UNIT -1: ERROR ANALYSIS AND NUMERICAL COMPUTATIONS

1.1 Learning Objectives
1.2 Introduction
1.3 Accuracy of Numbers
1.3.1 Approximate Numbers
1.3.2 Significant Figures
1.3.3 Rounding of Numbers
1.4 Errors
1.5 A General Formula for Errors
1.6 Computer Representation of Numbers
1.7 Floating Point Arithmetic
1.8 Propagation of errors
1.9 Illustrative Examples
1.10 Let Us Sum Up
1.11 Answers to Check Your Progress
1.12 Further Reading
1.13 Model Questions

1.1 LEARNING OBJECTIVES

After studying this unit, you should be able to:

- Understand different types of errors in numerical computations,
- Distinguish between accuracy and precision of numbers,
- Calculate significant figures and round off numbers,
- Explain the computer representation of numbers,
- Perform floating-point arithmetic operations,
- Describe how errors propagate in numerical computations.

1.2 INTRODUCTION

Numerical computations are a fundamental part of scientific and engineering analysis. Computers are extensively used to perform these calculations. However, due to limitations in representing real numbers and executing arithmetic operations, the results obtained from numerical computations may contain errors. These errors may arise due to approximations, rounding off, or truncation of numbers. This unit aims to provide a

comprehensive understanding of such errors, their types, sources, and the way they affect numerical computations.

1.3 ACCURACY OF NUMBERS

1.1.1 Approximate numbers

There are two types of numbers: exact and approximate. Exact numbers are 2, 4, 9, 13, 7/2, 1.25, … etc. But there are numbers such as 4/3 (= 1.33333...), $\sqrt{2}$ (= 1.414213...) and π (= 3.141592...) which cannot be expressed by a finite number of digits. These may be approximated by numbers 1.3333, 1.4142, and 3.1416, respectively. Such numbers, which represent the given numbers to a certain degree of accuracy, are called approximate numbers.

1.1.2 Significant Figures

The digits used to express a number are called significant digits (figures). Thus, each of the numbers 7845, 3.589, and 0.4758 contains four significant figures, while the numbers 0.00386, 0.000587, and 0.0000296 contain only three significant figures since zeros only help to fix the position of the decimal point. Similarly, the numbers 45000 and 7300.00 have two significant figures only.

1.1.3 Rounding of Numbers

There are numbers with large number of digits, e.g., 22/7 = 3.142857143. In practice, it is desirable to limit such numbers to a manageable number of digits, such as 3.14 or 3.143. This process of dropping unwanted digits is called rounding off.

Rule to round off a number to n significant figures:

(a) Discard all digits to the right of the nth digit.
(b) If this discarded number is
  (i) less than half a unit in the nth place, leave the nth digit unchanged;
  (ii) greater than half a unit in the nth place, increase the nth digit by unity;
  (iii) exactly half a unit in the nth place, increase the nth digit by unity if it is odd, otherwise leave it unchanged.

For example, the following numbers rounded off to three significant figures are:

7.893 to 7.89
12.865 to 12.9
6.4356 to 6.44
3.567 to 3.57
84767 to 84800
5.8254 to 5.82

Note: The numbers thus rounded off to n significant figures (or n decimal places) are said to be correct to n significant figures (or n decimal places).

1.4 ERRORS

We find the following kinds of errors in any numerical calculation:

1. **Inherent Error:** Errors that are already present in the statement of a problem before its solution, are called inherent errors. Such errors arise either because the given data is approximate or due to the limitations of mathematical tables, calculators, or digital computers. Inherent errors can be minimized by taking better data or by using high-precision computing aids.
2. **Rounding Errors:** Rounding errors occur when numbers are approximated during calculations, typically due to the limitations of computing tools. While such errors are often unavoidable, especially in digital computations, they can be minimized as follows:
   (i) by changing the calculation procedure to avoid subtraction of nearly equal numbers or division by a small number; or
   (ii) by retaining at least one more significant figure at each step than that given in the data and rounding off at the last step.
3. **Truncation Errors**: Truncation errors are caused by using approximate results or by replacing an infinite process with a finite one. If we are using a decimal computer having a fixed word length of four digits, rounding off 13.658 gives 13.66 whereas truncation gives 13.65.

   For example, if $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots \infty = X$ (say) is replaced by

   $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} = X'$ (say), then the truncation error is $X - X'$.

   Truncation error is a type of algorithm error.
4. **Absolute, Relative, and Percentage Errors:** If X is the true value of a quantity and X' is its approximate value, then $|X - X'|$ i.e., $|\text{Error}|$ is called the absolute error (E_a).

   The relative error is defined by, $E_r = \left|\frac{X - X'}{X}\right|$ i.e., $\left|\frac{\text{Error}}{\text{True value}}\right|$

   and the percentage error is $E_p = 100E_r = 100\left|\frac{X - X'}{X}\right|$.

   If $\overline{X}$ be such a number that $|X - X'| \le \overline{X}$, then $\overline{X}$ is an upper limit on the magnitude of absolute error and measures the absolute accuracy.

Note:

1. The relative and percentage errors are independent of the units used, while the absolute error is expressed in terms of these units.
2. If a number is correct to n decimal places, then the error $=\frac{1}{2}10^{-n}$. For example, if the number is 3.1416 correct to 4 decimal places, then the error $=\frac{1}{2}10^{-4}=0.00005$.

1.5 A GENERAL FORMULA FOR ERRORS

The following rules help estimate the errors that arise when numbers in a calculation are truncated or rounded to a specific number of digits.

If the approximate value of a number X having n decimal digits is X', then

1. Absolute error due to truncation to k digits $=|X-X'|<10^{n-k}$
2. Absolute error due to rounding off to k digits $=|X-X'|<\frac{1}{2}10^{n-k}$
3. Relative error due to truncation to k digits $=\left|\frac{X-X'}{X}\right|<10^{1-k}$
4. Relative error due to rounding off to k digits $=\left|\frac{X-X'}{X}\right|<\frac{1}{2}10^{1-k}$.

Note:

(i) If a number is correct to n significant digits, then the maximum relative error $\leq\frac{1}{2}10^{-n}$. If a number is correct to d decimal places, then the absolute error $\leq\frac{1}{2}10^{-d}$.

(ii) If the first significant figure of a number is k and the number is correct to n significant figures, then the relative error $<1/\left(k\times10^{n-1}\right)$.

Example 1.1: Round off the numbers

(i) 865250

(ii) 37.46235

to four significant figures and compute E_a, E_r, E_p in each case.

Solution:

(i) The number 865250, rounded off to four significant figures, is 865200.

$$E_a=|X-X'|=|865250-865200|=50$$

$$E_r=\left|\frac{X-X'}{X}\right|=\left|\frac{865250-865200}{865250}\right|=\frac{50}{865250}=5.78\times10^{-5}$$

$$E_p=E_r\times100=5.78\times10^{-3}.$$

(ii) Same as (i).

1.6 COMPUTER REPRESENTATION OF NUMBERS

The basic arithmetic operations performed by the computer are addition, subtraction, multiplication, and division. The decimal numbers are first converted to the machine numbers consisting only of the digits 0 and 1 with a base or radix depending on the computer. If the base is two, eight, or sixteen, the number system is called binary, octal, or hexadecimal, respectively. The decimal number system has the base of 10. For example, in the decimal system,

$(4987)_{10} = 4\times10^3 + 9\times10^2 + 8\times10 + 7\times10^0$, which represents a polynomial in the base 10.

$(0.6251)_{10} = 6\times10^{-1} + 2\times10^{-2} + 5\times10^{-3} + 1\times10^{-4}$, which represents a polynomial in 10^{-1} .

$(4987.6251)_{10} = 4\times10^3 + 9\times10^2 + 8\times10^1 + 7\times10^0 + 6\times10^{-1} + 2\times10^{-2} + 5\times10^{-3} + 1\times10^{-4}$.

Thus, a number $(N)_{10} = d_{n-1}10^{n-1} + d_{n-2}10^{n-2} + ... + d_1 10^1 + d_0 10^0 + d_{-1}10^{-1} + d_{-2}10^{-2} + ... + d_{-m}10^{-m}$, where $d_{n-1}, d_{n-2}, ..., d_{-m}$ are any digits between 0 and 9.

Binary Number System:

The binary number system has base 2 with digits 0 and 1 called bits, and any number N can be written as $(N)_2 = b_{n-1}b_{n-2}\ ...\ b_1 b_0 b_{-1} b_{-2}\ ...\ b_{-m}$, where $b_{n-1}, b_{n-2}, \ ... \ , b_{-m}$ are binary bits 0 or 1, and the point is called the binary point. The corresponding decimal number is easily calculated by using the formula

$(N)_{10} = b_{n-1}2^{n-1} + b_{n-2}2^{n-2} + ... + b_1 2^{-1} + b_0 2^0 + b_{-1}2^{-1} + b_{-2}2^{-2} + ... \ b_{-m}2^{-m}$.

Octal and Hexadecimal System:

The octal system has a base of 8 and uses the digits 0, 1, 2, 3, 5, 6, 7. Similarly, the hexadecimal system has base 16 and uses the digits 0 to 9 and A, B, C, D, E, F to represent 10, 11, 12, 13, 14, 15, respectively.

1.7 FLOATING POINT ARITHMETIC

The first step in using digital computers for calculations is to change decimal numbers into another number system (with a base like β) that the computer can understand, and then store these numbers in its memory. The memory of the digital computer is divided into separate cells called words. Each word can hold the same number of digits called bits, with respect to its base plus a sign. Negative numbers are stored as absolute values

plus a sign or in complement form. The number of digits which can be stored in a computer word is called its word length. The word length varies from one computer to another. The numbers in the computer word can be stored in two forms:

(a) Fixed-point form.
(b) Floating-point form.

In fixed-point form, a *t* digit number is assumed to have its decimal point at the left-hand end of the word. This implies that all numbers are supposed to be less than 1 in magnitude. The fixed-point number with base β and *t* digits word length may be written as $\pm\sum_{k=1}^{t} a_k \beta^{-k}$, where $0 \le a_k \le \beta$.

If *y* is any real number and y^* is its machine representation, then the error in y^* is at most $a_{t+1}\beta^{-(t+1)}$ or $\left|y - y^*\right| \le \beta^{-t}$.

To overcome the challenge of maintaining all numbers below 1 in magnitude during calculations, most computers represent real numbers using floating-point notation. A floating-point number is characterized by four parameters: the base β, the number of digits *t*, and the exponent range (*m*, *M*).

***Definition* 1.1:** A floating-point number is a number represented in the form

$$.d_1 d_2 \ldots d_t \times \beta^e, \tag{4.1}$$

where $d_1, d_2, \ldots, d_t$ are integers and satisfy $0 \le d_i \le \beta$ and the exponent *e* is such that $m \le e \le M$.

The fractional part $.d_1 d_2 \ldots d_t$ is called the mantissa and it lies between +1 and -1. The number 0 is written as

$$+.00 \ldots 0 \times \beta^e. \tag{4.2}$$

***Definition* 1.2:** A non-zero floating-point number as defined in (4.1) is in normal form if the value of the mantissa lies in the interval $(-1, -1/\beta)$ or in the interval $[1/\beta, 1)$.

***Definition* 1.3:** A non-zero floating-point number as defined in (4.1) is in t-digit-mantissa standard form if it is normalized and its mantissa consist of exactly *t*-digits.

If a number *x* has the representation in the form

$$x = .d_1 d_2 \ldots d_t d_{t+1} \ldots \times \beta^e \tag{4.3}$$

then the floating-point number $fl(x)$ in *t*-digit-mantissa standard form can be obtained in the following two ways.

(a) Chopping: Here, we neglect $d_{t+1}, d_{t+2}, \ldots$ in (4.3) and obtain $fl(x) = .d_1 d_2 \ldots d_t \times \beta^e$.

(b) Rounding: Here, the fractional part in (4.3) is written as $.d_1 d_2 \ldots d_t d_{t+1} + \frac{1}{2}\beta$

and the first t digits are taken to write the floating-point number.

Example 1.2: Find the sum of $.123\times10^3$ and $.456\times10^2$ and write the results in three-digit mantissa form.

Solution: We have $.1230\times10^3 + .0456\times10^3 = .1686\times10^3 = \begin{cases} .168\times10^3, \text{ for chopping,} \\ .169\times10^3, \text{ for rounding.} \end{cases}$

1.8 PROPAGATION OF ERRORS:

Several computational steps are performed to solve a problem. It is necessary to understand the way the error propagates with progressive computation.

If the approximate values of two numbers X and Y be X' and Y', respectively, then the absolute error

$E_{ax} = X - X'$ and $E_{ay} = Y - Y'$.

1. **Absolute error in the addition operation:**
   $X + Y = (X' + E_{ax}) + (Y' + E_{ay}) = X' + Y' + E_{ax} + E_{ay}$
   So, $|(X+Y) - (X'+Y')| = |E_{ax} + E_{ay}| \le |E_{ax}| + |E_{ay}|$
   Thus, the absolute error in taking $(X'+Y')$ as an approximation to $(X+Y)$ is less than or equal to the sum of the absolute errors in taking X' as an approximation to X and Y' as an approximation to Y.
2. **Absolute error in subtraction operation:**
   $X - Y = (X' + E_{ax}) - (Y' + E_{ay}) = X' - Y' + E_{ax} - E_{ay}$
   So, $|(X-Y) - (X'-Y')| = |E_{ax} - E_{ay}| \le |E_{ax}| + |E_{ay}|$
   Thus, the absolute error in taking $(X'-Y')$ as an approximation to $(X-Y)$ is less than or equal to the sum of the absolute errors in taking X' as an approximation to X and Y' as an approximation to Y.
3. **Absolute error in multiplication operation:**
   To find the absolute error E_a in the product of two numbers X and Y, we write
   $E_a = (X + E_{ax})(Y + E_{ay}) - XY$

where E_{ax} and E_{ay} are the absolute errors in X and Y, respectively. Then

$$E_a = XE_{ay} + YE_{ax} + E_{ax}E_{ay}$$

Assuming E_{ax} and E_{ay} are reasonably small so that $E_{ax}E_{ay}$ can be ignored. Thus, $E_a = XE_{ay} + YE_{ax}$ approximately.

4. **Absolute error in division operation:**
Similarly, the absolute error E_a in the quotient of two numbers X and Y is given by

$$E_a = \frac{X + E_{ax}}{Y + E_{ay}} - \frac{X}{Y} = \frac{YE_{ax} - XE_{ay}}{Y\left(Y + E_{ay}\right)} = \frac{YE_{ax} - XE_{ay}}{Y^2\left(1 + E_{ay}/Y\right)} = \frac{YE_{ax} - XE_{ay}}{Y^2},$$ assuming E_{ay}/Y to be small.

Thus, $E_a = \frac{X}{Y}\left(\frac{E_{ax}}{X} - \frac{E_{ay}}{Y}\right)$.

Error in the Approximation of a Function

Let $y = f\left(x_1, x_2\right)$ be a function of two variables x_1, x_2. If $\delta x_1, \delta x_2$ be the errors in x_1, x_2, then the error δy in y is given by

$$y + \delta y = f\left(x_1 + \delta x_1, x_2 + \delta x_2\right)$$

Expanding the right-hand side by Taylor's series, we get

$$y + \delta y = f\left(x_1, x_2\right) + \left(\frac{\partial f}{\partial x_1}\delta x_1 + \frac{\partial f}{\partial x_2}\delta x_2\right) + \text{terms involving higher powers of } \delta x_1 \text{ and } \delta x_2. \quad \text{(i)}$$

If the errors δx_1 and δx_2 are so small that their squares and higher powers can be neglected, then (i) gives

$$\delta y = \frac{\partial f}{\partial x_1}\delta x_1 + \frac{\partial f}{\partial x_2}\delta x_2 \text{ approximately.}$$

Hence, $\delta y \approx \frac{\partial f}{\partial x_1}\delta x_1 + \frac{\partial f}{\partial x_2}\delta x_2$.

In general, the error δy in the function $y = f\left(x_1, x_2, \ldots, x_n\right)$ corresponding to the errors δx_i in $x_i\left(i = 1, 2, \ldots, n\right)$ is given by

$$\delta y \approx \frac{\partial f}{\partial x_1}\delta x_1 + \frac{\partial f}{\partial x_2}\delta x_2 + \ \ldots\ + \frac{\partial f}{\partial x_n}\delta x_n$$

and the relative error in y is $E_r = \frac{\delta y}{y} \approx \frac{\partial y}{\partial x_1}\frac{\delta x_1}{y} + \frac{\partial y}{\partial x_2}\frac{\delta x_2}{y} + \ \ldots\ + \frac{\partial y}{\partial x_n}\frac{\delta x_n}{y}$.

1.9 ILLUSTRATIVE EXAMPLES:

Example 1.9.1: Find the absolute error and relative error in $\sqrt{6}+\sqrt{7}+\sqrt{8}$ correct to 4 significant digits.

Solution:

We have $\sqrt{6} = 2.449,\ \sqrt{7} = 2.646,\ \sqrt{8} = 2.828$

Therefore, $S = \sqrt{6}+\sqrt{7}+\sqrt{8} = 7.923$

Then, the absolute error E_a in S is

$E_a = 0.0005 + 0.0007 + 0.0004 = 0.0016$.

Hence, S is correct to 3 significant digits only. Thus, we have $S = 7.92$ and the relative error E_r is

$$E_r = \frac{0.0016}{7.92} = 0.0002.$$

Example 1.9.2: Find the product of the numbers 3.7 and 52.378, both of which are correct to the given significant digits.

Solution: Since the absolute error is greatest in 3.7, we round off the other number to 3 significant figures, i.e., 52.4.

The product, $P = 3.7 \times 52.4 = 193.88 = 1.9388 \times 10^2$.

Since the first number contains only two significant figures, therefore retain only two significant figures in the product and we get, $P = 1.9 \times 10^2$.

CHECK YOUR PROGRESS

(a) Round off the following numbers correct to four significant figures: 3.26425, 35.46735, 4985561, 0.70035, 0.00032217, and 18.265101.

(b) Round off the number 75462 to four significant digits and then calculate the absolute error and percentage error.

1.10 LET US SUM UP

- Discussed various types of numerical errors in computations, including: Inherent errors, Rounding errors, and Truncation errors
- Introduced the concepts of: Significant figures, rounding rules, and Error estimates
- Explained how numbers are represented in digital computers using binary, octal, and hexadecimal number systems
- Covered floating-point arithmetic and its operations in detail
- Highlighted the importance of understanding:
  - Error propagation in computational procedures
  - Approximation of functions, supported with illustrative examples

1.11 ANSWERS TO CHECK YOUR PROGRESS

(a) Try yourself.
(b) Try yourself.

1.12 FURTHER READING

1. Jain, M. K., Iyengar, S. R. K. and Jain, R. K. Numerical Methods for Scientific & Engineering Computation. New Age International, 2012.
2. Gupta, R. K. Numerical Methods: Fundamentals and Applications. 1st edition, Cambridge University Press, 2019.

1.13 MODEL QUESTIONS

1. Find the absolute error if the number X= 0.00545828 is
   (i) Truncated to three decimal digits.
   (ii) Rounded off to three decimal digits.
2. Find the relative error if the number X=0.004997 is
   (i) Truncated to three decimal digits.
   (ii) Rounded off to three decimal digits.
3. Find the percentage error if 625.483 is approximated to three significant figures.
4. Find the relative error in taking n = 3.141593 as 22/7.
5. If 0.333 is the approximate value of 1/3, find the absolute and relative errors.
6. The area of cross-section of a rod is desired up to 0.2% error. How accurately should the diameter be measured?

7. If $u = 4x^2y^3/z^4$ and errors in x, y, z are 0.001, compute the relative maximum error in u when $x = y = z = 1$.
8. Find the relative error in the function $y = ax_1^{m_1} x_2^{m_2} \ldots x_n^{m_n}$.

*** *** ***