Solution: i. $0.5453\,E12 \times 0.3111\,E15 = 0.16964283\,E3$ (D*)
$= 0.1696\,E3$

ii. $0.2222\,E10 \times 0.1234\,E15 = 0.02741948\,E25$
In normalized floating point, the mantissa is ≥ 0.1.
So, the result is $0.2741948\,E24 = 0.2741\,E24$ (D)

iii. $0.2121\,E51 \times 0.3334\,E50 = 0.07071414\,E101$
$= 0.7071414\,E100$
The result overflows.

iv. $0.1234\,E\text{-}48 \times 0.1111\,E\text{-}55 = 0.01370974\,E\text{-}103$
$= 0.1370974\,E\text{-}104$ (D)
The result overflows.

1.3.4 Division

In case of division, the mantissa of the numerator is divided by that of the denominator. The denominator exponent is subtracted from the numerator exponent. After the division of mantissas, the resulted mantissa is normalized as in addition or subtraction operation and the exponent appropriately adjusted.

Example 5: Solve the following floating point numbers:

i. $0.8998\,E1 \div 0.1000\,E46$
ii. $0.8989\,E5 \div 0.1000\,E97$
iii. $0.1000\,E4 \div 0.9999\,E2$
iv. $0.9432\,E2 \div 0.1000\,E98$

Solution:

i. $0.8998\,E1 \div 0.1000\,E46 = 8.998\,E47$
$= 0.8998\,E48$

ii. $0.8989\,E5 \div 0.1000\,E97 = 8.989\,E102$
$= 0.8989\,E101$
The result underflows.

iii. $0.1000\,E4 \div 0.9999\,E2 = 0.1000\,E2$

iv. $0.9432\,E2 \div 0.1000\,E98 = 9.432\,E100$
$= 0.9432\,E101$
The result overflows.

Example 6: Find the solution of the following equation using floating point arithmetic with 4-digit mantissa

$$x^2 - 1000x + 25 = 0$$

Give comments on the results so obtained.

Solution: We have

$$x^2 - 1000x + 25 = 0$$

$$x = \frac{10^3 \pm \sqrt{(10)^6 - (10)^2}}{2}$$

* D stands for discarded

or $$\frac{x_a}{x} > 0.98 \text{ or } x_a > 0.98x$$

or $$x < \frac{x_a}{0.98} = \frac{35.25}{0.98} = 35.9693877551$$

Therefore, we have

$$34.5588235294 < x < 35.9693877551$$

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

$$34.5588 < x < 35.9694.$$

EXERCISE 1.2

1. If 0.333 is the approximate value of 1/3, find absolute, relative and percentage errors. [Ans. 0.000333, 0.000999, 0.099%]
2. If true value = 10/3, approximate value = 3.33, then find the absolute and relative errors. [Ans. $E_A = 0.003333$, $E_R = 0.000999$]
3. Round off the numbers correct to four significant digits: 2.26325, 35.46735, 4585561, 0.60035, 0.000023317. [Ans. 2.263, 35.47, 458600, 0.6004, 0.00002332]
4. If $x = 2.536$, find the absolute error and relative error when (i) x is rounded off, ii x is truncated to two decimal digits. [Ans. 0.004, 0.0015772]
5. Define absolute, relative and percentage error.
6. What do you mean by truncation error? Explain with example.
7. Find the relative error of the number 8.6 if both of its digits are correct. [Ans. $E_A = 0.05$, $E_R = 0.0058$]
8. Find the percentage error if 625.483 is approximated to three significant digits. [Ans. E_A =0.483, $E_R = 0.000772$, $E_p = 0.077\%$]
9. Rounded off the number 75462 to four significant digits and then calculate the absolute error and percentage error. [Ans. $E_A = 2$, $E_R = 0.0000265$, $E_p = 0.00265$]
10. The height of an observation tower was estimated to be 47 metre whereas its actual height was 45 metre. Calculate the percentage relative error in the measurement. [Ans. 4.44%]
11. If the number x is correct to four decimal places, what will be the error. [Ans. 0.00005]

1.11 MACHINE EPSILON

A floating point number system within a computer is always limited by the finite word length of computers. This means that only a finite number of digits can be represented. As a result, numbers that are too large or too small can not be represented. Hence even in string an exact decimal number in its converted form in the computer memory, an error is occurred. This error is machine dependent and is called machine epsilon.

Error = True value – Approximate value

1.12 GENERAL FORMULA FOR TERMS

Let $u = g(x_1, x_2, x_3, ..., x_n)$ be a function of n variables $x_1, x_2, ..., x_n$ and let the error in any x_i be Δx_i, where $i = 1, 2, ..., n$. Then, the error Δu in u is given by

$$u + \Delta u = g(x_1 + \Delta x_1, x_2 + \Delta x_2, ..., x_n + \Delta x_n)$$

Solution: If we retain n terms then $(n+1)^{\text{th}}$ term $= (-1)^n \dfrac{x^{2n+1}}{2n+1}$

For $x = 1$, $(n+1)^{\text{th}}$ term $= \dfrac{(-1)^n}{2n+1}$

For the determination of $\tan^{-1}(1)$ correct to eight significant digit accuracy

$$\left|\frac{(-1)^n}{2n+1}\right| < \frac{1}{2} \times 10^{-8}$$

$$\Rightarrow \quad (2n+1) > 2 \times 10^8$$

which is satisfied by $n = 10^8 + 1$.

Example 32: Find the number of terms of the exponential series such that their sum gives the value of e^x correct to six decimal places at $x = 1$.

Solution:
$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^{n-1}}{(n-1)!} + R_n(x)$$

where
$$R_n(x) = \frac{x^n}{n!} e^w, \; 0 < w < x$$

Maximum absolute error at $(w = x) = \dfrac{x^n}{n!} e^x$ and maximum relative error $= \dfrac{x^n}{n!}$

Hence, $(E_R)_{\text{max.}}$ at $x = 1$ is $\dfrac{1}{\text{n!}}$

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

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

which gives $n = 10$.

Hence we need 10 terms of series (1) in order that its sum is correct to 6 decimal places.

EXERCISE 1.3

1. If $u = \dfrac{4x^2y^3}{z^4}$ and errors in x, y, z be 0.001, compute the relative maximum error in u when $x = y = z = 1$. [Ans. 0.009]
2. 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. [Ans. 0.5]
3. Compute the percentage error in the time period $T = 2\pi\sqrt{l/g}$ for $l = 1$ m if the error in the measurement of l is 0.01. [Ans. 0.5%]
4. If $y = 3x^7 - 6x$, find the percentage error in y at $x = 1$ if the error in x is 0.05. [Ans. −0.25%]

2. $f_0 \leftarrow f(x_0)$
3. $f_1 \leftarrow f(x_1)$
4. for $i = 1$ to n in steps of 1 do
5. $x_2 \leftarrow (x_0 f_1 - x_1 f_0)/(f_1 - f_0)$
6. $f_2 \leftarrow f(x_2)$
7. If $|f_2| \leq e$ then
8. begin write 'convergent solution', x_2, f_2
9. Stop end
10. If sign $(f_2) \neq$ sign (f_0)
11. then begin $x_1 \leftarrow x_2$
12. $f_1 \leftarrow f_2$ end
13. else begin $x_0 \leftarrow x_2$
14. $f_0 \leftarrow f_2$ end
    end for
15. Write 'Does not converge in n iterations'
16. Write x_2, f_2
17. Stop

Example 3: Find the root of the equation $\tan x + \tan hx = 0$ which lies in the interval (1.6, 3.0) correct to four significant digits using method of false position.

Solution: Let $f(x) = \tan x + \tan hx = 0$

Since $f(2.35) = -0.03$ and $f(2.37) = 0.009$

Hence, the root lies between 2.35 and 2.37. Take $x_0 = 2.35$ and $x_1 = 2.37$. Using Regula-Falsi method

$$x_2 = x_0 - \left(\frac{x_1 - x_0}{f(x_1) - f(x_0)}\right) f(x_0)$$

$$= 2.35 - \left(\frac{2.37 - 2.35}{0.009 + 0.03}\right)(-0.03)$$

$$= 2.35 + \frac{0.02}{0.039}(0.03) = 2.365$$

Now $f(x_2) = f(2.365) = -0.00004$, i.e. –ve

Hence, the root lies between 2.365 and 2.37.

$$x_3 = 2.365 - \left(\frac{2.37 - 2.365}{0.009 + 0.00004}\right)(-0.00004)$$

$$= 2.365 + \frac{0.005}{0.00904} \times 0.00004 = 2.365$$

Hence the required root is 2.365 correct to four significant digits.

EXERCISE 2.2

Find a real root of the following equations correct to three decimal places by using Regula-Falsi method.

1. $xe^x = 2$ [Ans. 0.852]
2. $2x - \log_{10} x = 7$ [Ans. 3.789]
3. $x \tan x = -1$ [Ans. 2.798]