Set-I

ERRORS

INTRODUCTION:

Most of the numerical methods give answers that are approximations to the desired solutions. In this situation, it is important to measure the accuracy of the approximate solution compared to the actual solution. To find the accuracy we must have an ideai of the possible errors that can arise in computational procedures. Now we shall introduce different forms of errors which are common in numerical computations.

Numbers and their accuracy:

There are two kinds of numbers – exact and approximate numbers.

The numbers $1, 2, 3, \ldots, \frac{1}{2}, \frac{3}{2}, \frac{3}{5}, \ldots$ etc are all exact and $\pi, \sqrt{2}, e, \ldots$ etc; written in this manner are also exact.

Approximate numbers are those that represent the numbers to a certain degree of accuracy. i.e, an approximate number x' is a number that differs but slightly from an exact number x. The approximate value of π is 3.1416 and to a better approximation it is 3.14159265 but not exact value.

Significant digits (figures)

The digits that are used to express a number are called significant digits or significant figures.

— A significant digit of an approximate number is any non-zero digit in its decimal representation, or any zero lying between significant digits or used as place holder to indicate a retained place.

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

for eg:, in the number 0.0005010 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.

Two notational conventions which make clear how many digits of a given number are significant ~~are~~ given below.

1. The significant figure in a number in positional notation consists of:

a) All non zero digits
b) zero digits which (i) lie between significant figures

(ii) lie to the right of decimal point, and at the same time to the right of a non-zero digit.

(iii) are specifically indicated to be significant.

2. The significant figure in a number written in scientific notation ($M \times 10^n$) consists of all the digits explicitly in M.

– Significant figures are counted from left to right starting with the left most non zero digit.

Number	significant figures	No. of significant figures
37.89	3, 7, 8, 9	4
0.00082	8, 2	2
0.000620	6, 2, 0	3
3.506×10	3, 5, 0, 6	4
8×10^{-3}	8	1
3.14167	3, 1, 4, 1, 6, 7	6
2.35698	2, 3, 5, 6, 9, 8	6

Rounding-off numbers

Sometimes, we come across numbers with a large number of digits and in making calculations it might be necessary to cut them to a useable number of figures. This process is known as rounding-off and will be done by the following rule

To round-off a number to a significant digit, we shall discard all digits right of the n^{th} digit. If discarded number is

a) greater than '$\frac{1}{2}$' a unit, in the n^{th} place, the n^{th} digit would be increased by unity.

b) less than '$\frac{1}{2}$' a unit, in the n^{th} place, the n^{th} digit would be left unaltered.

c) exactly half a unit, in the n^{th} place, the n^{th} digit would be increased by unity, if add otherwise left unchanged.

Ex-1

Number	Round-off to three figures	four figures	five figures
00.522341	00.522	00.5223	00.52234
93.2155	93.2	93.22	93.216
00.66666	00.667	00.6667	00.66667

Ex-2

Number	Round-off to four significant figures
9.6782	9.678
29.1568	29.16
8.24159	8.242
30.0567	30.06

→ In numerical analysis, the analysis of error is of great importance. Errors may occur at any stage of the process of solving a problem. By the error we mean the difference between the true value and the approximate value.

$\therefore$ Error = True value − Approximate value.

Ex: The true value of π is 3.14159265.....

In some mensuration problems the value $\frac{22}{7}$ is commonly used as an approximation to π. what is the error in this approximation?

Sol^n: The true value of π is 3.14159265.....

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

$$\frac{22}{7} = 3.1428571 4$$

$\therefore$ Error = True value − approximate value

$$= -0.0012644 9.$$

NOTE: In this case the error is negative.

Error can be positive or negative. We shall in general be interested in absolute value of the error which is defined as

$$|\text{error}| = |\text{True value} - \text{approximate value}|$$

In the above example, the absolute Error is

$$|error| = |-0.00126449\ldots|$$
$$= 0.001264\ldots$$

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

$$|\text{Relative error}| = \left|\frac{\text{True value} - \text{approximate value}}{\text{True value}}\right|$$

i.e $$|\text{Relative error}| = \left|\frac{\text{error}}{\text{True value}}\right|$$

→ The percentage error is 100 times the relative error. It is denoted by Ep. ∴ Ep = Er × 100

The errors classified into 3 types.

1) Inherent error 2) Round off - error
3) Truncation error.

(1) → The Inherent error is that quantity which is already present in the statement of the problem before its solution.

The inherent error arises either due to the simplified assumptions in the mathematical formulation of the problem or due to the physical measurments of the parameters of the problem.

(2) → Round-off error: when depicting even rational numbers in decimal system or some other positional system, there may be

Round-off error:

Now for example

The true value of π is 3.14159265

- The value $\frac{22}{7}$ is commonly used as an approximation to π. $\left(\frac{22}{7} = 3.14285714\right)$.
- If we approximate π using 2 digits after the decimal point, say.

  Chopping off the other digits, then

  we have $\pi = 3.14$

  $\therefore$ The error in this approximation is 0.00159265....
- If we use 3 digits after the decimal point, then using chopping,

  we have $\pi = 3.141$

  $\therefore$ The error in this approximation is 0.00059265....
- Now suppose we consider the approximate value rounded-off to three decimal places. Already we know how to round-off a number which has non-terminal decimal expansion.

The value of π rounded-off to 3 digits is 3.142.

$\therefore$ The error in this approximation is -0.00040734....

which is smaller, in absolute value than 0.00059265....

$\therefore$ in general whenever we want to use only a certain number of digits after the decimal point, then it is always better to use the value rounded-off to that many digits because in this case the error is usually small.

The error involved in a process where we use rounding-off method is called round-off error.

→ We now discuss the concept of floating point arithmetic

- In scientific computations a real number x is usually represented in the form $x = \pm(.d_1 d_2 \cdots d_n)10^m$ where $d_1, d_2, \ldots, d_n$ are natural numbers between 0 & 9 and m is an integer called exponent. Writing a number in this form is known as "floating point representation".

We denote this representation by $fl(x)$.

- Such a floating point number is said to be normalized if $d_1 \neq 0$.
- To translate a number into floating point representation we adopt any of the two methods – rounding & chopping.

for example:

Suppose we want to represent the number 537 in the normalized floating point representation. with $n=1$, then we get

$fl(537) = .5 \times 10^3$ Chopped

$= .5 \times 10^3$ rounded

In this case we are getting the same representation in rounding and chopping.

- Now if we take $n=2$, then we get

$fl(537) = .53 \times 10^3$ Chopped

$= .54 \times 10^3$ rounded.

In this case, the representations are different.

- Now if we take $n=3$, then we get

$fl(537) = .537 \times 10^3$ Chopped

$= .537 \times 10^3$ rounded

- The number n in the floating point representation is called precision.

The difference between the true value of a number x and rounded $fl(x)$ is called round-off error. From the above discussion it is clear that the round-off error decreases when precision increases.

Def Let x be a real number and x^* be a real number having non-terminal decimal expansion, then we say that x^* represents x rounded to k decimal places if $|x - x^*| \leq \frac{1}{2} 10^{-k}$, where $k > 0$ is a positive integer.

Def Let x be a real number and x^* be an approximation to x. Then we say that x^* is accurate to k decimal places if

$$\frac{1}{2} 10^{-(k+1)} \leq |x - x^*| \leq \frac{1}{2} 10^{-k}. \quad \text{--- (1)}$$

for example:

find out to how many decimal places the value of $\frac{22}{7}$ is accurate as an approximation to π = 3.141592653 ?

Sol Now we have $|\pi - \frac{22}{7}| = 0.00126499\ldots$

i.e $|x - x^*| = 0.00126499\ldots$

$\pi = 3.14159265$

$\frac{22}{7} = 3.14285714$

Now $0.0005 < 0.00126 < 0.005$

i.e $\frac{5}{10000} < 0.00126 < \frac{5}{1000}$

i.e $\frac{1}{2} 10^{-3} < 0.00126 < \frac{1}{2} 10^{-2}$

i.e $\frac{1}{2} 10^{-(2+1)} < 0.00126 < \frac{1}{2} 10^{-2}$

$\therefore$ The eq (1) is satisfied for $k = 2$.

$\therefore$ we conclude that the approximation is accurate to '2' decimal places.

2.

Note: Round-off errors can create serious difficulties in lengthy computations. suppose we have a problem which involves a long calculations.

In the course of these computations many rounding errors (some positive, and some negative) may occur in a number of ways. At the end of the calculations these errors will get accumulated and we don't know the magnitude of this error. Theoretically it can be large. But, in reality, some of these errors (b/w +ve & -ve errors) may get cancelled so that the accumulated error will be much smaller.

*(3) Truncation Error:

In the evaluation of certain quantities, which involve infinite mathematical process, we simply work with finite terms only, this leads to error in the final results and this error is called Truncation error.

Eg: while representing any function by an infinite Taylor's series, the remainder term gives us the measure of Truncation error.

Let $f(x) = \sum_{n=0}^{\infty} a_n (x-x_0)^n$ denote the Taylor's series of $f(x)$ about x_0.

In practical situations, we can not, in general find the sum of an infinite

number of terms, so we must stop after a finite number of terms, say N.

i.e $f(x) = \sum_{n=0}^{N} a_n (x-x_0)^n$

and ignoring the rest of the terms

i.e $\sum_{n=N+1}^{\infty} a_n (x-x_0)^n$

$\therefore$ Truncating error is given by

$$TE = f(x) - \sum_{n=0}^{N} a_n (x-x_0)^n$$

$$= \sum_{n=N+1}^{\infty} a_n (x-x_0)^n.$$

i.e the magnitude of the error in the value of the function due to cutting (truncating) of its series is equal to the sum of all the discarded terms.

Miscellaneous information regarding Significant figures:

There are certain rules for determining the number of significant figures. These are stated below.

(1) All non-zero digits are significant.

for example: In 285 cm, there are three significant figures and in 0.25 mL, there are two significant figures.

(2) All the zeros between two non-zero digits are significant, no matter where the decimal point is, if at all.

for example: 2.005 has four significant figures.

(3) Zeros preceding to first non-zero digit are not significant. Such zero indicates the position of decimal point. (or)

If the number is less than 1, the zero(s) on the right of the decimal point but to the left of the first non-zero digit are not significant.

for example (1) 0.002308, here the underlined zeros are not significant.

(2) 0.03 has one significant figure and 0.0052 has two significant figure.

(4) → Zeros at the end or right of a number are significant provided they are on the right-side of the decimal point.

For example, 0.200g has three significant figures.

But, if otherwise, the zeros are not significant for example, 100 has only one significant figure.

(or)

→ The terminal or trailing zero(s) in a number without a decimal point are not significant

For example

$123m = 12300cm = 123000mm$ has three significant figures 1, 2, 3; the trailing zero(s) being not significant.

However, the trailing zero(s) in a number with a decimal point are significant.

for example: 3.500 or 0.6900 have four significant figures each.

(5) → Exact numbers have an infinite number of significant figures.

for example, In 2 balls or 20 eggs, there are infinite significant figures as these are

exact numbers and can be represented by writing infinite number of zeros after placing a decimal i.e, $2 = 2.00000$

(or) $20 = 20.00000$

* There can be some confusion regarding trailing zero(s).

Suppose a length is reported to be 4.700m. It is evident that the zeros here are meant to convey the precision of the measurement and are, therefore significant.

[If these were not, it would be superfluous (unnecessary) to write them explicitly, the reported measurement would have been simply 4.7m)

Now suppose we change the units, then

$4.700\,m = 470.0\,cm = 4700\,mm = 0.004700\,km$

Since 4700 has trailing zeros in a number with no decimal, so here 4 & 7 are the only two significant figures, while infact it has four significant figures and mere change of units cannot change the no. of significant figures.

* To remove such ambiguities in determining the number of significant figures, the best way is to report every measurement in scientific notation (in power of 10).

In this notation, every number can be expressed as $M \times 10^n$, where M is a number between 1 and 10; and n is any +ve or −ve exponent (or power) of 10.

for example:

(1) 4.01×10^2 has three significant figures

(2) 8.256×10^3 has four significant figures.

(3) $4.700\,m = 4.700 \times 10^2\,cm = 4.700 \times 10^3\,mm$
$= 4.700 \times 10^{-3}\,km$

has four significant figures.

Here the power of 10 is irrelevant to the determination of significant figures. However, all zeros appearing in the base number(M) in the scientific notation are significant. Each number in this case has 4 significant figures.