Floating Point Arithmetic

Floating point arithmetic, also known as real point arithmetic, uses the numbers with fractional parts as operands and it is used in most of the computations. In computers, the numbers are stated in two forms:

- Fixed point
- Floating point

Fixed point is used to represent integers and the floating point is used to represent real numbers. Floating point number system, $F(\beta, k, m, M)$ can be defined as a subset of the real number system which is characterized by the parameters:

b : The base
k : The number of digits in the base b expansion.
m : The minimum exponent.
M : The maximum exponent.

Elements of $F(\beta, k, m, M)$ can also be expressed a $x = \pm (0.\ d_1\ d_2 \ldots d_n)_\beta\ \beta^e$ where $d_1, d_2, \ldots, d_n$ are all digits in the base β and all d's lie between 0 and β, e is an integer called the exponent and it always lies between m and M such that $m \le e \le M$ where m and M vary with computers.

The fractional parts could also be written as $(0.\ d_1\ d_2\ d_3\ \ldots..\ d_n)_\beta = d_1 \times \beta^{-1} + d_2 \times \beta^{-2} + \ldots + d_n \times \beta^{-n}$ which is called mantissa which always lies between –1 and 1 and it also restricts the size of number. The leading digit in the mantissa is always made non-zero by shifting and adjusting the value of the exponent accordingly. Shifting the mantissa to the left till the non-zero digit is called normalization. For instance, if we have to normalize 42.53×10^6 then after normalization this could be written as 0.4253×10^8 or this could be written as 0.4253E8 (where E8 represents 10^8). Therefore, 0.4253 is the mantissa and E8 is the exponent. Also the number 0.004678 in normalized floating point could be stored as 0.4678E-2.

Note: A floating point $x = \pm (0.\ d_1 d_2\ \ldots..d_n)_\beta\ \beta^e$ is said to be normalized if $d_1 \neq 0$ or else $d_1 = d_2 = \ldots\ d_n = 0$.

Arithmetic Operations

(a) Addition of Normalized Floating Point: For adding two normalized floating points, we first have to make their exponents equal by shifting the mantissa.

For Example: Add 0.7642*E*4 and 0.4253*E*6.
Solution: The exponent of a number with the smallest exponent is increased by 2 so that 0.7642*E*4 becomes 0.0076 *E*6.

Then 0.7642*E*4 + 0.4253*E*6 = 0.0076*E*6 + 0.4253*E*6
= 0.4329E6.

(b) Subtraction of Normalized Floating Point: This operation is performed by adding negative normalized floating points.

For Example: Subtract 0.4673*E*-4 from 0.8542*E*-5.

Solution: The smallest exponent is *E*-5 so we increase the exponent of 0.8542*E*-5 by 1 and it becomes 0.0854*E*-4, therefore 0.4673*E*-4 – 0.0854*E*-4 = 0.3819*E*-4.

(c) Multiplication of Normalized Floating Point: In order to multiply two floating points, we multiply their mantissa and add their exponents. The mantissa is only four digits of the resulting mantissa which are retained by dropping the rest of the digits.
Note: We multiply mantissa × mantissa and Exponent × Exponent.

For Example: Multiply 0.5634*E*11 × 0.1532*E*-14.

Solution: 0.5634 × 0.1532 = 0.08631288 and *E*11 × *E*-14 = *E*-3
Therefore, 0.5634*E*11 × 0.1532*E*-14 = 0.08631288 *E*-3.
Now the leading digit of mantissa should be non-zero, therefore 0.08631288*E*-3 becomes 0.8631288*E*-4 = 0.8631 *E*-4.

(d) Division of Normalized Floating Point: In this operation, the mantissa of numerator is divided by the mantissa of the denominator and the exponent of denominator is subtracted from the exponent of the numerator. The quotient mantissa obtained is normalized by retaining 4 digits and the exponent is suitably adjusted.

For Example: Divide 0.2000*E*5 by 0.8883*E*3.

$$\textbf{Solution: } \frac{0.2000}{0.8883} = 0.2251 \text{ and } \frac{E5}{E3} = E2$$

$$\text{Therefore } \frac{0.2000\ E2}{0.8883\ E3} = 0.2251E2.$$

Example 1. Multiply the following floating point numbers 0.1222*E*10 and 0.2143E15.

Solution: 0.1222 × 0.2143 = 0.02618746 and *E*10 × *E*15 = *E*25.
Therefore, 0.1222*E*10 × 0.2143*E*15 = 0.02618746*E*25
= 0.2618746*E*24
= 0.2618*E*24 (Retaining 4 digits in mantissa)

Example 2. Apply the procedure for the following multiplications :

(i) $(0.3554 \times 10^{9}) \times (0.1123 \times 10^{-25})$

(ii) $(0.1111 \times 10^{74}) \times (0.3000 \times 10^{80})$.

Solution: (i) Firstly 0.3554×10^{9} can be written as $0.3554E9$ and 0.1123×10^{-25} can be written as $0.1123E$-25.

Now, $0.3554 \times 0.1123 = 0.03991142$ and $E9 \times E$-$25 = E$-16

Therfore , $0.3554\ E9 \times 0.1123\ E$-$25 = 0.03991142E$-$16 = 0.3991142E$-17

$= 0.3991\ E$-17. (Retaining 4 digit in mantissa)

(ii) 0.1111×10^{74} and 0.3000×10^{80} can be written as $0.1111E74$ and $0.3000E80$ respectively.

Now, $0.1111 \times 0.3000 = 0.03333$ and $E74 \times E80 = E\ 154$

Therefore, $0.1111E74 \times 0.3000E80 = 0.03333E154 = 0.3333E153$.

Note: Overflows -The result overflows if the exponent is greater than 99.

Underflows-The result underflows if the exponent is less than 99.

In (ii) part the result overflows as the exponent is greater than 99.

Example 3. For $x = 0.8454$ and $y = 0.8400$ calculate the value of $\dfrac{x^2 - y^2}{x + y}$ using normalized floating point arithmetic. Compare with the value of $(x - y)$. Indicate the error in the former.

Solution: As $x = 0.8454 = 0.8454E0$

$y = 0.8400 = 0.8400E0$

Therefore, $x + y = 0.8454E0 + 0.8400E0$

$= 1.6854E0$

$= 0.16854E1$ (Normalizing)

$= 0.1685E1$ (Retaining 4 digits)

$x - y = 0.8454E0 - 0.8400E0$

$= 0.0054E0$

$= 0.5400E - 2$ (Normalizing)

Also, $x^2 = x \times x = 0.8454E0 \times 0.8454E0$

$= 0.71470116E0.$

$= 0.7147E0$ (Retaining 4 digits)

$$y^2 = y \times y = 0.8400E0 \times 0.8400E0 = 0.7056E0.$$

Now $$x^2 - y^2 = 0.7147E0 - 0.7056E0 = 0.0091E0$$
$$= 0.9100E - 2.$$

Therefore, $$\frac{x^2 - y^2}{x + y} = \frac{0.9100\ E - 2}{0.1685E1} = 5.40059347\ \text{E–3}$$
$$= 0.5400\ E\text{-}2 \text{ (Retaining 4 digits in mantissa)}$$

Now, as we are considering only the 4 digits in mantissa

Therefore, the value of $(x - y)$ and $\frac{x^2 - y^2}{x + y}$ are same .

$$\text{Relative error} = \frac{\text{Value of } (x - y) - \text{Value of } (x^2 - y^2)/(x + y)}{\text{Value of } (x - y)}$$

$$= \frac{0.500E\text{-}2 - 0.500E\text{-}2}{0.5400E\text{-}2}$$

Thus relative error = 0% .

Example 4. Given $e = 2.1738$, calculate the value of e^x when $x = 0.2505E1$. The expression for e^x is given by $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$

Solution: As we know that $e^{0.2505E1} = e^{2.505} = e^2 \times e^{0.505}$

Also, $$e = 2.1738 = 2.1738E0 = 0.2173E1$$

As $$e^2 = e \times e$$
$$= (0.2173\ E1) \times (0.2173\ E1)$$
$$= 0.04721929\ E2$$
$$= 0.4721\ E1$$

Also $$e^{0.505} = 1 + (0.505) + \frac{(0.505)^2}{2!} + \frac{(0.505)^3}{3!}$$
$$= 1.505 + 0.1275125 + 0.0214646$$
$$= 1.6539771$$
$$= 1.653 = 0.1653\ E1$$

Therefore, $$e^{0.2505\ E1} = e^2 \times e^{0.505}$$
$$= 0.4721\ E1 \times 0.1653\ E1$$
$$= 0.07803813\ E2$$
$$= 0.7803\ E1. \quad \text{(Normalized)}$$

Example 5. In case of normalized floating points representation, associative and distributive laws are not always valid. Give example to prove this statement.

Solution: Consider $a = 0.6556\ E1$

$b = 0.6555\ E\text{-}1$

$c = 0.6144\ E1$

$$a + b = 0.6556\ E1 + 0.6555\ E\text{-}1 = 0.6556E1 + 0.0065E1$$
$$= 0.6621\ E1$$

Now, $$(a + b) - c = 0.6621\ E1 - 0.6144\ E1$$
$$= 0.0477\ E1$$
$$= 0.4770\ E0$$
$$a - c = 0.0412\ E1$$
$$= 0.4120\ E0$$

Now $$(a - c) + b = 0.4120\ E0 + 0.6555\ E\text{-}1$$
$$= 0.4120E0 + 0.0655E0$$
$$= 0.4775\ E0$$

Obviously $(a + b) - c \neq (a - c) + b$.

This shows that associative law does not hold.

Again consider

$$a = 0.5555\ E1$$
$$b = 0.5454\ E1$$
$$c = 0.5435\ E1$$
$$b - c = 0.5454\ E1 - 0.5435\ E1$$
$$= 0.0019\ E1 = 0.1900\ E\text{-}1$$

Now $$a(b - c) = 0.5555\ E1\ (0.1900\ E\text{-}1)$$
$$= 0.105545\ E0$$
$$= 0.1055\ E0$$

$$ab = 0.5555\ E1 \times 0.5454\ E1$$
$$= 0.3029\ E2$$

$$ac = 0.5555\ E1 \times 0.5435\ E1$$
$$= 0.3019\ E2$$

$$ab - ac = 0.3029\ E2 - 0.3019\ E2$$
$$= 0.001\ E2$$
$$= 0.1000\ E0$$

Clearly $a(b - c) \neq ab - ac$

This shows that distributive law does not hold.

SIGNIFICANT FIGURES

The digits 1, 2, 3, 4, ... 8, 9 which are used to express a number are known as significant digits or figures. Zero also plays a role of significant digit except when it is used to fix the decimal point to fill the places of unknown or discarded digits.

When we have to find significant digits, the followings points should always be kept in mind:

1. If the number is in positional notation, then the significant figures in the number consists of

(i) All non-zero digits

(ii) The zero digit which lies between significant digits and lies to the right of decimal point and at the same time, to the right of a non-zero digit.

For instance if we have find the significant digits of 0.00789 then the significant digits will be 7, 8, 9.

2.

If it is in scientific notation (i.e. $k \times 10^n$), then the significant digits are all digits explicitly in k.

For instance, significant digits in 6×10^{-4} is 6.

Note: Significant digits are counted from left to right starting with the left most non zero digits.

ERRORS

In any numerical computation, there are several types of error. Now we will be discussing those errors.

1. Inherent Error: Errors which are already present in the problem even before its solution are called inherent errors. These errors arise because of the given data which are being approximated or because of the limitation of the computing aids such as mathematical tables, disk calculators, or the digital computers.

There are some ways by which we can minimize inherent error. Some of them are by taking better data, by correcting obvious errors in the data or using computing aids of high precision which in fact is closely related to the significant digits, i.e. precision refers to the number of decimal position or order of magnitude of the last digit. For instance : In 3.265431, precision is 10^{-6} and in 3.45, precision is 10^{-2}.

2. Rounding off Error: Errors which arises in the process of rounding off the numbers during computation. Firstly we will understand how to round off numbers. The process of cutting off unwanted digits, and retaining as many as desired is called round-off.

There are certain rules while rounding off any number. To round off a number to n significant digits, discard all digits to the right of nth digit according to the following rule.

(i) If the discarded number is less than 5 at the $(n + 1)$th place, leave the nth digit as such, for example 1.864 becomes 1.86 and 2.383 becomes 2.38.
(ii) If the discarded number is greater than 5 at $(n + 1)$th place, add 1 to the nth digit, for example 3.679 becomes 3.68 and 1.867 becomes 1.87.
(iii) If the discarded number is exactly 5 at $(n + 1)$ th place, leave the nth digit unchanged if it is even, for example 58.125 becomes 58.12.
(iv) If the discarded number is exactly 5 at $(n + 1)$ th place, add 1 to the nth digit if it is odd, for example 4.3775001 becomes 4.378.

These errors are unavoidable in most of the calculations due to the limitation of the computing aids. However, there are some ways by which we can reduce them by the following rules:
(i) Rounding off error can be reduced by retaining at least one more significant digit at each step than that given in the data and rounding off at the last step.
(ii) Rounding off error can also be reduced by changing the calculation procedure so as to avoid subtraction of nearly equal numbers or division by a small number.

3. Absolute error: Absolute error is the numerical difference between the true value of a quantity and its approximate value. If Y is the true value of a quantity and $Y¢$ be its approximate value, then $|Y - Y'|$ is the absolute error.
It is donated by Ea. Here $Ea = |Y - Y'|$.

Example: Let $Y = 1.253$ and $Y' = 1.25$

$$Ea = |Y - Y'| = |1.253 - 1.25| = 0.003$$

4. **Relative error:** The relative error is the absolute error divided by the magnitude of the exact value. It is given by $Er = \left|\frac{Y - Y'}{Y}\right|$.

   Example: Let $Y = 1.253$ and $Y' = 1.25$

$$Er = \left|\frac{Y - Y'}{Y}\right| = \left|\frac{0.003}{1.253}\right| = 0.0023.$$

5. **Percentage error:** The percentage error is the relative error expressed in terms of per 100 and is given by $Ep = 100\ Er = 100\left|\frac{Y - Y'}{Y}\right|$.

   Example: $Ep = 100\left|\frac{0.003}{1.253}\right| = 0.23.$

6. **Truncation Error:** Truncation error is defined as the error which is created on approximating a mathematical procedure.

   Let us take exponential series $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots\infty = X(\text{say})$

If it is replaced by $1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} = X'(\text{say})$ (here we are approximating a mathematical procedure), then the truncation error $= E_T = X - X'$.

Here, we are truncating a mathematical procedure. When we convert infinite mathematical procedure to finite mathematical procedure, then the error which arises is called truncation error.

Consider, $\int_a^b f(x)\,dx$ in Figure 1. We have to find the area of the curve.

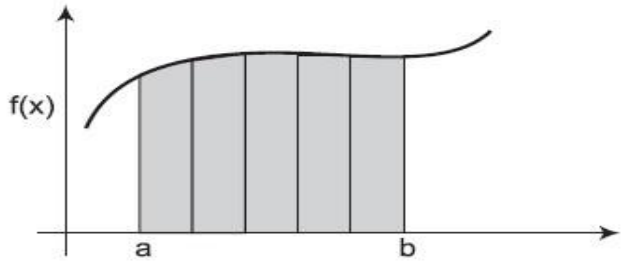

Fig. 1

In this figure, these rectangles do not represent exact area of curve from *a* to *b*, we have to draw infinite numbers of rectangles to find the exact area. Now in converting infinite numbers of rectangles to finite numbers, the error which arises is truncation error. Taylor series is used to solve many numerical methods and the error which is produced here by neglecting higher order terms is known as Truncation Error. Now we will further understand Local and Global Truncation Error.

(a) Local Truncation Error: Consider the trapezoidal rule for numerical integration which can be written as

$$\int_a^{a+h} f(x)dx \approx \frac{h}{2}[f(a)+f(a+h)]$$

Define the function $F(t)=\int_a^t f(x)dx$

We represent F by its Taylor Polynomial with the remainder as follows:

$$F(a+h)=F(a)+hF^{I}(a)+\frac{h^2}{2!}F^{II}(a)+\frac{h^3}{3!}F^{III}(C_1)\ \forall\ C_1\in[a,a+h]$$

Since $F^{I}=f$, $F^{II}=f^{I}$, $F^{III}=f^{II}$ and $F(a)=0$

We have

$$\int_a^{a+h} f(x)dx = F(a+h)=hf(a)+\frac{h^2}{2!}f^{I}(a)+\frac{h^3}{3!}f^{II}(C_1) \qquad ...(1)$$

Also, the Taylor polynomial with remainder f is

$$f(a+h)=f(a)+hf^{I}(a)+\frac{h^2}{2!}f^{II}(C_2)\forall C_2\in[a,a+h]$$

which gives

$$\frac{h}{2}[f(a+h)+f(a)]=hf(a)+\frac{h^2}{2}f^{I}(a)+\frac{h^3}{4}f^{II}(C_2) \qquad ...(2)$$

The error in the trapezoidal rule is the difference of eq. (1) and (2)

$$\int_a^{a+h} f(a)dx-\frac{h}{2}[f(a+h)+f(a)]=\frac{h^2}{6}f^{II}(C_1)-\frac{h^3}{4}f^{II}(C_2)$$

Now we have to show that is sufficiently smooth that is continuous and bounded on $[a, a+h]$. We can combine these two terms.

If $m \le f^{II} \le M\ \forall\ x\in[a,a+h]$

Then $$\frac{h^3}{6}m\le\frac{h^3}{6}f^{II}(C_1)\le\frac{h^3}{6}M$$

and $$\frac{h^3}{4}m\le\frac{h^3}{4}f^{II}(C_2)\le\frac{h^3}{4}M$$

So, $$\frac{-h^3}{12}m\le\frac{h^3}{6}f^{II}(C_1)-\frac{h^3}{4}f^{II}(C_2)\le\frac{-h^3}{12}M\ .$$

By applying intermediate value theorem to f^{II}, there is some point n in the interval $[a, a+h]$ s.t.

$$f^{II}(n)=\frac{-12}{h^3}\left[\frac{h^3}{6}f^{II}(C_1)-\frac{h^3}{4}f^{II}(C_2)\right]$$

Thus, we have

$\int_a^{a+h} f(x)dx - \frac{h}{2}[f(a+h) + f(a)] = -\frac{h^3}{2} f^{II}$(n) for some n ∈ [$a$, $a + h$].

This is known as Local Truncation Error which comes from truncating the Taylor series expansions, for one step of the trapezoidal rule.

(b) Global Truncation Error: It is used to to improve the results. Subdivide the interval into n equal subintervals [a, x_1], [x, x_2], ... [x_{n-1}, b] and apply the method in each region. The length of each subinterval is $h = (b - a)/n$ which gives the more general trapezoidal rule.

$$\int_a^b f(x)dx \approx \frac{h}{2}[f(a) + 2f(x_1) + \ldots\ldots 2f(x_{n-1}) + f(b)]$$

The global error is the result of adding the local error in each of the regions. If f is sufficiently smooth, the global error can represented as

$$-\frac{(b-a)h^2}{12} f^{II}(n) \text{ for some } a \le n \le b$$

Thus, the global error for the trapezoidal rule is proportional to h^2, and the method is of order h^2. This means that if the step size is cut in half, the bound on the global truncation error is reduced by a factor of one fourth.

Example 1. Find the significant digits in the following:
(i) 9353 (iii) 53.07 (iii) 0.0460

Solution: (i) 9353 The significant digits are 9, 3, 5, 3.
(ii) 53.07 The significant digits are 5, 3, 0, 7.
(iii) 0.0460 The significant digits are 4, 6, 0.

Example 2. Round off the following numbers to three significant digits.
(i) 8.894 (ii) 23.865 (iii) 9.4356 (iv) 5.8254

Solution: (i) 8.894 becomes 8.89
(ii) 23.865 becomes 23.9
(iii) 9.4356 becomes 9.44
(iv) 5.8254 becomes 5.82

Example 3. Round off the numbers 665250 and 27.46235 to four significant digits and compute

Solution: (i) 665250 is rounded off to four significant digits = 665200

Here $Y = 665250$ and $Y' = 665200$

$$Ea = |Y - Y'| = |665250 - 665200| = 50$$

$$Er = \left|\frac{Y - Y'}{Y}\right| = \frac{50}{665250} = 7.52 \times 10^{-5}$$

Ea, Er, Ep in each case.

and $$Ep = 100\, Er = 100 \times 7.52 \times 10^{-5}$$
$$= 7.52 \times 10^{-3}$$

(ii) 27.46235 is rounded off to four significant digits = 27.46.

In this case $Y = 27.46235$ and $Y' = 27.46$.

Therefore, $$Ea = |Y - Y'| = |27.46235 - 27.46|$$
$$= 0.00235$$

$$Er = \left|\frac{Y - Y'}{Y}\right| = \frac{0.00235}{27.46235} = 8.56 \times 10^{-5}$$

$$Ep = 100\, Er = 100 \times 8.56 \times 10^{-5} = 8.56 \times 10^{-3}$$

Example 4. If 0.333 is the approximate value of $\frac{1}{3}$, find absolute, relative and percentage error.

Solution: Here $$Y = \frac{1}{3} = 0.333333 \text{ and } Y' = 0.333$$

$$Ea = |Y - Y'| = |0.333333 - 0.333| = 0.000333.$$

$$Er = \left|\frac{Y - Y'}{Y}\right| = \frac{0.000333}{0.333333} = 0.000999\,.$$

$$Ep = 100\, Er = 100 \times 0.000999 = 0.0999\%.$$

Example 5. Evaluate the sum $S = \sqrt{5} + \sqrt{7} + \sqrt{8}$ to 4 significant digits and find its absolute and relative error.

Solution: As $\sqrt{5} = 2.236$ (After rounding off)

$$\sqrt{7} = 2.646$$

$$\sqrt{8} = 2.828$$

$$S = \sqrt{5} + \sqrt{7} + \sqrt{8} = 2.236 + 2.646 + 2.828$$
$$= 7.710$$

But the true value of $\sqrt{5} + \sqrt{7} + \sqrt{8} = 7.71024641331057$

Therefore, $$Ea = |7.71024641331057 - 7.710| = 0.00024641331057$$

And $$Er = \frac{Ea}{S} = \frac{0.00024641331057}{7.710} = 0.000031960$$
$$= 3.196 \times 10^{-5}$$

Example 6. Suppose 1.732 is used as an approximation to $\sqrt{3}$ then find the absolute and relative error.

Solution: Here Y (True value) = 1.732050807 and $Y' = 1.732$

Therefore, $$Ea = |Y - Y'| = |1.732050807 - 1.732| = 0.000050807$$

and $$Er = \left|\frac{Y - Y'}{Y}\right| = \frac{0.000050807}{1.732050807} = 0.000029333$$

$$= 2.933 \times 10^{-5}$$

Example 7. Obtain a second order degree polynomial approximation to $f(x)=(1+x)^{\frac{1}{2}}$, where $x\in[0,0.1]$ by using the Taylor series expansion about $x=0$. Use the expansion to approximate $f(0.05)$ and find a bound of the truncation error.

Solution: We have

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

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

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

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

Thus, the Taylor series expansion with remainder term may be written as

$$(1+x)^{1/2} = 1+\frac{x}{2}-\frac{x^2}{8}+\frac{1}{16}\frac{x^3}{\left[(1+\xi)^{\frac{1}{2}}\right]^5},\ 0<\xi<0.1$$

The truncation term is given by

$$T = (1+x)^{\frac{1}{2}} - \left(1 + \frac{x}{2} - \frac{x^2}{8}\right)$$

$$= \frac{1}{16} \frac{x^3}{\left[(1+\xi)^{\frac{1}{2}}\right]^5}$$

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

The bound of the truncation error, for $x \in [0, 0.1]$ is

$$|T| \leq \frac{\max}{0 \leq x \leq 0.1} \frac{(0.1)^3}{16\left[(1+x)^{\frac{1}{2}}\right]^5}$$

$$\leq \frac{(0.1)^3}{16} = 0.625 \times 10^{-4}.$$