UNIT 1 FLOATING POINT ARITHMETIC AND ERRORS

Structure **Page Nos.**

1.0 INTRODUCTION

Numerical Analysis is the study of computational methods for solving scientific and engineering problems by using basic arithmetic operations such as addition, subtraction, multiplication and division. The results obtained by using such methods, are usually approximations to the true solutions. These approximations to the true solutions introduce errors but can be made more accurate up to some extent. There can be several reasons behind this approximation, such as the formula or method used to solve a problem may not be exact. i.e., the expression of $\sin x$ can be evaluated by expressing it as an infinite power series. This series has to be truncated to the finite number of terms. This truncation introduces an error in the computed result. As a student of computer science you should also consider the computer oriented aspect of this concept of approximation and errors, say the machine involved in the computation doesn't have the capacity to accommodate the data or result produced by calculation of a numerical problem and hence the data is to be approximated in to the limitations of the machine. When this approximated data is to be further utilized in successive calculations, then it causes the propagation of error, and if the error starts growing abnormally then some big disasters may happen. Let me cite some of the well-known disasters caused because of the approximations and errors.

Instance 1: On February 25, 1991, during the Gulf War, an American Patriot Missile battery in Dhahran, Saudi Arabia, failed to intercept an incoming Iraqi Scud Missile. The Scud struck an American Army barracks and killed 28 soldiers. A report of the General Accounting office, GAO/IMTEC-92-26, entitled *Patriot Missile Defense: Software Problem Led to System Failure at Dhahran, Saudi Arabia* reported on the cause of the failure. It turns out that the cause was an inaccurate calculation of the time since boot due to computer arithmetic errors.

Instance 2: On June 4, 1996, an unmanned Ariane 5 rocket launched by the European Space Agency exploded just forty seconds after lift-off. The rocket was on its first voyage, after a decade of development costing $7 billion. A board of inquiry investigated the causes of the explosion and in two weeks issued a report. It turned out that the cause of the failure was a software error in the inertial reference system.

Specifically, a 64-bit floating point number relating to the horizontal velocity of the rocket with respect to the platform was converted to a 16-bit signed integer. The number was larger than 32,768, the largest integer storeable in a 16-bit signed integer, and thus the conversion failed.

In this Unit, we will describe the concept of number approximation, significant digits, the way, the numbers are expressed and arithmetic operations are performed on them, types of errors and their sources, propagation of errors in successive operations etc. The *Figure 1* describes the stages of Numerical Computing.

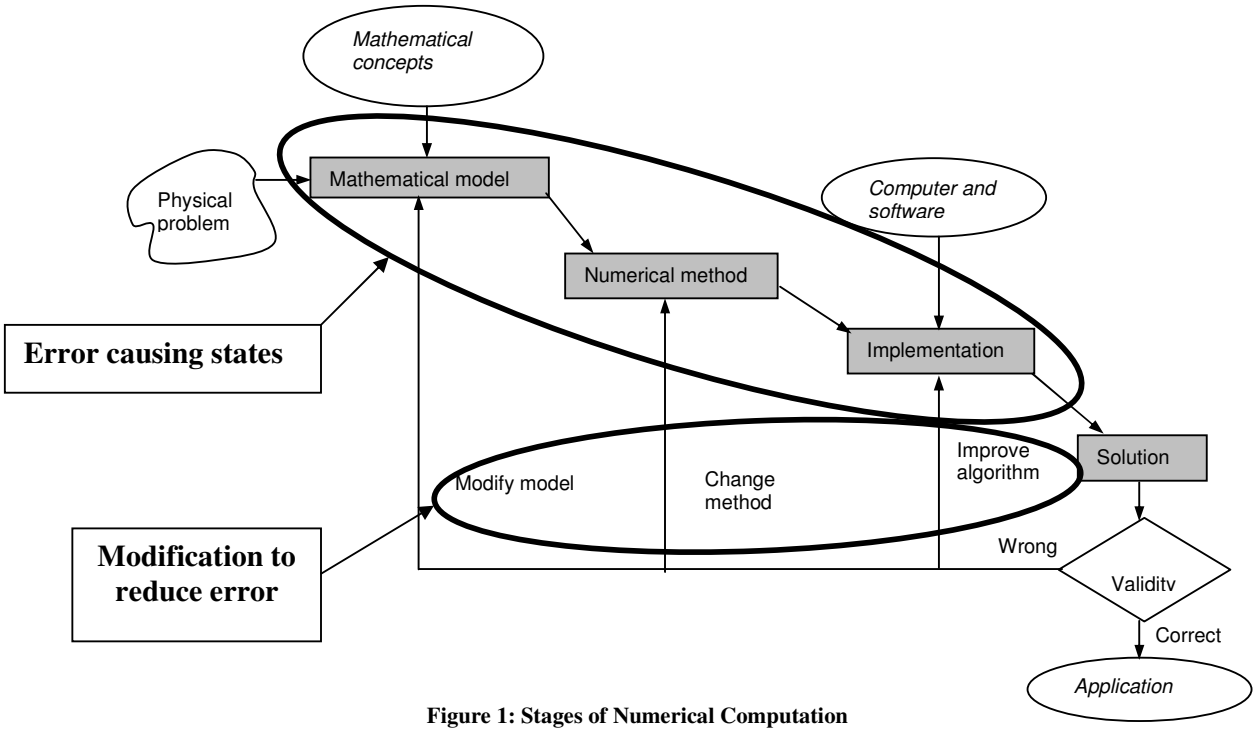


Figure 1: Stages of Numerical Computation

1.1 OBJECTIVES

After studying this unit, you should be able to:

- describe the concept of fixed point and floating point numbers representations;
- discuss rounding-off errors and the rules associated with round-off errors;
- implement floating-point arithmetic on available data;
- conceptual description of significant digits, and
- analysis of different types of errors – absolute error, relative errors, truncation error.

1.2 FLOATING POINT REPERESENTATIONS

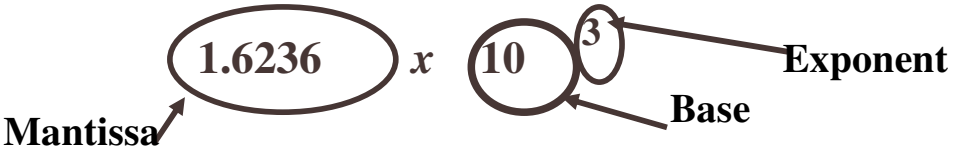
In scientific calculations, very large numbers such as velocity of light or very small numbers such as size of an electron occur frequently. These numbers cannot be satisfactorily represented in the usual manner. Therefore, scientific calculations are usually done by floating point arithmetic.

This means that we need to have two formats to represent a number, which are fixed point representation and floating point representation. We can transform data of one

format in to another and *vice versa*. The concept of transforming fixed point data into floating point data is known as normalisation, and it is done to preserve the maximum number of useful information carrying digits of numbers. This transformation ultimately leads to the calculation errors. Then, you may ask what is the benefit of doing this normalisation when it is contributing to erroneous results. The answer is simply to proceed with the calculations keeping in mind the data and calculation processing limitation of machine.

Fixed-Point numbers are represented by a fixed number of decimal places. Examples are 62.358, 1.001, 0.007 all correctly expressed up to 3rd decimal place.

Floating-Point numbers have a fixed number of significant places. Examples are 6.236×10^3 1.306×10^{-3} which are all given as four significant figures. The position of the decimal point is determined by the powers of base (in decimal number system it is 10) 1.6236×10^3.



Let us first discuss what is a floating-point number. Consider the number 123. It can be written using exponential notation as:

1.23×10^2, 12.3×10^2, 123×10^2, 0.123×10^2, 1230×10^2, etc.

Notice how the decimal point "floats" within the number as the exponent is changed. This phenomenon gives floating point numbers their name. The representations of the number 123 above are in kind of standard form. The first representation, 1.23×10^2, is in a form called "scientific notation".

In scientific computation, a real number x is usually represented in the form

$$x = \pm(d_1 d_2 .....d_n) x 10^m \qquad (1)$$

where $d_1, d_2, ...d_n$ are decimal digits and m is an integer called **exponent.** $(d_1 d_2 .....d_n)$ is called **significand** or **mantissa**. We denote this representation by **fl(x)**. A floating-point number is called a nomalised floating-point number if $d_1 \neq 0$ or else $d_2 = d_3 = \ldots. = d_n = 0$. The exponent m is usually bounded in a range

$$-M < m < M \qquad (2)$$

In scientific notation, such as 1.23×10^2 in the above example, the significand is always a number greater than or equal to 1 and less than 10. We may also write 1.23E2.

Standard computer normalisation for floating point numbers follows the fourth form namely, 0.123×10^3 in the list above.

In the standard normalized floating-point numbers, **the significand is greater than or equal to 0.1, and is always less than 1.**

In floating point notation (1), if $fl(x) \neq 0$ and $m \geq M$ (that is, the number becomes too large and it cannot be accommodated), then x is called an **over-flow number** and if

$m \leq -M$ (that is the number is too small but not zero) the number is called an **under-flow number.** The number n in the floating-point notation is called its **precision**.

1.2.1 Floating Point Arithmetic

When arithmetic operations are applied on floating-point numbers, the results usually are not floating-point numbers of the same length. For example, consider an operation with 2 digit precision floating-point numbers (i.e., those numbers which are accurate up to two decimal places) and suppose the result has to be in 2 digit floating point precision. Consider the following example,

$x = 0.30x10^{1}, \quad y = 0.66x10^{-6}, \quad z = 0.10x10^{1}$

then
$$\begin{aligned} x + y &= 0.300000066 \times 10^{1} = 0.30 \times 10^{1} \\ x \times y &= 0.198 \times 10^{-5} \quad = 0 \\ z/x &= 0.333\ldots \times 10^{0} \quad = 0.33 \times 10^{0} \end{aligned} \tag{3}$$

Hence, if θ is one of the arithmetic operations, and θ^* is corresponding floating-point operation, then we find that

$$x\ \theta^*\ y \neq x\ \theta\ y$$

However, $x\ \theta\ y = fl(x\ \theta\ y)$ (4)

1.2.2 Properties of Floating Point Arithmetic

Arithmetic using the floating-point number system has two important properties that differ from those of arithmetic using real numbers.

Floating point arithmetic is not associative. This means that in general, for floating point numbers x, y, and z:

- $(x + y) + z \neq x + (y + z)$
- $(x \,.\, y) \,.\, z \neq x \,.\, (y \,.\, z)$

Floating point arithmetic is also not distributive. This means that in general,

- $x \,.\, (y + z) \neq (x \,.\, y) + (x \,.\, z)$

Therefore, the order in which operations are carried out can change the output of a floating-point calculation. This is important in numerical analysis since two mathematically equivalent formulas may not produce the same numerical output, and one may be substantially more accurate than the other.

Example 1: Let $a = 0.345 \times 10^{0}$, $b = 0.245 \times 10^{-3}$ and $c = 0.432 \times 10^{-3}$. Using 3-digit decimal arithmetic with rounding, we have

$$\begin{aligned} b + c &= 0.000245 + 0.000432 = 0.000677 \text{ (in accumulator)} \\ &= 0.677 \times 10^{-3} \end{aligned}$$

$$\begin{aligned} a + (b + c) &= 0.345 + 0.000677 \text{ (in accumulator)} \\ &= 0.346 \times 10^{0} \text{ (in memory) with rounding} \end{aligned}$$

$$\begin{aligned} a + b &= 0.345 \times 10^{0} + 0.245 \times 10^{-3} \\ &= 0.345 \times 10^{0} \text{ (in memory)} \end{aligned}$$

$$\begin{aligned} (a + b) + c &= 0.345432 \text{ (in accumulator)} \\ &= 0.345 \times 10^{0} \text{ (in memory)} \end{aligned}$$

Hence, we see that,
$(a + b) + c \neq a + (b + c)$.

Example 2: Suppose that in floating point notation (1) given above, n = 2 and m = 11. Consider $x = 0.10x10^{10}$, $y = -0.10x10^{10}$ and $z = 0.10x10^{1}$. Then,

$(x + y) + z = 0.1x10^{1}$ while $x + (y + z) = 0.0$.

Hence, $(x + y) + z \neq x + (y + z)$.

From the above examples, we note that in a computational process, every floating-point operation gives rise to some error, which may then get amplified or reduced in subsequent operations.

Check Your Progress 1

1) Let a = 0.41, b = 0.36 and c = 0.70. Prove $\frac{(a-b)}{c} \neq \frac{a}{c} - \frac{b}{c}$.

……

……

……

2) Let a = .5665E1, b = .5556E – 1, c = .5644E1. Verify the associative property for the floating point numbers i.e., prove (a + b) – c # (a – c) + b.

……

……

……

3) Let a = .5555E1, b = .4545E1, c = .4535E1. Verify the distributive property for these floating point numbers, i.e., prove a(b – c) # ab –ac.

……

……

……

1.2.3 Significant Digits

The concept of significant digits has been introduced primarily to indicate the accuracy of a numerical value. For example, if, in the number y = 23.40657, only the digits 23406 are correct, then we may say that y has given significant digits and is correct to only three decimal places.

The number of significant digits in an answer in a calculation depends on the number of significant digits in the given data, as discussed in the rules below.

When are Digits Significant?

Non-zero digits are always significant. Thus, 22 has two significant digits, and 22.3 has three significant digits. The following rules are applied when zeros are encountered in the numbers,

a) Zeros placed before other digits are not significant; 0.046 has two significant digits.

b) Zeros placed between other digits are always significant; 4009 kg has four significant digits.

c) Zeros placed after other digits but behind a decimal point are significant; 7.90 has three significant digits.

d) Zeros at the end of a number are significant only if they are behind a decimal point as in (c). For example, in the number 8200, it is not clear if the zeros are significant or not. The number of significant digits in 8200 is at least two, but could be three or four. To avoid uncertainty, we use scientific notation to place significant zeros behind a decimal point.

$8.200 * 10^3$ has four significant digits,
$8.20 * 10^3$ has three significant digits,
$8.2 * 10^3$ has two significant digits.

Note: Accuracy and precision are closely related to significant digits. They are related as follows:

1) Accuracy refers to the number of significant digits in a value. For example, the number 57.396 is accurate to five significant digits.

2) Precision refers to the number of decimal positions, i.e. the order of magnitude of the last digit in a value. The number 57.396 has a precision of 0.001 or 10^{-3}.

Example 1: Which of the following numbers has the greatest precision?

a) 4.3201, b) 4.32, c) 4.320106.

Solution:

a) 4.3201 has a precision of 10^{-4}
b) 4.32 has a precision of 10^{-2}
c) 4.320106 has a precision of 10^{-6}

The last number has the greatest precision.

Example 2: What is the accuracy of the following numbers?

a) 95.763, b) 0.008472, c) 0.0456000, d) 36 e) 3600.00.

Solution:

a) This has five significant digits.
b) This has four significant digits. The leading or higher order zeros are only place holders.
c) This has six significant digits.
d) This has two significant digits.
e) This has six significant digits. Note that the zeros were made significant by writing .00 after 3600.

Significant digits in Multiplication, Division, Trigonometry functions, etc.

In a calculation involving multiplication, division, trigonometric functions, etc., the number of significant digits in an answer should equal the least number of significant digits in any one of the numbers being multiplied, divided, etc.

Thus, in evaluating sin(kx), where k = 0.097 m^{-1} (two significant digits) and x = 4.73 m (three significant digits), the answer should have two significant digits.

Note that whole numbers have essentially an unlimited number of significant digits. As an example, if a hairdryer uses 1.2 kW of power, then 2 identical hairdryers use 2.4 kW.

1.2 kW{2 significant digit} *x* 2{unlimited significant digit) = 2.4 kW
{2 significant digit}

Significant digits in Addition and Subtraction

When quantities are being added or subtracted, the number of *decimal places* (not significant digits) in the answer should be the same as the least number of decimal places in any of the numbers being added or subtracted.

Keep one extra digit in Intermediate Answers

When doing multi-step calculations, *keep at least one or more significant digits in intermediate results* than needed in your final answer.

For instance, if a final answer requires two significant digits, then carry at least three significant digits in calculations. If you round-off all your intermediate answers to only two digits, you are discarding the information contained in the third digit, and as a result the *second* digit in your final answer might be incorrect. (This phenomenon is known as "round-off error.")

This truncation process is done either through rounding off or chopping, leading to round off error.

Example 3: Let x = 4.5 be approximated to $x^* = 4.49998$. Then,

$x^x - x = -0.00002$,

$$\frac{|x - x^x|}{x} = 0.0000044 \le 0.000005 \le \frac{1}{2}\ (.00001) = \frac{1}{2}\ 10^{-5} = \frac{1}{2} \times 10^{1-6}$$

Hence, x^* approximates x correct to 6 significant decimal digits.

Wrong way of writing significant digits

1) Writing more digits in an answer (intermediate or final) than justified by the number of digits in the data.

2) Rounding-off, say, to two digits in an intermediate answer, and then writing three digits in the final answer.

Example 4: Expressions for significant digits and scientific notation associated with a floating point number.

Number	**Number of Significant Figures**	**Scientific Notation**	
0.00682	3	$6.82 * 10^{-3}$	Leading zeros are not significant.
1.072	4	$1.072 (* 10^{0})$	Embedded zeros are always significant.
300	1	$3 * 10^{2}$	Trailing zeros are significant only if the decimal point is specified.
300	3	$3.00 * 10^{2}$	
300.0	4	$3.000 * 10^{2}$	

Loss of Significant Digits

One of the most common (and often avoidable) ways of increasing the importance of an error is known as loss of significant digits.

Loss of significant digits in subtraction of two nearly equal numbers:

Subtraction of two nearly equal number gives the relative error

$$r_{x-y} = r_x \frac{x}{x-y} - r_y \frac{y}{x-y}$$

which becomes very large. It has largest value when r_x and r_y are of opposite signs.

Suppose we want to calculate the number z = x – y and x^* and y^* are approximations for x and y respectively, accurate to *r* digits and assume that x and y do not agree in the most left significant digit, then $z^* = x^* - y^*$ is as good an approximation to x – y as x^* and y^* to x and y.

But, if x^* and y^* agree at left most digits (one or more), then the left most digits will cancel and there will be loss of significant digits.

The more the digits on left agrees, the more loss of significant digits. A similar loss in significant digits occurs when a number is divided by a small number (or multiplied by a very large number).

Remark 1: To avoid this loss of significant digits in algebraic expressions, we must rationalise these numbers. If no alternative formulation to avoid the loss of significant digits is possible, then we can carry more significant digits in calculation using floating-point numbers in double precision.

Example 5: Solve the quadratic equation $x^2 + 9.9\,x - 1 = 0$ using two decimal digit arithmetic with rounding.

Solution:

Solving the quadratic equation, we have one of the solutions as

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-9.9 + \sqrt{(9.9)^2 - 4.1.(-1)}}{2}$$

$$= \frac{-9.9 + \sqrt{102}}{2} = \frac{-9.9 + 10}{2} = \frac{0.1}{2} = 0.05$$

while the true solutions are – 10 and 0.1. Now, if we rationalize the expression, we obtain

$$= \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-4ac}{2a(b + \sqrt{b^2 - 4ac})}$$

$$= \frac{-2c}{b + \sqrt{b^2 - 4ac)}} = \frac{2}{9.9 + \sqrt{102}} = \frac{2}{9.9 + 10} = \frac{2}{19.9} = \frac{2}{20} \cong 0.1 \;.\; (0.1000024)$$

which is one of the true solutions.

1.3 ERROR - BASICS

What is Error?

An error is defined as the difference between the actual value and the approximate value obtained from the experimental observation or from numerical computation. Consider that x represents some quantity and xa is an approximation to x, then

Error = actual value – approximate value = x – xa

How errors are generated in computers?

Every calculation has two parts, one is operand and other is operator. Hence, any approximation in either of the two contributes to error. Approximations to operands causes propagated error and approximation to operators causes generated errors. Let us discuss how the philosophy behind these errors is related to computers.

Operand Point of View: Computers need fixed numbers to do processing, which is mostly not available. Hence, we need to transform the output of an operation to a fixed number by performing truncation of series, rounding, chopping etc. This contributes to difference between exact value and approximated value. These errors get further amplified in subsequent calculations as these values and the results produced are further utilized in subsequent calculations. Hence, this error contribution is referred to as **propagated error.**

Operator Point of View: Computers need some operation to be performed on the operands available. Now, the operations that occur in computers are at bit level and complex operations are simplified. There are, hence, small changes in actual operations and operations performed by computer. This difference in operations produces errors in calculations, which get further amplified in subsequent calculations. This error contribution is referred to as **generated error.**

What are the sources of error?

The sources of error can be classified as (i) data input errors, (ii) errors in algorithms and (iii) errors during computations.

Sources of Error?

Data Input	Algorithms	Computations
Input Error: The input information is rarely exact since it comes from experiments and any experiment can give results of only limited accuracy. Moreover, the quantity used can be represented in a computer for only a limited number of digits.	**Algorithmic Errors:** If direct algorithms based on finite sequence of operations are used, errors due to limited steps don't amplify the existing errors. But if algorithms with infinite steps are used, the algorithm has to be stopped after a finite number of steps	**Computational Errors:** Even when elementary operations, such as multiplication or division are used, the number of digits increases greatly so that the number cannot be held fully in register available in a given computer. In such a case, a certain number of digits must be discarded, and this again leads to error

Type of Errors?

We list below the types of errors that are encountered while carrying out numerical calculations to solve a problem.

1) Round off errors arise due to floating point representation of initial data in the machine. Subsequent errors in the solution due to this are called propagated errors.
2) Due to finite digit arithmetic operations, the computer produces generated errors or rounding errors.
3) Error due to finite representation of an inherently infinite process. For example, consider the use of a finite number of terms in the infinite series expansions of Sin x, Cos x or f(x) by Maclaurin's or Taylor Series expression. Such errors are called truncation errors.

Remark 2: Sensitivity of an algorithm for a numerical process used for computing f(x): if small changes in the initial data x lead to large errors in the value of f(x), then the algorithm is called *unstable*.

How error measures accuracy?

The two terms "error" and " accuracy" are inter-related, one measures the other, in the sense less the error is, more the accuracy is and vice versa. In general, the errors which are used for determination of accuracy are categorized as:

a) Absolute error b) Relative error c) Percentage error

Now, we define these errors.

a) **Absolute Error**: Absolute error is the magnitude of the difference between the true value x and the approximate value xa. Therefore, absolute error = | x – xa |.

b) **Relative Error:** Relative error is the ratio of the absolute error and actual value. Therefore, relative error = |x – xa | / x .

c) **Percentage Error:** Percentage error is defined as,
percentage error = 100er= 100 * |x – xa|/x.

Now, we discuss each of the errors defined above, and its propagation in detail.

1.3.1 Rounding-off Error

There are two ways of translating a given real number *x* into floating-point number f(x) – rounding and chopping. For example, suppose we want to represent the number 5562 in the normalized floating point representation. The representations for different values of *n* are as follows:

$$n = 1, \quad fl(5562) = .5 * 10^4 \text{ chopped}$$
$$= .6 * 10^4 \text{ rounded}. \quad (5)$$

$$n = 2, \quad fl(5562) = .55 * 10^4 \text{ chopped}$$
$$= .56 * 10^4 \text{ rounded}. \quad (6)$$

$$n = 3, \quad fl(5562) = .556 * 10^4 \text{ chopped}$$
$$= .556 * 10^4 \text{ rounded}. \quad (7)$$

Rules for rounding-off: Whenever, we want to use only a certain number of digits after the decimal point, then number is rounded-off to that many digits. A number is rounded-off to n places after decimal by seeing (n+1)th place digit $d_{n+1,}$ as follows:

i) If $d_{n+1} < 5$, then it is chopped
ii) If $d_{n+1} > 5$, then $d_n = d_n + 1$
iii) If $d_n = 5$, and d_n is odd then $d_n = d_n + 1$ else the number d_{n+1} is chopped.

The difference between a number x and fl(x) is called the **round-off error**. It is clear that the round-off error decreases when precision increases. The round-off error also depends on the size of x and is therefore represented relative to x as

$$fl(x) = x(1+\delta). \tag{8}$$

It is not difficult to show that

$|\delta| < .5 * 10^{-(n-1)}$ in rounding

while, $-10^{-(n-1)} < \delta \le 0$ in chopping . (9)

Definition 1: Let x be a real number and x* be a real number having non-terminal decimal expansion, then we say x* represents x **rounded to k decimal places** if $|x - x^*| \le \frac{1}{2}10^{-k}$, where k is a positive integer.

Example 6: If p = 3.14159265, then find out to how many decimal places the approximate value of 22/7 is accurate?

Solution: We find that

$$\left|p - \frac{22}{7}\right| = 0.00126449$$

Since, $0.00126449 < 0.005 = \frac{1}{2}10^{-2}$. Hence, k = 2, and we conclude that the approximation is accurate to 2 decimal places or three significant digits.

Check Your Progress 2

1) Round off the following numbers to four significant digits.

(i) 450.92, (ii) 48.3668, (iii) 9.3265, (iv) 8.4155,
(v) 0.80012, (vi) 0.042514, (vii) 0.0049125, (viii) 0.00020215

……

……

……

……

2) Write the following numbers in floating-point form rounded to four significant digits.

(i) 100000, (ii) –0.0022136, (iii) –35.666

……

……

……

3) The numbers 28.483 and 27.984 are both approximate and are correct up to the last digit shown. Compute their difference. Indicate how many significant digits are present in the result and comment.

……

……

……

……

4) Consider the number 2/3. Its floating point representation rounded to 5 decimal places is 0.66667. Find out to how many decimal places the approximate value of 2/3 is accurate?

……

……

……

……

5) Find out to how many decimal places the value 355/133 is accurate as an approximation to π ?

……

……

……

……

1.3.2 Absolute and Relative Errors

We shall now discuss two types of errors that are commonly encountered in numerical computations. You are already familiar with the rounding off error. These rounded-off numbers are approximations of the actual values. In any computational procedure, we make use of these approximate values instead of the true values. How do we measure the goodness of an approximation fl(x) to x ? The simplest measure which naturally comes to our mind is the difference between x and fl(x). This measure is called the **error**. Formally, we define error as a quantity which satisfies the identity

$$x = \text{fl}(x) + e, \tag{10}$$

If error e is considerably small, then we say that fl(x) is a good approximation of x. Error can be positive or negative. We are in general interested in the magnitude or absolute value of the error which is defined as follows

$$|e| = |x - \text{fl}(x)| \tag{11}$$

Sometimes, when the true value x is very large or very small, we prefer to study the error by comparing it with the true value. This is known as relative error and we define this error as

$$\text{relative error} = r_x = \frac{x - f(x)}{x}$$

and

$$\left|\text{relative error}\right| = \left|\frac{x - \text{fl}(x)}{x}\right| = \left|\frac{e}{x}\right| \tag{12}$$

Note that in certain computations, the true value may not be available. In that case, we replace the true value by the computed approximate value in the definition of relative error.

Theorem: If fl(x) is the n-digit floating point representation in base β of a real number x, then r_x the relative error in x, satisfies the following:

i) $|r_x| < \frac{1}{2}\beta^{1-n}$ if rounding is used.

ii) $0 \le |r_x| \le \beta^{1-n}$ if chopping is used.

For proving i), you may use the following:

Case 1. $d_{n+1} < \frac{1}{2}\beta$, *then* $fl(x) = \pm(.d_1d_2\ldots d_n)\beta^e$

$$|x\text{-}fl(x)| = d_{n+1}, d_{n+2}\ldots\beta^{e-n-1}$$

$$\le \frac{1}{2}\beta.\beta^{e-n-1} = \frac{1}{2}\beta^{e-n}$$

Case 2. $d_{n+1} \ge \frac{1}{2}\beta$,

$$fl(x) = \pm\{(.d_1d_2\ldots d_n)\beta^e + \beta^{e-n}\}$$

$$|x\text{-}fl(x)| = .\left|-d_{n+1}, d_{n+2}.\beta^{e-n-1} + \beta^{e-n}\right|$$

$$= \beta^{e-n-1}\left|d_{n+1}.\, d_{n+2} - \beta\right|$$

$$\le \beta^{e-n-1} \times \frac{1}{2}\beta = \frac{1}{2}\beta^{e-n}$$

Example 7: The true value of p is 3.14159265… In menstruation problems the value 22/7 is commonly used as an approximation to p. What is the error in this approximation?

Solution: The true value of p is p = 3.14159265.

Now, we convert 22/7 to decimal form, so that we can find the difference between the approximate value and true value. Then, the approximate value of p is $\frac{22}{7} = 3.14285714$

Therefore, absolute error = 0.00126449 and relative-error = 0.00040249966.

The round-off error of computer representation of the number p depends on how many digits are left out. Make sure that you understand each line of the following rounding off of the number p :

Number of digits	**Approximation for p**	**absolute error**	**relative error**
1	3.100	0.041593	0.0132%
2	3.140	0.001593	0.0507%
3	3.142	0.000407	0.0130%

Round-off errors may accumulate, propagate and even lead to catastrophic cancellations leading to loss of accuracy of numerical calculations.

Check Your Progress 3

1) Let x^* = .3454 and y^* = .3443 be approximations to x and y respectively correct to 3 significant digits. Further, let $z^* = x^* - y^*$ be the approximation to x – y. Then show that the relative error in z^* as an approximation to x – y can be as large as 100 times the relative error in x or y.

……

……

……

……

2) Round the number x = 2.2554 to three significant figures. Find the absolute error and the relative error.

……

……

……

……

3) If π = 3.14 instead of 22/7, find the relative error and percentage error.

……

……

……

……

4) Determine the number of correct digits in s = 0.2217, if it has a relative error, $\varepsilon_r = 0.2 * 10^{-1}$.

……

……

……

……

5) Round-off the number 4.5126 to four significant figures and find the relative percentage error.

……

……

……

……

1.3.3 Truncation Error

Truncation error is a consequence of doing only a finite number of steps in a calculation that would require an infinite number of steps to do exactly. A simple example of a calculation that will be affected by truncation error is the evaluation of an infinite sum. The computer uses only a finite number of terms and the terms that are left out lead to truncation error.

Numerical integration is another example of an operation that is affected by truncation error. A quadrature formula works by evaluating the integrand at a finite number of points and using smooth functions to approximate the integrand between those points. The difference between those smooth functions and the actual integrand leads to truncation error.

Taylor series represents the local behaviour of a function near a give n point. If one replaces the series by the n-th order polynomial, the truncation error is said to be **order of n**, or $O(h^n)$, where *h* is the distance to the given point. Consider the irrational number *e*

$e = 2.71828182845905\ldots$

and compare it with the Taylor series of the function *exp(x)* near the given point $x = 0$.

$$\exp(x) = 1 + x + x/2 + x^3/6 + ...$$

Let us check a few Taylor series approximations of the number $e = exp(1)$:

order of n	approximation for e	absolute error	relative error
3	2.500000	0.218282	8.030140%
4	2.666667	0.051615	1.898816%
5	2.708333	0.009948	0.365984%

Example 8: Find the value of *e* correct to three decimal places.

Solution: Recall that $e = 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + ...$

The series is to be truncated such that the finite sum equals *e* to three decimal places. This means the must be less than 0.0005. Suppose that the tail starts at n = k+1. Then,

$$\sum_{n=k+1}^{\infty} \frac{1}{n!} = \frac{1}{(k+1)!} + \frac{1}{(k+2)!} ..... + ...$$

$$< \frac{1}{(k+1)!}[1 + \frac{1}{(k+1)} + \frac{1}{(k+1)^2} + ...$$

$$= \frac{1}{(k+1)!}\left[\frac{(k+1)}{1 - /(k+1)}\right] = \frac{1}{k!k} < 0.0005$$

For k = 6, This expression is satisfied and the truncated value of $e = 2.7181$.

1.4 SUMMARY

In this unit, we have defined the floating point numbers and their representation for usage in computers. We have defined accuracy and number of significant digits in a given number. We have also discussed the sources of errors in computation. We have defined the round-off and truncation errors and their propagation in later computations

using these values, which contains errors. Therefore, care must be taken to analyise the computations, so that we are sure that the output of computations is meaningful.

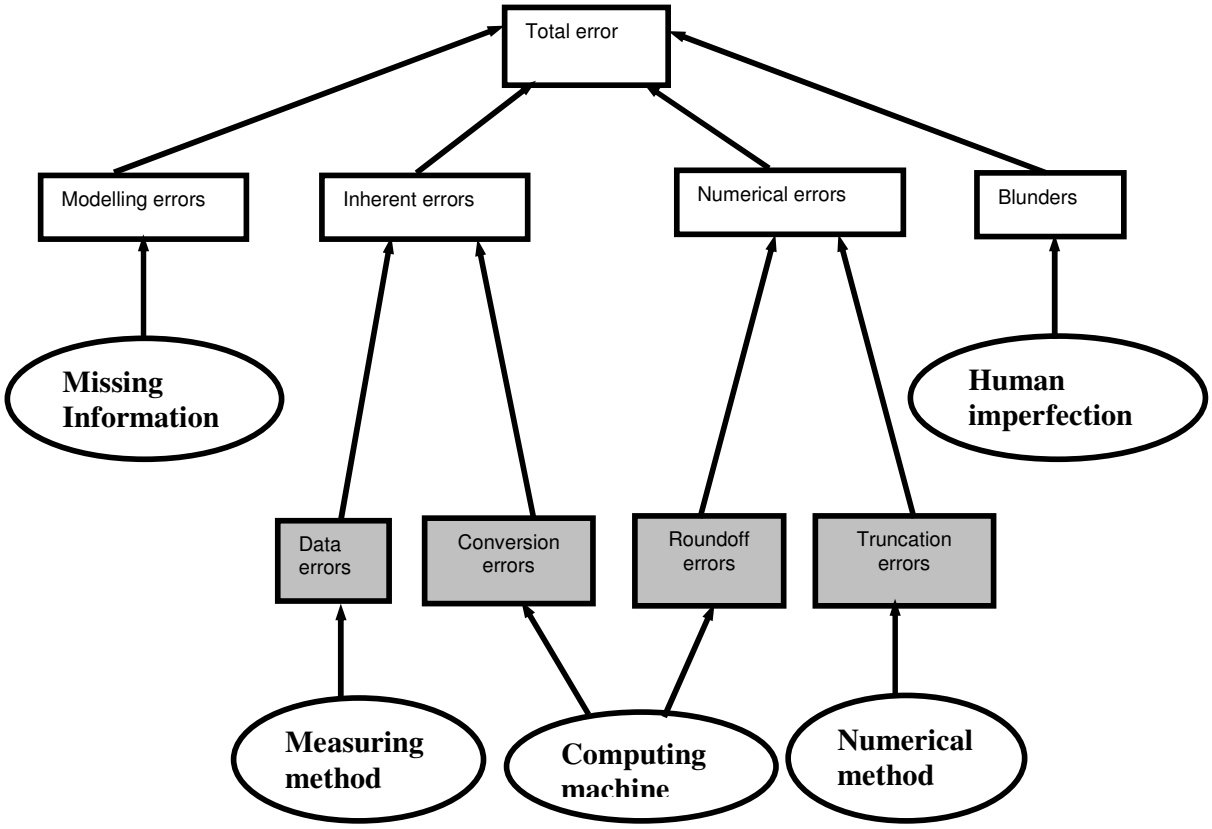


Figure 2: Types of error and their contribution to total errors

1.5 SOLUTIONS/ANSWERS

Check Your Progress 1

1) Using two decimal digit arithmetic with rounding we have,

$$\frac{(a-b)}{c} = .71 \times 10^{-1} \text{ and } \frac{a}{c} - \frac{b}{c} = .59 - .51 = .80 \times 10^{-1}$$

while true value of $\frac{(a-b)}{c} = 0.071428 \ldots$.

Therefore, $\frac{(a-b)}{c} \neq \frac{a}{c} - \frac{b}{c}$

2) Do as 1) above.
3) Do as 1) above.

Check Your Progress 2

1) (i) 50.9 (ii) 48.37 (iii) 9.326 (iv) 8.416 (v) 0.8001 (vi) 0.04251 (vii) 0.004912 (viii) 0.0002022

2) (i) $1000 * 10^2$ or $0.1000 * 10^6$ (ii) $-0.2214 * 10^{-2}$ (iii) $-0.3567 * 10^2$

3) We have 28.483 – 27.984 = 00.499. The result has only three significant digits. This is due to the loss of significant digits during subtraction of nearly equal numbers.

4) We find that $|2/3 - 0.66667| = 0.0000033... < \frac{1}{2}10^{-5}$

We find, k = 5. Therefore, the approximation is accurate to 5 decimal places.

5) Left as an exercise.

Check Your Progress 3

1) Given, $|r_x|, |r_y|, \leq \frac{1}{2}10^{1-3}$

$z^* = x^* - y^* = 0.3454 - 0.3443 = 0.0011 = 0.11 \times 10^{-2}$.

This is correct to one significant digit since last digits 4 in x^* and 3 in y^* are not reliable and second significant digit of i^* is derived from the fourth digits of x^* and y^*.

Max. $|r_z| = \frac{1}{2}\ 10^{1-1} = \frac{1}{2} = (100).\left(\frac{1}{2}\right).10^{-2} \geq 100\ |r_x|, 100\ |r_y|$

2) The rounded-off number is 2.25. The absolute error is 0.0054.

The relative error is $\approx \frac{0.0054}{2.25} = 0.0024$. The percentage error is 0.24%.

3) Relative error = $\left(\frac{22}{7} - 3.14\right) \Big/ \frac{22}{7} = 0.00093$. Percentage error = 0.093 %.

4) Absolute error = $0.2 * 10^{-1} * 0.2217 = 0.04493$. Hence x has only one correct digit $x \approx 0.2$.

5) The number 4.5126 round-off to four significant figures is 4.153.

Relative percentage error = $\frac{-0.0004}{4.5126} * 100 = -0.0088\%$.

1.6 EXERCISES

E1) Give the floating-point representation of the following numbers in 2 decimal digit and 4 decimal digit floating point number using (i) rounding and (ii) chopping.

(a) 37.21829
(b) 0.022718
(c) 3000527.11059

E2) Show that a(b – c) ≠ ab – ac, where, a = .5555 × 10^1, b = .4545 × 10^1, c = .4535 × 10^1.

E3) How many bits of significance will be lost in the following subtraction? 37.593621 – 37.584216. Assume each number is correct to seven significant digits.

E4) What is the relative error in the computation of x – y, where x = 0.3721448693 and y = 0.3720214371 with five decimal digit of accuracy?

E5) Find the smaller root in the magnitude of the quadratic equation $x^2 + 111.11x + 1.2121 = 0$, using five-decimal digit floating point chopped arithmetic.

1.7 SOLUTIONS TO EXERCISES

E1) a)

	rounding	**chopping**
	$.37 \times 10^2$	$.37 \times 10^2$
	$.3722 \times 10^2$	$.3721 \times 10^2$

b)

	$.23 \times 10^{-1}$	$.22 \times 10^{-1}$
	$.2272 \times 10^{-1}$	$.2271 \times 10^{-1}$

c)

	$.31 \times 10^2$	$.30 \times 10^2$
	$.3056 \times 10^2$	$.3055 \times 10^2$

E2) Let, $a = .5555 \times 10^1$, $b = .4545 \times 10^1$, $c = .4535 \times 10^1$
$b - c = .0010 \times 10^1 = .1000 \times 10^{-1}$

$a(b - c) = (.5555 \times 10^1) \times (.1000 \times 10^{-1}) = .05555 \times 10^0 = .5550 \times 10^{-1}$

$ab = (.5555 \times 10^1)(.4545 \times 10^1) = (.2524 \times 10^2)$

$ac = (.5555 \times 10^1)(.4535 \times 10^1) = (.2519 \times 10^2)$

and $ab - ac = .2524 \times 10^2 - .2519 \times 10^2 = .0005 \times 10^2 = .5000 \times 10^{-1}$

Hence $a(b - c) \neq ab - ac$.

E3) $37.593621 - 37.584216 = (0.37593621)10^2 - (0.37584216)10^2$
$= x^* - y^* = (0.00009405)10^2$

The numbers are, correct to seven significant digits. Then, in eight digit floating-point arithmetic, the number can be written as $z^* = x^* - y^* = (0.94050000)10^{-2}$ But as an approximation to $z = x - y$, z^* is good only to three digits, since the fourth significant digit of z^* is derived from the eighth digits of x^* and y^*, and both possibly contains errors. Here, while the error in z^* as an approximation to $z = x - y$ is at most the sum of the errors in x^* and y^*, the relative error in z^* is possibly 10,000 times the relative error in x^* or y^*. Loss of significant digits is, therefore, dangerous only if we wish to keep the relative error small.

Given $|r_x|, |r_y| < \frac{1}{2}10^{1-7}$, $z^* = (0.9405)10^{-2}$, is correct to three significant digits.

Max $|r_z| = \frac{1}{2}10^{1-3} = 10000.\frac{1}{2}10^{-6} \geq (1000)|r_z|(10000)|r_y|$

E4) With five decimal digit accuracy $x^* = 0.37214 \times 10^0$, $y^* = 0.37202 \times 10^0$, $x^* - y^* = 0.00012$ while $x - y = 0.0001234322$.

$$\frac{\left|(x-y)-(x^*-y^*)\right|}{|x-y|} = \frac{0.0000034322}{0.0001234322} \approx 3 \times 10^{-2}.$$

The magnitude of this relative error is quite large when compared with the relative errors of x^* and y^* (which cannot exceed 5×10^{-5} and in this case it is approximately 1.3×10^{-5})

E5) Using the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \text{ we get, } \quad x_1 = \frac{-111.11+111.09}{2} = -0.01000$$

while the true solution is $x_1 = -0.010910,$ correct to the number of digits shown.

However, if we calculate x_1 as $x_1 = \frac{2c}{b + \sqrt{b^2 - 4ac}}$, we get

$$x_1 = \frac{-2 \; x \; 1.2121}{111.11+111.09} = \frac{-2.4242}{222.20}$$

$$= -\frac{24242}{2222000} = -0.0109099 \quad = -.0109099$$

which is accurate to five digits.