## Errors And Their Analysis

Following are the broad sources of errors in numerical analysis

### (i) Input errors

The input information is rarely exact since it comes from the experiments and any experiment can gives ~~results~~ results of only limited accuracy. Moreover, the quantity used can be represented in a computer for only a limited number of digits.

### (ii) Algorithmic errors

If direct algorithms based on a finite ~~sq~~ sequence of operations are used, errors due to limited steps does not amplify the existing errors but if infinite algos are used, ideally exact results are expected only after an infinite number of steps. As this cannot be done in practice, the algorithm has to be stopped after a finite number of steps and as a consequence the result are not exact.

### (iii) Computational errors

Even when elementary operations, such as multiplication and division are used, the number of digits increases greatly so the results cannot be held fully in register available in a given computer; In such cases, a certain number of digits must be discarded. Furthermore, the errors here accumulates one after another from operation to operations, changing during the process and producing new errors.

## Accuracy of Number

### 1. Approximate numbers →

There are two types of numbers : Exact and Approximate

Exact numbers are $2, 4, 9, \frac{7}{2}, 6.45, \ldots$ etc.

But there are numbers such that

$\frac{4}{3} = 1.3333\ldots$, $\sqrt{2} = 1.414213\ldots$, $\pi = 3.141592\ldots$

which can not be expressed by a finite number of digits. These may be approximated by numbers 1.333, 1.4141 and 3.1416 respectively such numbers which represent the given numbers to a certain degree of accuracy are called approximate numbers.

### 2. Significant digits.

The digits used to express a number are called significant digits.

The digits 1, 2, 3, 4, 5, 6, 7, 8, 9, are significant digits. '0' is also a significant digit except when it is used to fix the decimal point or to fill the places of unknown or discarded digits.

For example number 7845, 3.589, .4758 contains 4 significant figures, while the number .00386, .000587, .0000296 contain only three significant figures. Since zeros only help to fix the position of decimal point.

Similarly, in the number .0003090, the first four '0's are not significant digits since they serve only to fix the position of the decimal point and indicate the place values of the other digits. The other two '0's are significant.

Rounding off

# Accuracy of Number

## 1. Approximate numbers →

There are two types of numbers : Exact and Approximate

Exact numbers are $2, 4, 9, \frac{7}{2}, 6.45, \ldots$ etc.

But there are numbers such that

$\frac{4}{3} = 1.3333\ldots$, $\sqrt{2} = 1.414213\ldots$, $\pi = 3.141592\ldots$

which can not be expressed by a finite number of digits. These may be approximated by numbers 1.333, 1.4141 and 3.1416 respectively such numbers which represent the given numbers to a certain degree of accuracy are called approximate numbers.

## 2. Significant digits.

The digits used to express a number are called significant digits.

The digits 1, 2, 3, 4, 5, 6, 7, 8, 9, are significant digits. '0' is also a significant digit except when it is used to fix the decimal point or to fill the places of unknown or discarded digits.

For example number 7845, 3.589, .4758 contains 4 significant figures, while the number .00386, .000587, .0000296 contain only three significant figures. Since zeros only help to fix the position of decimal point.

Similarly, in the number .0003090, the first four '0's are not significant digits since they serve only to fix the position of the decimal point and indicate the place values of the other digits. The other two '0's are significant.

Rounding off

Note:

1. The significant figure in a number in positional notation consists of
   (i) All non-zero digits
   (ii) Zero digits which
      (a) Lie between significant digits
      (b) Lie to the right of decimal point and at the same time, to the right of non-zero digits.
      (c) Are specifically indicated to be significant.
2. The significant figure in a number written in scientific notation (eg $\pm M \times 10^K$) consists of all the digits explicitly in M. Here $\frac{1}{10} \leqq M < 1$

| Number | significant digits | No. of significant digits |
|---|---|---|
| 3969 | 3, 9, 6, 9 | 04 |
| 3060 | 3, 0, 6 | 03 |
| 3900 | 3, 9 | 02 |
| 39.69 | 3, 9, 6, 9 | 04 |
| .3969 | 3, 9, 6, 9 | 04 |
| 39.00 | 3, 9, 0, 0 | 04 |
| .00039 | 3, 9 | 02 |
| .00390 | 3, 9, 0 | 03 |
| 3.0069 | 3, 0, 0, 6, 9 | 05 |
| $3.9 \times 10^6$ | 3, 9 | 02 |
| $3.909 \times 10^5$ | 3, 9, 0, 9 | 04 |
| $6 \times 10^{-2}$ | 6 | 01 |

## Rounding off

There are numbers with large number of digits e.g $\frac{22}{7} = 3.142857143$. In Practice it is desirable to limit such numbers to a manageable number of digits such as 3.14 or 3.143. This Process of dropping unwanted digits is called rounding-off.

Numbers are rounded-off according to following rule:

To round off a number to n significant digits, discard all digits to the right of nth digit and if this discarded number is

(i) Less than 5 in $(n+1)^{th}$ place, leave the nth digit unaltered.
e.g 7.893 to 7.89

(ii) Greater than 5 in $(n+1)^{th}$ place, increase the nth digit by Unity e.g 6.3456 to 3.346.

(iii) Exactly 5 in $(n+1)^{th}$ place, increase the nth digit by Unity if is odd otherwise leave it unchanged
e.g $12.675 \approx 12.68$
$12.685 \approx 12.68$

| Number | Rounded-off to | | |
|---|---|---|---|
| | Three digits | Four digits | Five digits |
| .543241 | .543 | .5432 | .54324 |
| 39.5255 | 39.5 | 39.52 | 39.526 |
| 69.4155 | 69.4 | 69.42 | 69.416 |
| .667676 | .668 | .6677 | .66768 |

3

## Errors

Errors = True Value - Approximate Value

In Any numerical ~~Analysis~~ computation, We come across following type of errors:

1. Inherent errors
2. Rounding errors
3. Truncation errors
4. Absolute errors
5. Relative errors
6. Percentage errors.

### 1. Inherent Errors :

Errors which are already Present in the statement of a Problem before its solution are called inherent errors. Such errors arise either due to the given data being approximate or due to limitations of mathematical tables, calculators or the digital computer.

Inherent errors can be minimized by taking better data or using High Precision.

Accuracy refers to the no. of significant digit in a value eg. 53.965 is accurate to 5 significant digits. ~~Prec~~ Precision Refers no. of decimal Position or order of magnitude of the last digit in value. eg in 53.965, Precision is $10^{-3}$.

### 2. Rounding Errors :

They arise from the Process of rounding off the numbers during the computation. It is also called Procedual error or numerical error. such errors are unavoidable in most of the calculations due to limitations of computing aids

These errors can be reduced however by

(i) Changing the calculation Procedure so as to avoid subtraction of nearly numbers or division by a small number

(ii) Retaining atleast one more significant digit at each step and rounding off at last step.

Rounding off may be executed in two ways

(a) **Chopping**: In it extra digits are dropped by truncation of ...

(b) **Symmetric round off**: In it, the last retained significant digit is rounded up by unity if the first discarded digit is $\geq 5$ otherwise the last retained digit is unchanged.

## (3) Truncation Errors

They are caused by using approximate results or on replacing an infinite process by a finite one.

For Example If $S = \sum_{i=1}^{\infty} a_i x_i$ is replaced by or truncated to $S = \sum_{i=1}^{n} a_i x_i$ then error developed is called Truncation error.

Truncation error is a type of algorithm error. Also

If $e^x = 1 + x + \frac{x^2}{\lfloor 2} + \frac{x^3}{\lfloor 3} + \frac{x^4}{\lfloor 4} + \cdots + \infty = X$ is truncated to

$1 + x + \frac{x^2}{\lfloor 2} + \frac{x^3}{\lfloor 3} = x'$ (say), then Truncation error $= X - x'$

## (4) Absolute Error

Absolute Error is the numerical difference between the true of a quantity and its approximate value.

Thus if $x$ is the true value of a quantity and $x'$ is the approximate value then $|x - x'|$ is called absolute error $e_a$.

$$e_a = |x - x'| = |\text{Error}|$$

(5) Relative Error:

Relative error is denoted by $e_r$ and defined as

$$e_r = \left|\frac{X - X'}{X}\right|$$

Where X is true value and $X - X'$ is error.

(6) Percentage Error:

Percentage Error is denoted by $e_p$ and defined as

$$e_p = \left|\frac{X - X'}{X}\right| \times 100$$

Note:

1. If a number is ~~used~~ correct to n decimal places, Then

~~Error = $\frac{1}{2}(10^{-n})$~~

$$\text{Error} = \frac{1}{2}(10^{-n})$$

For Example If the number 3.1416 is correct to 4 decimal places, then Error $= \frac{1}{2}(10^{-4}) = .00005$

2. If the first significant digit of a number is k and the number is correct to n significant digits then,

$$\text{Relative error} < \frac{1}{(k \times 10^{+n-1})}$$

Note: 2

To estimate the errors which creep in when the numbers in a calculation are truncated or rounded off to certain number of digits, the following rules are useful

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

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

(4) Relative error due to Rounding off to k digits.

$$= \left|\frac{x - x'}{x}\right| < \frac{1}{2} 10^{1-k}$$

**Note**

(1) If a number is correct to n significant digits, then maximum relative error $\leq \frac{1}{2}(10^{-n})$

If a number is correct to d decimal places, then absolute error $\leq \frac{1}{2} 10^{-d}$

(2) If the first significant figure of a number is k and the number is correct to n significant figures, then relative error $< \frac{1}{k(10^{n-1})}$

(1) Suppose 1.414 is used as an approximation to $\sqrt{2}$. Find the absolute and relative errors.

Sol$^n$

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

Approximate value $= 1.414$

Error = True value − Approximate value

$= 1.41421356 - 1.414$

$= .00021356$

$\therefore$ Absolute Error $= e_a = |\text{Error}| = |.00021356|$

$= .21356 \times 10^{-3}$

Relative Error $= e_r =$ [illegible]

$= \frac{e_a}{\text{True value}}$

$= \frac{.21356 \times 10^{-3}}{1.41421356} = .151 \times 10^{-3}$

(2) If .333 is approximate relative and

(2) If .333 is approximate value of $\frac{1}{3}$, find absolute, relative and percentage errors.

Sol$^n$

Given True Value $X = \frac{1}{3}$

Approximate value $X' = .333$

Absolute error $e_a = |X - X'|$

$= \left|\frac{1}{3} - .333\right| =$

$= |.33333 - .333| = .000333$

Relative error $= e_r = \frac{e_a}{X} = \frac{.000333}{\text{True Value}}$

$= \frac{.000333}{\frac{1}{3}} = .000999$

Percentage Error $e_p = e_r \times 100$

$=$ Relative error $\times 100$

$= .000999 \times 100$

$= .099\%$

(3) An approximate value of $\pi$ is given by 3.1428571 and its true value is 3.1415926. Find absolute and relative errors.

True Value $= 3.1415926$

Approximate Value $= 3.1428571$

Error $=$ True Value $-$ Approximate value

$= 3.1415926 - 3.1428571$

$= -.0012645$

Absolute error $= |\text{Error}| = |-.0012645| = .0012645$

Relative Error $= \frac{\text{Absolute Error}}{\text{True Value}} = \frac{.0012645}{3.1415926} = .00040250\overline{2}$

Ans

Q.4 Three approximate values of number $\frac{1}{3}$ are given as .30, .33 and .34 which of these three is the best approximation?

Sol$^n$ The best Approximation will be the one which has least absolute error.

$$\text{True value} = \frac{1}{3} = .3333$$

Case-I Approximate value = .30

$$\text{Absolute Error} = |\text{True value} - \text{Approximate value}|$$
$$= |.3333 - .30| = .03333$$

Case-II Approximate value = .33

$$\text{Absolute error} = |\text{True value} - \text{Approximate value}|$$
$$= |.3333 - .33| = .00333$$

Case-III Approximate value = .34

$$\text{Absolute error} = |\text{True value} - \text{Approximate value}|$$
$$= |.33333 - .034| = .00667$$

Since absolute error is least in case II. Hence .33 is the best approximation.

Q. Find the relative error if $\frac{2}{3}$ is approximated to .667

Sol$^n$ True value $= \frac{2}{3} = .66666$

Approximate value = .667

$\therefore$ Absolute error $e_a = |\text{True value} - \text{Approximate value}|$
$= |.66666 - .667| = .000334$

Relative Error $e_r = \frac{.000334}{.66666} = .0005$.

Q.1 Suppose 1.414 is used as an approximation to $\sqrt{2}$. Find the absolute and Relative errors.

Sol$^n$

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

Approximate Value $= 1.414$

Error = True Value − Approximate Value

$= \sqrt{2} - 1.414$

$= 1.41421356 - 1.414 = .00021356$

Absolute error = |Error|

$= |.00021356| = .21356 \times 10^{-3}$

$\therefore$ Relative Error $= \frac{e_a}{\text{True value}} = \frac{.21356 \times 10^{-3}}{\sqrt{2}} = .151 \times 10^{-3}$ Ans

Q.2 If 0.333 is the approximate value of $\frac{1}{3}$, find absolute, relative and Percentage errors.

Sol$^n$

True value $x = 1/3$

Approximate value $= x' = .333$

$\therefore$ Absolute error $\cdot e_a = |x - x'| = \left|\frac{1}{3} - .333\right| = .000333$

Relative Error $\cdot e_r = \frac{e_a}{x} = \frac{.000333}{.333} = .000999$

Percentage error $e_p = e_r \times 100 = .000999 \times 100 = .099\ \%$

Q.3 An Approximate value of $\pi$ is given by 3.1428571 and its true value is 3.1415926. Find Absolute and relative errors.

Given True value $= 3.1415926$

Approximate value $= 3.1428571$

$\therefore$ Error = True value − Approximation value

$= 3.1415926 - 3.1428571 = -.0012645$

$\therefore$ Absolute Error $= e_a = |\text{Error}| = .0012645$

Relative Error $= \frac{e_a}{\text{True value}} = \frac{.0012645}{3.1415926} = .000402502$ Ans

Find the

Q. Round off the number 865250 and 37.46235 to four significant figures and compute $e_a$, $e_r$, $e_p$ in each case.

(i) Number rounded off to four significant digits = 865200

$\therefore X = 865250$

$X' = 865200$

$\therefore$ Error $= X - X' = 865250 - 865200 = 50$

$\therefore$ Absolute error $= e_a = |\text{error}| = 50$

Relative error $= e_r = \frac{e_a}{X} = \frac{50}{865250} = 5.77 \times 10^{-5}$

Percentage error $= e_p = e_r \times 100 = 5.77 \times 10^{-5} \times 100$

$= 5.77 \times 10^{-3}$ Ans

(ii) Number rounded off to four significant digits = 37.46

Given True Value $= 37.46235 = X$

Approximate Value $= X' = 37.46$

Error = True Value − Approximate Value

$= 37.46235 - 37.46 = .00235$

$\therefore$ Absolute Error $= e_a = |\text{Error}|$

$e_a = .00235$

Relative error $= \frac{\text{Absolute Error}}{\text{True Value}} = \frac{.00235}{37.46235} = 6.2729 \times 10^{-5}$

Percentage Error = Relative Error $\times 100$

$= 6.2729 \times 10^{-5} \times 10^{+2} = 6.2729 \times 10^{-3}$ Ans

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

Sol$^n$ Number Rounded off to Four significant digits = 75460

Absolute error $= |\text{True Value} - \text{App. Value}|$

$= |75462 - 75460| = 2$

Relative error $= e_r = \frac{\text{Absolute error}}{\text{True value}}$

$e_r = \frac{2}{75462} = .0000265$

Percentage error $= e_p = e_r \times 100$

$= .0000265 \times 100$

$= .00265$ Ans

Q. Find the absolute, relative and Percentage errors if $x$ is rounded off to three decimal digits. Given $x = .005998$.

Sol$^n$ The number rounded off to three decimal digits

$= .006$

Error $= .005998 - .006 = -.000002$

Absolute error $= e_a = |\text{Error}| = .000002$

Relative Error $= \frac{\text{Absolute error}}{\text{True value}} = \frac{.000002}{.005998} = .0033344$

Percentage Error $= e_p = e_r \times 100$

$= .0033344 \times 100$

$= .33344$

Ans

Q. Evaluate the sum $S = \sqrt{3} + \sqrt{5} + \sqrt{7}$ to 4 significant digits and find its absolute and relative errors.

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

$S = 1.732 + 2.236 + 2.646 = 6.614$

$\because \sqrt{3} = 1.732050808$ (True value)

$\sqrt{3}$ to 4 significant digit $= 1.732$ (Approximate value)

$\therefore$ error $= |1.732050 - 1.732| = .0005$

Similarly $\sqrt{5} = 2.236067977$

$\sqrt{5}$ to 4 significant digit $= 2.236$

error $= .0005$

Similarly error in $\sqrt{3}$ after 4 significant digits

$= .0005$

$\therefore$ Total error $= .0005 + .0005 + .0005$

$= .0015$

The total absolute errors shows that the sum is correct to 3 significant figure only

Now $S = \sqrt{3} + \sqrt{5} + \sqrt{7}$

$= 6.61$

$e_r = \frac{.0015}{6.61} = .0002$ Ans

Q1. Find the absolute error if the number $X = .00545828$ is

(i) Truncated to three decimal digits

(ii) Rounded off to three decimal digits.

Sol$^n$

$$X = .00545828$$
$$= .545828 \times 10^{-2}$$

(i) After Truncation to three decimal places its approximate value

$$x' = .545 \times 10^{-2}$$

$\therefore$ Absolute error $= |X - x'|$

$$= |.545828 \times 10^{-2} - .545 \times 10^{-2}|$$
$$= .000828 \times 10^{-2}$$

$\because$ Absolute error due to truncation to k digit $= |X - x'| < 10^{n-k}$

$\therefore$ $.000828 \times 10^{-2} < 10^{-2-3}$

$.828 \times 10^{-5} < 10^{-5}$ This Proves Rule - I

(ii) After rounding off to three decimal places, its approximate value $x' = .546 \times 10^{-2}$

$\therefore$ Absolute error $= |X - x'|$

$$= |.545828 - .546|\, 10^{-2}$$
$$= .000172 \times 10^{-2} = .172 \times 10^{-5}$$

Which is $< .5 \times 10^{-2-3}$. This proves Rule - II

hon x = 1.00545820

## Errors in Numerical Computations

① If $u = \frac{4x^2y^3}{3t}$ and error in $x, y, z$ be .001, compute the relative max. error in $u$ when $x=y=z=1$

Sol $u = \frac{4x^2y^3}{3t}$

$$\delta u = \frac{\partial u}{\partial x}\delta x + \frac{\partial u}{\partial y}\delta y + \frac{\partial u}{\partial z}\delta z$$

$$= \frac{8xy^3}{3t}\delta x + \frac{12x^2y^2}{3t}\delta y + \frac{16x^2y^3}{3t}\delta z$$

$$(\delta u)_{max} = \left|\frac{8xy^3}{3t}\delta x\right| + \left|\frac{12x^2y^2}{3t}\delta y\right| + \left|\frac{16x^2y^3}{3t}\delta z\right|$$

$$= \frac{8(1)(1)}{1}\times.001 + \frac{12(1)(1)}{(1)}(.001) + \frac{16(1)(1)}{(1)}(.001)$$

$$= .008 + .012 + .016$$

$$\delta u = .036$$

$\therefore$ Max Relative error $= \frac{\delta u}{u}$

$$= \frac{.036}{4}$$

$$= .009 \text{ Ans}$$

$u = \frac{4x^2y^3}{3t}$

$x=y=z=1$

$= 4$

② Compute the percentage error in the time period $T = 2\pi\sqrt{\frac{l}{g}}$ for $l = 1$ m if the error in the ~~magnitude~~ measurement of $l$ is .01. Here $g$ is constant

$T = 2\pi\sqrt{l/g}$

$\text{Log}\, T = \text{Log}\, 2\pi + \frac{1}{2}[\text{Log}\, l - \text{Log}\, g]$

$\frac{1}{T}\delta T = 0 + \frac{1}{2}\cdot\frac{1}{l}\delta l - 0$

$= \frac{1}{2}\cdot\frac{1}{1}(.01) = \frac{.01}{2} = .005$

$\frac{1}{T}\delta T \times 100 = .005 \times 100 = 1.5\%$ Ans

Q. If $u = 2v^6 - 5v$, find the percentage error in $u$ at $v = 1$ if error in $v$ is .05.

$u = 2v^6 - 5v$

diff. it

$\delta u = 12v\,\delta v - 5\,\delta v$

$\delta u = 12v\,\delta v - 5\,\delta v$

$\delta u = 12 \times 1 (.05) - 5(.05)$

$= (12-5).05 = .35$

$\therefore \frac{\delta u}{u} = \frac{.35}{2v^6 - 5v} = \frac{.35}{2-5} = -.1166$

$\frac{\delta u}{u} \times 100 = -.1166 \times 100 = -11.66\%$.

Percentage error $= -11.66$ Ans

Q. If $r = 3h(h^6 - 2)$, find the percentage error in $r$ at $h = 1$, if the percentage error in $h$ is 5.

$r = 3h(h^6 - 2)$

$r = 3h^7 - 6h$

$\delta r = 21h^6 \delta h - 6\delta h$

$\delta r = 21(1)(5) - 6(5) = 105 - 30 = 75$

$\frac{\delta r}{r} = \frac{75}{r} = \frac{75}{3h(h^6-2)} = \frac{75}{3(1-2)} = -25$

$\frac{\delta r}{r} \times 100 =$

$\delta r = (21h^6 - 6)\delta h$

$\frac{\delta r}{r} = \frac{(21h^6 - 6)}{3h(h^6-2)} \cdot \delta h$

Given Percentage error in $h$ i.e $\frac{\delta h}{h} \times 100 = 5$

$\frac{\delta r}{r} \times 100 = \frac{21h^6 - 6}{3(h^6-2)} \cdot \frac{\delta h}{h} \times 100$

$= \frac{21-6}{3(-1)}(5) = \frac{15}{-3} \times 5 = -25\%$.

Q. The ~~distance~~ 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 .15 cm is possible in its measurement, estimate the percentage error in computing the discharge.

Sol$^n$

$Q = kH^{5/2}$, Here k is constant

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

$\frac{1}{Q}\delta Q = 0 + \frac{5}{2}\cdot\frac{1}{H}\delta H$

$\frac{\delta Q}{Q} \times 100 = \frac{5}{2}\frac{1}{H}\delta H \times 100$

$= \frac{5}{2}\cdot\frac{1}{75}(.15)\times 100 = .5\%$

# Floating Point Representation of Number

There are two type of arithmetic operations available in Computer

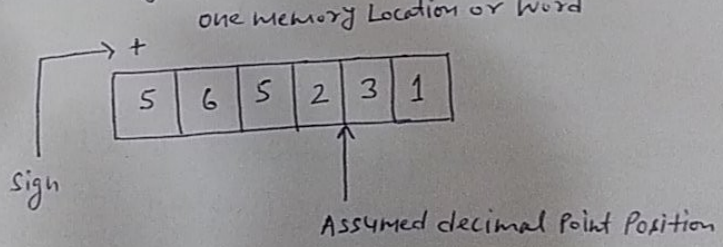
(i) Integer Arithmetic

Integer arithmetic deals with integer operands and used mainly in counting and as subscripts.

(ii) Real or ~~Fixed~~ Floating Point Arithmetic →

Real arithmetic uses numbers with fractional parts as operands and is used in most computations. ~~Computers~~ Computers are usually designed such that each location, called Word, in memory stores only a finite number of digits.

Let us assume a hypothetical computer having memory in which each location can store 6 digits and having provision to store one or more signs. One method of representing real numbers in that computer would be to assume a fixed position for the decimal point and store all numbers after appropriate shifting if necessary with an assumed decimal point.




(A memory location storing number 5652.31)

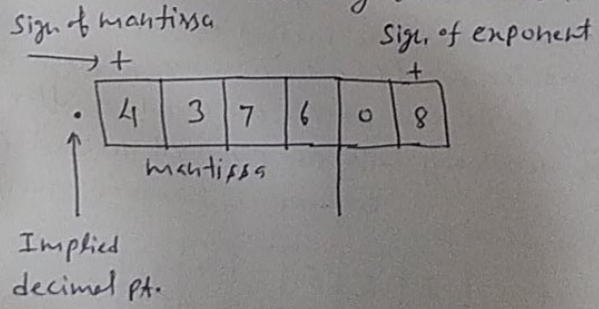
In such ~~a~~ convention, the max. and minimum possible numbers to be stored are 9999.99 and 0000.01 respectively in magnitude.

For this a new convention is adopted which aims to preserve the maximum number of significant digits in a real number and also increase the range of values of real numbers stored. This representation is called the normalized floating point mode representation.

In this mode, a real number is ~~expressed~~ expressed as combination of a mantiss and an exponent. The mantiss is made less than 1 or $\geq .1$ and the exponent is the power of 10 which multiplies the mantissa.

For example: The number $43.76 \times 10^6$ is represented in this notation as .4376 E8, where E8 is used to represent $10^8$. The mantissa is .4376 and Exponent is 8.

The number is stored in memory location as:




Moreover, the shifting of mantissa to the left till its most significant digits is non-zero is called normalization

e.g: The number .006031 may be stored as ~~.6031X~~ .6031 E-2.

Q.1 Add the following floating point number

(i) ~~·4546E5~~ ·4546 E5 and ·5433E5

Here Exponents are equal

∴ mantissa are added

∴ sum = ·9979E5 Ans

(ii) ·4546 E5 and ·5433 E7

Here Exponents are not equal.

The operand with the larger exponents is kept as it is

·5433 E7
·0045 E7
·5478 E7

(iii) ·4546 E3 and ·5433 E7

·5433 E7
·0000 E7
·5433 E7

(iv) ·6434 E3 and ·4845 E3

·6434E3 + ·4845E3 = 1.1279E3
= ~~·1127E4~~
= ·1127 E4 Ans

(v) ·6434 E99 and ·4845 E99

Here ·6434 + ·4845 = 1.1279E99 = ·1127 E100

Here Exponents part cannot store more than two digits.

This condition is called an overflow conditions and the arithmetic unit will intimate an error condition.

Q2. Subtract the following floating Point numbers:

(i) .9432E-4 from .5452E-3

.5452E-3
.0943E-3
.4509 E-3

(ii) .5424 E3 from .5452 E3

.5452 E3
.5424 E3
.0028 E3

, In normalized Floating Point, the mantis is ≥ .1 so Result in .28 E1

(iii) .5424 E-99 from .5452 E-99

.5452 E-99
.5424 E-99
.0028 E-99 or .28 E-101

But exponent becomes -100
But Exponent Part can not store more than two digit
Such condition is called underflow.

Q3

(i) .4546 E3 + .5454 E8

.4546E3
.5454E8

Sol'

.5454E8
+ .0000E8
.5454E8 Ans

(ii) $\cdot 9432E-4 - \cdot 6363E-5$

Sol$^n$

$$\begin{array}{r} \cdot 9432\ E-4 \\ -\ \cdot 0636\ E-4 \\ \hline \cdot 8796\ E-4 \end{array}$$

## Multiplication :—

Two numbers are multiplied in the normalized floating point mode by multipling the mantissas and adding the exponents After the multiplication of the mantissas, the result mantiss is normalized as in addition or subtraction operation and the exponent appropriately adjusted.

Q.1 Multiply the following point number

(i) $\cdot 5543\ E12$ and $\cdot 4111\ E-15$

Sol$^n$ $\cdot 5543\ E+12 * \cdot 4111\ E-15 = \cdot 2278\ E-3$ Ans

(ii) $\cdot 1111\ E10$ And $\cdot 1234\ E15$

Sol$^n$ $\cdot 1111\ E10 * \cdot 1234\ E15 = \cdot 013709\ E25$

$= \cdot 1370\ E24$ Ans

(iii) $\cdot 1111\ E51$ and $\cdot 4444\ E50$

Sol $\cdot 1111\ E51 * \cdot 4444\ E50 = \cdot 04937284\ E101$

$= \cdot 49372\ E100$

overflow

(iv) $\cdot 1234\ E-49$ And $\cdot 1111\ E-54$

Ans $\cdot 1234\ E-49 * \cdot 1111\ E-54 = \cdot 01370974\ E-103$

$= \cdot 1370\ E-102$

The result underflow

## Division

In division, the mantissa of the numerator is divided by that of the denominator. The denominator exponent is subtracted from the numerator exponent. The quotient mantissa is normalized to make the most significant digit non-zero and the exponent appropriately adjusted. The mantissa of the result is chopped down to occupy 4 digits.

Q1 (i) $.9998\ E1 \div .1000\ E-99$ ??

Sol$^n$ $\dfrac{.9998\ E1}{.1000\ E-99} = 9998\ E100$

(ii) $.1000\ E5 \div .9999\ E3$

Sol$^n$ $\dfrac{.1000\ E5}{.9999\ E5} = .1000\ E2$ Ans

(iii) $.9998\ E-5 \div .1000\ E98$

Sol$^n$ $\dfrac{.9998\ E-5}{.1000\ E98} = .9998\ E-104$

Hence Result is underflow Ans

Q2 For $x = .4845$ and $y = .4800$, Calculate the value of $\dfrac{x^2-y^2}{x+y}$ using normalized floating point airthmetic compare with the value of $x-y$.

Sol$^n$ $x^2 = .4845\ E0 \times .4845\ E0 = .2347\ E0$

$y^2 = .4800\ E0 \times .4800\ E0 = .2304\ E0$

$x^2-y^2 = .2347\ E0 - .2304\ E0 = .0043\ E0$

$x+y = .4845\ E0 + .4800\ E0 = .9645\ E0$

$$\therefore \frac{x^2-y^2}{x+y} = \frac{.0043\,E0}{.9645\,E0} = .004458\,E0$$

$$= .4458\,E-2$$

Now $x-y = .4845\,E0 - .4800\,E0$

$$= .0045\,E0 = .4500\,E-2$$

Relative Error $= \frac{.4500-.4458}{.4500} = .0093$ or

$= .93\%$.

# Loss of Significance

① Let $p = .54617$ and $q = .54601$. Use 4 digit arithmetic to approximate $p-q$ and determine the absolute and relative error using rounding.

Sol

$p = .54617$, $q = .54601$

After 4 digit rounding

$\bar{p} = .5462$, $\bar{q} = .5460$

$$\text{Error}_{\text{Relative}} = \frac{(p-q)-(\bar{p}-\bar{q})}{p-q} = \frac{(.54617-.54601)-(.5462-.5460)}{.54617-.54601}$$

$$e_x = \left|\frac{.00016-.0002}{.00016}\right| = .25$$

$$e_a = .00016-.0002 = .00004$$

② Compute $\sqrt{x+1}-1$, when $x = .12345678\times10^{-5}$ use 8 digit rounding. Rewrite the expression to avoid the subtraction

Soln $f(x) = \sqrt{x+1}-1$

Here $\sqrt{x+1} = \sqrt{1+.12345678\times10^{-5}} = \sqrt{1.00001234}$

$= 1.0000062$

(Rounding 8 significance digit)

$\therefore \sqrt{1+x}-1 = 1.0000062-1 = .0000062$

Here only 2 two significance digit

i.e Loss of significant digit

So to fix this problem rationalize the expression

$$\frac{\sqrt{1+x}-1}{1} \times \frac{\sqrt{1+x}+1}{\sqrt{1+x}+1} = \frac{x}{\sqrt{1+x}+1} = 6.17281995 \times 10^{-7}$$

i.e 8 significant digit