1 Computer Arithmetic and Errors

1.1 INTRODUCTION

When we solve a mathematical problem on a computer, a step-by-step procedure utilising the characteristics of a computer should be evolved. It should be observed that only the arithmetic operations may be used even when solving problems involving the operations of calculus.

In numerical analysis, the analysis of error is of great importance. So far we have used various data on the assumptions that they are pure and the techniques of their computations are perfect, but this is not the case all the time, either the data can be impure or there can be error in the computational procedure. In this chapter we shall discuss different types of errors, their determination along with the computer arithmetic. This chapter is divided into two sections namely

(i) Computer arithmetic

(ii) Errors in Numerical Computations

Now, before discussing the above sections, let us recall some mathematical preliminaries.

1.2 SOME MATHEMATICAL PRELIMINARIES

In this section we state some certain mathematical results which are very useful.

1. **Intermediate Theorem.** If $f(x)$ is a continuous function in a closed interval $[a, b]$ *i.e.* $a \leq x \leq b$ and if $f(a)$ and $f(b)$ are of opposite signs then there must exist a number c lies between a and b such that $f(c) = 0$.

2. **Rolle's Theorem.** Let $f(x)$ be a function defined on $[a, b]$ such that it is

   (i) continuous in a closed interval $[a, b]$

   (ii) differentiable in the open interval (a, b)

   (iii) $f(a) = f(b)$

   Then there exists at least one value of x say c in (a, b) such that $\mathrm{f}'(\mathrm{c}) = 0$

3. **Lagrange's mean value Theorem.** Let $f(x)$ be a function defined on (a, b) such that it is

   (i) continuous in a closed interval $[a, b]$

   (ii) differentiable in the open interval (a, b)

   Then there exists at least one value of x say c between a and b such that

$$\frac{f(b) - f(a)}{b - a} = f'(c)$$

4. **Taylor's series for a function of one variable.** If $f(x)$ is continuous and possesses continuous derivatives of order n in an interval including $x = a$, then in that interval.

$$f(x) = f(a) + (x - a)f'(a) + \frac{(x - a)^2}{2!} f''(a) + \ldots + \frac{(x - a)^{n-1}}{(n - 1)!} f^{n-1}(a) + R_n(x)$$

where $R_n(x)$ is the remainder term given by

$$R_n(x) = \frac{(x-a)^n}{n!} f^n(c), \quad a < c < b$$

5. **Maclaurin's series.** Taking $a = 0$ in the above Taylor's series we have the following Maclaurin's series

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2!} f''(0) + \ldots + \frac{x^n}{n!} f^n(0) + \ldots$$

6. **Taylor's series for a function of two variables.** We have

$$f(x_1 + \Delta x_1, x_2 + \Delta x_2) = f(x_1, x_2) + \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2$$
$$+ \frac{1}{2}\left[\frac{\partial^2 f}{\partial x_1^2}(\Delta x_1)^2 + 2\frac{\partial^2 f}{\partial x_1 \partial x_2}(\Delta x_1)(\Delta x_2) + \frac{\partial^2 f}{\partial x_2^2}(\Delta x_2)^2\right] + \ldots$$

7. **Some important expansions**

(i) $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$

(ii) $(1+x)^n = 1 + nx + \frac{(n)(n-1)}{2!}x^2 + \frac{(n)(n-1)(n-2)}{3!}x^3 + \ldots$

(iii) $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$

(iv) $\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots$

(v) $\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \ldots$

(vi) $a^x = 1 + x\log_e a + \frac{(x\log_e a)^2}{2!} + \ldots$

(vii) $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots$

(viii) $\sin^{-1} x = \frac{x^3}{6} + \frac{3}{40}x^5 + \ldots$

Section A : Computer Arithmetic

1.3 NUMBER SYSTEM

It is a machine language which provides a facility to make the numbers. We may define a number system as a system which consists of,

- a set of symbol used for formation of numbers.
- a set of rules which may be used to form numbers from these symbols and assign values to them.
- a set of rules performing common arithmetic operations on this system.

There are many types of number system. Some important number systems are as follows:

(i) Decimal number system **(ii)** Binary number system.

(iii) Octal number system **(iv)** Hexadecimal number system

1.3.1 DECIMAL NUMBER SYSTEM

This number system has a base of 10, *i.e.*, 0, 1,2,3,4,5,6,7,8,9, number of digits needed to represent a number are changed after every 10*n* intervals, where *n* is an integer. A number can be written in expanded notation form by breaking every digit according to its place value.

For examples

1. The number 456 can be written as $4 \times 10^2 + 5 \times 10^1 + 6 \times 10^0$
2. The number 6428.31 can be written as

$$6 \times 10^3 + 4 \times 10^2 + 2 \times 10^1 + 8 \times 10^0 + 3 \times 10^{-1} + 1 \times 10^{-2}$$

1.3.2 BINARY NUMBER SYSTEM

In binary number system, numbers can be represented using 2 digits only so the base of binary numbers system is 2. The two digits that are used in binary number system are 0 and 1. A binary number can be written in expanded notation form by breaking the number into digits according to their place value.

e.g.,
$$\begin{aligned} 1010 &= (1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (0 \times 2^0) \\ &= 1 \times 8 + 0 \times 4 + 1 \times 2 + 0 \times 1 = 8 + 2 = 10 \end{aligned}$$

This means $(1010)_2 = (10)_{10}$

1.3.3 OCTAL NUMBER SYSTEM

Octal number system is the number system with base 8. This means; in this number system, there are 8 symbols or digits which are used for formation of the numbers. These symbols are 0, 1, 2, 3, 4, 5, 6 and 7. The place value in octal number system are the power of 8. Consider, a number $(156)_8$. This can be written in the expanded form as,

$$\begin{aligned} (156)_8 &= 6 \times 8^0 + 5 \times 8^1 + 1 \times 8^2 \\ &= 6 \times 1 + 5 \times 8 + 1 \times 64 = 6 + 40 + 64 \end{aligned}$$

The means, $(156)_8 = (110)_{10}$

1.3.4 HEXADECIMAL NUMBER SYSTEM

This number system is number system with base 16. Using the symbols 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E and F. In number system, in addition to decimal digits 0 to 9, the symbols A, B, C, D, E and F are used to represent the numbers 10, 11, 12 ,13, 14 and 15 respectively.

Consider a number $(13BD)_{16}$. This number can be written in expanded form as

$$\begin{aligned} (13BD)_{16} &= 1 \times 16^3 + 3 \times 16^2 + B \times 16^1 + D \times 16^0 \\ &= 4096 + 768 + 11 \times 16 + 13 \times 1 \\ &= 4096 + 768 + 176 + 13 = (5053)_{10}. \end{aligned}$$

1.4 BASE CONVERSION

1.4.1 DECIMAL TO BINARY (TO CONVERT THE INTEGER PART)

To convert the number in decimal number system to the number in binary number system, we apply the method of repeated division. The division is done by 2.

WORKING PROCEDURE

STEP 1. Divide the given number by 2.

STEP 2. Note the quotient and remainder. Remainder will be either 0 or 1.

STEP 3. If quotient is not 0, then divide the quotient by 2 and go to step 2.

STEP 4. If quotient is 0, then stop the process of division.

STEP 5. The process of first remainder is called least significant digit (LSD) and last remainder is called most significant digit (MSD).

STEP 6. Arrange all the remainders from MSD to LSD in a sequence from left to right.

Then the combination of 0 and 1 thus obtained is the required binary equivalent of given number.

For example: *Convert* $(45)_{10}$ *into binary number system.*

Solution: Performing repetitive division by 2.

2	45	remainder	
2	22	1	LSD
2	11	0	
2	5	1	
2	2	1	
2	1	0	
	0	1	MSD

Thus $(45)_{10} = (101101)_2$

To convert the fractional part: For converting a fractional decimal number in binary, we use the method of repeated multiplication. The multiplier is 2.

WORKING PROCEDURE

STEP 1. Multiply the given number by 2 and separate the integral part.

STEP 2. Multiply the fractional part again by 2 and separate the integral part.

STEP 3. Continue this process, till the fractional part reduces to zero.

STEP 4. Write the integral parts and prefix the binary point.

This will be the desired binary fraction.

Solved Examples

EXAMPLE 1. ***Convert $(0.8176)_{10}$ to binary number system.***

SOLUTION.

	0	0.8176 × 2
MSD	1	0.6352 × 2
	1	0.2704 × 2
	0	0.5408 × 2
LSD	1	0.0816 × 2
	0	0.1632 × 2

$(08176)_{10} = (0.11010\ ...)_2$

EXAMPLE 2. ***Convert $(67.25)_{10}$ to binary number system.***

SOLUTION. First we convert the integral part into binary equivalent.

2	67	remainders
2	33	1
2	16	1
2	8	0
2	4	0
2	2	0
2	1	0
	0	1

Now we convert the decimal part

MSD	0	0.25 × 2
	0	0.50 × 2
LSD	1	0.00 × 2

Thus $(67.25)_{10} = (1000011.01)_2$

1.4.2 BINARY TO DECIMAL

To convert the binary number to decimal number, use the following procedure

WORKING PROCEDURE

STEP 1. Multiply the digit of whole binary number with powers of 2. The power for integral part of number are positive and negative for fractional part of number.

STEP 2. Add the total result which are obtained by multiplying the power of digits.

We obtain the final result after addition.

For example: *Convert the following binary numbers to decimal number*

(i) $(1100111)_2$ **(ii)** $(11001101.01)_2$

Solution. **(i)** $(1100111)_2$

$= 1\times2^6+1 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$

$= 64 + 32 + 0 + 0 + 4 + 2 + 1 = (103)_{10}$

(ii) $(11001101.01)_2 = 1 \times 2^7 + 1 \times 2^6 + 0 \times 2^5 + 0 \times 2^4 + 1 \times 2^3 + 1 \times 2^2$
$+ 0 \times 2^1 + 1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2}$

$= 128 + 64 + 0 + 0 + 8 + 4 + 0 + 1 + 0 + 0.25 = (205.25)_{10}$

1.4.3 BINARY TO OCTAL

To convert a binary number into octal number system, use the following procedure

WORKING PROCEDURE

STEP 1. Firstly we convert binary number to decimal and then decimal to octal. We make the groups of three digits. We start the grouping from right to left.

STEP 2. Now each group of three digits converts the decimal number system. After that written the decimal numbers combinedly.

The group of three binary digits from an octal number as shown the table given below:

0	1	2	3	4	5	6	7
000	001	010	011	100	101	110	111

For example: *Convert the following binary number to octal*

(i) $(101011101)_2$ **(ii)** $(111100011)_2$

(iii) $(10011011101010)_2$

Solution. (i) Grouping these into three bits each we get

101	011	101	Group of three bits from right octal equivalent.
III	II	I	
5	3	5	

Thus $(101011101)_2 = (535)_8$

(ii)

111	100	011	Group of three bits from right octal equivalent.
III	II	I	
7	4	3	

$\Rightarrow \quad (111100011)_2 = (743)_8$

(iii)

010	011	011	101	010	Group of three bits from right octal equivalent.
V	IV	III	II	I	
2	3	3	5	2	

$\Rightarrow \quad (10011011101010)_2 = (23352)_8$

1.4.4 BINARY TO HEXADECIMAL (TO CONVERT AN INTEGER)

WORKING PROCEDURE

STEP 1. For this conversion we divide all binary digit of the number to be converted in the groups of four bits each and start the grouping from right to left.

STEP 2. Now each of these groups of four bit each will be converted to decimal number system and written below the groups.

A group of four binary digits forms one hexadecimal as shown in the table below:

Hexadecimal digit	Binary equivalent
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	1111
8	1000
9	1001
10 or A	1010
11 or B	1011
12 or C	1100
13 or D	1101
14 or E	1110
15 or F	1111

For example: *Convert* $(1110101101)_2$ *to hexadecimal equivalent.*

Solution. Grouping these into four bits each we each

11	1010	1101

Here, we see that 11 is alone so we have written two zero's to its lefts.
Now we have four groups as

0011	1010	1101
III	II	I
3	10 or A	13 or D

Thus $(1110101101)_2 = (3AD)_{16}$

To convert a fraction:

WORKING PROCEDURE

STEP 1. For this conversion we divide all binary digit of the fraction part to be converted in the groups of four bits each. Start the grouping from left to right.

STEP 2. Now each of these groups of four bits each will be converted to decimal number system. After that these numbers written in groups.

For example: *Convert* $(100011.01)_2$ *to hexadecimal equivalent.*

Solution. After grouping of 100011.01, we get

0100	0011	0100
III	II	I
4	3	4

Thus $(100011.01)_2 = (43.4)_{16}$

1.4.5 DECIMAL TO OCTAL (TO CONVERT THE INTEGER)

For converting the decimal number to octal we apply the following process step by step as

WORKING PROCEDURE

STEP 1. Divide the number by 8.

STEP 2. Note down the quotient and remainder. Remainder will be any digit from 0 to 7.

STEP 3. If quotient is not 0, then divide the quotient again by 8 and go to step 2.

STEP 4. If quotient is 0, then stop the process of division.

STEP 5. Write all remainder from left to right.

The combination of digit 0 to 7 thus obtained is the required octal equivalent of number.

For example: *Convert* $(8765)_{10}$ *to octal number system.*

Solution.

8	8765	remainders
8	1095	5
8	136	7
8	17	0
8	2	1
	0	2

Thus $(8765)_{10} = (21075)_8$

To convert the fraction: To convert a fractional decimal number in octal, use the method of repeated multiplication. The multiplier is 8.

WORKING PROCEDURE

STEP 1. Multiply the number by 8.

STEP 2. Note down the integer part and fractional part of the result separately.

STEP 3. If the fractional part of the result satisfies any two conditions, stop the process of multiplication.

Conditions are:

(i) fractional part is 0.

(ii) fractional part achieved has already appeared before that position.

STEP 4. If the resultant fraction does not satisfy any of the above conditions, then go to step 1.

Write all carries from left to right. The combination of digit 0 to 7 thus obtained is the required result.

Solved Examples

EXAMPLE 1. ***Convert $(0.1015625)_{10}$ to octal number system.***

SOLUTION. Multiply repeated by 8.

MSD	0	0.1015625 × 8
↓	0	0.8125000 × 8
↓	6	0.5000000 × 8
LSD	4	0.0000000 × 8

Thus $(0.1015625)_{10} = (0.064...)_8$.

EXAMPLE 2. ***Convert $(1093.21875)_{10}$ to octal number system.***

SOLUTION. Converting both integral part and fractional part separately

8	1093	remainders
8	136	5
8	17	0
8	2	1
8	0	2

0	0.21875 × 8
1	0.75000 × 8
6	0.00000 × 8

$\Rightarrow$ $(1093)_{10} = (2105)_8$ $\Rightarrow$ $(0.21875)_{10} = (0.16)_8$

Thus $(1093.21875)_{10} = (2105.16)_8$.

1.4.6 DECIMAL TO HEXADECIMAL

To convert an integer: For converting the number in decimal number system to the number in hexadecimal number system, use the method of repeated division.

WORKING PROCEDURE

STEP 1. Divide the number by 16.

STEP 2. Note down the quotient and remainder. Remainder will be from 0 to 9 or A to F.

STEP 3. If quotient is not 0, then divide the quotient by 16, and go the step 2.

STEP 4. If quotient is 0 or any digit or symbol less than 16 then stop the process of division.

STEP 5. Write all remainder from left to right. The combination obtained is the desired Hexadecimal number.

For example: *Convert* $(198275)_{10}$ *to hexadecimal equivalent.*

Solution.

16	198275	remainders
16	12392	3
16	774	8
16	48	6
16	3	0
	0	3

Thus $(198275)_{10} = (30683)_{16}$.

To convert a fraction: To convert a fraction decimal number in hexadecimal, use the method of repeated multiplication.

WORKING PROCEDURE

STEP 1. Multiply the fraction part by 16.

STEP 2. Note down the integer part (carry) and fractional part of the result separately.

STEP 3. If the fractional part is 0 or achieved has already appeared before that position, stop the process of multiplication.

STEP 4. If the resultant fraction, does not satisfy the condition of step 3, then go to step 1.

After this process we write first carry to last carry in the sequence. This sequence obtained is the required result.

For example: *Convert* 0.6875875 *to hexadecimal number system.*

Solution.

0	0.6875875×16
11	0.00110000×16
0	0.0176×16
0	0.2816×16
4	0.5056×16
8	0.896×16

Thus $(0.68756875)_{10} = (0.110049)_{16} = (B0049)_{16}$

1.4.7 OCTAL TO DECIMAL

For the conversion of octal number to decimal number, multiply the whole octal number with power of 8. These powers are positive for integral part of number and negative for fractional part of number.

For example: *Convert* $(1727)_8$ *to decimal equivalent.*

Solution. $(1727)^8 = 1 \times 8^3 + 7 \times 8^2 + 2 \times 8^1 + 7 \times 8^0 = 512 + 448 + 16 + 7 = (983)_{10}$

Example: *Convert* $(3027.105)_8$ *to decimal equivalent.*

Solution. $(3027.105)8 = 3\times8^3 +0\times8^2+2\times8^1+7\times 8^0 +1\times 8^{-1} +0\times8^{-2}+5\times8^{-3}$

$= (1559.124765625)_{10}$

1.4.8 OCTAL TO BINARY

The conversion octal to binary is very easy. Every digit of the number which is to be converted from octal to binary, is individually converted to the 3-bit binary equivalent. The combination of 0 and 1 is our desired result.

For example: *Convert* $(103.2)_8$ *to binary equivalent.*

Solution.

	$(103.2)_8$	=	1	0	3	2	
			001	000	011	010	Binary equivalent
Thus	$(103.2)_8$	=	$(001000011.010)_2$.				

1.4.9 OCTAL TO HEXADECIMAL

For converting an octal number to hexadecimal number, use the following procedure

WORKING PROCEDURE

STEP 1. Convert the octal number to binary equivalent.

STEP 2. Now convert this binary equivalent to hexadecimal number system. The number obtained is the required result.

For example: *Convert* $(72232321)_8$ *to hexadecimal equivalent.*

Solution. Firstly we convert the given octal number to Binary equivalent.

$(72232321)_8$ = 7 → 111
2 → 010
2 → 010
3 → 011
2 → 010
3 → 011
2 → 010
1 → 001

Thus, $(72232321)_8 = (111010010011010011010001)_2$

Now we convert this number into hexadecimal equivalent. Grouping these into four bits each we get

1110	1001	0011	0100	1101	0001
14 or E	9	3	4	13 or D	1

Thus $(111010010011010011010001)_2 = (E934D1)_{16}$.

1.4.10 HEXADECIMAL TO BINARY

For converting an hexadecimal number to binary equivalent, we individually convert to the 4-bit binary equivalent. Then the combination of 0 and 1 thus obtained the desired result.

For example: *Convert* $(A92)_{16}$ *to Binary equivalent.*

Solution. $(A92)_{16} = A \times 16^2 + 9 \times 16^1 + 2 \times 16^0$

$= 10 \times 256 + 9 \times 16 + 2 \times 1 = 2560 + 144 + 2 = (2706)_{10}$

Now

2	2706	Remainder
2	1353	0
2	676	1
2	338	0
2	169	0
2	84	1
2	42	0
2	21	0
2	10	1
2	5	0
2	2	1
2	1	0
	0	1

Thus $(2706)_{10} = (101010010010)_2$

Hence $(A92)_{16} = (101010010010)_2$

Alternate Method:

A	9	2	
10	9	2	
1010	1001	0010	Binary equivalent

$\Rightarrow \quad (A92)_{16} = (101010010010)_2.$

1.4.11 HEXADECIMAL TO DECIMAL

For converting hexadecimal to decimal equivalent. We individually separate the number and multiply the whole number with power of 16. After this process add the total resultant numbers, which will be desired decimal number.

Solved Examples

EXAMPLE 1. ***Convert* $(\mathbf{5009B})_{16}$ *to Decimal equivalent.***

SOLUTION.
$$(5009B)_{16} = 5 \times 16^4 + 0 \times 16^3 + 0 \times 16^2 + 9 \times 16^1 + B \times 16^0$$
$$= (327680 + 0 + 0 + 144 + 11) \quad = (327835)_{10}$$

Thus $(5009B)_{16} = (327835)_{10}$

EXAMPLE 2. ***Convert* $(\mathbf{BCD})_{16}$ *to Decimal equivalent.***

SOLUTION.
$$(BCD)_{16} = B \times 16^2 + C \times 16^1 + D \times 16^0$$
$$= B \times 256 + C \times 16 + D = 11 \times 256 + 12 \times 16 + 13$$
$$= 2816 + 192 + 13 = (3021)_{10}$$

1.4.12 HEXADECIMAL TO OCTAL

For converting the hexadecimal number to octal number system, firstly convert the hexadecimal number to binary equivalent. After this process, convert this binary equivalent to octal number system. The number thus obtained is the desired result.

For example: *Convert* $(E934D1)_{16}$ *to hexadecimal number system.*

SOLUTION.

$(E934D1)_{16}$	=	E	9	3	4	D	1
		1110	1001	0011	0100	1101	0001

$\Rightarrow$ $(E934D1)_{16} = (111010010011010011010001)_2$

Now we convert this binary number to octal equivalent.

Grouping these into three bits each, we get

111	010	010	011	010	011	010	001
7	2	2	3	2	3	2	1

Therefore $(111010010011010011010001)_2 = (72232321)_8$

This implies $(E934D1)_{16} = (72232321)_8$

A conversion table between decimal, hexadecimal octal and binary relation is given below:

Decimal $()_{10}$	**Hexadecimal** $()_{16}$	**Octal** $()_8$	**Binary** $()_2$
0	0	00	0000
1	1	01	0001
2	2	02	0010
3	3	03	0011
4	4	04	0100
5	5	05	0101
6	6	06	0110
7	7	07	0111
8	8	10	1000
9	9	11	1001
10	A	12	1010
11	B	13	1011
12	C	14	1100
13	D	15	1101
14	E	16	1110
15	F	17	1111

1.5 BINARY ARITHMETIC

Arithmetic operations additions, subtraction, multiplication and division on binary numbers constitute binary arithmetic.

1.5.1 BINARY ADDITION

The rules of binary addition are

$0 + 0 = 0$

$0 + 1 = 1$

$1 + 0 = 1$

$1 + 1 = 10$ Sum 0 with carry 1.

Like in decimal system when the sum of two digits exceed the highest digit, 1 is carried to the next higher bit position in binary system when the sum exceeds 1 a one is carried to the next higher bit position.

Solved Examples

EXAMPLE 1. ***Add the binary numbers*** $\mathbf{(10110)_2}$ ***and*** $\mathbf{(1101)_2}$**.**

SOLUTION.

$$\begin{array}{r} \_\,\_\,\_\,1\,\_ \leftarrow \text{carry} \\ 10110 \\ +1101 \\ \hline 100011 \\ \hline \end{array}$$

EXAMPLE 2. ***Add the binary numbers*** $\mathbf{(11001)_2}$ ***and*** $\mathbf{(10011)_2}$.

SOLUTION.

$$\begin{array}{r} 11001 \\ +10011 \\ \hline 101100 \\ \hline \end{array}$$

1.5.2 BINARY SUBTRACTION

The rules for binary subtraction are

$$0 - 0 = 0$$
$$1 - 0 = 1$$
$$0 - 1 = 1$$
$$1 - 1 = 0$$

with borrow or 1 from the next column to the left.

If we need to borrow from a digit which is 0, then two or more borrows must be made toward the left. We borrow from the first non zero digit to the left and each. intervening 0 becomes 1 in the process.

EXAMPLE 3. ***Subtract*** $\mathbf{(1111)_2}$ ***from*** **1110101.**

SOLUTION.

$$\begin{array}{r} 1110101 \\ -1111 \\ \hline 1100110 \\ \hline \end{array} \quad \textit{i.e.}, (1100110)_2$$

EXAMPLE 4. ***Subtract*** $\mathbf{(101111)_2}$ ***from*** $\mathbf{(110101)_2}$.

SOLUTION.

$$\begin{array}{r} 110101 \\ -101111 \\ \hline 000110 \\ \hline \end{array} \quad \textit{i.e.},(110)_2.$$

EXERCISE 1.1

1. Convert the following binary numbers to decimal equivalent:
(i) 110111 (ii) 0.101
(iii) 11010111.1101

2. Convert the following decimal numbers to binary:
(i) 5233 (ii) 0.8125
(iii) 9342.982

3. Convert the following number into octal system:
(i) $(9786)_{10}$ (ii) $(8765.27)_{10}$
(iii) $(100000000)_2$ (iv) $(1110111011)_2$

4. Convert the following number into hexadecimal:
(i) $(19)_{10}$ (ii) $(286)_{10}$
(iii) $(100110101111)_2$
(iv) $(360.13)_8$

5. Convert the following number into octal:
(i) $(1011101)_2$ (ii) $(A985B)_{16}$
(iii) $(5834E.B93)_{16}$

6. Fill in the blanks:
(i) $(FA9B)_{16} = (\_\_\_\_\_)_{10}$
(ii) $(217)_{10} = (\_\_\_\_\_)_8$
(iii) $(1046.25)_{10} = (\_\_\_\_\_)_{16}$
(iv) $(A92)_{16} = (\_\_\_\_\_)_{10}$
(v) $(1100110)_2 = (\_\_\_\_\_)_{10}$
(vi) $(42.25)_{10} = (\_\_\_\_\_)_2$

7. What is the decimal equivalent to the hexadecimal number $(BCDE)_{16}$?

8. Find the sum of following binary numbers:
(i) 1001, 101010
(ii) 10110, 1101
(iii) 110101, 101111
(iv) 111011, 10111000
(v) 1001011, 1101001

9. Find the difference of following binary numbers:
(i) 1000 – 1 (ii) 11010 – 101
(iii) 1110001 – 100110
(iv) 11011 – 1101100 (v) 110.110 – 1.1011

10. Calculate the following:
(i) $(100111)_2 - (111010)_2$
(ii) $(111111)_2 + (10101)_2 + (11011)_2$

Answers

1. (i) $(55)_{10}$ (ii) $(0.625)_{10}$ (iii) $(215.15)_{10}$

2. (i) 1010001110001 (ii) 0.1101 (iii) 10011001110010.11111011

3. (i) $(23072)_8$ (ii) $(21075.212...)_8$ (iii) $(400)_8$ (iv) $(733)_8$

4. (i) $(13)_{16}$ (ii) $(AF9)_{16}$ (iii) $(9AF)_{16}$

5. (i) $(135)_8$ (ii) $(2514133)_8$ (iii) $(3701516.5623)_8$

6. (i) $(64155)_{10}$ (ii) $(330)_8$ (iii) $(416.4)_{16}$ (iv) $(2706)_{10}$

6. (v) $(102)_{10}$ (vi) $(101010.01)_2$ **7.** (3021.875)10

8. (i) 110011 (ii) 100011 (iii) 1100100 (iv) 11110011 (v) 10110100

9. (i) 111 (ii) 10101 (iii) 1001011 (iv) 1010001 (v) 101.0001

10. (i) 1101 (ii) 1011111

Section B : Errors in Numerical Computations

1.6 APPROXIMATIONS AND ERRORS

Approximations and errors are on integral part of our life. These are exist everywhere, and sometime are unavoidable. A number of different types of errors arise during the process of numerical computing. These errors contribute to the total error in the final result.

Also the numerical data used in solving the problems are usually not exact, and the numbers expressing such data are therefore not exact. They are merely approximations, two to three, four or more figures. Not only are the data of practical problems usually result is to be obtained are also approximate. Therefore, an approximate calculation is one which involves approximate data, or approximate methods or both. Therefore, it is evident that the error in a computed result may be due to one or both sources, *i.e.*,

(i) error in data and (ii) error in calculation.

The first type of error can not be decrease, but the second type can be made as small as we please, by taking the number to as many figure as we desired. Therefore, we can assume that the calculations are always carried out in such a manner as to make the errors of calculation negligible.

In this section, we examine the sources of various types of computational errors and their subsequent propagation.

1.7 ACCURACY OF NUMBERS

1.7.1 EXACT NUMBERS

The numbers in which, there is no uncertainty and no approximation, is said to be exact numbers.

For example: $5, 6, 7, \frac{8}{2}, \frac{1}{5}, ...$ are exact numbers.

1.7.2 APPROXIMATE NUMBERS

These are numbers which are not exact.

For example: 1.41421 .... 3.141592.... are not exact numbers, since they contains infinitely many digits, are called approximate numbers.

☛ REMARKS

- The approximate number is a number which can not be expressed by a finite number of digits.
- Although, the numbers $\pi, \sqrt{2}$, etc. are exact numbers, they can not be expressed exactly by a finite number of digits. But when we expressed these numbers in digital form 3.141592, 1.41421, etc. such numbers are therefore only approximation to the true values and in such cases are called approximate numbers.
- Some authors always insist that one must say "approximate value" of a number in place of approximate number.
- Here, we used the symbol $\simeq$ for approximately equal to.
- Such numbers which represents the given numbers to a certain degree of accuracy are called approximate numbers.

1.7.3 ROUNDING-OFF A NUMBER

If we divide 22 by 7 we get $\frac{22}{7} = 3.142857143...$ a quotient which a non-terminating decimal fraction. For use this type of number in practical computation, it is to be cut-off to a manageable size such as 3.14, 3.143,... . *The process of cutting-off superfluous digits and retaining as many digits as desired is known as rounding off a number.*

☛ REMARK

- To round off a number is to retain a certain number of digits, counted from the left and dropped the others. Thus, to round off π to three, four or five and six figures respectively, we have 3.14, 3.142, 3.1416, 3.14159.

WORKING PROCEDURE

To rounding off a number or digit to n significant figures, discard all digits to the right of the nth place using the following concepts.

STEP 1. If this number is less than half a unit in the nth place, leave the nth digits as it is.

STEP 2. If the discarded number is greater than half a unit in the nth place, add 1 to the nth digit.

STEP 3. If the discarded number is exactly half a unit in the nth place, leave the nth digit unchanged.

For Example : The following numbers are rounded off correctly to four significant figures

(i) 38.63243 becomes 38.63 **(ii)** 91.8773 becomes 91.88

(iii) 21.64489 becomes 21.64 **(iv)** 87.495 becomes 87.50.

> The old rule of rounding off the number says that when a 5 is dropped the preceding digit should always be increased by 1. It is not a good exercise and give inaccuracy in computations. Since, it is obvious that when a 5 is cut off, the preceding digit should be increased by one in only half the cases and should be left unchanged in the other half.

☛ REMARK

- The numbers rounded off to n significant figures are said to be correct to n significant figures.

1.7.4 SIGNIFICANT FIGURES

Here, all the digits 1, 2 ... upto 9 are significant figures and 0 is a significant figure except when it is used to fix the decimal point or to fill the places of unknown digits, *i.e.*, 0 may or may not be a significant figure. It depends upon the position in which zero has been used. As discussed earlier when zero is used to fixup the decimal point or to fill up the places of discarded digits, it is not a significant figure.

For example: Consider the numbers 0.00086 and 5800, correct to two significant figures. Then all zeros, which are used are insignificant. On the other hand, zero used in 430 correct to three significant figures, is a significant figure.

☛ REMARKS

- The zeroes used between two non-zero digits are always significant figure e.g. 408.
- To round off a number or figure to r significant digits, discard all the digits or replace by zeros to the right of r^{th} digit according as the number to be rounded off is a decimal fraction or whole number. Then r^{th} digit to be increased by 1 or to be left unaltered, according as the portion to be discarded or replaced by zeroes as greater than or less than half of the unit at the r^{th} places (counted from the left). In case the discarded portion is exactly half of the r^{th} unit, then the r^{th} unit is to be increased by 1, if it is odd, otherwise it is left unchanged.

WORKING PROCEDURE

STEP 1. Significant digits are counted from left to right starting with the left most non-digits.

STEP 2. The significant figure in a number in positional notations consists of

(a) all non-zero digits

(b) zero digits which

– lie between significant digits

– lie to the right of decimal points and at the same time, to the right of a non-zero digit.

– are specifically indicated to be significant

STEP 3. The significant figure in a number written in scientific notation e.g. $M \times 10^k$ consists of all the digits explicitly in M.

For Example

(1) The number 8.3678235, when rounded to three places of decimal, we get it as 8.368. Because, we leave the portion 0.0008235 which is more than half of 0.001.

(2) The number 83988235, when rounded to five significant digits, we get as 83988. Because the portion left out is 235, which is less than half of 1000.

(3) The number 8.6325 when rounded to three decimal places, we get 8.632 as the rounded number.

(4) 83675, rounded to four significant figures as obtained as 83680. Here the fourth place, when we counted from the left is 7 which is odd and the portion left out is exactly half of the unit at this place. Therefore we increase 7 by one.

Solved Examples

EXAMPLE 1. ***Round-off the following numbers correct to four significant figures***

68.3643, 878.367, 8.7265, 56.395

SOLUTION. Here, we have to retain first four significant figures. Therefore

(i) 68.3643 becomes 68.36

(ii) 878.367 becomes 878.4

(iii) 8.7265 becomes 8.726 (Because the digit in the fourth place is even).

(iv) 56.395 becomes 56.40 (Because the digit in fourth place is odd).

EXAMPLE 2. ***Find the sum of the following approximate numbers, each being correct to its last figures***

396.56, 657.2, 758.9826, 3.052

SOLUTION. Since the number 657.2 is correct to one decimal place. Therefore, it is not worth while to retain digits beyond two decimal places. Hence, we rounded off the given numbers to two decimal places, and then found the sum.

Therefore, the required sum

$$= 396.56 + 657.20 + 758.98 + 3.05 = 1815.79 \simeq 1815.8$$

☛ REMARKS

➠ When we deal with the approximate numbers of unequal accuracies, retain one more significant figure is more accurate numbers then are contained in the least accurate number as it being done in above example. In the end the sum has been rounded to one decimal place.

The concept of accuracy and precision are closely related to the significant digits, as follows:

(a) Accuracy refers to the number of significant digits in a value. For example, the number 86.498 is accurate to five significant digits.

(b) Precision refers to the number of decimal positions, i.e., the order of magnitude of the last digit in a value. Here the number 86.498 has a precision of 0.001 or 10^{-3}.

1.8 ERRORS AND THEIR ANALYSIS

Definition : *The difference between true value and approximate value is called the error.*

1.8.1 SOURCES OF ERRORS

Following are some sources of error in numerical computations.

(i) Input Errors: The input information is rarely exact. It comes from the experiments and any experiment can give results of any limited accuracy.

(ii) Algorithmic Errors: Sometimes, the direct algorithms based on a finite sequence of operations are used. Errors due to limited steps don't amplify the existing errors. Since the application of some formula is not possible for a infinite number of times, algorithm has to be stopped after a finite number of steps. Hence, the obtained results are not exact.

(iii) Computational Errors: Sometimes, when we performing elementary operations, the number of digits increases greatly. Therefore, the result can not be held fully in a register available in the given system.

1.8.2 TYPES OF ERROR

(i) Absolute error: If x^A is the approximate value of exact number x^T, then the absolute error denoted by E_a is defined by

$$E_a = \Delta x = |x^T - x^A|$$

$$\Rightarrow \qquad E_a = |x^T - x^A|$$

☛ **REMARK**

➠ In error analysis, the magnitude of the error is not important, not the sign of error. Therefore, we consider the absolute error generally.

(ii) Relative Error: In many cases, absolute error may not reflect its influence correctly as it does not take into account the order of magnitude of the value under consideration.

For example: An error of 1 gram is much more significant in the weight of 10 grams Gold, that in the weight of a bag of sugar. Due to this reason the concept of relative error is introduced.

The relative error is the absolute error divided by the true value of the given quantity.

It is denoted by E_r and defined as

$$E_r = \left|\frac{x^T - x^A}{x^T}\right| = \frac{\text{Absolute error}}{\text{True value}}$$

> The relative error of a product of *n* numbers is approximately equal to the algebraic sum of their relative errors.

(iii) Percentage Error: The percentage error in x^A, which is the approximate value of x^T is

$$E_p = 100 \times E_r = 100 \times \left|\frac{x^T - x^A}{x^T}\right|$$

☛ **REMARKS**

➠ The relative error is also known as normalized absolute error.

➠ If $\bar{x}$ be a number such that $\left|x^T - x^A\right| \le \bar{x}$, then $\bar{x}$ is said to be an upper limit on the magnitude of absolute error and measures the absolute accuracy.

➠ If a number is correct to n significant figures then its absolute error can not be greater than half a unit in a n^{th} places.

➠ If a number is correct to n decimal places then the error $= \frac{1}{2} \cdot 10^{-n}$.

For example: If the number 8.869 correct to three decimal points its absolute error is not greater than $0.001 \times \frac{1}{2} = \frac{1}{2} \times 10^{-3} = 0.0005$.

➠ The relative and percentage errors are independent of the units of measurement, while absolute errors are expressed in terms of unit used.

Solved Examples

EXAMPLE 1. ***Find the sum of* 392, 780.56, 64320, 72300, 23657 *assuming that the number* 72300 *is known to only three significant figures.***

SOLUTION. Since the number 72300 is known to hundred places.

Therefore, we round off other numbers correct to tens places and then find the sum, *i.e.*,

$$\text{Sum, } S = 390 + 780 + 64320 + 72300 + 23660$$
$$= 161450 \simeq 161400$$

Here, we observe that, the last significant digit (counting from left) is 4 which is uncertain by one unit of this place.

THEOREM 1. ***If the first significant figure of a number is r and the number is correct to n significant figures, then the relative error is less than*** $\dfrac{1}{r \times 10^{n-1}}$.

PROOF. Let us suppose that N be any given exact number which contains n significant figures and m denotes the number of correct decimal places.

Then, there are following three cases :

Case (i): If $m < n$

In this case the number of digits in the integral part of N is given by $(n - m)$. Let us denote the first significant figure of N by r. Then, we have

Absolute error $\quad E_a \le \dfrac{1}{10^m} \times \dfrac{1}{2}$

and $\quad N \ge r \times 10^{n-m-1} - \dfrac{1}{10^m} \times \dfrac{1}{2}$

which gives

$$E_r \le \frac{\dfrac{1}{10^m} \times \dfrac{1}{2}}{r \times 10^{n-m-1} - \dfrac{1}{10^m} \times \dfrac{1}{2}}$$

$$E_r = \frac{10^{-m}}{2r \times 10^{n-1} \times 10^{-m} - 10^{-m}}$$

$$= \frac{1}{2r \times 10^{n-1} - 1} = \frac{1}{2\left(r \times 10^{n-1} - \dfrac{1}{2}\right)}$$

Now, since n is any positive integer and r stands for any digits 0, 1, ..., 9. Then we have $2r \times 10^{n-1} > r \times 10^{n-1}$ in all cases except $r = 1$ and $n = 1$. (We can ignore this case, because it is a trivial case when $N = 1, 0.001, 0.0001$ etc., *i.e.*, the case in which N contains only one digit different from zero, which would not occur in common practice). Therefore, we may assume that

$$2r \times 10^{n-1} - 1 > r \times 10^{n-1} \text{ for all cases}$$

Then, the relative error $E_r < \dfrac{1}{r \times 10^{n-1}}$

Case (II): If $m = n$

Here we have N is a decimal and r is the first decimal figure, then we have

the absolute error $\; E_a \le \dfrac{1}{10^m} \times \dfrac{1}{2}$

and $$N \geq r \times 10^{-1} - \frac{1}{10^m} \times \frac{1}{2}$$

$\Rightarrow$ $$E_r \leq \frac{10^{-m} \times \frac{1}{2}}{r \times 10^{-1} - 10^{-m} \times \frac{1}{2}}$$

$$= \frac{10^{-m}}{2r \times 10^{-1} - 10^{-m}} = \frac{1}{2r \times 10^{m-1} - 1}$$

$$= \frac{1}{2r \times 10^{m-1} - 1} < \frac{1}{r \times 10^{m-1}}$$

Case (III): If $m > n$

Here we have $m > n$, therefore, r occupies the $(m - n + 1)^{\text{th}}$ decimal place.

$\Rightarrow$ $$N \geq r \times 10^{-(m-n+1)} - \frac{1}{10^m} \times \frac{1}{2} \text{ and } E_a \leq \frac{1}{10^m} \times \frac{1}{2}$$

Therefore, $$E_r \leq \frac{10^{-m} \times \frac{1}{2}}{r \times 10^{-m} \times 10^{n-1} - 10^{-m} \times \frac{1}{2}}$$

$$= \frac{10^{-m}}{2r \times 10^{-m} \times 10^{n-1} - 10^{-m}}$$

$$= \frac{1}{2r \times 10^{n-1} - 1} < \frac{1}{r \times 10^{n-1}}$$

Here, we can say that the theorem is true in all the three possible cases.

REMARKS

- Except in the case of approximate numbers of the form $r(1.000...) \times 10^k$, in which r is the only digit from zero, the relative error is less than $\frac{1}{2r \times 10^{n-1}}$.
- If $r \geq 5$ then the given approximate number is not of the form $r(1.000...) \times 10^k$, then $E_r < \frac{1}{10^n}$; for in the case $2r \geq 10$ and therefore $2r \times 10^{n-1} \geq 10^n$.

THEOREM 2. ***If the relative error in an approximate number is less than $\left[\frac{1}{(r+1) \times 10^{n-1}}\right]$, the number is correct to n significant figures or at least is in error by less than a unit in the n^{th} significant figures.***

PROOF. Let us assume

N = The given number, *i.e.*, the exact value,

n = number of correct significant figure in N,

r = first significant figure in N,

k = number of digits in the integral part of N.

Then, we have

$n - k$ = number of decimal in N,

Also, given $N \leq (r + 1) \times 10^{k-1}$

Now, let the relative error

$$E_r < \frac{1}{(r+1) \times 10^{n-1}}$$

Then, we have the absolute error

$$E_a < (r+1)\times 10^{k-1}\times\frac{1}{(r+1)\times 10^{n-1}}=\frac{1}{10^{n-k}}$$

Now, $\frac{1}{10^{n-k}}$ is one unit in $(n–k)^{\text{th}}$ decimal places or in the n^{th} significant figure.

Therefore, the absolute error E_a is less than a unit in the n^{th} significant figure.

Now, let us suppose that the given number is pure decimal number. Also let k be the number of zero between the decimal point and the first significant figure. Then $(n + k)$ is equal to the number of decimals in N.

and $$N \le \frac{(r+1)}{10^{k+1}}$$

Therefore, if $E_r < \frac{1}{(r+1)\times 10^{n-1}}$ then, we have

$$E_a < \frac{(r+1)}{10^{k+1}}\times\frac{1}{(r+1)\times 10^{n-1}}=\frac{1}{10^{n+k}}$$

Now, $\frac{1}{10^{n+k}}$ is one unit in $(n+k)^{\text{th}}$ decimal places or in the n^{th} significant figure. Hence the absolute error E_a is less than a unit in the n^{th} significant figure.

REMARKS

- If $E_r < \frac{1}{[2(r+1)\times 10^{n-1}]}$, then E_a is less than half a unit in the n^{th} significant figures and the given number is correct to n^{th} significant figures.
- If the relative error of any number is not greater than $\frac{1}{(2\times 10^n)}$, the number is certainly correct to n significant figures.

The absolute error is always connected with the number of decimal places, whereas the relative error is connected with the number of significant figures.

Solved Examples

EXAMPLE 1. ***Verify the theorem* (1) *for the number* 875.32 *correct to five significant figures.***

SOLUTION. The given number $N = 875.32$

We observe that $r = 8$ and $n = 5$

Since, we have the absolute error $E_a \ngtr 0.01\times\frac{1}{2}=0.005$.

Therefore, the relative error $\le \frac{0.005}{875.32}=\frac{5}{875320}$

$$=\frac{1}{2\times 87532}<\frac{1}{2\times 80000}=\frac{1}{2\times 8\times 10^4}$$

$$<\frac{1}{8\times 10^4}=\left(\frac{1}{r\times 10^{n-1}}\right)$$

Hence, the theorem is verified.

EXAMPLE 2. ***Round off the numbers* 865250 *and* 37.46235 *to four significant figures and compute* E_a, E_r *and* E_p.** [MEERUT–2018; DELHI–2007]

SOLUTION. Here, the given numbers are (i) 865250 and (ii) 37.46235

(i) 865250

If we rounded off the given number to four significant figures, then we get 865200. Therefore, the absolute error

$$E_a = \left|x^T - x^A\right| = |865250 - 865200| = 50$$

Now, the relative error

$$E_r = \frac{E_a}{x^T} = \frac{50}{865250} = 5.78 \times 10^{-5}$$

Also, the percentage error

$$E_p = E_r \times 100 = 5.78 \times 10^{-5} \times 100 = 5.78 \times 10^{-3}.$$

(ii) 37.46235

If we rounded off the given number to four significant figures, then we get 37.46.

Then $$E_a = |37.46235 - 37.46| = 0.00235$$

$$E_r = \frac{E_a}{x^T} = \frac{0.00235}{37.46235} = 6.27 \times 10^{-5}$$

and $$E_p = E_r \times 100 = 6.27 \times 10^{-3}$$

EXAMPLE 3. ***If 0.333 is the approximate value of $\frac{1}{3}$, find the absolute, relative and percentage errors.*** [MEERUT–2011]

SOLUTION. Here, we have

Therefore, $$x^T = \frac{1}{3}, x^A = 0.333$$

(i) Absolute error

$$E_a = \left|x^T - x^A\right| = \left|\frac{1}{3} - 0.333\right| = \left|\frac{1}{3} - \frac{333}{1000}\right| = \frac{1}{3000} = 0.00033$$

(ii) Relative Error

$$E_r = \frac{E_a}{x^T} = \frac{0.00033}{1/3} = 0.00099$$

(iii) Percentage error

$$E_p = 100 \times E_r = 100 \times 0.00099 = 0.099$$

EXAMPLE 4. ***Let x = 0.005998. Find the relative error if x is truncated to three decimal digits.*** [UPTU MCA–2006; UPTU B.TECH.–2004]

SOLUTION. Given that $x = 0.005998 = 0.5998 \times 10^{-2}$.

Now, $x_a = 0.599 \times 10^{-2}$ (after truncating to three decimal places)

$$\text{Relative error} = \left|\frac{x - x_a}{x}\right| = \left|\frac{0.5998 \times 10^{-2} - 0.599 \times 10^{-2}}{0.5998 \times 10^{-2}}\right|$$

$$= 0.001337 = 0.133 \times 10^{-2}.$$

EXAMPLE 5. ***If 1.414 is used as an approximation to $\sqrt{2}$. Find the absolute and relative errors.*** [MEERUT–2012]

SOLUTION. We have

True value = $\sqrt{2}$ = 1.41421356

and approximate value = 1.414

Therefore, Error = True value – Approximate value

$= \sqrt{2} - 1.414 = 1.41421356 - 1.414 = 0.00021356$

Then, absolute error = $|0.00021356| = 0.21356 \times 10^{-3}$.

Finally, the relative error $= \dfrac{\text{Absolute error}}{\text{True value}} = \dfrac{0.21356\times10^{-3}}{\sqrt{2}} = 0.151\times10^{-3}$.

EXAMPLE 6. ***Find the sum $S = \sqrt{3} + \sqrt{5} + \sqrt{7}$ to 4 significant digits and find its absolute and relative errors.*** [MEERUT–2015]

SOLUTION. It is known that

$$\sqrt{3} = 1.732, \sqrt{5} = 2.236, \sqrt{7} = 2.646$$

$$\therefore \qquad S = 1.732 + 2.236 + 2.646 = 6.614$$

Now, absolute error $E_a = 0.0005 + 0.0005 + 0.0005 = 0.0015$

The total absolute error shows that the sum is correct to 2 decimal places, so S is correct to 3 significant figures only.

Thus, we take $S = 6.61$

Then, we have relative error $= \dfrac{0.0015}{6.61} = 0.0002$

EXAMPLE 7. ***It is required to obtain the roots of $x^2 - 2x + \log_{10}2$ to four decimal places. To what accuracy should $\log_{10}2$ be given?***

SOLUTION. The roots of the given equation are

$$x = \frac{2 \pm \sqrt{4 - 4\log_{10} 2}}{2} = 1 \pm \sqrt{1 - \log_{10} 2}$$

Then $$|\Delta x| = \frac{1}{2}\frac{\Delta(\log 2)}{\sqrt{1 - \log_{10} 2}} < 0.5\times10^{-4}$$

$$= \Delta(\log 2) < 2\times0.5\times10^{-4}(1 - \log 2)^{1/2} < 0.83604\times10^{-4}$$

$$= 8.3604 \times 10^{-5}$$

EXAMPLE 8. ***If $a = 10.00 \pm 0.05$, $b = 0.0356 \pm 0.0002$, $c = 15300 \pm 100$, $d = 62000 \pm 500$. Find the maximum value of absolute error in $a + b + c + d$.*** [MDU(BE)–2005]

SOLUTION. We have

Absolute error in $a = |\pm 0.05| = 0.05$
Absolute error in $b = |\pm 0.0002| = 0.0002$
Absolute error in $c = |\pm 100| = 100$
Absolute error in $d = |\pm 500| = 500$

Hence, the maximum absolute error in $a + b + c + d$

$$= 0.05 + 0.0002 + 100 + 500 = 600.0502$$

EXAMPLE 9. ***Three approximated values of number $\frac{1}{3}$ are given as 0.30, 0.33 and 0.34. Which of these three is the best approximation?***

SOLUTION. We know that the best approximation will be the one which has the least absolute error.

Here, true value $= \dfrac{1}{3} = 0.33333$

Case I. Approximate value = 0.30

$\therefore$ Absolute error = |True value – Approximate value| = $|0.33333 - 0.30|$
$= 0.03333$

Case II. Approximate value = 0.33

$\therefore$ Absolute error = |True value – Approximate value| = $|0.33333 - 0.33|$
$= 0.00333$

Case III. Approximate value = 0.34

$\therefore$ Absolute error = |True value – Approximate value| = |0.33333 – 0.34|
= |– 0.00667| = 0.00667

We observe that, absolute error is least in case II. Hence, 0.33 is the best approximation.

EXAMPLE 10. ***Given the solution of a problem as x_A = 35.25 with the relative error in the solution atmost 2%. Find, to four decimal digits, the range of values within which the exact value of the solution must lie.*** (UPTU MCA–2002)

SOLUTION. It is given that

(i) Maximum relative error in the solution = 2% = 0.02

(ii) Approximate value of the solution is x_A = 35.25.

Let x be the exact value of the solution, then as per given, we have

$$\left|\frac{x - x_A}{x}\right| < 0.02\,,\ i.e.,\left|1 - \frac{x_A}{x}\right| < 0.02$$

$$\Rightarrow \quad -0.02 < \left(1 - \frac{x_A}{x}\right) < 0.02$$

If $\left(1 - \frac{x_A}{x}\right) > -0.02$ then

$$-\frac{x_A}{x} > -1 - 0.02 \quad \Rightarrow \quad -\frac{x_A}{x} > -1.02$$

$$\Rightarrow \quad \frac{x_A}{x} < 1.02 \quad \Rightarrow \quad x_A < 1.02x\,.$$

$$\Rightarrow \quad x > \frac{x_A}{1.02} = \frac{35.25}{1.02} = 34.558823594$$

Also, if $\left(1 - \frac{x_A}{x}\right) < 0.02$, then we have

$$-\frac{x_A}{x} < -1 + 0.02 \quad \Rightarrow \quad -\frac{x_A}{x} > -0.98$$

$$\Rightarrow \quad \frac{x_A}{x} > 0.98 \quad \Rightarrow \quad x_A > 0.98x$$

$$\Rightarrow \quad x < \frac{x_A}{0.98} = \frac{35.25}{0.98} = 35.9693877551$$

Thus, we have

34.558823594 < x < 35.9693877551

Hence, the range of values within which the exact value of the solution lies, correct to four decimal places is given by

34.5588 < x < 35.9694.

EXERCISE 1.2

1. Round off the following numbers correct to four significant figures :

(i)	58.3643	(ii)	979.267
(iii)	7.7265	(iv)	0.065738
(v)	3.26425	(vi)	35.46735
(vii)	7326583000	(viii)	18.265101

2. Find the relative error if 2/3 is approximated to 0.667. [MEERUT–2013]

3. If the number r is correct to 3 significant digits, what will be the maximum relative error.

4. A carpenter measures a 10-foot beam to the nearest eighth of an inch and a mechanist measures a $\frac{1}{2}$ inch bolt to the nearest thousandth of an inch. Which measurement is more correct ?

5. The following numbers are all approximate and are correct as far as their last digit only.

Find their sum 136.421, 28.3, 321, 68.243, 17.482.

6. If the number p is correct to three significant digits, what will be the maximum relative error ?

7. The height of an observation tower was estimated to be 47 m whereas it's actual height was 45 m. Find the percentage relative error in the measurement.

8. If true value $= \dfrac{10}{3}$, approximate value = 3.33. Then, find absolute and relative errors.

9. Round off the number 75462 to four significant digits and then calculate the absolute error and percentage error. (UPTU–2004)

10. Find the relative error in taking $\pi = 3.141593$ as 22/7. (VTU–2007)

11. Suppose that you have a task of measuring the lengths of a bridge and a rivet, and come up with 9999 and 9 am, respectively. If the true values are 10,000 and 10 cm. respectively, compute the percentage relative error in each case. (PUNE–2004)

12. Given $a = 9.00 \pm 0.05$, $b = 0.0356 \pm 0.0002$, $c = 15300 \pm 100$, $d = 62000 \pm 500$. Find the maximum value of absolute error in $a + b + c + d$. (PTU–2001)

13. Find the absolute error and the relative error in the product of 432.8 and 0.12584 using four digit mantissa. (KERALA–2003)

Answers

1. (i) 58.36 (ii) 979.3 (iii) 7.726 (iv) 0.06574 (v) 3.264 (vi) 35.45 (vii) 7327×10^6 (viii) 18.26 **2.** 0.0005 **3.** 0.0005

4. Beam measurement **5.** 571 **6.** 0.0005 **7.** 4.44%

8. 0.003333, 0.000999 **9.** 0.7546; -0.0002×10^5; 0.00265 **10.** – 0.0004

11. 0.01%; 10% **12.** 600.0002 **13.** 0.17312; 0.0003178

1.9 INHERENT ERRORS

The errors which are already present in the statement of a problem before its solution are called Inherent errors. These types of errors arise either due to the given data being approximated or due to limitations of the mathematical measurements.

The inherent error contains two components.

1.9.1 DATA ERRORS

The data error arises when data are obtained by some experimental methods with limited accuracy and precision. This may be due to some special limitations in instrument or in reading.

1.9.2 CONVERSION ERRORS

The conversion error arise due to the limitations of the computer to store the data exactly. Generally, it occurs in the floating- point representation which retains only a specified number of digits. The digits which are not retain gives the round off error.

☛ REMARKS

- The inherent errors is also known as input errors.
- Data errors is also known as empirical errors.
- Conversion errors are also known as representation errors.

1.10 ROUNDING OFF ERROR

It occurs from the process of rounding off the numbers during the computations, *i.e.*, it occur when a fixed number of digits are used to represent exact numbers. Such types of errors are unavoidable in most of the calculations due to the limitations of the computing aids.

If a number x has the floating point representation of the form

$$x = d_1 d_2 \dots d_t d_{t+1} \dots \times B^e \qquad \dots(1)$$

where d_1, d$_2$,..., d_t ... are integers and satisfies $0 \le d_i \le B$ and e is the exponent. Then

Rounding a number can be done by the following two ways :

1.10.1 CHOPPING

Here, we neglect d_{t+1}, d_{t+2} ... in (1) and obtain the number = $d_1d_2 \ldots d_t \times B^e$

1.10.2 SYMMETRIC ROUNDING

Here the fractional part in (1) is written as $d_1d_2...d_td_{t+1}+\frac{1}{2}B$ and the first t digits are taken to write the floating point number.

For Example- *Find the sum of* 0.223×10^3 *and* 0.556×10^2 *and write the result in three digit mantissa.*

Solution. Here, the number of the smaller magnitude is adjusted so that its exponent is same as that of the number of larger magnitude. We have

$$\begin{array}{r} 0.2230\times10^3 \\ \underline{0.0556\times10^3} \\ 0.2786\times10^3 \end{array}$$

$$\Rightarrow \qquad \begin{cases} 0.278\times10^3, & \text{for chopping} \\ 0.279\times10^3, & \text{for rounding} \end{cases}$$

☛ **REMARKS**

- In chopping, the extra digits are dropped, which is called truncating the number.
- In symmetric round off method, the last retained significant digit is rounded up by 1 if the first discarded digit is larger or equal to 5, otherwise the last retained digits is unchanged.

  For example: The numbers 83.8893 becomes 83.89 and the number 86.6431 would become 86.64.

> The rounded off error can be reduced by retaining at least one more significant figure at each step than that given in the data and rounded off at the last step.

1.11 TRUNCATION ERROR

The truncation errors arises by using some approximations in place of an exact mathematical procedure.

For example- When we calculate the sine of an angle using the following series

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

Then, we can not use the infinite terms of above series. After a certain number of terms, we terminate the process. Then, an error which is introduced here, is called truncation error.

☛ **REMARKS**

- Truncation error is a type of algorithm error.
- In numerical computing, we used many iterative procedures, which are infinite. Therefore, a knowledge of the truncation error is very much important.
- This error can be reduced by using a better numerical model which increases the number of arithmetic operations.
- When we use a number of discrete steps in the solution of a differential equation, then the error which is introduced here, is called discretisation error.

Solved Examples

EXAMPLE 1. ***Obtain a second degree polynomial approximation to***

$$f(x) = (1 + x)^{1/2},\ x \in [0, 0.1]$$

Using the Taylor series expansion about $x = 0$. Use the expansion to approximate $f(0.05)$ and found the truncation error.

SOLUTION. Here, the given function is

$$f(x) = (1+x)^{1/2}$$

Then, we get

$$f(x) = (1+x)^{1/2} \Rightarrow f(0) = 1$$

$$f'(x) = \frac{1}{2}(1+x)^{-1/2} \Rightarrow f'(0) = \frac{1}{2}$$

$$f''(x) = -\frac{1}{4}(1+x)^{-3/2} \Rightarrow f''(0) = -\frac{1}{4}$$

$$f'''(x) = \frac{3}{8}(1+x)^{-5/2} \Rightarrow f'''(0) = \frac{3}{8}$$

Now, using the Taylor series expansion, we get

$$(1+x)^{1/2} = 1 + \frac{x}{2} - \frac{x^2}{8} + R_n$$

where R_n is the remainder term and given by

$$R_n = \frac{1}{16} \cdot \frac{x^3}{[(1+\theta)^{1/2}]^5}, 0 < \theta < 0.01$$

Then the truncation error is given by

$$T = (1+x)^{1/2} - \left(1 + \frac{x}{2} - \frac{x^2}{8}\right) = \frac{1}{16} \cdot \frac{x^3}{[(1+\theta)^{1/2}]^5}$$

Now,

$$f(0.05) = 1 + \frac{0.05}{2} - \frac{(0.05)^2}{8} = 0.10246875 \times 10^1$$

Then, the bound of the truncation error for $x \in [0, 1]$ is given by

$$|T| \le \frac{(0.1)^3}{16[(1+8)^{1/2}]^5} \le \frac{(0.1)^3}{16} = 0.625 \times 10^{-4}$$

EXAMPLE 2. ***Find the truncation error in the result of the following functions for $x = \frac{1}{5}$ when we use***

(a) ***First three terms*** **(b)** ***First four terms***

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!}$$

SOLUTION. (a) Let T denote the truncation error. If we add first three terms then

$$T = \left(1 + x + \frac{x^2}{2!} + \ldots + \frac{x^6}{6!}\right) - \left(1 + x + \frac{x^2}{2!}\right) = \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!}$$

Now, T at $x = \frac{1}{5} = \frac{(0.2)^3}{6} + \frac{(0.2)^4}{24} + \frac{(0.2)^5}{120} + \frac{(0.2)^6}{720} = 0.1402755 \times 10^{-2}$

(b) Now, we find the truncation error, when first four terms are added

$$T = \left(1 + x + \frac{x^2}{2!} + \ldots + \frac{x^6}{6!}\right) - \left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}\right) = \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!}$$

Now, T at $x = \frac{1}{5} = \frac{(0.2)^4}{24} + \frac{(0.2)^5}{120} + \frac{(0.2)^6}{720} = 0.694222 \times 10^{-4}$

1.12 THE GENERAL FORMULA FOR ERRORS

Let $Y = f(x_1, x_2, ..., x_n)$ be a function of n variables $x_1, x_2, ..., x_n$. Suppose, ΔY is the error in Y due to the errors $\Delta x_1,\ \Delta x_2, ..., \Delta x_n$ in $x_1, x_2, ..., x_n$ respectively.

Then we have

$$Y + \Delta Y = f(x_1 + \Delta x_1, x_2 + \Delta x_2, ..., x_n + \Delta x_n) \qquad \ldots(1)$$

Expanding by Taylor series, we get

$$Y + \Delta Y = f(x_1, x_2, ..., x_n) + \left(\Delta x_1 \frac{\partial Y}{\partial x_1} + \Delta x_2 \frac{\partial Y}{\partial x_2} + \ldots + \Delta x_n \frac{\partial Y}{\partial x_n}\right)$$
$$+ \frac{1}{2}\left[(\Delta x_1)^2 \frac{\partial^2 Y}{\partial x_1^2} + (\Delta x_2)^2 \frac{\partial^2 Y}{\partial x_2^2} + \ldots + (\Delta x_n)^2 \frac{\partial^2 Y}{\partial x_n^2} + 2\Delta x_1 \Delta x_2 \frac{\partial^2 Y}{\partial x_1 \partial x_2} + \ldots\right] + \ldots \qquad \ldots(2)$$

Now, since the errors $\Delta x_1, \Delta x_2, ..., \Delta x_n$ all are very small. So, that we can neglect $(\Delta x_i)^2$ and higher order terms of Δx_i.

Then, we have

$$Y + \Delta Y = f(x_1, x_2, ..., x_n) + \left(\Delta x_1 \frac{\partial Y}{\partial x_1} + \Delta x_2 \frac{\partial Y}{\partial x_2} + \ldots + \Delta x_n \frac{\partial Y}{\partial x_n}\right) \qquad \ldots(3)$$

$$\Rightarrow \qquad \Delta Y = \Delta x_1 \frac{\partial Y}{\partial x_1} + \Delta x_2 \frac{\partial Y}{\partial x_2} + \ldots + \Delta x_n \frac{\partial Y}{\partial x_n} \qquad \ldots(4)$$

$[\because\ Y = f(x_1, x_2, ..., x_n)]$

Now, divide the equation (4) by Y, we get the relative error as

$$\frac{\Delta Y}{Y} = \frac{\partial x_1}{Y} \cdot \frac{\partial Y}{\partial x_1} + \frac{\partial x_2}{Y} \cdot \frac{\partial Y}{\partial x_2} + \ldots + \frac{\partial x_n}{Y} \cdot \frac{\partial Y}{\partial x_n} \qquad \ldots(5)$$

Now, taking the modulus of (4) and (5), the maximum absolute error and relative error are given by

$$|\Delta Y| \le \left|\Delta x_1 \frac{\partial Y}{\partial x_1}\right| + \left|\Delta x_2 \frac{\partial Y}{\partial x_2}\right| + \ldots + \left|\Delta x_n \frac{\partial Y}{\partial x_n}\right|$$

and
$$\left|\frac{\Delta Y}{Y}\right| \le \left|\frac{\Delta x_1}{Y} \cdot \frac{\partial Y}{\partial x_1}\right| + \left|\frac{\Delta x_2}{Y} \cdot \frac{\partial Y}{\partial x_2}\right| + \ldots + \left|\frac{\Delta x_n}{Y} \cdot \frac{\partial Y}{\partial x_n}\right|$$

Solved Examples

EXAMPLE 1. ***In a $\triangle ABC$, $a = 6$ cm, $c = 15$ cm and $\angle B = 90°$. Find the possible error in the computed value of A, if the errors in the measurement of a and c are 1 mm and 2 mm respectively.***

SOLUTION. Here, we have $a = 6$ cm

$c = 15$ cm

$\angle B = 90°$

Then, we have the triangle given by fig. 1.

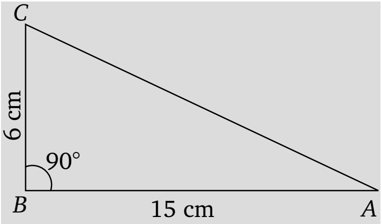


Fig. 1

From figure 1, we have $A = \tan^{-1}\frac{a}{c}$

$$\Rightarrow \qquad \Delta A = \Delta a \frac{\partial A}{\partial a} + \Delta c \frac{\partial A}{\partial c}$$

$$= (\Delta a) \cdot \frac{c}{(a^2 + c^2)} - \frac{a}{(a^2 + c^2)} \cdot \Delta c \qquad \ldots(1)$$

or $$|\Delta A| \le \left|\Delta a \cdot \frac{c}{a^2+c^2}\right| + \left|\Delta c \cdot \frac{a}{a^2+c^2}\right|$$

Given that $\Delta a = 1$ mm $= 0.1$ cm, $\Delta c = 2$ mm $= 0.2$ cm, $a = 6$ cm and $c = 15$ cm. Putting all these values in equation (1), we get

$$|\Delta A| \le \left|\frac{0.1\times 15}{(6)^2+(15)^2}\right| + \left|\frac{0.2\times 6}{(6)^2+(15)^2}\right| = \frac{1.5+1.2}{261} = \frac{2.7}{261} = 0.0103 \text{ Radians}$$

$\Rightarrow$ $|\Delta A| \le 0.0103$ radians

or $|\Delta A| \le 35'25''$

EXAMPLE 2. ***If $u = \frac{4x^2y^3}{z^4}$ and $\Delta x = \Delta y = \Delta z = 0.001$, compute the relative maximum error in u when $x = y = z = 1$.*** [MEERUT–2018; PURVANCHAL–2012]

SOLUTION. Here, we have $a = 6$ cm

$$u = \frac{4x^2y^3}{z^4} \qquad \ldots(1)$$

From eq. (1), we have

$$\frac{\partial u}{\partial x} = \frac{8xy^3}{z^4}, \frac{\partial u}{\partial y} = \frac{12x^2y^2}{z^4} \text{ and } \frac{\partial u}{\partial z} = -\frac{16x^2y^3}{z^5}$$

Now, we have

$$\Delta u = \frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y + \frac{\partial u}{\partial z}\Delta z \qquad \ldots(2)$$

Now, putting the values of $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}$ and $\frac{\partial u}{\partial z}$ in eq. (2), we get

$$\Delta u = \frac{8xy^3}{z^4}\Delta x + \frac{12x^2y^2}{z^4}\Delta y - \frac{16x^2y^3}{z^5}\Delta z$$

Now, $$(\Delta u)_{\max} = \left|\frac{8xy^3}{z^4}\Delta x\right| + \left|\frac{12x^2y^2}{z^4}\Delta y\right| + \left|\frac{16x^2y^3}{z^5}\Delta z\right|$$

$$= 8(0.001) + 12(0.001) + 16(0.001) = 0.036$$

Therefore, the maximum relative error is

$$= \frac{(\Delta u)_{\max}}{(u)_{\text{at } x=y=z=1}} = \frac{0.036}{4} = 0.009$$

EXAMPLE 3. ***In a ΔABC, $a = 30$ cm, $b = 80$ cm, $\angle B = 90°$. Find the maximum error in the computed value of A, if possible errors in a and b are $\frac{1}{3}$% and $\frac{1}{4}$% respectively.***

SOLUTION. Here, we have

In ΔABC, $a = 30$ cm, $b = 80$ cm, $\angle B = 90°$

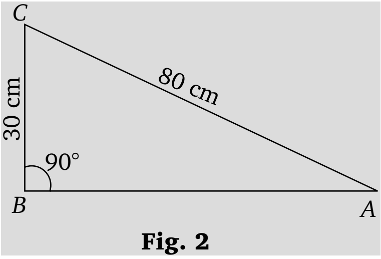


Fig. 2

From figure 2, we have

$$\sin A = \frac{a}{b}$$

$$\Rightarrow \quad A = \sin^{-1}\frac{a}{b} \qquad \ldots(1)$$

Therefore, we have

$$|\Delta A| < \left|\Delta a \cdot \frac{\partial A}{\partial a}\right| + \left|\Delta b \cdot \frac{\partial A}{\partial b}\right| \qquad \ldots(2)$$

Now, we have the possible errors in a and b are 1/3% and 1/4% respectively, then

$$\frac{\Delta a}{a}\times 100=\frac{1}{3} \quad \Rightarrow \quad \Delta a = 0.1$$

and $\frac{\Delta b}{b}\times 100=\frac{1}{4} \quad \Rightarrow \quad \Delta b = 0.2$

Also, from equation (1)

$$\frac{\partial A}{\partial a}=\frac{1}{\sqrt{b^2-a^2}} \text{ and } \frac{\partial A}{\partial b}=\frac{a}{b\sqrt{b^2-a^2}}.$$

Putting all these values in equation (2), we get

$$|\Delta A| < |0.00135 + 0.00100| = 0.00235 \text{ radians}$$

$$\Rightarrow \quad \Delta A < 8'5''$$

EXAMPLE 4. ***Find the relative error in the function*** $y = ax_1^{m_1}x_2^{m_2}\ldots x_n^{m_n}$

SOLUTION. Here, we have

$$y = ax_1^{m_1}x_2^{m_2}\ldots x_n^{m_n} \qquad \ldots(1)$$

Taking log of both sides, we get

$$\log y = \log a + m_1 \log x_1 + m_2 \log x_2 + \ldots + m_n \log x_n \qquad \ldots(2)$$

Now, differentiating eq.(2), we get

$$\frac{1}{y}\cdot\frac{\partial y}{\partial x_1}=\frac{m_1}{x_1}$$

$$\frac{1}{y}\cdot\frac{\partial y}{\partial x_2}=\frac{m_2}{x_2},\ldots\frac{1}{y}\cdot\frac{\partial y}{\partial x_n}=\frac{m_n}{x_n}$$

Therefore, the error

$$E_r=\frac{\partial y}{\partial x_1}\cdot\frac{\Delta x_1}{y}+\frac{\partial y}{\partial x_2}\cdot\frac{\Delta x_2}{y}+\ldots+\frac{\partial y}{\partial x_n}\cdot\frac{\Delta x_n}{y}$$

$$=m_1\frac{\Delta x_1}{x_1}+m_2\frac{\Delta x_2}{x_2}+\ldots+m_n\frac{\Delta x_n}{x_n}$$

Hence, $(E_r)_{\max} \le m_1\left|\frac{\Delta x_1}{x_1}\right|+m_2\left|\frac{\Delta x_2}{x_2}\right|+\ldots+m_n\left|\frac{\Delta x_n}{x_n}\right|$

☛ **REMARK**

➠ The relative error of a product of n numbers is approximately equal to the algebraic sum of their relative errors. This result can be verified easily by taking $a = 1$, $m_1 = m_2 = \ldots = m_n = 1$, then

$$E_r=\frac{\Delta x_1}{x_1}+\frac{\Delta x_2}{x_2}+\ldots\frac{\Delta x_n}{x_n}.$$

1.13 FLOATING POINT ARITHMETIC AND ERRORS

Generally, there are two types of numbers, which we used in calculations

(i) Integers : 0, ±1, ±2, ±3, ...

(ii) Real numbers : Such as numbers with decimal.

Since, we used finite digit arithmetic in computers, therefore all the integers can be represented easily with finite digits. On the other hand, all real numbers can not be represented as a finite digits numbers like $\left(\frac{2}{3}\right)= 0.666\ldots$ Hence, we use floating point representation.

(iii) Floating Point Numbers:

An n digit floating point number β has the form

$$x = \pm(d_1d_2\ldots d_n)_\beta\cdot\beta^e,\ 0 \le d_i < \beta,\ m \le e \le M$$

where $(d_1d_2...d_n)_\beta$ is a β fraction called mantissa and its value is given by

$$(d_1d_2...d_n)_\beta = d_1 \times \frac{1}{\beta} + d_2 \times \frac{1}{\beta^2} + ... + d_n \times \frac{1}{\beta^n}$$

Also e is called the exponent.

☛ **REMARKS**

- A floating point number is said to be normalised if $d_1 \neq 0$ or else $d_1 = d_2 = ... = d_n = 0$.
- The precision or length n of floating-point numbers on any computer is usually determined by the word length of the computer. **For example:** IBM 1130, in single precision 6 decimal digits and inextend precision, *i.e.*, double precision, nine decimal digits are used.
- Calculation in double precision usually doubles the storage requirements and running time as compared with single precision.
- The exponent e is also limited to range $m<e<M$, where m and M are integers varying from computer to computer.

1.14 COMPUTER STORAGE

Computer storage has its own limitations. Storage is provided into locations. Each location or word has a storage capacity which means a finite number of digits. The limitation causes errors and concept of floating point becomes more important. To discuss it, we must keep in mind the constants of number of digits that can be stored in one word or location *i.e.*, it would be very difficult to store a number as 1, 2, 3, 4, ..., 10.

The solution to this problem to some extent can be used of floating point, *i.e.*, representation of this number to same digits of accuracy and with power of 10. For example, say representing this number to 4 digits of accuracy as 1.234×10^9.

Although, these two are not same, yet second option will be significantly accurate for most application purpose.

To convert to floating point, the major concern is number of digits of accuracy to return.

To discuss this concept let us assume that each location can store 6 digits:

•	•	•	•	•	•

Location or word

Initially we can assume, first 3 digits represents integer portion of a fractional number and last 3 as fractional part. **For example:** to store 123.456

1	2	3	4	5	6

↑assumed decimal position

Decimal point is assumed in middle and this sign does not exist physically. In this system range is very limited. Tracking of decimal point will be more difficult in this system as we perform mathematical operations like +, –, *, /.

Range is ±999.999 to 000.001.

To improve this range concept, most usual representation is to use 4 digits for integers and 2 for floating, *i.e.*, 1234.56 is stored as

1	2	3	4	5	6

↑assumed decimal position

Range is increased from 9999.99 to 0000.01 still is very inadequate for most of computations. To remove this problem we use concept of floating point in power notation form.

For example : 1234.56 is *represented* as 0.1234×10^7 *and written as* 1234 E07 *is i.e.*,

1	2	3	4	07

↑

Clearly range is increased

$$0.9999 \times 10^{99} \text{ to } 0.1000 \times 10^{-99}$$

This is much larger. Problem still arise as sign is not a available. If sign bit is used then representation of negative numbers will be reduced to 10^{-9} only as one bit will be consumed as sign bit. To avoid this a concept of Excess method is used. This is a split range of exponent with 50 as base from 00 to 99.

50 is centre so all exponent > 50 are positive and < 50 are negative. Range will be from –50 to 49.

Excess –50. Method says add 50 to exponent.

For example: 0.123456×10^3 *will be stored as*

1	2	3	4	5	6

↑

1.15 CONCEPT OF NORMALIZED FLOATING POINT

Consider a number 0.001234×10^{-5}, which is to be stored. It will be stored as

0	0	1	2	45

↑

We loose 2 significant digits. If we represent this number as 0.1234×10^{-7}, the storage will be which is much reliable representation.

1	2	3	4	43

↑

So removing zeroes in beginning is termed as normalized floating. In normalized floating range is further increased.

1.16 PITFALLS OF FLOATING POINT REPRESENTATION

We know that mantissa have to be truncated to four digits in order to fit into the normalized floating-point format of the hypothetical system.

For example.

$$4x = x + x + x + x \qquad ...(1)$$

When arithmetic is performed using normalized floating point representation, equation (1) may not hold true.

Solved Examples

EXAMPLE 1. ***(i) Add 0.1234 × 10^{-3} and 0.5678 × 10^{-3} using concept of normalized floating point.***

SOLUTION. We have $0.1234 \times 10^{-3} + 0.5678 \times 10^{-3}$

$$\begin{array}{r} \Rightarrow 0.1234\ E3 \\ +\ 0.5678\ E3 \\ \hline 0.6912\ E3 \end{array} \quad \Rightarrow \quad 0.6912 \times 10^{-3}$$

6	9	1	2	47

↑

(***ii***) ***Add*** $0.2315 \times 10^{2} + 0.9443 \times 10^{2}$

$$\begin{array}{r} \Rightarrow 0.2315\ E02 \\ +\ 0.9443\ E02 \\ \hline 1.1758\ E02 \end{array} \quad \Rightarrow \quad 0.1175 \times 10^{3}$$

1	1	7	5	53

↑

(***iii***) ***For different base*** $0.1234 \times 10^{3} + 0.4567 \times 10^{2}$

$$\begin{array}{r} \Rightarrow 0.1234\ E3 \\ +\ 0.4567\ E2 \\ \hline 0.5801\ E3 \end{array}$$

Make base as same

$$\begin{array}{r} \Rightarrow 0.1234\ E3 \\ +\ 0.0456\ E3 \\ \hline 0.1690\ E3 \end{array} \quad \Rightarrow \quad 0.1690 \times 10^{3}$$

1	6	9	0	53

↑

EXAMPLE 2. ***Subtract the following :***

(i) $0.4567 \times 10^{8} - 0.1234 \times 10^{8}$

$$\begin{array}{r} 0.4567\ E8 \\ \underline{0.1234\ E8} \\ 0.3333\ E8 \end{array} \Rightarrow \quad 0.3333 \times 10^{8}$$

(ii) Different base $0.4567 \times 10^{8} - 0.1234 \times 10^{7}$

$$\begin{array}{l} 0.4567\ \ E8 \\ 0.1234\ \ E7 \Rightarrow \begin{array}{r} 0.4567\ E8 \\ \underline{0.0123\ E8} \\ 0.4444\ E8 \end{array} \quad \Rightarrow \quad 0.4444 \times 10^{8} \end{array}$$

(iii) Normalized answer $0.4567 \times 10^{8} - 0.4566 \times 10^{8}$

$$\begin{array}{r} 0.4567\ E8 \\ \underline{0.4566\ E8} \\ 0.0001\ E8 \end{array} \Rightarrow \ 0.1 \times 10^{5}$$

(iv) Condition of overflow :

$$0.4568 \times 10^{49}$$

$$0.7767 \times 10^{49}$$

$$\begin{array}{r} 0.4568\quad E49 \\ \underline{0.7767\quad E49} \\ 0.12335\quad E49 \end{array} \Rightarrow \quad 0.1233 \times 10^{50} \text{ over flow}$$

(v) Condition of underflow:

$$\begin{array}{r} 0.4567\ E52 \\ \underline{0.4500\ E52} \\ 0.0067\ E52 \end{array} \Rightarrow \quad 0.67 \times 10^{-52}$$

↓ under flow

☛ **REMARKS**

- In multiplication, exponents are added and mantissa multiplied. If added expanded >99 overflow

  For example: Multiply 0.55432 * 0.4111 E7

  = 0.22787273*E9

  decreased

  = 0.22789 E9

- In division, exponents are subtracted

  For example: Divide 0.9380 E5 by 0.3500 E2

  $$= \frac{0.9380\,E5}{0.3500\,E5}$$

  = 0.2680 E3

EXAMPLE 3. ***Apply the procedure of multiplication of two floating point numbers for the following multiplications*** :

$$(\mathbf{0.5334 \times 10^{9}}) \times (\mathbf{0.1132 \times 10^{-25}})$$

and $$(\mathbf{0.1111 \times 10^{74}}) \times (\mathbf{0.2000 \times 10^{80}})$$

indicate if the result is overflow or underflow.

SOLUTION. The procedure for multiplication of two floating point numbers is

(i) multiply the mantissas of the two normalized floating point numbers.

(ii) and their exponents.

(iii) Resultant mantissa is normalized.

Therefore, $(0.5334 \times 10^{9}) \times (0.1132 \times 10^{-25})$

$$= (0.5334) \times (0.1132) \times (10^{9} \times 10^{-25})$$

$$= 0.06038038 \times 10^{-16}$$

$$= (0.6038\,E\,{-17})$$

and $(0.1111 \times 10^{71}) \times (0.20000 \times 10^{80})$

$$= (0.1111) \times 9.20000 \times (10^{74} \times 10^{80})$$

$$= (0.02222) \times 10^{154}$$

$$= (0.2222\,E\,153)$$

Since exponent is greater than 99, therefore, the result is "overflow".

EXAMPLE 4. ***In normalized floating point mode, carry out the following mathematical operations***

(i) **(0.4546 *E*3) + (0.5454 *E*8)**

(ii) **(0.9432 *E* – 4) – (0.6353 *E* – 5)**

SOLUTION. We have

(i)
$$\begin{array}{r} 0.5454\,E8 \\ +0.0000\,E8 \\ \hline 0.5454\,E8 \end{array}$$
$(\because\ 4546\,E\,3 = 0.0000\,E8)$

(ii)
$$\begin{array}{r} 0.9432\,E\ -4 \\ -0.0635\,E\ -4 \\ \hline 0.8797\,E\ -4 \end{array}$$
$(\because\ 6353\,E\,{-5} = 0.0635\,E - 4)$

EXAMPLE 5. ***Multiplying the following floating point number*** **0.1111 *E*10 *and* 0.1234 *E*15.**

SOLUTION. We have 0.1111 E 10 × 0.1234 E 15 = 0.1370 E 24.

EXAMPLE 6. ***For e = 2.7183 calculate the value of e^x when x = 0.5250 E1, where***

$$\boldsymbol{e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots}$$

SOLUTION. We have $e^{0.5250\,E1} = e^5 \times e^{0.25}$

Now $e^5 = (0.2718\,E1)\times(0.2718\,E1)\times(0.2718\,E1)\times(0.2718\,E1)\times(0.2718\,E1)$

$= 0.1484\,E3.$

Also, $e^{0.25} = 1 + (0.25) + \frac{(0.25)^2}{2!} + \frac{(0.25)^3}{3!}$

$= 1.25 + 0.03125 + 0.002604 = 0.1284\,E1$

Therefore,

$$e^{0.5250\,E1} = (0.1484\,E3)\times(0.1284\,E1) = (0.1905\,E3)$$

EXAMPLE 7. ***Find the smallest root of equation x^2 – 400x + 1 = 0 using four digit arithmetic.***

SOLUTION. It is known that, roots of equation $ax^2 - bx + c = 0$ are

$$\frac{b - \sqrt{b^2 + 4ac}}{2a} \text{ and } \frac{b - \sqrt{b^2 - 4ac}}{2a}$$

Also, product of roots are $\frac{c}{a}$.

∴ smaller root is

$$\frac{c/a}{\left(\dfrac{b + \sqrt{b^2 - 4ac}}{2a}\right)} = \frac{2c}{b + \sqrt{b^2 - 4ac}}$$

Here, $a = 1 = 0.1000\,E1, b = 400 = 0.4000\,E3, c = 1 = 0.1000\,E1$

Now, $b^2 - 4ac = 0.1600\,E6 - 0.4000\,E1 = 0.1600\,E6$

$\Rightarrow \sqrt{b^2 - 4ac} = 0.4000\,E3$

Hence, smaller root $= \frac{2\times(0.1000\,E1)}{0.4000\,E3 + 0.4000\,E3} = \frac{0.2000\,E1}{0.8000\,E3} = 0.25\,E-2 = 0.0025$

EXAMPLE 8. ***Determine the number of terms of the exponential series.***

$$\boldsymbol{e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \ldots}$$

such that this gives the values of e^x correct to six decimal places for 0 ≤ x ≤ 1. [ROHILKHAND–2004, 10]

SOLUTION. Given that $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^{n-1}}{(n-1)!} + R_n(x)$

Where $R_n(x) = \frac{x^n}{n!} e^{\theta}, 0 < \theta < x$

Max, absolute error (at $\theta = x$) $= \frac{x^n}{n!} e^x$

and the maximum relative error $= \frac{x^n}{n!}$

Hence $(E_r)_{\max}$ at $x = 1 = \frac{1}{n!}$

For a six decimal accuracy at $x = 1$ we have

$$\frac{1}{n!} < \frac{1}{2} \times 10^{-6} \text{ or } n! > 2 \times 10^{6}$$

Which gives $n = 10$

EXAMPLE 9. ***In case of normalized floating point representations, associative and distributive laws are not always valid. Give example to prove the statement.***

Or

If the normalization on floating point is carried out at each stage, prove the following

(i) $a(b - c) + ab - ac$, where $a = 0.5555\ E1$, $b = 0.4545\ E1$, $c = 0.4535\ E1$.

(ii) $(a + b) - c \neq (a - c) + b$, where $a = 0.5565\ E - 1$, $b = 0.5556\ E - 1$, $c = 0.5644\ E1$.

SOLUTION. In normalized floating point representations, the associative and the distributive laws of arithmetic are not always valid.

Consider the following examples:

Non-distributivity of Arithmetic

Since $a = 0.5555\ E1$, $b = 0.4545\ E1$, $c = 0.4535\ E1$

$\therefore \quad (b - c) = 0.0010\text{ E1} = 0.1000\ E{-1}$

$\Rightarrow \quad a(b - c) = (0.5555\ E1) \times (0.1000\ E{-1})$

$= (0.0555\ E0) = 0.5550\ E{-1}$

Also, $\quad ab = (0.5555\ E1) \times (0.4545\ E1)$

$= 00.2524\ E2$

$ac = (0.555\ E1) \times (0.4535\ E1)$

$= 0.2519\ E2$

$\Rightarrow \quad a(b - c) \neq ab - ac$

Non-Associativity of Arithmetic

Let $a = 0.5665\ E1$, $b = 0.5556\ E{-1}$, $c = 0.5644\ E1$

Therefore, $\quad (a + b) = 0.5665\ E1 + 0.5556\ E{-1}$

$= 0.5665\ E1 + 0.0055\ E1 = 0.572\ E1$

$\therefore \quad (a + b) - c = 0.5720\ E1 - 0.5644\ E1$

$= 0.0076\ E1$

$= 00.7600\ E{-1}$

$(a - c) = 0.5665\ E1 - 0.5644\ E1$

$= 00.0021\ E1 = 0.2100\ E{-1}$

$\Rightarrow \quad (a - c) + b = 0.2100\ E{-1} + 0.5556\ E{-1}$

$= 0.7656\ E{-1}$

$\Rightarrow \quad (a + b) - c \neq (a - c) + b$

EXAMPLE 10. ***Calculate the value of polynomial $x^3 - 4x^2 + 0.1x - 0.5$ for $x = 4.011$, using floating point arithmetic with 4 digit mantissa in two different ways. Find the relative errors in the two methods.***

SOLUTION. We have $x = 4.011$

Value of x in floating point representation is

$$x = 0.4011\,E1$$

Now value of given polynomial in real arithmetic is

$$\begin{aligned} x^3 - 4x^2 + 0.1x - 0.5 &= (4.011)^3 - 4(4.011)^2 + 0.1(4.011) - 0.5 \\ &= 64.529453 - 4(16.088121) + (0.4011) - 0.5 \\ &= 0.0780693 \qquad \ldots(1) \end{aligned}$$

Now, in normalised floating point

$$\begin{aligned} x^3 - 4x^2 + 0.1x - 0.5 &= x{\cdot}x{\cdot}x - 4{\cdot}x{\cdot}x + 0.1x - 0.5 \\ &= (0.4011\,E1)(0.4011\,E1)(0.4011\,E1) - 4(0.4011\,E1) \\ &\qquad (0.4011\,E1) + 0.1(0.4011\,E1) - 0.5000\,E0 \\ &= 0.6452\,E2 - 0.6435\,E2 + 0.4011\,E0 - 0.5000\,E0 \\ &= 0.0017\,E2 - 0.0989\,E0 \\ &= 0.1700\,E0 - 0.0989\,E0 \\ &= 0.0711\,E0 \qquad \ldots(2) \end{aligned}$$

Now relative error in two methods

$$= (1) - (2) = 0.0780 - 0.0711 = 0.0069$$

1.17 ERROR IN A SERIES APPROXIMATION

The Taylor's series for $f(x)$ at $x = a$ is

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \ldots + \frac{(x-a)^{n-1}}{(n-1)!}f^{n-1}(a) + R_n(x)$$

where $R_n(x)$ is the remainder term and given by

$$R_n(x) = \frac{(x-a)^n}{n!}f^n(\theta), a < \theta < x$$

Here, if the series is convergent, $R_n(x) \to 0$ as $n \to \infty$. Now, if $f(x)$ is approximated by the first n terms of this series, then the maximum error will be given by the $R_n(x)$. Also if the accuracy required in a series approximation is preassigned, then we can find the number of terms which gives the desired accuracy.

1.17.1 SERIES WITH REMAINDER TERMS

(1) The Binomial series

$$(1+x)^m = 1 + m\cdot x + \frac{m(m-1)}{2!}x^2 + \frac{m(m-1)(m-2)}{3!}x^3 + \ldots + \frac{m(m-1)\ldots(m-n+2)}{(n-1)!}x^{n-1} + R_n$$

where

(a) $R_n = \dfrac{m(m-1)(m-2)\ldots(m-n+1)}{n!}x^n(1+\theta x)^{m-n}, 0 < \theta < 1$

(b) If $x > 0$ then $R_n < \left|\dfrac{m(m-1)\ldots(m-n+1)}{n!}\cdot x^n\right|$

(c) If $x < 0$ and $n > m$ then $R_n < \left|\frac{m(m-1)(m-2)\ldots(m-n+1)}{n!}\cdot\frac{x^n}{(1+x)^{n-m}}\right|$

(2) Exponential Series

(a) $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^{n-1}}{(n-1)!} + R_n$ with $R_n = \frac{x^n}{n!}e^{\theta x}$ [MDU(BE)–2005]

In general $e < 3$ and $\theta \le 1$

$$\Rightarrow \qquad R_n < \frac{3}{n!}$$

(3) Logarithmic Series

$$\log_e(m+1) = \log_e m + 2\left(\frac{1}{2m+1} + \frac{1}{3(2m+1)^3} + \frac{1}{5(2m+1)^5} + \ldots + \frac{1}{(2n-1)(2m+1)^{2n-1}}\right) + R_n$$

where $R_n = 2\left[\frac{1}{(2n+1)(2m+1)^{2n+1}} + \frac{1}{(2n+3)(2m+1)^{2n+3}} + \ldots\right]$

Also, we have $R_n < \frac{1}{2}\cdot\frac{1}{m(m+1)(2n+1)(2m+1)^{2n-1}}$

(4) Series a^x

$$a^x = 1 + x\log a + \frac{(x\log a)^2}{2!} + \ldots + \frac{(x\log a)^{n-1}}{(n-1)} + R_n \quad \text{where} \quad R_n = \frac{(x\log a)^n}{n!}a^{\theta x}$$

1.18 ERROR IN DETERMINANTS

If the elements of a determinant are not exact due to rounding or otherwise, then the value of the determinant may be seriously affected, due to the loss of some important significant figures. The amount of such type of losses can not be determined in advance. Here we determine the upper limit of the error in a determinant as follows:

Let us define a determinant as

$$D = \begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{vmatrix} \qquad \ldots(1)$$

Now, let Δx_i, Δy_i and Δz_i are the errors in x_i, y_i and z_i respectively and ΔD is the error in D, then we have

$$D + \Delta D = \begin{vmatrix} x_1+\Delta x_1 & x_2+\Delta x_2 & x_3+\Delta x_3 \\ y_1+\Delta y_1 & y_2+\Delta y_2 & y_3+\Delta y_3 \\ z_1+\Delta z_1 & z_2+\Delta z_2 & z_3+\Delta z_3 \end{vmatrix} \qquad \ldots(2)$$

From eq.(1), we have

$$dD = \begin{vmatrix} dx_1 & x_2 & x_3 \\ dy_1 & y_2 & y_3 \\ dz_1 & z_2 & z_3 \end{vmatrix} + \begin{vmatrix} x_1 & dx_2 & x_3 \\ y_1 & dy_2 & y_3 \\ z_1 & dz_2 & z_3 \end{vmatrix} + \begin{vmatrix} x_1 & x_2 & dx_3 \\ y_1 & y_2 & dy_3 \\ z_1 & z_2 & dz_3 \end{vmatrix}$$

$$\Rightarrow \qquad \begin{aligned} dD = {} & (y_2z_3 - y_3z_2)dx_1 - (x_2z_3 - x_3z_2)dy_1 + (x_2y_3 - x_3y_2)dz_1 \\ & - (y_1z_3 - y_3z_1)dx_2 + (x_1z_3 - x_3z_1)dy_2 - (x_1y_3 - x_3y_1)dz_2 \\ & + (y_1z_2 - y_2z_1)dx_3 - (x_1z_2 - x_2z_1)dy_3 - (x_1y_2 - x_2y_1)dz_3 \end{aligned} \qquad \ldots(3)$$

Here, we observe that, the maximum possible error would occur when the signs of the elements and the signs of the errors are such that all the eighteen terms in equation (3) are of the same sign.

Now, equation (3) shows that the error in a determinant composed of non-exact elements may be anything from zero upto a number of sufficient magnitude.

1.19 APPLICATION OF ERROR FORMULA TO THE FUNDAMENTAL OPERATIONS OF ARITHMETICS

1.19.1 ERROR IN ADDITION OF NUMBERS

Let $y = x_1 + x_2 + \dots x_n$ be a function.

Let us suppose Δx_i to denote the error in x_i. Then we have

$$y + \Delta y = (x_1 + \Delta x_1) + (x_2 + \Delta x_2) + \dots + (x_n + \Delta x_n)$$
$$= (x_1 + x_2 + \dots + x_n) + (\Delta x_1 + \Delta x_2 + \dots + \Delta x_n)$$
$$\therefore \quad \Delta y = \Delta x_1 + \Delta x_2 + \dots + \Delta x_n$$

Now, dividing by y, we get

$$\frac{\Delta y}{y} = \frac{\Delta x_1}{y} + \frac{\Delta x_2}{y} + \dots + \frac{\Delta x_n}{y}$$
$$\Rightarrow \quad \left|\frac{\Delta y}{y}\right| \le \left|\frac{\Delta x_1}{y}\right| + \left|\frac{\Delta x_2}{y}\right| + \dots + \left|\frac{\Delta x_n}{y}\right|$$

Then, the absolute error is obtained by the relation given by

$$\Delta y = \left|\frac{\Delta y}{y}\right| \cdot y = \text{Product of Relative error and the number } y.$$

1.19.2 ERROR IN SUBTRACTION OF NUMBERS

Let $y = x_1 - x_2$ be given.

Let us suppose Δy, Δx_1 and Δx_2 denote the errors in y, x_1 and x_2 respectively.

Then, we have

$$y + \Delta y = (x_1 + \Delta x_1) - (x_2 + \Delta x_2) = (x_1 - x_2) + (\Delta x_1 - \Delta x_2)$$
$$\Rightarrow \quad \Delta y = \Delta x_1 - \Delta x_2 \qquad (\because y = x_1 - x_2)$$
$$\Rightarrow \quad \frac{\Delta y}{y} = \frac{\Delta x_1}{y} - \frac{\Delta x_2}{y}$$

But, we have

$$|\Delta y| \le |\Delta x_1| + |\Delta x_2| \Rightarrow \left|\frac{\Delta y}{y}\right| \le \left|\frac{\Delta x_1}{y}\right| + \left|\frac{\Delta x_2}{y}\right|$$

Therefore, the relative error and absolute errors are given by

$$\text{Relative error} = \left|\frac{\Delta y}{y}\right| \le \left|\frac{\Delta x_1}{y}\right| + \left|\frac{\Delta x_2}{y}\right|$$

and Absolute error $= |\Delta y| \le |\Delta x_1| + |\Delta x_2|$

1.19.3 ERROR IN PRODUCT OF NUMBERS

Let $\quad y = x_1 x_2 \dots x_n$

Now, suppose that $\Delta y, \Delta x_1, \Delta x_2, \dots, \Delta x_n$ denote the errors in $y, x_1, x_2, \dots, x_n$ respectively.

Then, we have

$$\frac{\Delta y}{y} = \frac{\Delta x_1}{y} \cdot \frac{\partial y}{\partial x_1} + \frac{\Delta x_2}{y} \cdot \frac{\partial y}{\partial x_2} + \dots + \frac{\Delta x_n}{y} \cdot \frac{\partial y}{\partial x_n}$$

Now $\quad \dfrac{1}{y} \cdot \dfrac{\partial y}{\partial x_1} = \dfrac{x_2 x_3 \dots x_n}{x_1 x_2 x_3 \dots x_n} = \dfrac{1}{x_1}$

$$\frac{1}{y} \cdot \frac{\partial y}{\partial x_2} = \frac{x_1 x_3 \dots x_n}{x_1 x_2 x_3 \dots x_n} = \frac{1}{x_2}$$

....

....

$$\frac{1}{y}\cdot\frac{\partial y}{\partial x_n}=\frac{x_1x_2...x_{n-1}}{x_1x_2...x_n}=\frac{1}{x_n}$$

$$\therefore \qquad \frac{\Delta y}{y}=\frac{\Delta x_1}{x_1}+\frac{\Delta x_2}{x_2}+...+\frac{\Delta x_n}{x_n}$$

Therefore, the Relative error and absolute error are given by

$$\text{Relative error } =\left|\frac{\Delta y}{y}\right|\le\left|\frac{\Delta x_1}{x_1}\right|+\left|\frac{\Delta x_2}{x_2}\right|+...+\left|\frac{\Delta x_n}{x_n}\right|$$

$$\text{Absolute error } =\left|\frac{\Delta y}{y}\right|\cdot y=\left|\frac{\Delta y}{y}\right|\cdot(x_1x_2...x_n)$$

1.19.4 ERROR IN DIVISION OF TWO NUMBERS

Let $y=\frac{x_1}{x_2}$. Since, we have

$$\frac{\Delta y}{y}=\frac{\Delta x_1}{y}\cdot\frac{\partial y}{\partial x_1}+\frac{\Delta x_2}{y}\cdot\frac{\partial y}{\partial x_2}=\frac{\Delta x_1}{x_1/x_2}\times\frac{1}{x_2}+\frac{\Delta x_2}{x_1/x_2}\left(\frac{-x_1}{x_2^2}\right)=\frac{\Delta x_1}{x_1}-\frac{\Delta x_2}{x_2}$$

$$\therefore \qquad \left|\frac{\Delta y}{y}\right|\le\left|\frac{\Delta x_1}{x_1}\right|+\left|\frac{\Delta x_2}{x_2}\right|$$

Thus, the relative error is given by

$$\text{Relative Error } \le\left|\frac{\Delta x_1}{x_1}\right|+\left|\frac{\Delta x_2}{x_2}\right|$$

1.19.5 ERROR IN EVALUATING x^k

Let $y = x^k$, where k is any integer or a fraction. Then, we have the relative error

$$=\left|\frac{\Delta y}{y}\right|<\frac{\Delta x}{y}\cdot\frac{dy}{dx}$$

i.e., $$\left|\frac{\Delta y}{y}\right|<\frac{\Delta x}{x^k}\cdot k\cdot x^{k-1}=k\cdot\frac{\Delta x}{x}$$

Thus, relative error in evaluating $x^k=k\cdot\left|\frac{\Delta x}{x}\right|$

1.19.6 INVERSE PROBLEM

Let $y = f(x_1, x_2, ..., x_n)$ be a function, which have a desired accuracy, *i.e.*, if Δy is error in y. Then we have to determine errors $\Delta x_1, \Delta x_2, ..., \Delta x_n$ in $x_1, x_2, ..., x_n$.

Since, we have

$$\Delta y=\Delta x_1\cdot\frac{\partial y}{\partial x_1}+\Delta x_2\cdot\frac{\partial y}{\partial x_2}+...+\Delta x_n\cdot\frac{\partial y}{\partial x_n}$$

Now using the principle of equal effects, we have

$$\Delta x_1\cdot\frac{\partial y}{\partial x_1}=\Delta x_2\cdot\frac{\partial y}{\partial x_2}=...=\Delta x_n\cdot\frac{\partial y}{\partial x_n}$$

$$\Delta y=\Delta x_1\cdot\frac{\partial y}{\partial x_1}+\Delta x_1\cdot\frac{\partial y}{\partial x_1}+...+\Delta x_1\cdot\frac{\partial y}{\partial x_1}=n\Delta x_1\cdot\frac{\partial y}{\partial x_1}$$

$$\therefore \qquad \Delta x_1=\frac{\Delta y}{n\frac{\partial y}{\partial x_1}}$$

Similarly $\Delta x_2 = \dfrac{\Delta y}{n\dfrac{\partial y}{\partial x_2}} \dots \quad \Delta x_n = \dfrac{\Delta y}{n\dfrac{\partial y}{\partial x_n}}$

Thus $\Delta x_1 = \dfrac{\partial y}{n\dfrac{\partial y}{\partial x_1}}, \Delta x_2 = \dfrac{\partial y}{n\dfrac{\partial y}{\partial x_2}}, \dots, \Delta x_n = \dfrac{\partial y}{n\dfrac{\partial y}{\partial x_n}}$

Solved Examples

EXAMPLE 1. ***Find the possible relative error and absolute error in the sum of 0.1429 and 0.0909, where 0.1429 and 0.0909 are the approximate values of 1/7 and 1/11, correct to four decimal places.***

SOLUTION. Since, we consider the approximation in four decimal places, therefore in each case, the maximum error is

$$\frac{1}{2} \times 0.0001 = 0.00005$$

Now

(i) The relative error $= \left|\dfrac{\Delta y}{y}\right| < \left|\dfrac{0.00005}{0.2338}\right| + \left|\dfrac{0.00005}{0.2338}\right|$

$(\because y = x_1 + x_2 = 0.1429 + 0.0909 = 0.2338)$

$$\therefore \qquad \left|\frac{\Delta y}{y}\right| < \frac{0.0001}{0.2338} = 0.00043$$

(ii) The absolute error $= \left|\dfrac{\Delta y}{y}\right| y = \dfrac{0.0001}{0.2338} \times 0.2338 = 0.0001$

EXAMPLE 2. ***Find the relative error in the difference of following two numbers, given by $\sqrt{5.5} \approx 2.345$ and $\sqrt{6.1} \approx 2.470$, correct to four significant figures.***

SOLUTION. Here we have $\Delta x_1 = \Delta x_2 = \frac{1}{2}(0.001) = 0.0005$

$(\because$ we consider the approximation into four significant figures)

$$\therefore \qquad \text{The relative error} < \left|\frac{\Delta x_1}{y}\right| + \left|\frac{\Delta x_2}{y}\right|$$

$$= 2\left|\frac{\Delta x_1}{y}\right| = 2\left|\frac{0.0005}{2.470 - 2.345}\right| \qquad (\because y = x_1 - x_2)$$

$$= 2\left|\frac{0.0005}{0.125}\right| = \frac{0.001}{0.125} = 0.0008$$

Hence, the possible maximum error is = 0.0008.

EXAMPLE 3. ***Find the product of 346.1 and 865.2 and state how many figures of the results are trustworthy, given that the numbers are correct to four significant figures.***

SOLUTION. Since we consider the approximation in one decimal place, therefore

$$\Delta x_1 = \frac{1}{2}(0.1) = \Delta x_2 = 0.05$$

and $y = 346.1 \times 865.2 = 299446$

which is correct to six significant figures.

Then, the relative error $\leq \left|\dfrac{\Delta x_1}{x_1}\right| + \left|\dfrac{\Delta x_2}{x_2}\right| = \left|\dfrac{0.05}{346.1}\right| + \left|\dfrac{0.05}{865.2}\right|$

$$= 0.000144 + 0.000058 = 0.000202$$

Therefore, the absolute error = Relative error $\leq 0.000202 + 299446 \approx 60$

The true value of the product of the numbers lies between

$$299446 - 60 = 299386 \text{ and } 299446 + 60 = 299506$$

Now, the mean of these values is $\dfrac{299386+299506}{2} = 299446$ which can be written as 299.4×10^2 correct to four significant figures.

EXAMPLE 4. ***Find the number of trustworthy figures in (0.491)^3 assuming that the number is 0.491 is correct to last figure.***

SOLUTION. Since, we know that the relative error $E_r = \dfrac{\Delta y}{y} < k\dfrac{\Delta x}{x}$

And we consider the approximation of given number up to three decimal places

$$\therefore \quad \Delta x = \frac{1}{2}(0.001) = 0.0005$$

Also, here $k = 3$

$$\Rightarrow \quad k\frac{\Delta x}{x} = \frac{3\times 0.0005}{(0.491)^3} = \frac{3\times 0.0005}{0.118371} = 0.01267$$

$$\therefore \quad \text{The absolute error} = E_r \cdot y$$
$$< 0.01267 \times (0.491)^3$$
$$= 0.01267 \times 0.118371 = 0.0015$$

Since the error affects the third decimal places, therefore, $(0.491)^3 = 0.1183$ is correct to second decimal places.

EXAMPLE 5. ***The error in the measurement of the area of circle is not allowed to exceed 0.1%. How accurately should the diameter be measured?***

SOLUTION. Let d be the diameter of the circle.

Then area is given by $A = \dfrac{\pi d^2}{4}$

$$\Rightarrow \quad \frac{\partial A}{\partial d} = \frac{\pi d}{2}$$

$$\Delta A = \Delta d \cdot \frac{\partial A}{\partial d}, \qquad \therefore \quad \Delta d = \frac{\Delta A}{\frac{\partial A}{\partial d}}$$

Now percentage error in $A = \dfrac{\Delta A}{A} \times 100 = 0.1$

$$\therefore \quad \Delta A = \frac{0.1 \times A}{100} = 0.001 \times A = \frac{0.001 \times \pi d^2}{4}$$

$$\therefore \text{ The percentage error in } d = \frac{\Delta d}{d} \times 100 = \frac{100}{d} \times \frac{\Delta A}{\partial A / \partial d}$$

$$= \frac{100}{d}\left(\frac{0.001 \times \pi d^2}{4}\right)\frac{\pi d}{2} = \frac{0.1\pi d^2}{4d} \times \frac{2}{\pi d} = \frac{0.1}{2} = 0.05$$

EXAMPLE 6. ***The percentage error in R, which is given by $R = \dfrac{r^2}{2h} + \dfrac{h}{2}$, is not allowed to exceed 0.2%. Find allowable error in r and h when r = 4.5 cm and h = 5.5cm.*** [MEERUT–2011]

SOLUTION. The percentage error in R

$$= \frac{\Delta R}{R} \times 100 = 0.2$$

$$\therefore \qquad \Delta R = \frac{0.2}{100} \times R = \frac{0.2}{100} \times \left[\frac{(4.5)^2}{2 \times 5.5} + \frac{5.5}{2}\right] \qquad \left(\because R = \frac{r^2}{2h} + \frac{h}{2}\right)$$

$$= \frac{0.2}{100} \times \frac{50.5}{11} = \frac{0.002 \times 50.5}{11} \qquad \ldots(1)$$

(i) Percentage error in $r = \frac{\Delta r}{r} \times 100$

$$= \frac{100}{r}\left(\frac{\Delta R}{2\frac{\partial R}{\partial r}}\right) \qquad \left(\because \Delta r = \frac{\Delta R}{2\frac{\partial R}{\partial r}}\right)$$

$$= \frac{100}{r} \times \frac{\Delta R}{2\left(\frac{r}{h}\right)} = \frac{100(\Delta R) \cdot h}{2r^2} \qquad \ldots(2)$$

Put $r = 4.5$ and value of ΔR from equation (1), in equation (2), we get

Percentage error $= \frac{100}{2 \times (4.5)^2} \times \frac{0.002 \times 50.5}{11} \times h$

$$= \frac{0.1 \times 50.50 \times 5.5}{11 \times 20.25} = 0.12$$

(ii) Percentage error in $h = \frac{\Delta h}{h} \times 100$

$$= \frac{100}{h} \times \frac{\Delta R}{2\frac{\partial R}{\partial h}} = \frac{100}{h} \cdot \frac{\Delta R}{2\left(\frac{-r^2}{2h^2} + \frac{1}{2}\right)}$$

$$= \frac{100 \Delta R}{\left(\frac{-r^2}{h^2} + h\right)} = \frac{100}{20/11} \times \frac{50.5 \times 0.002}{11} = 0.505$$

EXAMPLE 7. ***Use the Series*** $\log_e\left(\frac{1+x}{1-x}\right) = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots\right)$

to compute the value of* log (1.2) *correct to seven decimal place and find the number of terms retained. [MEERUT–2015]

SOLUTION. Let $\frac{1+x}{1-x} = 1.2 \qquad \Rightarrow \qquad x = \frac{1}{11}$

If we retains n terms, then $(n+1)^{\text{th}}$ term $= \frac{x^{2n+1}}{2n+1} = \frac{\left(\frac{1}{11}\right)^{2n+1}}{2n+1}$

For seven decimal accuracy, we have

$$\frac{1}{2n+1} \cdot \left(\frac{1}{11}\right)^{2n+1} < \frac{1}{2} \times 10^{-7} \;\Rightarrow\; (2n+1)(11)^{2n+1} > 2 \times 10^7$$

$$\Rightarrow \qquad n \geq 3$$

Hence, retaining the first three terms of the given series, we get

$$\log_e(1.2) = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5}\right)_{\text{at } x=\frac{1}{11}} = 0.1823215.$$

EXAMPLE 8. ***For x = 0.4845 and y = 0.4800. Calculate the value of $\frac{x^2 - y^2}{x + y}$ by using normalized floating point arithmetic. Compare with the value of (x – y) indicate error in the former.***

SOLUTION. Given that $x = 0.4845, y = 0.4800$

Now, $(x^2 - y^2) = (0.4845\ E0 \times 0.4845\ E0) - (0.4800\ E0 \times 0.4800\ E0)$

$= (0.0043\ E0) = (0.4300\ E - 2)$

$(x + y) = (0.4845\ E0 + 0.4800\ \text{E}0) = (0.9645\ E0)$

So, $\frac{(x^2 - y^2)}{(x + y)} = \frac{(0.4300\ E - 2)}{(0.9645\ E0)}$

$x - y = (0.4845\ E0) - 0.4800\ E0$

$= (0.0045\ E0) = (0.4500\ E - 2)$

Hence in normalized floating point arithmetic, the value of $\frac{(x^2 - y^2)}{x + y} \neq x - y$

The error is $(0.4500\ E -2) - (0.4458\ E -2) = (0.0042\ E -2) = (0.4200\ E -4)$

EXAMPLE 9. ***Compare the percentage error in the time period $T = 2\pi\sqrt{\frac{l}{g}}$ for l = 1 m if the error in measurement of l is 0.01.***

SOLUTION. We have $T = 2\pi\sqrt{\frac{l}{g}}$

Taking log of both the sides, we get

$$\log T = \log 2\pi + \frac{1}{2}\log l - \frac{1}{2}\log g$$

$$\Rightarrow \quad \frac{1}{T}\delta T = \frac{1}{2}\cdot\frac{\delta l}{l}$$

$$\Rightarrow \quad \frac{\delta T}{T}\times 100 = \frac{\delta l}{2l}\times 100 = \frac{0.01}{2\times 1}\times 100 = 5\%$$

EXAMPLE 10. ***The discharge Q over a notch for head H is calculated by the formula $Q = kH^{5/2}$, where k is a given constant. If the head is 75 cm and an error of 0.15 cm is possible in its measurement, estimate the percentage error in computing the discharge.***

SOLUTION. Here, we have $Q = kH^{5/2}$

Taking log of both the sides, we get

$$\log Q = \log k + \frac{5}{2}\log H$$

On differentiating, we get

$$\frac{\delta Q}{Q} = \frac{5}{2}\cdot\frac{\delta H}{H}$$

$$\therefore \quad \frac{\delta Q}{Q}\times 100 = \frac{5}{2}\times\frac{0.15}{75}\times 100 = \frac{1}{2} = 0.5$$

EXAMPLE 11. ***If r = 3h(h*^{6} *– 2). Find the percentage error in r at h = 1 if the percentage error in h is 5.*** [MEERUT–2008, 10]

SOLUTION. We have $\delta r = \dfrac{\partial r}{\partial h}\cdot \delta h = (21h^6 - 6)\delta h$

$$\therefore \quad \frac{\delta r}{r}\times 100 = \left(\frac{21h^6-6}{3h^7-6h}\right)\delta h\times 100 = \left(\frac{21-6}{3-6}\right)\left(\frac{\delta h}{h}\times 100\right) = \frac{15}{-3}\cdot 5\% = -25\%$$

Now, percentage error is $= \left|\dfrac{\delta r}{r}\times 100\right| = 25\%$

EXAMPLE 12. ***If*** $\sqrt{29} = 5.385$ ***and*** $\sqrt{\pi} = 3.317$ ***correct to four significant figures, find the relative error in their sum and differences.*** [MEERUT–2017; KANPUR–2011]

SOLUTION. The numbers 5.385 and 3.317 are correct to four significant figures. Therefore. Maximum error in each case is

$$\frac{1}{2}\times 10^{-3} = 0.0005$$

$$\therefore \qquad \Delta x_1 = \Delta x_2 = 0.0005$$

Now, relative error in their sum is

$$\left|\frac{\Delta X}{X}\right| \le \left|\frac{\Delta x_1}{x}\right| + \left|\frac{\Delta x_2}{x}\right| \qquad (\because X = x_1 + x_2 = 8.702)$$

$$\le \left|\frac{0.0005}{8.702}\right| + \left|\frac{0.0005}{8.702}\right| < 1.149\times 10^{-4}$$

Also, relative error in their difference is

$$\left|\frac{\Delta X}{X}\right| \le \left|\frac{\Delta x_1}{x}\right| + \left|\frac{\Delta x_2}{x}\right| \text{ where } X = x_1 - x_2 = 2.068$$

$$\le \left|\frac{0.0005}{2.068}\right| + \left|\frac{0.0005}{2.068}\right| < 4.835\times 10^{-4}$$

EXERCISE 1.3

1. Find the number of terms of the exponential series such that their sum gives e^x correct to six decimal places at $x = 1$.

2. If $R = \dfrac{4xy^2}{z^3}$ and errors in x, y, z be 0.001.

Show that the maximum relative error at $x = y = z = 1$ is 0.006.

3. If $R = \dfrac{1}{2}\left(\dfrac{r^2}{h} + h\right)$ and error in R is at the most 0.4%. Find the percentage error allowable in r and h when $r = 5.1$ cm and $h = 5.8$ cm.

4. Determine the number of terms required in the series for $\log(1 + x)$ to evaluate log 1.2 correct to six decimal places.

5. Find the relative error in calculation of $\dfrac{7.342}{0.241}$, where the number 7.342 and 0.241 are correct to three decimal places. Determine the smallest interval in which true result lies.

6. Find the number of trustworthy figures in $(367)^{1/5}$ where 367 is correct to three significant figures.

7. How accurately, the length and time of vibration of a pendulum should be measured in order that the computed value of g be correct to 0.01%.

8. Let n_0 be the approximate cube root of n and let $x = \dfrac{n}{n_0^3} - 1$, show that cube root of n is given by

$$n^{1/3} = n_0\left[1 + \frac{x}{3} - \frac{x^2}{9} + \frac{5x^3}{81} - \frac{10x^4}{243} + \dots\right]$$

Hence, find the value of $(6)^{1/3}$ correct to four significant figures.

9. If n_0 is the approximate value of the square root of n and $x = \frac{n}{n_0^2} - 1$, show that the square root of n is given by

$$n^{1/2} = n_0\left[1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \frac{5x^4}{128} + \ldots\right]$$

Hence, find the square root of 5 correct to three decimal places.

10. Write a short note on 'Error in Numerical computations'.

11. Let x^* approximate x correct upto n significant figures if e^x is evaluated for x, $-8 \le x \le 9$. Then, what should be the relative error.

12. If $R = 4x^2y^3z^{-4}$, find the maximum absolute error and maximum relative error in R when errors in $x = 1$, $y = 2$, $z = 3$ respectively are equal to 0.001, 0.002, 0.003. (UPTU–2003)

13. Represent 44.85×10^6 in normalized floating point mode. (UPTU–2004)

14. If $r = h(4h^5 - 5)$, find the percentage error in r at $h = 1$, if the error in h is 0.04. (WBTU–2005)

Answers

1. $n = 10$ **3.** 0.23, 0.14 **4.** $n = 10$ **5.** 0.0021, (30.4647 – 0.0639)

6. 3.26, correct to three figures

7. (i) Percentage error in length = 0.005 (ii) Percentage error in time = 0.0025 **9.** 1.817

9. 2.236 **12.** 0.00355, 0.0089 **13.** 0.4485 $E8$ **14.** 76

1.20 ORDER OF APPROXIMATIONS

Let us suppose $f(h)$ be a function with approximation $g(h)$ and the error bound is known to be $\mu(h^n)$ where n is a positive integer so that

$$|f(h) - g(h)| \le \mu|h^n|$$

where h is sufficiently small.

Then, we say that $g(h)$ approximate the function $f(h)$ with order of approximation $O(h^n)$ and write

$$f(h) = g(h) + O(h^n)$$

For example: (i) Consider $(1 - h)^{-1} = 1 + h + h^2 + h^3 + h^4 + \ldots$

$$\Rightarrow \qquad (1 - h)^{-1} = 1 + h + h^2 + h^3 + O(h^4)$$

Similarly

$$\cosh = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} - \frac{h^6}{6!} + \ldots = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6)$$

1.20.1 ORDER OF APPROXIMATION FOR SUM AND PRODUCT

(i) Approximation for Sum: Consider, from the previous example

$$(1 - h)^{-1} = 1 + h + h^2 + h^3 + O(h^4) \qquad \ldots(1)$$

and

$$\cosh = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6) \qquad \ldots(2)$$

Then, for the approximation of sum of eq. (1) and (2), we get

$$[(1+h)^{-1} + \cosh] = 2 + h + \frac{h^2}{2!} + h^3 + O(h^4) + \frac{h^4}{4!} + O(h^6) \qquad \ldots(3)$$

Now since

$$O(h^4) + \frac{h^4}{4!} = O(h^4)$$

and

$$O(h^4) + O(h^6) = O(h^4)$$

Therefore, from eq. (3), we get

$$[(1+h)^{-1} + \cosh] = 2 + h + \frac{h^2}{2!} + h^3 + O(h^4)$$

a fourth order approximation.

(ii) Approximation for Product:

For the approximation of product of (1) and (2), we get

$$[(1+h)^{-1}\cosh] = (1+h+h^2+h^3)\left[1-\frac{h^2}{2!}+\frac{h^4}{4!}\right]+(1+h+h^2+h^3)O(h^6)$$

$$+\left(1-\frac{h^2}{2!}+\frac{h^4}{4!}\right)O(h^4)+O(h^4)O(h^6)$$

$$=1+h+\frac{h^2}{2}+\frac{h^3}{2}-\frac{11h^4}{24}+\frac{11h^5}{24}+\frac{h^6}{24}+\frac{h^7}{24}+O(h^4)$$

$$+O(h^6)+O(h^4)O(h^6) \qquad \text{...(4)}$$

Now since

$$O(h^4)O(h^6) = O(h^{10})$$

$$\Rightarrow \quad -\frac{11h^4}{24}+\frac{11}{24}h^5+\frac{h^6}{24}+\frac{h^7}{24}+O(h^4)+O(h^6)+O(h^{10})=O(h^4)$$

Therefore, from eq. (4), We get

$$[(1-h)^{-1}\cosh]=1+h+\frac{h^2}{2}+\frac{h^3}{2}+O(h^4)$$

which is of the first order approximation.

1.21 PROPAGATION OF ERROR

Let us suppose $g(n)$ represents the growth of error after n steps of a computation process. Then, we have the following observations

(i) If $|g(n)| \sim n\varepsilon$ then, the growth of error is linear.

(ii) If $|g(n)| \sim \delta^n\varepsilon$ then, the growth of the error is exponential.

(iii) If $\delta > 1$ then the exponential will grow indefinitely as $n \to \infty$ and

(iv) If $0 < \delta < 1$ then exponential error decrease to zero as $n \to \infty$

1.21.1 SOME IMPORTANT OBSERVATIONS ON ERRORS

- If C_1 and C_2 are the first significant figures of two numbers which are each correct to n significant figures and if neither number is of the form $C(1.00...) \times 10^P$, then their product or quotient is correct to :

  (a) $(n-1)$ significant figures if $C_1 \geq 2$ and $C_2 \geq 2$.

  (b) $(n-2)$ significant figures if either $C_1 = 1$ or $C_2 = 1$.

- If C is the first significant figure of a number which is correct to n significant figures, and if this number contains more one digits different from zero, then its p^{th} power is correct to:

  (a) $(n-1)$ significant figures if $p \leq C$

  (b) $(n-2)$ significant figures if $p \leq 10C$.

  and its r^{th} root is correct to

  (a) n significant figures if $rC \leq 10$.

  (b) $(n-1)$ significant figures if $rC \leq 10$.

- If C is the first significant figures of a number which is correct to n significant figures and if this number contains more than one digit different from zero, then for the absolute error in its common logarithms we have

$$E_a < \frac{1}{4C \times 10^{n-1}}$$

- If a logarithm (base 10) is not in error by more than two units in the m^{th} decimal places, the antilog is certainly correct to $(m-1)$ significant figures.

1.21.2 PROPAGATED ERROR

In any numerical problem, the true value of numbers may not be used exactly, *i.e.*, in place of true values of the numbers, some approximate values like floating point numbers are used initially. The error arising in the problem due to those inexact/approximate values is called propagated error.

Let x^A, y^A be approximation to x and y respectively and w be arithmetic operation. Then,

$$\text{The propagated error} = xwy^A - x^Awy^A$$

$$\text{The relative propagated error } = \frac{xwy - xw^Ay^A}{xwy}$$

$$\text{Total relative error } = \frac{xwy - x^Aw^Ay^A}{xwy}$$

$$= \frac{xwy - x^Awy^A}{xwy} + \frac{x^Awy^A - x^Aw^Ay^A}{xwy}$$

☛ **REMARK**

➠ For the first approximation.

Total relative error = relative propagated error + relative generated error.

1.21.3 PROPAGATION OF ERROR IN FUNCTION EVALUATION OF A SINGLE VARIABLE

Let $f(x)$ be evaluated and x^A be an approximation to x. Then, the absolute error in evaluation of $f(x)$ is $f(x) - f(x^A)$ and relative error is

$$\gamma_{f(x)} = \frac{f(x) - f(x^A)}{f(x)}$$

Let us suppose $\quad x = x^A + \rho_x$

Then, by Taylor's series expansion, we get

$$f(x) = f(x^A) + \rho_x f'(x^A) + \ldots$$

$$\Rightarrow \qquad \gamma_{f(x)} = \frac{\rho_x f'(x^A)}{f(x)} \qquad \text{(By neglecting the higher order terms)}$$

$$= \frac{\rho_x}{f(x)} \approx \frac{\gamma f'(x^A)}{f(x)} = \gamma_x \frac{xf'(x^A)}{f(x)}$$

$$|\gamma_{f(x)}| = |\gamma_x| \left|\frac{xf'(x^A)}{f(x)}\right|$$

☛ **REMARKS**

➠ For evaluation of $f(x)$ in denominator of R.H.S. after simplification, $f(x)$ must be replaced by $f(x^A)$ in some cases so

$$|\gamma_{f(x)}| = |\gamma_x| \left|\frac{xf'(x^A)}{f(x)}\right|$$

The expression $\left|\dfrac{xf'(x^A)}{f(x)}\right|$ is called condition number $f(x)$ at x.

➠ If the condition number is very large, then function is said to be more ill-conditioned.

Solved Examples

EXAMPLE 1. ***Let f(x) = $x^{1/10}$ and x^A approximates x correct to n significant decimal digit. Show that f(x^A) approximates f(x) correct to (n + 1) significant decimal digits.***

SOLUTION. We have

$$\gamma_{f(x)} = \gamma_x \cdot \frac{xf'(x^A)}{f(x)}$$

$$= \gamma_x \cdot \frac{x \cdot \frac{1}{10} x^{A} \frac{9}{10}}{x^{1/10}} = \left(\frac{1}{10}\right)\gamma_x$$

$$\therefore \qquad \left|\gamma_{f(x)}\right| = \left(\frac{1}{10}\right)\left|\gamma_x\right| \le \frac{1}{10}\cdot\frac{1}{2}\cdot 10^{1-n} = \frac{1}{2}\cdot 10^{1-(n+1)}$$

$\Rightarrow$ $f(x^A)$ approximates $f(x)$ correct to $(n + 1)$ significant digits.

EXAMPLE 2. ***The function f(x) = cos(x) can be explained as***

$$\mathbf{\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots}$$

compute the number of terms requires to estimate $\mathbf{\cos\left(\frac{\pi}{4}\right)}$ ***so that the result is correct to least two significant digits.***

SOLUTION. We know that the pre-specified tolerance e_s can be obtained by using

$$e_s = (0.5 \times 10^{2-n})\%$$

Therefore, we have

$$e_s = 0.5 \times 10^{-m} = 0.5 \times 10^{-2}$$

The remainder term R_n is given by $R_n = \frac{x^{2n}}{(2n)!}\cos\xi$

Then, maximum relative error $= \frac{(\pi/4)^{2n}}{(2n)!}$

Therefore,

$$0.5\times 10^{-2} \ge \frac{(\pi/4)^{2n}}{2n!}$$

$$i.e., \frac{1}{0.5\times 10^{-2}} \le \frac{2n!}{(\pi/4)^{2n}}$$

$$\Rightarrow \qquad 200 \le \frac{2n!}{(\pi/4)^{2n}}$$

$\boldsymbol{n}$	$\mathbf{\frac{(2n)!}{(\pi/4)^{2n}}}$
1	3.24
2	63.074
3	3067.561

Thus $\qquad n = 3$

EXAMPLE 3. ***The function f(x) =*** $\mathbf{\tan^{-1}}$***x can be expanded as follows***:

$$\mathbf{\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \ldots + (-1)^{n-1}\frac{x^{2n-1}}{(2n-1)} + \ldots}$$

Compute number of terms n such that the series determines $\mathbf{\tan^{-1}1}$ ***correct to eight significant figures.*** [MDU(BE)–2006]

SOLUTION. Proceed same as above, we get

$$e_s = 0.5 \times 10^{-m} = 0.5 \times 10^{-8}$$

Also, the remainder term after n terms is given by

$$R_n = \frac{x^{2n+1}}{2n+1}\tan^{-1}\xi, \qquad 0 < \xi \le x$$

Therefore, the maximum relative error is given by

$$\left(\frac{x^{2n+1}}{2n+1}\right)_{x=1} = \frac{1}{2n+1}$$

Since, the error must be less than e_s, therefore

$$0.5 \times 10^{-8} \geq \frac{1}{2n+1}$$

$$\Rightarrow \qquad \frac{1}{0.5 \times 10^{-8}} \leq 2n+1$$

$$\Rightarrow \qquad 2 \times 10^{8} \leq 2n+1$$

Therefore $\qquad n = 10^{8+1}$.

EXAMPLE 4. ***Determine the number of terms of the exponential series***

$$\boldsymbol{e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \ldots}$$

such that their sum gives the values of e^x correct to six decimal places for $0 \leq x \leq 1$. [UPTU(MCA)–2002]

SOLUTION. Here $\qquad e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^{n-1}}{(n-1)!} + R_n(x) \qquad \ldots(1)$

Where $\quad R_n(x) = \frac{x^n}{n!} e^{\theta}, 0 < \theta < x$

Max, absolute error (at $\theta = x$) $= \frac{x^n}{n!} e^x$ and the max. relative error $= \frac{x^n}{n!}$

Hence, $(E_r)_{\max}$ at $x = 1 = 1/n!$

For a six decimal accuracy at $x = 1$ we have $\frac{1}{n!} < \frac{1}{2} \times 10^{-6}$ or $n! > 2 \times 10^6$, which gives $n = 10$.

1.22 BLUNDERS

Blunders are errors which arises due to human imperfection. Since these errors are due to human mistakes, it should be possible to avoid them. These types of errors can occur at any stage of the numerical processing due to the

(i) lack of understanding of the problem
(ii) wrong assumptions
(iii) selecting a wrong method
(iv) wrong guessing the initial values.

The solution have its care, coupled with a careful examination of the results for reasonableness. Sometimes a test run with known results is worthwhile, but it is no guarantee of freedom from foolish error. When hand computation was more common check sums were usually computed. They were designed to reveal the mistake and permits its correction.

1.23 NUMERICAL INSTABILITY

We know that every arithmetic operation performed during computations, gives some errors, which may grow or decay in subsequent calculations. In some cases errors may grow so large as it make the computed result totally redundant. Such a procedure is called numerically unstable.

On the other hand, in some cases it can be avoided by changing the calculation procedure, which avoids subtractions of nearly equal numbers or division by a small number or by retaining more digits in the mantissa.

There are the following types of instability:

1.23.1 INHERENT INSTABILITY

This instability may arise due to the ill-condition ness of the problem. Here, we can not avoid the inherent instability by changing the method of solution. It is the property of the problem itself. We can avoid this instability by suitable reformulation of the problem.

1.23.2 INDUCED INSTABILITY

The induced instability may arise due to the wrong choice of the method of solution. Although, the problem is well conditioned in this case. Induced instability can be avoided by a suitable modification or change of the method of solution.

1.24 SENSITIVITY ANALYSIS

Investigation to see how small changes (or perturbations or disturbances) in input parameters influence the output are termed as sensitivity analysis, when problem is sensitive to small changes in its parameter, it is impossible to make a numerically stable method for its solution.

1.25 MACHINE COMPUTATIONS

When we solve any problem using computers, then to obtain meaningful results, we have the following phases:

(i) Choice of a method: A method is defined by a mathematical formula for finding the solution of the given equation. In some cases, there may be more than one methods are available to solve the same problem. Choose the method which suits the given problem best. The assumptions and limitations of the method must be studied carefully.

(ii) Designing the Algorithm: Since, we know that the computer do not solve problem rather they are used to implement the solution to problems.

The logical and concise list of procedure for solving a problem is called an algorithm. It describes the steps that lead to required results in a finite number of operations. Here, it may be noted that the computer is concerned with the algorithm and not with the method. The algorithm tells the computer where to start, what information to use, what operation to be carried out and in which order, what information to be printed and when to stop. An algorithm should also include steps to identify and abnormal data or results and take corrective measures. In case of large problem we use the modular approach. A module is a program unit or entity that is responsible to a single task. It is also known as sub-programs.

An algorithm has five important properties:

(i) Algorithm should be completed after a finite number of steps.

(ii) Every step of algorithm should be well defined.

(iii) Algorithm should clearly specify which quantities are to be read.

(iv) Algorithm should clearly specify which quantities are to be displayed.

(v) In an algorithm, all operations should be executable.

Algorithm to find the root of a quadratic Equation: If we design an algorithm to find the real roots of the equation.

$$ax^2 + bx + c = 0 \qquad a, b, c \in \boldsymbol{R}$$

for 10 set of values, using the usual method $x = \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

Then, we have the following computational steps:

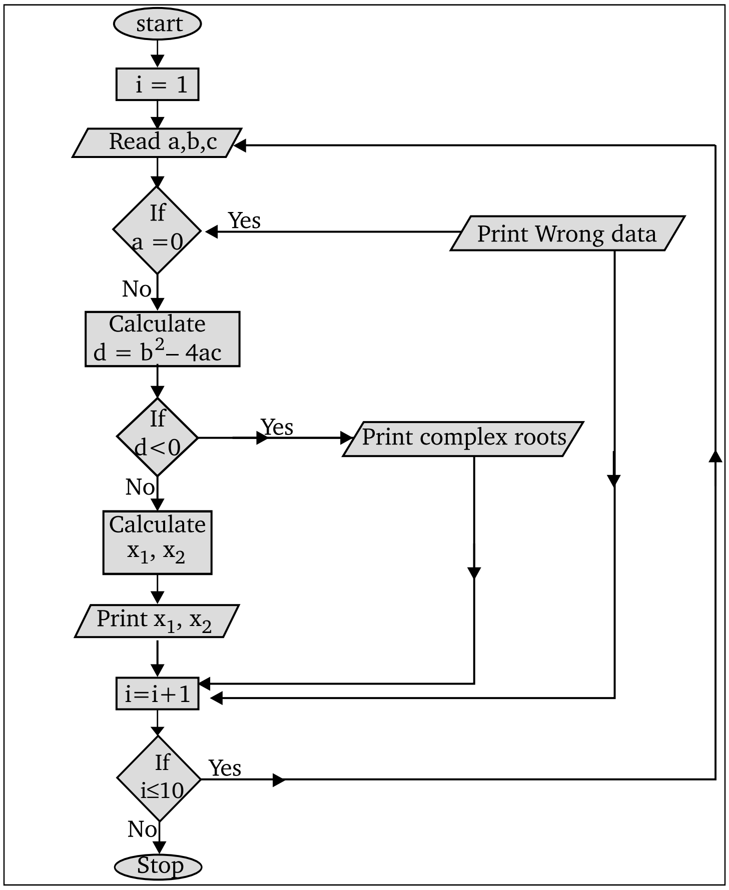
```
1.  Set $I = I$
2.  Read $a, b, c$
3.  Check is $a = 0$? If yes print wrong data and go to step 9
4.  Calculated $d = b^2 - 4ac$
5.  Check : is $d < 0$? If yes, print, roots and complex and goto step 9
6.  Calculate $e = \sqrt{d}$
7.  Calculate $x_1$ and $x_2$ using the usual method
8.  Print $x_1$ and $x_2$
9.  $I = I + 1$
10. Check : is $I \leq 10$? If yes, goto step 2, otherwise goto step (11)
11. Stop
```

(iii) Flowchart: A flowchart is a graphical representation of a specific number of sequences of steps (algorithm) to be followed by the computer to produce the solution of a given problem. It maks use of flowchart symbols to represent the basic operations to be carried out and the arrow indicate the flow of information and processing.

Flow chart symbols:

	Symbols	Meaning		Symbols	Meaning
1.	⬭	Start or End	4.	◇	Decision making and branching
2.	▭	Computational steps	5.	○	Connector
3.	▱	Input or output	6.	⟶	Flow of control

FLOW CHART FOR FINDING REAL ROOTS OF THE QUADRATIC EQUATION



☛ REMARKS

- Flowchart provide a graphic representation of the problem so it is easy to understand the plan of the solution.
- It helps in reviewing and correcting the program.
- It provides a convenient aid to writing computer instructions.

(iv) Programming: In this phase we write the program into any computer language.

Program for the roots of a Quadratic equation in 'C' Language

```
# include <math.h>
main()
{
float a, b, c d;
float root 1, root 2;
printf ("Input the values a, b, c\n")
else
{
root 1= (–b + sqrt (d))/(2.0*a);
root 2= (–b – sqrt (d))/(2.0*a);
printf ("/n/nRoot = % f/n/nRoot 2 = % f\n" root 1, root 2);
}
}
```

(v) Computer Execution : After writing the program of instruction for the computer in a suitable computer language, check the errors in program and remove. After that, prepare the data in the required form. Then, the computations are performed by the computer and the results are given out.

1.25 COMPUTER SOFTWARE

The computer software provide a useful computational tool for users. The writing of a computer software requires a good understanding of the programming. A good computer software must contain some criteria of:

(i) Self starting
(ii) Accuracy and reliability
(iii) Minimum number of levels
(iv) Good documentation
(v) Criteria of portability

(i) Self starting: A good computer software should be self starting as far as possible. Since, any numerical method involves some parameters, whose values are to be determined. The program will be more acceptable, if it can be made automatic in the sense that the program will select the initial approximation itself rather than requiring the user to specify them.

(ii) Accuracy and reliability: Accuracy and reliability are measures of the performance of an algorithm on all similar problem. Fixed the error criteria and get the solutions of all similar problems to that accuracy. The program must be able to prevent most of the exceptional conditions.

(iii) Minimum number of levels: A good software must contain the minimum number of levels, because if the number of levels are increased, then there is a wasted of time due to the interlinking and transfer of parameters.

(iv) Good documentation: A good documentation should clarify what kind of problems can be solved using the software, what parameters are to be supplied, what accuracy can be achieved, which method has been used and other useful details. It should be noted that the program must have some comments lines at various places giving more explanation about the method and steps.

(v) Criteria of portability: The software should be made independent of the computer being used as far as possible. Therefore we have that the software must be machine independent, *i.e.*, the same program should be able to run on any machine with minimum modifications.

EXERCISE 1.4

1. Obtain polynomial approximation to $f(x) = (1 - x)^{1/2}$ over [0, 1] by means of Taylor series about $x = 0$. Find the number of terms required in the expansion of obtain results correct to 5×10^{-1} for $0 \le x \le 1/2$.

2. Obtain a second degree polynomial approximation to $f(x) = (1 + x)^{1/2}, x \in [0, 0.1]$ using Taylor series expansions about $x = 0$. Use the expansions to approximate $f(0.5)$ and found to truncation error.

Answers

2. Truncation error = 0.625×10^{-4}

MULTIPLE CHOICE QUESTIONS (CHOOSE THE MOST APPROPRIATE ONE)

1. The number in which, there is no uncertainity and no approximation, is said to be:
(a) exact number
(b) approximate number
(c) both (a) and (b) are true
(d) none of these

2. The error in x^A, which is the approximate value of x^T is $E_p = 100 \times E_r = 100 \times \left|\frac{x^T - x^A}{x^T}\right|$ is called :
(a) absolute error
(b) relative error
(c) percentage error
(d) none of these

3. The error $E_r = \left|\frac{x^T - x^A}{x^T}\right|$ is known as :
(a) absolute error
(b) relative error
(c) percentage error
(d) none of these

4. The normalized absolute error is known as :
(a) relative error (b) relative error
(b) percentage (d) none of these

5. Inherent error is also known as :
(a) input error
(b) empirical error
(c) representation error
(d) none of these

6. Conversion error is also known as :
(a) input error
(b) empirical error
(c) representation error
(d) none of these

7. Addition of binary numbers (10110)2 and (1101)2 is :
(a) 110010 (b) 100011
(c) 101011 (d) 110011

Answers

1. (a) **2.** (c) **3.** (b) **4.** (a) **5.** (a) **6.** (c) **7.** (b)

ARCHIVE

1. The height of an observations tower was estimated to be 47m whereas its actual height was 45m. Find the percentage error in the measurement. [AGRA–2000]

2. If $u = \frac{3xy}{z^2} = f(x, y, z)$. Find maximum relative error. [LUCKNOW–2011; ROHTAK–2008; AVADH–2004, 08]

3. Find the relative error in the function we have taking log on both sides we get. [COIMBATORE(BE)–2002, 05]

4. Suppose that you have a task of measuring the length of a bridge and a river and come up with 9999 and 9cm respectively. If the true values are 10000 and 10 cm respectively compute the percentage error in each case. [PUNE(B.Tech.)–2004]

Answers

1. 4.44% **2.** 0.0004 **3.** $(E_R)_{\max} \le m_1\left|\frac{\delta x_1}{x_1}\right| + m_2\left|\frac{\delta x_2}{x_2}\right| + \ldots + m_n\left|\frac{\delta x_n}{x_n}\right|$

4. 0.01%, 10%

❑❑❑❑