.

THEOREM 3.4: Eigenvectors corresponding to distinct eigenvalues of a hermitian matrix are orthogonal.

PROOF: Let $AX_1 = \lambda_1 X_1$,

and $AX_2 = \lambda_2 X_2$,

where $\lambda_1 \neq \lambda_2$, and A is a hermitian matrix. Proceeding as above we have

$$\overline{X}_1 A X_2 = \lambda_1 \overline{X}_1 X_2$$

where we used the fact that λ_1 is real. The above equation implies

$$(\lambda_2 - \lambda_1) \overline{X}_1 X_2 = 0.$$

Since $\lambda_1 \neq \lambda_2$, we have

$$\overline{X}_1 X_2 = 0,$$

which proves the theorem.

An important corollary of the above theorem is as follows.

COROLLARY: A symmetric real matrix has only real eigenvalues and the eigenvectors corresponding to distinct eigenvalues are orthogonal.

It is easily shown [8], (page 84) that if $X_1, X_2, \ldots, X_n$ are eigenvectors of A corresponding to distinct eigenvalues then the set $\{X_1, \ldots, X_n\}$ is linearly independent. Hence every n×1 column vector **X** can be expressed as a linear combination of these vectors. However if A is hermitian, it can be shown that [1] even if the eigenvalues are not distinct an orthogonal set of linearly independent eigenvectors $\{X_r\}_{r=1}^{n}$ can always be found.

EXAMPLE 3.4: The unit matrix of order 3 has an eigenvalue repeated three times. However

$$X_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad X_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

are three linearly independent orthonormal vectors such that

$$IX_i = X_i, \qquad i = 1, 2, 3.$$

EXAMPLE 3.5: The matrix

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

has the eigenvalue unity repeated three times. However there is only one linearly independent eigenvector

$$X = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

3.3 GERSHGORIN'S THEOREM

Now we consider some estimates for the absolute value $|\lambda|$ where λ is any eigenvalue of a square matrix $A = \{a_{ij}\}$. If $X = (x_1, x_2, \ldots, x_n)^T$ is an eigenvector corresponding to λ, then we have

$$\lambda x_i = \sum_{r=1}^{n} a_{ir} x_r, \qquad i = 1, 2, \ldots, n. \tag{3.5}$$

If $|x_N| = \max_r |x_r|$, then we have

$$|\lambda x_N| = \left| \sum_{r=1}^{n} a_{Nr} x_r \right| \le \sum_{r=1}^{n} |a_{Nr}| \, |x_r|$$

$$\le \sum_{r=1}^{n} |a_{Nr}| \, |x_N|$$

$$= |x_N| \sum_{r=1}^{n} |a_{Nr}|$$

Hence $$|\lambda| \le \sum_{r=1}^{n} |a_{Nr}|.$$

Since N is, in general, not known we replace the above estimate by a weaker one

$$|\lambda| \leq \max_i \sum_{r=1}^{n} |a_{ir}| \tag{3.6}$$

Also the transposed matrix has the same set of eigenvalues, hence

$$|\lambda| \leq \max_i \sum_{r=1}^{n} |a_{ri}|. \tag{3.7}$$

From (3.5) we have

$$(\lambda - a_{NN})x_N = \sum_{r \neq N}^{n} a_{Nr} x_r .$$

Thus $$|\lambda - a_{NN}|\ |x_N| \leq \sum_{r \neq N}^{n} |a_{Nr}|\ |x_r|$$

$$\leq |x_N| \sum_{r \neq N}^{n} |a_{Nr}|$$

or $$|\lambda - a_{NN}| \leq \sum_{r \neq N}^{n} |a_{Nr}|.$$

As before we replace the above result by a weaker one:"λ" lies in one of the discs with centre at a_{ii} and radius

$$\sum_{\substack{r=1 \\ r \neq i}}^{n} |a_{ir}|,\ i = 1, 2, \ldots, n.$$

$$|\lambda - a_{ii}| \leq \sum_{r \neq i}^{n} |a_{ir}|,\ i = 1, 2, \ldots, n. \tag{3.8}$$

and from the transposed matrix, we have, "λ" lies in one of the discs,

$$|\lambda - a_{ii}| \leq \sum_{r \neq i}^{n} |a_{ri}|,\ i = 1, 2, \ldots, n. \tag{3.9}$$

All the above estimates hold simultaneously, therefore the best of them is preferred. The above results are known as Gershgorin's

Theorem:

THEOREM 3.5: If λ is an eigenvalue of $A = \{a_{ij}\}$ then it satisfies the estimates (3.6) - (3.9).

EXAMPLE 3.6: Consider

$$A = \begin{bmatrix} -2 & 6 & -24 \\ 0 & -3 & 10 \\ 1 & -4 & 13 \end{bmatrix}$$

Here

$$\sum_{r=1}^{3} |a_{1r}| = 2 + 6 + 24 = 32$$

$$\sum_{r=1}^{3} |a_{2r}| = 0 + 3 + 10 = 13$$

$$\sum_{r=1}^{3} |a_{3r}| = 1 + 4 + 13 = 18$$

Therefore (3.6) gives

$$|\lambda| \leq 32. \qquad (3.10)$$

Similarly (3.7) gives

$$|\lambda| \leq 47 \qquad (3.11)$$

From (3.8) we get

$$|\lambda + 2| \leq 30 \quad \text{or} \quad |\lambda + 3| \leq 10 \quad \text{or} \quad |\lambda - 13| \leq 5. \qquad (3.12)$$

Finally (3.9) gives

$$|\lambda + 2| \leq 1 \quad \text{or} \quad |\lambda + 3| \leq 10 \quad \text{or} \quad |\lambda - 13| \leq 34 \qquad (3.13)$$

In (3.12) the disc $|\lambda + 2| \leq 30$ contains the other two and in (3.13) $|\lambda - 13| \leq 34$ does the same. Thus

$|\lambda + 2| \leq 30$ is the best estimate.

EXAMPLE 3.7: Consider

$$A = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 3 & -2 \\ -5 & 3 & 8 \end{bmatrix}$$

Here we have $|\lambda| \leq 16$ and $|\lambda| \leq 12$.

Also $|\lambda - 1| \leq 4$ or $|\lambda - 3| \leq 4$ or $|\lambda - 8| \leq 8.$

and $|\lambda - 1| \leq 7$ or $|\lambda - 3| \leq 5$ or $|\lambda - 8| \leq 4.$

Thus all eigenvalues lie in the union of $|\lambda-1| \leq 7$ and $|\lambda-8| \leq 4.$
The matrix has eigenvalues 3, 4 and 5.

3.4 THE POWER METHOD

The power method is an iterative technique to find the dominant eigenvalue of a given matrix A. Assume that a matrix has eigenvalues which are so arranged that $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$ and $u_1, u_2, \ldots, u_n$ are n linearly independent eigenvectors. If u is any arbitrary n x 1 vector then it can be expressed as a linear combination of the eigenvectors and we write

$$u = a_1 u_1 + a_2 u_2 + \ldots + a_n u_n.$$

Then
$$Au = a_1 Au_1 + a_2 Au_2 + \ldots + a_n Au_n,$$
$$= a_1 \lambda_1 u_1 + a_2 \lambda_2 u_2 + \ldots + a_n \lambda_n u_n.$$

It is clear

$$A^k u = a_1(\lambda_1^k)u_1 + a_2(\lambda_2^k)u_2 + \ldots + a_n(\lambda_n^k)u_n$$

$$= \lambda_1^k \left[a_1 u_1 + a_2\left(\frac{\lambda_2}{\lambda_1}\right)^k u_2 + \ldots + a_n\left(\frac{\lambda_n}{\lambda_1}\right)^k u_n\right] \qquad (3.14)$$

Since $\left|\frac{\lambda_r}{\lambda_1}\right| < 1, \quad r = 2, 3, \ldots n,$

the vector in brackets will tend to $a_1 u_1$ as k is taken large. If $(V)_i$ denotes the i-th component of the column vector V, we have

$$\frac{(A^{k+1}u)_i}{(A^k u)_i} = \lambda_1 \left[\frac{[a_1 u_1 + a_2\{\frac{\lambda_2}{\lambda_1}\}^{k+1} u_2 + \ldots + a_n\{\frac{\lambda_n}{\lambda_1}\}^{k+1} u_n]_i}{[a_1 u_1 + a_2\{\frac{\lambda_2}{\lambda_1}\}^{k} u_2 + \ldots + a_n\{\frac{\lambda_n}{\lambda_1}\}^{k} u_n]_i} \right]$$

Hence
$$\lim_{k \to \infty} \frac{(A^{k+1}u)_i}{(A^k u)_i} = \lambda_1 \frac{a_1 (u_1)_i}{a_1 (u_1)_i} = \lambda_1 \qquad (3.15)$$

From (3.15) we have for large k

$$A^{k+1}u \approx \lambda_1 A^k u$$

Thus
$$A(A^k u) \approx \lambda_1 (A^k u)$$

Therefore $A^k u$ is an approximate eigenvector corresponding to the eigenvalue λ_1. The method based on the above analysis, for finding the dominant (in absolute value) eigenvalue of a matrix, is called the power method. We proceed as follows:

a) Choose a suitable arbitrary column vector u. Usually u is chosen as $(1, 1, \ldots, 1)^T$.
b) Form $\mathbf{v} = \mathbf{Au}$.

c) Multiply **v** by a number k_i so as to make the largest element equal to 1. If one is working with an automatic machine, it is easier to multiply, at every step, with

$$k_i = \left(|v_1|^2 + |v_2|^2 + \ldots + |v_n|^2 \right)^{-1/2}$$

We do this because if v is an eigenvector then so is kv. This is advisable to stop the numbers from getting too large.

d) Iterate $v_{i+1} = Ak_i v_i \qquad i = 1, 2, \ldots.$

EXAMPLE 3.8: Find the dominant eigenvalue of

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

SOLUTION: Let

$$u = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \text{ then } v = Au = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad u_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

Denote by u_i the normalized column vector $k_i v_i$ i.e. the vector whose largest element in absolute value has been made equal to 1.

$$v_2 = A u_1 = \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \qquad u_2 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

$$v_3 = A u_2 = \begin{bmatrix} 3 \\ -4 \\ 3 \end{bmatrix} \qquad u_3 = \begin{bmatrix} -0.75 \\ 1 \\ -0.75 \end{bmatrix}$$

$$v_4 = A u_3 = \begin{bmatrix} -2.5 \\ 3.5 \\ -2.5 \end{bmatrix} \qquad u_4 = \begin{bmatrix} -0.7083 \\ 1 \\ -0.7083 \end{bmatrix}$$

$$v_5 = A u_4 = \begin{bmatrix} -2.428 \\ 3.428 \\ -2.428 \end{bmatrix} \qquad u_5 = \begin{bmatrix} -0.7073 \\ 1 \\ -0.7073 \end{bmatrix}$$

$$v_6 = A u_5 = \begin{bmatrix} -2.4166 \\ 3.4166 \\ -2.4166 \end{bmatrix} \qquad u_6 = \begin{bmatrix} -.70714 \\ 1 \\ -.70714 \end{bmatrix}$$

$$v_7 = A u_6 = \begin{bmatrix} -2.4146 \\ 3.4146 \\ -2.4146 \end{bmatrix} \qquad u_7 = \begin{bmatrix} -0.707112 \\ 1 \\ -0.707112 \end{bmatrix}$$

$$v_8 = A u_7 = \begin{bmatrix} -2.41428 \\ 3.41428 \\ -2.41428 \end{bmatrix} \qquad u_8 = \begin{bmatrix} -0.70711 \\ 1 \\ -0.70711 \end{bmatrix}$$

$$v_9 = A u_8 = \begin{bmatrix} -2.414224 \\ 3.414224 \\ -2.414224 \end{bmatrix} \qquad u_9 = \begin{bmatrix} -0.70711 \\ 1 \\ -0.70711 \end{bmatrix}$$

Hence to 5 decimal places

$$A u_8 = 3.414224 \begin{bmatrix} -0.70711 \\ 1 \\ -0.70711 \end{bmatrix} = 3.414224\, u_9$$

Therefore the dominant eigenvalue is 3.414224 and the corresponding eigenvector is $(-0.70711, 1, -0.70711)^T$.

The exact value is $2+\sqrt{2} = 3.414214$.

The smallest and the largest eigenvalues:

We can modify the power method to find the smallest and

largest eigenvalues.

Let λ be an eigenvalue of a matrix A and X be the corresponding eigenvector. Then

$$AX = \lambda X$$

Subtracting αX from both sides of the above equation we get

$$AX - \alpha X = \lambda X - \alpha X$$
$$(A - \alpha I)X = (\lambda - \alpha)X$$

Thus $\lambda-\alpha$ is an eigenvalue of the matrix $A-\alpha I$ and X is the associated vector. In otherwords, the effect of subtracting α from each element of the leading diagonal of a matrix has the effect of subtracting α from each eigenvalue but leaves the eigenvectors unchanged. Therefore, we can make, by a proper choice of α, the largest or the smallest eigenvalue the dominant one in absolute value.

Suppose that α is taken equal to or greater than the dominant eigenvalue of A. Then, if the dominant eigenvalue of $A - \alpha I$ is found, adding α to it will give the numerically smallest eigenvalue of A. But the correspondig eigenvector will be the same as that found for the dominant eigenvalue of $A - \alpha I$.

EXAMPLE 3.9: Find the smallest eigenvalue of

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

SOLUTION: It is clear that all eigenvalues satisfy $|\lambda - 2| \leq 2$. Thus they must lie in the interval [0, 4]. Hence the largest eigenvalue is also the dominant (in absolute value). To find the smallest eigenvalue we consider the matrix

$$B = A - 4I = \begin{bmatrix} -2 & -1 & 0 \\ -1 & -2 & -1 \\ 0 & -1 & -2 \end{bmatrix}$$

Apply the power method to B.

Let $u = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, then $v = Au = \begin{bmatrix} -3 \\ -4 \\ -3 \end{bmatrix} = -4 \begin{bmatrix} 0.75 \\ 1 \\ 0.75 \end{bmatrix} = -4u_1$

$$v_2 = Au_1 = \begin{bmatrix}-2.5\\-3.5\\-2.5\end{bmatrix} = -3.5\begin{bmatrix}0.7143\\1\\0.7143\end{bmatrix} = -3.5\ u_2$$

$$v_3 = Au_2 = \begin{bmatrix}-2.4286\\-3.4286\\-2.4286\end{bmatrix} = -3.4286\begin{bmatrix}0.70833\\1\\0.70833\end{bmatrix} = -3.4286u_3$$

$$v_4 = Au_3 = \begin{bmatrix}-2.41667\\-3.41667\\-2.41667\end{bmatrix} = -3.41667\begin{bmatrix}0.70732\\1\\0.70732\end{bmatrix} = -3.41667u_4$$

$$v_5 = Au_4 = \begin{bmatrix}-2.41464\\-3.41464\\-2.41464\end{bmatrix} = -3.41464\begin{bmatrix}0.7071434\\1\\0.7071434\end{bmatrix} = -3.41464u_5$$

$$v_6 = Au_5 = \begin{bmatrix}-2.4142868\\-3.4142868\\-2.4142868\end{bmatrix} = -3.4142868\begin{bmatrix}0.7071131\\1\\0.7071131\end{bmatrix} = -3.4142868$$

$$v_7 = Au_6 = \begin{bmatrix}-2.414226127\\-3.414226127\\-2.414226127\end{bmatrix} = -3.414226\begin{bmatrix}0.707108\\1\\0.707108\end{bmatrix} = -3.4144226$$

Thus an approximate eigenvalue for B is -3.414226. The corresponding eigenvalue for A is -3.414226 + 4 = 0.585774.

The exact value is $2-\sqrt{2}$ = 0.58578644. The approximate corresponding eigenvector for this eigenvalue is

$$\begin{bmatrix}0.707108\\1\\0.707108\end{bmatrix}$$

The question arises that whether the smallest eigenvalue can found without finding the largest one. Suppose

$$AX = \lambda X$$

then $$X = A^{-1}\lambda X = \lambda A^{-1}X \quad \text{(As } \lambda \text{ is scalar)}$$

or $$A^{-1}X = \lambda^{-1}X.$$

This equation shows that $1/\lambda$ is an eigenvalue of A^{-1}. If power method is used to find the largest eigenvalue $1/\lambda$ of then the smallest eigenvalue of A is λ. Moreover, if X is eigenvector of A^{-1} corresponding to eigenvalue $1/\lambda$, then X also the eigenvector of A corresponding to eigenvalue λ.

EXAMPLE 3.10

By power meth
is 0.25. The
eigenvalue of
is $(1, -0.5)^T$.
4, 9. Thus both

Rayleigh Quot
$u_1, u_2, \ldots, u_n$ a
them orthonorm

$$\bar{u}u$$

$$v_{p+1}$$

Hence using the

$$\gamma = \frac{\bar{v}_{p+1} v_p}{\bar{v}_p v_p}$$

Thus we see that

much faster as
s called the sequ
xample 3.9,

$$\gamma_1 = -3.$$

EXAMPLE 3.10: Let $A = \begin{bmatrix} 5 & 2 \\ 2 & 8 \end{bmatrix}$, then

$$A^{-1} = \frac{1}{36}\begin{bmatrix} 8 & -2 \\ -2 & 5 \end{bmatrix} = \begin{bmatrix} 0.2222 & -0.0556 \\ -0.0556 & 0.1389 \end{bmatrix}$$

By power method we can find the largest eigenvalue of A^{-1} which is 0.25. The corresponding eigenvector is $(1, -0.5)^T$. The eigenvalue of A is 1/0.25 or 4 and the corresponding eigenvector is $(1, -0.5)^T$. The eigenvalues of A calculated analytically are 4, 9. Thus both the results agree.

Rayleigh Quotient: If a matrix is Hermitian its eigenvectors $u_1, u_2, \ldots, u_n$ are orthogonal. We suppose that we have also made them orthonormal. Let u be an arbitrary column vector such that

$\bar{u}u_1 \neq 0$. Now

$$u = c_1u_1 + c_2u_2 + \ldots + c_nu_n$$

$$v_p = A^pu = c_1(\lambda_1)^p u_1 + c_2(\lambda_2)^p u_2 + \ldots + c_n(\lambda_n)^p u_n$$

and $$v_{p+1} = c_1(\lambda_1)^{p+1}u_1 + \ldots + c_n(\lambda_n)^{p+1} u_n.$$

Hence using the orthonormality property of the eigenvectors i.e.

$\bar{u}_i u_j = \delta_{ij}$, we find

$$\gamma_p = \frac{\bar{v}_{p+1} v_p}{\bar{v}_p v_p} = \lambda_1 \frac{|c_1|^2+|c_2|^2\left(\frac{\lambda_2}{\lambda_1}\right)^{2p+1}+\ldots+|c_n|^2\left(\frac{\lambda_n}{\lambda_1}\right)^{2p+1}}{|c_1|^2+|c_2|^2\left(\frac{\lambda_2}{\lambda_1}\right)^{2p}+\ldots+|c_n|^2\left(\frac{\lambda_n}{\lambda_1}\right)^{2p}}$$

Thus we see that the convergence of the sequence $\{\gamma_p\}_{p=1}^{\infty}$ to λ_1 is much faster as compared with the sequence given by (3.15). This is called the sequence of **Rayleigh quotients**. For example in the example 3.9,

$$\gamma_2 = -3.4117647, \ \ldots, \ \gamma_6 = -3.414213473$$

and we see that γ_6 is much closer to the exact value -3.4142135 as compared with the value -3.414226 obtained by the earlier method.

3.5 DEFLATION

Suppose we have found an eigenvalue λ_1 and the corresponding eigenvector X_1 of an n×n matrix A. Deflation means finding a matrix of order (n-1)×(n-1) whose eigenvalues coincide with the rest of the eigenvalues of A. Suppose $X_1 = (x_1, x_2, \ldots, x_n)^T$. We assume that $x_1 \neq 0$ and multiply X_1 by $1/x_1$ to get an eigenvector X_1^* whose top element is unity i.e.

$$X_1^* = (1, x_2^*, \ldots, x_n^*)^T.$$

If the i-th row of A is denoted by R_i, we see that

$$AX_1^* = \begin{bmatrix} R_1 X_1^* \\ R_2 X_1^* \\ R_3 X_1^* \\ \vdots \\ R_n X_1^* \end{bmatrix}.$$

Since $AX_1^* = \lambda_1 X_1^* = (\lambda_1, \lambda_1 x_2^*, \ldots, \lambda_1 x_n^*)^T$, it is clear that

$$R_1 X_1^* = \lambda_1 \tag{3.16}$$

If rest of the eigenvectors have also been normalized in the same manner i.e. their top element is 1, then

$$R_i X_i = \lambda_i \tag{3.17}$$

If, however X_i is such that its top element is zero, then

$$R_1 X_i^* = 0. \qquad (3.16c)$$

Now we form the matrix B defined as

$$B = A - X_1^* R_1.$$

Note that $X_1^* R_1$ is a an n×n matrix whose first row is R_1, hence B has only zeros in its first row. We show that with the exception of λ_1, matrices A and B have all eigenvalues common.

THEOREM 3.6: If A has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ then B has eigenvalues $0, \lambda_2, \ldots, \lambda_n$.

PROOF: Suppose $AX_i^* = \lambda_i X_i^*$ $(i = 1, 2, \ldots, n)$

where X_1^* has its top element equal to unity and each of $X_2^*, X_3^*, \ldots, X_n^*$ has its top element either 1 or 0. Now

$$\begin{aligned} BX_1^* &= (A - X_1^* R_1)X_1^*, \\ &= AX_1^* - X_1^*(R_1 X_1^*), \\ &= \lambda_1 X_1^* - X_1^* \lambda_1, \\ &= 0. \end{aligned}$$

Hence 0 is an eigenvalue of B with eigenvector X_1^*. If X_1^* has its top element unity then

$$\begin{aligned} B(X_1^* - X_i^*) &= (A - X_1^* R_1)(X_1^* - X_i^*) \\ &= AX_1^* - AX_i^* - X_1^*(R_1 X_1^*) + X_1^*(R_1 X_1^*) \\ &= \lambda_1 X_1^* - \lambda_i X_i^* - \lambda_1 X_1^* + \lambda_i X_1^*, \\ &= \lambda_i (X_1^* - X_i^*). \end{aligned} \qquad (3.17)$$

In the above, we have made use of (3.16a) and (3.16b). Now it is clear from (3.17) that λ_i is an eigenvalue of B and $X_1^* - X_i^*$ is the

corresponding eigenvector. Finally if X_1^* has its top eleme
equal to zero, then

$$B X_J^* = AX_J^* - X_1^* R_1 X_J^*$$
$$= AX_J^*.$$

Since $R_1 X_J^* = 0$ by (3.16c). Thus again we have

$$BX_J^* = \lambda_J X_J^* \qquad (3.18$$

This completes the proof.

Now define an $(n-1)\times(n-1)$ matrix A_1 obtained by delet
the first row and the first column of B. Let

$$B = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ b_{21} & b_{22} & b_{23} & \dots & b_{2n} \\ b_{31} & b_{32} & b_{33} & \dots & b_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & b_{n3} & \dots & b_{nn} \end{bmatrix}.$$

Then

$$A_1 = \begin{bmatrix} b_{22} & b_{23} & b_{24} & \dots & b_{2n} \\ b_{32} & b_{33} & b_{34} & \dots & b_{3n} \\ b_{42} & b_{43} & b_{44} & \dots & b_{4n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{n2} & b_{n3} & b_{n4} & \dots & b_{nn} \end{bmatrix}$$

Since

$$|B - \lambda I_n| = -\lambda |A_1 - \lambda I_{n-1}|, \qquad (3$$

it is obvious that B and A_1 have all eigenvalues common with exception of $\lambda = 0$. In (3.19) I_n denotes the unit matrix order n. A_1 is the so called deflated matrix we were seeking.

EXAMPLE 3.11:

A has an eigenv
$X_1^* = (1,1,1/2)^T$ a
Now

and $B = A - X_1^* R_1$

B

The matrix A_1 ob
column of B is fou

A_1 =

Now $|A_1 - \lambda$

Thus A_1 has eigenv
eigenvalues of A.

We can recover
deflated matrix A_1 i
$(y_1, y_2, \dots, y_{n-1})$
$(0, y_1, y_2, \dots, y_{n-1})$
denote this vector b
is λ_1, i.e.

EXAMPLE 3.11: Let $A = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 3 & -2 \\ -5 & 3 & 8 \end{bmatrix}$.

A has an eigenvalue 4 with $(2,2,1)^T$ as an eigenvector. Here $\lambda_1 = 4$, $X_1^* = (1,1,1/2)^T$ and $R_1 = (1,2,2)$.

Now

$$X_1^* R_1 = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 2 \\ 1/2 & 1 & 1 \end{bmatrix}, \tag{3.18}$$

and $B = A - X_1^* R_1$ comes out as

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & -4 \\ \frac{-11}{2} & 2 & 7 \end{bmatrix}$$

The matrix A_1 obtained by deleting the first row and the first column of B is found to be

$$A_1 = \begin{bmatrix} 1 & -4 \\ 2 & 7 \end{bmatrix}$$

Now $$|A_1 - \lambda I| = \lambda^2 - 8\lambda + 15$$
$$= (\lambda - 5)(\lambda - 3).$$

Thus A_1 has eigenvalues 5 and 3. These are also the remaining eigenvalues of A.

We can recover eigenvectors of the matrix A from those of deflated matrix A_1 in a simple manner. First we note that if $(y_1, y_2, \ldots, y_{n-1})^T$ is an eigenvector of matrix A_1 then $(0, y_1, y_2, \ldots, y_{n-1})^T$ is an eigenvector of the matrix B. Let us denote this vector by Y_1 and suppose the corresponding eigenvalue is λ_1, i.e.

$$By_1 = \lambda_1 Y_1 \tag{3.19}$$

However from (3.17) and (3.18) we have

$$B(X_1^* - X_i^*) = \lambda_i(X_1^* - X_i^*) \tag{3.21}$$

if $R_1 X_i^* = \lambda_i \neq 0$

and $BX_i^* = \lambda_i X_i^*$

if $R_1 X_i^* = 0.$ (3.22)

If (3.21) holds then Y_i must be proportional to $X_1^* - X_i^*$ since these are eigenvectors of B corresponding to the same eigenvalue. Thus for some α

$$X_1^* - X_i^* = \alpha Y_i. \tag{3.23}$$

Multiply (3.23) with R_1, we get

$$R_1 X_1^* - R_1 X_i^* = \alpha R_1 Y_i,$$

or

$$\lambda_1 - \lambda_i = \alpha R_1 Y_i.$$

This gives $\alpha = (\lambda_1 - \lambda_i)/R_1 Y_i$ and (3.23) gives

$$X_i^* = X_1^* - (\lambda_1 - \lambda_i)Y_i/R_1 Y_i. \tag{3.24}$$

If $R_1 X_i^* = 0$, i.e. (3.22) holds, then

$$Y_i = \beta X_i^*$$

for some $\beta \neq 0$, since X_i^* and Y_i happen to be eigenvectors of corresponding to the same eigenvalue. In this case Y_i itself an eigenvector of A. Actually we do not know X_i^*, however $R_1 Y_i \neq 0$ implies $R_1 X_1^* \neq 0$ and $R_1 Y_i = 0$ implies $R_1 X_i^* = 0$.

EXAMPLE 3.1

We know t
matrix A_1 can
A_1 has an eig
value 3 with a

Since $R_1 Y_2 = 0$
let $\lambda_3 = 3$ and

Here $R_1 Y_3 = 2$ a

Thus we have fin
and 3 with re
$(1,0,1)^T$.

In the beginnin
element of X_1 is d
make the second e
suitable constant a
modifications.

In practice we c
ponding eigenvector

EXAMPLE 3.12: Let $A = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 3 & -2 \\ -5 & 3 & 8 \end{bmatrix}$.

We know that $\lambda_1 = 4$, $X_1^* = (1,1,1/2)^T$. We found that the deflated matrix A_1 came out to be

$$A_1 = \begin{bmatrix} 1 & -4 \\ 2 & 7 \end{bmatrix}.$$

A_1 has an eigenvalue 5 with an eigenvector $(1,-1)^T$ and an eigenvalue 3 with an eigenvector $(2,-1)^T$. Let us take $\lambda_2 = 5$ and

$$Y_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}.$$

Since $R_1 Y_2 = 0$, we conclude that Y_2 is an eigenvector of A. Now let $\lambda_3 = 3$ and

$$Y_3 = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix}.$$

Here $R_1 Y_3 = 2$ and (3.24) gives

$$X_3^* = \begin{bmatrix} 1 \\ 1 \\ 1/2 \end{bmatrix} - \frac{(4-3)}{2} \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix},$$

$$= \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Thus we have finally found that the matrix A has eigenvalues 4,5 and 3 with respective eigenvectors $(1,1,1/2)^T$, $(0,2,-1)^T$ and $(1,0,1)^T$.

In the beginning of this section we assumed that the top element of X_1 is different from zero. If it happens to be zero we make the second element from the top unity by multiplying with a suitable constant and carry through with the analysis with minor modifications.

In practice we can find the dominant eigenvalue and the corresponding eigenvector of a given matrix A by the power method, then

find the matrix A_1 by deflation, apply the power method to A_1 to find the next dominant eigenvalue and so on.

In the case of 3×3 matrix we can find the dominant eigenvalue and the corresponding eigenvector of the matrix by the power method, then the smallest eigenvalue and the corresponding vector. The third eigenvalue follows immediately as the sum of the eigenvalues is equal to the trace of the given matrix.

The power method fails if the dominant eigenvalues are a pair of complex conjugate numbers. Also the convergence is poor if ratio of the two dominant eigenvalues is close to unity. In such a case Jacobi's method may be used. For details see Fröberg [4].

EXERCISE 3

1. Find the eigenvalues and the corresponding eigenvectors of the following matrices analytically.

$$A = \begin{bmatrix} 1 & 2 & -8 \\ -2 & 2 & -2 \\ 1 & -4 & 10 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 2 & -10 \\ -3 & 1 & -3 \\ 1 & -5 & 11 \end{bmatrix}, \quad C = \begin{bmatrix} -2 & 2 & 2 \\ 3 & -1 & 3 \\ 1 & 1 & -3 \end{bmatrix}.$$

2. Find the dominant eigenvalue and the corresponding eigenvector for each of the following matrices. Give your answers correct to three decimal places.

(a) $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix}$.

3. Find all the eigenvalues of the matrix

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

by calculating the roots of its characteristic equation. Hence determine the associated eigenvectors and normalize them. Give your answers correct to four decimal places.

4. Taking $u = (0,0,1)^T$, use the power method to evaluate the dominant eigenvalue and corresponding eigenvector of the matrix A given in Exercise 3. Give your answer correct to four decimal places and compare them with those obtained previously.

5. Using the estimate of eigenvector obtained in Exercise 4, deflate the matrix A and hence calculate the remaining eigenvalues and associated eigenectors. Compare your results with those obtained in Exercise 3.

6. Using the estimate of eigenvector obtained in Exercise 4, find the smallest eigenvalue and associated eigenvector of A.

7. Consider the marix

$$A = \begin{bmatrix} 0 & 2 & -8 \\ -2 & 1 & -2 \\ 1 & -4 & 9 \end{bmatrix}$$

having $\lambda_1 = -1$ as an eigenvalue and $X_1 = (1, 3/2, 1/2)^T$ the corresponding eigenvector. Deflate the matrix and hence calculate the remaining eigenvalues and associated eigenvectors.

CHAPTER 4

INTERPOLATION WITH UNEQUALLY SPACED DATA

4.1 INTRODUCTION

Suppose we are given the values of a function $f(x)$ at certain points $x_0, x_1, \ldots, x_n$ and we want to estimate $f(\alpha)$ where α is any point in the interval $[x_0, x_n]$. We assume that the function can be approximated by a polynomial. This is called polynomial interpolation.

There are a number of interpolation formulae, most of which have certain advantages over the others in certain situations, but no one of which is preferable to all others in all respects. In the general case, when the abscissas are not equally spaced, the use of divided differences is convenient.

4.2 LAGRANGE'S FORMULA

We assume that the polynomial

$$P(x) = a_0 + a_1 x + \ldots + a_n x^n$$

is such that $P(x_r) = y_r$, $r = 0, 1, 2, \ldots, n$. Thus the curves representing graphs of the two functions $y = f(x)$ and $y = P(x)$ pass through the same $n+1$ points $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$. If the function $f(x)$ has continuous derivatives of all orders, it is reasonable to suppose that $f(x)$ and $P(x)$ will be fairly close at other points of the interval $[x_0, x_n]$. To determine $a_0, a_1, \ldots, a_n$ we have to solve the system of equations:

$$\left.\begin{aligned} a_0 + a_1x_0 + a_2x_0^2 + \ldots + a_nx_0^n &= y_0, \\ a_0 + a_1x_1 + a_2x_1^2 + \ldots + a_nx_1^n &= y_1, \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots &\\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots &\\ a_0 + a_1x_n + a_2x_n^2 + \ldots + a_nx_n^n &= y_n. \end{aligned}\right\} \tag{4.1}$$

The determinant of the system is

$$\begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix}$$

Since $x_0, x_1, \ldots, x_n$ are all distinct, this determinant is not zero and the system (4.1) has a unique solution. However one or more, but not all, of a_n may turn out to be zero. Thus we have shown that a polynomial of degree $\leq n$ exists such that the curve representing $y = P(x)$ passes through the given $n+1$ points. Also this polynomial is unique. Let us define

$$L_k(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_0)(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)}.$$

We see that

$$L_k(x_r) = \begin{cases} 0, & \text{if } r \neq k, \\ 1, & \text{if } r = k. \end{cases} \tag{4.2}$$

If we write

$$P(x) = L_0(x)y_0 + L_1(x)y_1 + \ldots + L_n(x)y_n, \tag{4.3}$$

we see that $P(x)$ is a polynomial of degree $\leq n$ (since $L_0(x), L_1(x)$, etc. are polynomials of degree n and coefficients of x^n, x^{n-1}

may add up to zero), also

$$P(x_0) = y_0,\ P(x_1) = y_1, \ldots,\ P(x_n) = y_n.$$

Thus P(x) is the required polynomial. Equation (4.3) is called **the Lagrange's interpolation formula**. Note that, although we proved the existence of such a polynomial by a different method, we seldom use the determinant method to actually evaluate the constants $a_0, a_1, \ldots, a_n$. Later we shall discuss other methods for finding the interpolation polynomial. Since such a polynomial is unique, it does not matter how we find it. The result is always the same.

EXAMPLE 4.1: Let values of $y = f(x)$ be as given in the following table. Find a polynomial which interpolates the given data

x	y
-1	-8
1	-2
2	4
3	28

SOLUTION: Here $x_0 = -1,\ x_1 = 1,\ x_2 = 2,\ x_3 = 3,$

$$y_0 = -8,\ y_1 = -2,\ y_2 = 4,\ y_3 = 28.$$

Thus $$L_0(x) = \frac{(x-1)(x-2)(x-3)}{(-1-1)(-1-2)(-1-3)},$$

$$L_1(x) = \frac{(x-(-1))(x-2)(x-3)}{(1-(-1))(1-2)(1-3)} \text{ etc.}$$

Here

$$P(x) = \frac{(x-1)(x-2)(x-3)}{(-2)(-3)(-4)}(-8) + \frac{(x+1)(x-2)(x-3)}{2(-1)(-2)}(-2)$$

$$+ \frac{(x-(-1))(x-1)(x-3)}{(2-(-1))(2-1)(2-3)}(4) + \frac{(x-(-1))(x-1)(x-2)}{(3-(-1))(3-1)(3-2)}(28)$$

$$= (x^3 - 6x^2 + 11x - 6)(1/3) + (x^3 - 4x^2 + x + 6)(-1/2)$$
$$+ (x^3 - 3x^2 - x + 3)(-4/3) + (x^3 - 2x^2 - x + 2)(7/2)$$

$$= 2x^3 - 3x^2 + x - 2.$$

4.3 THE ERROR OF THE INTERPOLATION POLYNOMIAL

Suppose we want to consider the error $R(x)=f(x)-P(x)$ at a point α in the interval $[x_0,x_n]$. Of course $R=0$ if $\alpha = x_0, x_1,\ldots,x_n$. So let α be a point distinct from them. Let us define

$$F(x) = (x - x_0)(x - x_1) \ldots (x - x_n)$$

and write

$$g(x) = f(x) - p(x) - KF(x). \qquad (4.4)$$

In (4.4) K is a constant to be chosen in a manner to make $g(\alpha)=0$. Note that $g(x_0) = g(x_1) = \ldots = g(x_n) = 0$.

Consider $g(x)$ on the interval $[x_0,x_1]$. Now $g(x_0)=g(x_1)= 0$, also g is continuous on $[x_0,x_1]$ and differentiable on (x_0,x_1) hence, by Rolle's theorem, there is a point β_1 in (x_0,x_1) such that $g'(\beta_1) = 0$. Applying the same argument to $g(x)$ on intervals $[x_1,x_2]$, ..., $[x_{n-1},x_n]$ we find that there exist points $\beta_2,\beta_3,\ldots,\beta_{n+1}$ such that the derivative of $g(x)$ vanishes at each of these points. Now we apply Rolle's theorem to $g'(x)$ and find that there exist points $\gamma_1,\gamma_2,\ldots,\gamma_n$ such that $g''(x)$ vanishes at each of these points. Proceeding in this manner, we find that there exists a point ξ in (x_0,x_n) such that $g^{(n+1)}(\xi)=0$. Now from (4.4)

$$g^{n+1}(\xi) = f^{(n+1)}(\xi) - P^{(n+1)}(\xi) - K(n+1)!$$

But $P^{n+1}(\xi)= 0$, because $P(x)$ is a polynomial of degree at most n. Hence

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

From (4.4) we get, on putting $x = \alpha$ (remember $g(\alpha) = 0$)

$$f(\alpha) = P(\alpha) + \frac{f^{(n+1)}(\xi)}{(n+1)!} F(\alpha). \qquad (4.5)$$

In the beginning we assumed that $\alpha \neq x_0, x_1, \ldots, x_n$. However (4.5) is obviously satisfied even if we remove this restriction. Since α is an arbitrary point we can write

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!} F(x), \quad x \in [x_0, x_n].$$

Note that ξ also depends on x. We have shown that the error $f(x) - P(x)$ committed by approximating the function $f(x)$ by the interpolation polynomial is

$$R = \frac{f^{n+1}(\xi)\ F(x)}{(n+1)!},\ x_0 < \xi < x_n. \tag{4.6}$$

4.4 DIVIDED DIFFERENCES

Let the values of a function f be given at n+1 points $x_0, x_1, \ldots, x_n$. We define the first divided difference of f(x) between x_0 and x_1 as

$$f(x_0,x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}. \tag{4.7}$$

Note that $f(x_0,x_1) = f(x_1,x_0)$. We define the second divided difference, $f(x_0,x_1,x_2)$, as

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

Similarly the n-th divided difference is defined as

$$f(x_0,x_1,\ldots,x_n) = \frac{f(x_1,x_2,\ldots,x_n) - f(x_0,x_1,\ldots,x_{n-1})}{x_n - x_0}. \tag{4.8}$$

Now we shall prove the following theorem:

THEOREM 4.1:

$$f(x_0,x_1,\ldots,x_n) = \sum_{r=0}^{n} \frac{f(x_r)}{(x_r-x_0)(x_r-x_1)\ldots(x_r-x_{r-1})(x_r-x_{r+1})\ldots(x_r-x_n)},$$

PROOF: Obviously the result is true for $n = 1$, (see 4.7). Let the result be true for $n = k$. Thus we have

$$f(x_0,x_1,\ldots,x_k)$$

$$= \sum_{r=0}^{k} \frac{f(x_r)}{(x_r-x_0)(x_r-x_1)\ldots(x_r-x_{r-1})(x_r-x_{r+1})\ldots(x_r-x_k)} \qquad (4.9)$$

Now by definition

$$f(x_0,x_1,\ldots,x_{k+1}) = \frac{f(x_1,\ldots,x_{k+1}) - f(x_0,\ldots,x_k)}{x_{k+1}-x_0}$$

$$= \frac{1}{(x_{k+1}-x_0)}\left[\sum_{r=1}^{k+1} \frac{f(x_r)}{(x_r-x_1)\ldots(x_r-x_{k+1})} - \sum_{r=0}^{k} \frac{f(x_r)}{(x_r-x_0)\ldots(x_r-x_k)}\right]$$

where we have used Eq. (4.9),

$$= \frac{1}{(x_{k+1}-x_0)}\left[\frac{f(x_{k+1})}{(x_{k+1}-x_1)\ldots(x_{k+1}-x_k)} - \frac{f(x_0)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_k)} + \sum_{r=1}^{k} \frac{[x_r-x_0-(x_r-x_{k+1})]f(x_r)}{(x_r-x_0)\ldots(x_r-x_{k+1})}\right]$$

$$= \frac{f(x_0)}{(x_0-x_1)\ldots(x_0-x_{k+1})} + \sum_{r=1}^{k} \frac{f(x_r)}{(x_r-x_0)\ldots(x_r-x_{k+1})} + \frac{f(x_{k+1})}{(x_{k+1}-x_0)\ldots(x_{k+1}-x_k)}$$

$$f(x_0,x_1,\ldots,x_{k+1}) = \sum_{r=0}^{k+1} \frac{f(x_r)}{(x_r-x_0)\ldots(x_r-x_{k+1})}.$$

Thus we see that, if the result is true for n = k, then it is true for n = k+1. Since it is true for n = 1, it is true for all integral n.

We see that the right hand side of Eq. (4.8) remains the same even if we interchange any two of the arguments $x_0, x_1, \ldots, x_n$. Thus the divided differences are symmetric functions of their arguments.

If any two of $x_0, x_1, \ldots, x_n$ happen to be equal, we define the divided differences as follows:

$$f(x_0, x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0).$$

Similarly, $$f(x_0, x_0, x_0) = f''(x_0).$$

4.4.1 DIVIDED-DIFFERENCE TABLE

The first, second and higher divided differences are usually arranged in the form of a table like the one given below:

x	f( )	f(,)	f(, ,)	f(, , ,)
x_0	$f(x_0)$			
		$f(x_0, x_1)$		
x_1	$f(x_1)$		$f(x_0, x_1, x_2)$	
		$f(x_1, x_2)$		$f(x_0, x_1, x_2, x_3)$
x_2	$f(x_2)$		$f(x_1, x_2, x_3)$	
		$f(x_2, x_3)$		$f(x_1, x_2, x_3, x_4)$
x_3	$f(x_3)$		$f(x_2, x_3, x_4)$	
		$f(x_3, x_4)$		
x_4	$f(x_4)$			

Thus if we draw a line starting at $f(x_1)$ and sloping towards the right, it passes through successive divided differences $f(x_1, x_2)$, $f(x_1, x_2, x_3)$ and so on.

EXAMPLE 4.2: Form the divided difference table for the data

x	1	2	4	7	8
f(x)	-9	-41	-189	9	523

SOLUTION: The divided-difference table will be as follows

x	f(x)	f(,)	f(,,)	f(,,,)	f(,,,,)
1	-9				
		-32			
2	-41		-14		
		-74		7	
4	-189		28		1
		66		14	
7	9		112		
		514			
8	523				

From the above table, we can easily write down any difference we want. For example to write down f(4,7,8) we draw a diagonal from f(4) = -189. Thus f(4,7) = 66 and f(4,7,8) = 112.

4.4.2 NEWTON'S INTERPOLATION POLYNOMIAL

From the definition and symmetry of $f(x_0,x_1,x_2)$ etc., we can write

$$f(x) = f(x_0) + (x - x_0)\, f(x,x_0)$$

$$f(x,x_0) = f(x_0,x_1) + (x - x_1)\, f(x,x_0,x_1)$$

$$f(x,x_0,x_1) = f(x_0,x_1,x_2) + (x - x_2)\, f(x,x_0,x_1,x_2)$$

... ...

... ...

... ...

$$f(x,x_0,x_1,\ldots,x_{n-1}) = f(x_0,x_1,\ldots,x_n) + (x - x_n) f(x,x_0,x_1,\ldots,x_n)$$

Multiply the second equation by $(x - x_0)$, the third by $(x - x_0)(x - x_1)$, the fourth by $(x-x_0)(x-x_1)(x-x_2)$ and so on, and add. We get

$$f(x) = f(x_0) + (x-x_0)f(x_0,x_1) + (x-x_0)(x-x_1)f(x_0,x_1,x_2) + \ldots + (x-x_0)(x-x_1)\ldots(x-x_{n-1})f(x,x_0,x_1,\ldots,x_n) + R_n \quad (4.10)$$

where $R_n = (x - x_0)(x - x_1)\ldots(x - x_n)f(x,x_0,x_1,\ldots,x_n)$ (4.11)

Let us write (4.10) as

$$f(x) = P_n(x) + R_n(x)$$

$$P_n = f(x_0)+(x-x_0)f(x_0,x_1)+\ldots+(x-x_0)\ldots(x-x_{n-1})f(x,x_0,x_1,\ldots,x_n) \quad (4.12)$$

It is clear that $P_n(x)$ is a polynomial of degree at most n, also $R_n(x_r) = 0$ for r=0,1,2,...,n. Hence we see that $P_n(x_r) = f(x_r)$, r=0,1,2,...,n. Therefore $P_n(x)$ is a polynomial of degree at most n that passes through the n + 1 given points $(x_r, f(x_r))$, r = 0, 1,...,n. We have already seen that such a polynomial is unique. Thus $R_n(x)$ must be identical with the error term found in (4.6)

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} F(x). \quad (4.13)$$

where, as before,

$$F(x) = (x - x_0)(x - x_1)\ldots(x - x_n).$$

From (4.11) and (4.13) we also see that

$$f(x,x_0,x_1,\ldots,x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

The polynomial $P_n(x)$ is called **Newton's interpolation polynomial** with divided differences. We emphasize the point that the Lagrangian and Newton's polynomials are identical. However calculations in the case of Newton's polynomial are simpler. Also x_r (r=1,2,...,n) does not appear in any of the differences which involve $x_0, x_1, \ldots,$ upto x_{r-1}. Suppose we have found the Newton's

polynomial $P_n(x)$ which passes through $n + 1$ given points and now wish to add a new point $(x_{n+1}, f(x_{n+1}))$ to the data. All we have to do is to calculate

$$(x-x_0)(x-x_1)\ldots(x-x_n)\, f(x_0,x_1,\ldots,x_{n+1})$$

and add this to $P_n(x)$:

$$P_{n+1}(x) = P_n(x) + (x - x_0)(x - x_1)\ldots(x - x_n)f(x_0,x_1,\ldots,x_{n+1}). \quad (4.12)$$

However, for the Lagrangian polynomial we have to calculate every term afresh and the labour is increased manyfold.

EXAMPLE 4.3: Find a 4th degree polynomial which passes through the 5 points given below.

x	f(x)
1.0	- 9
2.0	- 41
4.0	-189
7.0	9
8.0	523

SOLUTION: The table of divided differences is as follows:

x	f(x)	f(,)	f(,,)	f(,,)	f(,,,,)
1.0	-9				
		-32			
2.0	-41		-14		
		-74		7	
4.0	-189		28		1
		66		14	
7.0	9		112		
		514			
8.0	523				

$$P_4(x) = f(x_0)+(x-x_0)f(x_0,x_1)+(x-x_0)(x-x_1)f(x_0,x_1,x_2)+\ldots.$$
$$= -9+(x-1)(-32)+(x-1)(x-2)(-14)+(x-1)(x-2)(x-4)\ (7)$$
$$+(x-1)(x-2)(x-4)(x-7)\ (1)$$
$$= -9-32(x-1)-14(x^2-3x+2)+7(x^3-7x^2+14x-8)$$
$$+(x^4-14x^3+63x^2-106x+56)$$
$$= -5 + 2x + 0x^2 - 7x^3 + x^4.$$

So, $P_4(x) = x^4 - 7x^3 + 2x - 5.$

Now if we add a point (5,-173) to the data and ask for the polynomial which involves the 6 given points, all we have to do is to calculate one line in the divided difference table and add one term to the already calculated $P_4(x)$. The last line in the new table is

5 -173 232 141 29 5 1

Hence the new polynomial will be

$$P_5(x) = P_4(x) + (x - 1)(x - 2)(x - 4)(x - 7)(x - 8)(1)$$
$$= x^4-7x^3+2x-5+(x^5-22x^4+175x^3-610x^2+904x-448)$$
$$= x^5 - 21x^4 + 168x^3 - 610x^2 + 906x - 453.$$

EXERCISE 4

1. A function $y = f(x)$ is given by the following table.

x	0	2.5069	5.0154	7.52270
f(x)	0.3989423	0.3988169	0.3984408	0.3978138

a) Construct the divided differences table for the above data.
b) Find f(3.7608) using Newton's interpolation polynomial.

2. The following table gives the function $y = f(x)$. Calculate f(323.5) by Newton's interpolation polynomial.

x	321.0	322.8	324.2	325.0
f(x)	2.50651	2.50893	2.51081	2.51188

Compare your result with the exact value if $f(x) = \log_{10} x$.

3. The relation between steam pressure P and temperature T is given by the following table. Evaluate the pressure at temperature 372.1^0.

T	361^0	367^0	378^0	387^0	399^0
P	154.9	167.0	191.0	212.5	244.2

4. Compute the value of $\log_{10} 323.5$ by Lagrange's formula taking the data of problem 2. Compare your result with those obtained previously.

5. Construct the Lagrange interpolation polynomial for the function $y = \sin \pi x$, choosing the points

$$x_0 = 0,\ x_1 = 1/6,\ x_2 = 1/2.$$

6. Find the second degree Lagrange interpolating polynomial passing through the three points in the following table.

i	x_i	$f(x_i)$
0	0	-5
1	1	1
2	3	25

7. Use appropriate Newton's interpolating polynomials of degree one, two, and three to approximate f(2.5) if

f(2.0) = .5103757 f(2.6) = .4813306
f(2.2) = .5207843 f(2.8) = .4359160
f(2.4) = .5104147

8. Use the Newton's interpolating polynomial of degree 3 or less to approximate cos(.750) using the values given below.

cos(.698) = .7661 cos(.768) = .7193
cos(.733) = .7432 cos(.803) = .6946.