

FLOATING POINT ARITHMETIC

In digital computer a number is stored in its memory. The memory is divided into separate cells called words. Each word can hold the same number of digits called bits, with respect to its base plus a sign. Negative numbers are stored as absolute values plus a sign. One in complement form. The number of digits which can be stored in a computer word is called its word length. The word length varies from one computer to another. The numbers in the computer word can be stored in two forms

- (i) fixed point form
- (ii) Floating point form

In fixed point form a t -digit number is assumed to have its decimal point at the left hand end of the word. This implies that all numbers are assumed to be less than 1 in magnitude.

A floating point number is a number represented in this form

$$.d_1 d_2 \dots d_t \times \beta^e \dots (1)$$

where $d_1, d_2, \dots, d_t$ are integers and satisfy $0 \leq d_i < \beta$ and the exponent e is such that $m \leq e < M$. Here β is the base.

The fractional part $.d_1 d_2 \dots d_t$ is called the mantissa and it lies between $+1$ and -1 . The number 0 is written as $+ .000 \dots 0 \times \beta^e$

A non-zero floating point number (1) is in normal form if the value of the mantissa consist of exactly $[-1, -1/\beta]$ or in the interval $[-1/\beta, 1]$ i.e., $d_1 \neq 0$

A non-zero floating point number (1) is in t -digit mantissa standard form if it is normalized and its mantissa consists of exactly t -digits.

If a number x has the representation in the form $x = .d_1 d_2 \dots d_t d_{t+1} \dots \times \beta^e$ then the floating number $f(x)$ in t -digit mantissa standard form (2) can be obtained in the following two ways

(i) Chopping: - Here we neglected $d_{t+1}, d_{t+2}, \dots$ in (2) and obtained $f(x) = .d_1 d_2 \dots d_t$

(ii) Rounding: -

The following rules are followed to round off numbers. Suppose we desire to retain digits up to k th decimal place of a number ' x ' and let $x = a_n a_{n-1} \dots a_0 . a_{-1} a_{-2} \dots a_{-k} a_{-k+1} \dots$

The following alternation is to be done at the k th decimal place.

while rounding

- (a) if $a_{-k+1} < 5$, then a_{-k} is not changed.
- (b) if $a_{-k+1} \geq 5$, then a_{-k} is increased by one.

(c) if $a_{-k+1} = 5$, increase a_{-k} by one if a_{-k} is odd and leave it if it is even. The resulting approximation to x is said to be rounded off to k th decimal place.

The rule (a) is said to be rounding down while rules (b) and (c) are referred to as rounding up.

Significant Digits :-

The first non-zero and all the digits to its right of a number are defined to be the significant digits of the number.

For example the significant digits of the number 2.103 are 2, 1, 0, 3 and that of 0.010 are 1, 0 (between 1 and 3) and 3.

Rounding off a number to k -significant figures or digits simply means to round off the number at the k th significant digit following the rules of round off. For example, the number 0.0136 when rounded off to two significant digits become 0.014.

In chopping of a number, we discard all the digits after certain decimal place in the number for example, chopping the number 3.012466 at the 4th decimal place it becomes 3.0124.

Problem :-

Subtract following floating point numbers 0.36143447×10^7 and 0.36132346×10^7 . Express the result in normalized floating point form.

Solution :- we have

$$\begin{array}{r} 0.36143447 \times 10^7 \\ 0.36132346 \times 10^7 \\ \hline 0.00011101 \times 10^7 \end{array}$$

The result is a floating point number but not a normalized floating point number due to the presence of the three leading zeroes. Shifting the fractional part three places to the left we get the result 0.11101×10^4 which is a normalized floating point number.

Example :-

Find the sum of 0.123×10^3 and 0.456×10^2 and write the result in three digit mantissa.

Solution :-

The number of the smaller magnitude is adjusted so that its exponent is same as that of the number of larger magnitude we have

$$\begin{array}{r} 0.123 \times 10^3 \\ + 0.0456 \times 10^3 \\ \hline 0.1686 \times 10^3 \end{array}$$

$$= \begin{cases} 0.168 \times 10^3 & \text{for chopping} \\ 0.169 \times 10^3 & \text{for rounding} \end{cases}$$

Problem:-

compute the mid-point of the numbers $a = 4.968$ and $b = 6.762$ using the four digit arithmetic.

Solution:-

$$c = \frac{a+b}{2} = \frac{0.4968 \times 10^1 + 0.6762 \times 10^1}{2} = \frac{0.1133 \times 10^1}{2} = 0.5665 \times 10^1$$

However using the formula $c = a + \frac{b-a}{2}$ we have

$$c = 0.4968 \times 10^1 + \frac{0.6762 \times 10^1 - 0.4968 \times 10^1}{2} = 0.4968 \times 10^1 + 0.0977 \times 10^1 = 0.5665 \times 10^1$$

The result obtained by two different methods is same. so the correct result is 0.5665×10^1

Problem:-

Find the smaller root of the equation $x^2 - 400x + 1 = 0$ using four digit arithmetic.

Solution:-

For the equation $ax^2 - bx + c = 0$, $b > 0$ the smaller root is given by

$$x = \frac{b - \sqrt{b^2 - 4ac}}{2a}$$

Here $a = 1 = 0.1000 \times 10^1$, $b = 400 = 0.4000 \times 10^3$, $c = 1 = 0.1000 \times 10^0$

$$b^2 - 4ac = 0.1600 \times 10^6 - 0.0000 \times 10^6 = 0.1600 \times 10^6 \text{ (to four digit accuracy)}$$

$$\sqrt{b^2 - 4ac} = 0.4000 \times 10^3$$

putting the above formula we get $x = 0.0000$

However if we write the above formula in the form $x = \frac{2c}{b + \sqrt{b^2 - 4ac}}$

$$\text{we get } x = \frac{0.2000 \times 10^0}{0.4000 \times 10^3 + 0.4000 \times 10^3} = \frac{0.2000 \times 10^0}{0.8000 \times 10^3} = 0.0025$$

which is the exact root.

Example:-

using 5-digits floating point arithmetic, find the product of $x = 1/3$ and $y = 5/7$

Solution:-

Let $*$ denote the multiplication operation in 5-digit floating point arithmetic.

$$x = 1/3 = 0.33333 \dots \rightarrow fl(x) = 0.33333 \text{ (chopping being used)}$$

$$y = 5/7 = 0.7142857 \dots \rightarrow fl(y) = 0.71428 \text{ (chopping being used)}$$

(03)

Alpha Bismajit Team

$$\text{Now } f(x) \times f(y) = 0.83333 \times 0.4428 = 0.2380907$$

$$\Rightarrow f(x) \times f(y) = 0.23809 \text{ (chopping being used)}$$

$$\therefore x * y = 0.23809$$

This result agrees upto 5-decimal places with the exact value of the product of x and y .

Drawbacks of k -digit Arithmetic:-

The following problems are usually encountered in k -digit arithmetic

(i) Loss of Accuracy:-

This explained through the following examples

consider the addition of the numbers $x = 1/3$ and $y = 1234$ with $k = 4$ (chopping being used)

$$\text{Now } f(x) = 0.3333, \text{ and } f(y) = 0.1234 \times 10^4$$

In this case the two numbers in floating point form are having different exponents. The operand with large exponent is kept as it is and the mantissa of the operand with smaller exponent is adjusted so as to make the exponent equal to the larger exponent. (this is how the addition is carried out by machines)

$$\text{In this case } f(x) = 0.3333 = 0.00003333 \times 10^4 = 0.0000 \times 10^4$$

Since we are dealing with 4-digit floating point arithmetic.

$$\text{therefore } f(x) + f(y) = 0.1234 \times 10^4 \Rightarrow f(x + y) = 0.1234 \times 10^4$$

which is the final result of addition of two numbers performed by the computer. The final result shows that x is effectively value as zero compared to y ; hence there is a Loss of Accuracy.

(ii) Loss of significance:-

When two nearly equal numbers are subtracted, there is a loss of significant digits of figures

to illustrate this consider the following example

$$\text{Let } x = 0.623781 \text{ and } y = 0.6238567$$

$$\text{Then } f(x) = 0.6237 \text{ and } f(y) = 0.6238 \text{ (} k=4 \text{)}$$

$$\therefore f(x) - f(y) = 0.0001 = 0.1 \times 10^{-3}$$

which has only one significant digit whereas the numbers x and y have respectively six and seven significant digits thus there is a loss of significant digits.

This phenomenon is called Loss of significance or catastrophic cancellation.

Errors: - Error = True value - Approximate value

Relative Error = $\frac{|\text{Error}|}{|\text{True value}|}$

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

Inherent Error: -

The inherent error is that quantity which is already present in the statement of the problem before its solution.

Round-off Error: -

The round-off error is the quantity ϵ which must be added to the finite representation of a computed number in order to make it the true representation of that number.

Truncation Error: -

The truncation error is the quantity τ which must be added to the true representation of the quantity in order that the result be exactly equal to the quantity we are seeking to generate.

Example: -

Suppose $x = 0.025$ is approximated by $a = 0.02$, the absolute error in such approximation is given by
 $EA = 0.025 - 0.02 = 0.005$ and consequently the relative error $ER = 0.005/0.025 = 0.2$
The percentage of error $EP = 100 \times 0.2 = 20$

Example: -

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 find a bound on the truncation error.

Solution: - we have

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

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

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

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

Thus, Taylor series expansion with remainder term may be written as
 $(1+x)^{1/2} = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{1}{16} \frac{x^3}{[(1+\theta)^{5/2}]^5}, 0 < \theta < 0.1,$

The truncation term is given by $T = \frac{1}{16} \frac{x^3}{[(1+\theta)^{5/2}]^5}$
we have $f(0.05) = 1 + \frac{0.05}{2} - \frac{(0.05)^2}{8} = 0.10246875 \times 10^1$

The bound of the truncation error, for $x \in [0, 0.1]$ is $|T| \leq \max_{0 \leq x \leq 0.1} \frac{(0.1)^3}{16(1+x)^2} \leq \frac{(0.1)^3}{16}$

EXERCISE

Q1. convert the following decimal numbers to binary, octal and hexadecimal numbers :

(i) 43 (ii) 1256 (iii) 78.5 (iv) 81.25 (v) 189.3125

Q2. convert the following numbers to decimal form

(i) $(111.11)_2$ (ii) $(101010.0011)_2$ (iii) $(232.1)_8$ (iv) $(254)_8$

(v) $(A03.3)_{16}$ (vi) $(211)_{16}$

Q3. convert the following octal and Hexa decimal numbers to binary forms.

(i) $(4725)_8$ (ii) $(333)_8$ (iii) $(6458)_{16}$ (iv) $(A012)_{16}$ (v) $(1AB.9C)_{16}$

Q4. convert the following binary numbers to octal and Hexadecimal forms :

(i) $(10110111)_2$ (ii) $(1110001101)_2$ (iii) $(100011100110)_2$ (iv) $(1110010.11101)_2$

(v) $(110011011.1101011011)_2$

Q5. (a) chop the following numbers to k -decimal places :

(i) 2.0123, $k=3$ (ii) 50.52050, $k=4$
(iii) 12.310, $k=2$ (iv) 581.301085, $k=5$
(v) -31.0002, $k=3$ (vi) -3.000012, $k=4$

(b) chop the following numbers to k -significant digits

(i) 2.0212, $k=3$ (ii) -15.1205, $k=4$
(iii) 0.00512, $k=2$ (iv) $1/3$, $k=5$
(v) 10.00015, $k=3$ (vi) $2/3$, $k=3$

Q6. (a) Round off the following numbers to k -decimal places

(i) 3.01243, $k=4$ (ii) -9.19351, $k=3$
(iii) 13.01256, $k=4$ (iv) 2.9501, $k=1$
(v) 4.12531, $k=2$ (vi) 2.912, $k=2$

(b) Round off the following numbers to k -significant digits

(i) 14.2312, $k=2$ (ii) 0.00401, $k=2$

(iii) 10.3150 , $k=4$, (iv) 0.001101 , $k=1$

(v) 0.0046 ; $k=1$, (vi) 1.20451 , $k=4$

07. Suppose x is any approximation to a real number α . Answer the following's:

(a) If $x = 0.3000 \times 10$ and $\alpha = 0.3100 \times 10$, then find the absolute error and relative error.

(b) Find also the absolute error and relative error if $x = 0.3000 \times 10^3$ and $\alpha = 0.3100 \times 10^3$

(c) $x = 0.3000 \times 10^4$ and $\alpha = 0.3100 \times 10^4$. Find the absolute error, relative error and percentage of error

08. Find the product of 0.2431×10^4 and 0.4213×10^2 , using 4-digit arithmetic

09. Convert $(0.7)_{10}$ to the binary form. If only 7 bits are retained in the binary fraction then what will be round off error.

10. Convert $(764.36)_8$ to the binary form.