Course Title:	CONM	Semester:	2	Date:	02.01.2023
Subject Module:	1	Subject Code:	05BH0201	Faculty:	Keshavi Mehta

UNIT 1 : FLOATING POINT ARITHMETIC

Course Content:

1) Errors : Addition Operation, Subtraction Operation, Division Operation, Multiplication Operation

2) Types of Errors : Data Errors, Truncation Errors, Round off Errors, Computational Errors

3) Measure of Accuracy : Absolute Error, Relative Error

❖ Error :

An error is defined as the difference between the actual value (true value) and the approximate value obtained from the experimental observation or from numerical computation.

Let x denote the actual value of a quantity and x_a denote its approximated value. Then the Error is defined as,

Error = Actual Value – Approximated Value

∴ Error = $x - x_a$

Example : Find Error when actual value is 0.987642 and approximated value is 0.987630.

The errors in the computed results can be classified in the following categories:

(1) Errors in Input Data:

- Due to Approximate measurements, known as Data Errors.

- Due to Truncation of the digits, known as Truncation Errors.

- Due to Rounding off the digits, known as Round off Errors.

(2) Computational Errors:

- Due to pitfalls in computational algorithm.

- Due to truncation of digits during the arithmetic operation, because of limitations of storage.

- **Data Error**

These are errors that occur due to inaccurate measurements or observations that may be due to limitations of the measuring device. For Example, Vernier calliper, screw gauge etc. can measure the quantity accurate to certain smallest value. The accuracy of the measurements also depends on the experience of the person.

- **Truncation Error**

Truncation Errors occur when some digits from the number are discarded.

There are mainly two situations when truncation error occurred –

(i) During the representation of numbers in Normalized floating point form. Because in this only few digits in the mantissa can be accommodated, for Example only four digits in our hypothetical computer.

(ii) During the conversion of a number from one system to another.

Example : Truncate Last 4 digits from the following numbers.

(i) 0.90012346

Answer : 0.9001

(ii) 0.8912390821

Answer : 0.891239

(iii) 0.98567342

Answer: 0.9856

- **Round off Error**

The Errors occurr due to rounding off the digits is called Round off Errors.

- Rounding off the number is similar to the truncation but with some adjustment to the last digit of the remaining digits depending upon the first digit of truncation.

- Suppose a number is required to be rounded off to the n^{th} decimal place. Then 1 is added to the n^{th} decimal digit if the $(n+1)^{th}$ digit is from 5 to 9, and the n^{th} digit is kept unchanged if the $(n+1)^{th}$ digit is from 0 to 4.

Example : Round off Last 4 digits from the following numbers.

(i) 0.90012346 Answer : 0.9001

(ii) 0.8912390821 Answer : 0.891239

(iii) 0.98567342 Answer : 0.9857

Example : Find round off error and truncation error of the following.

(i) 0.789021567 (last 3 digits)

T.E = 0.789021

R.E. = 0.789022

(ii) 0.0098790123 (last 4 digits)

$= 0.98790123 \times 10^{-2}$

R.E. $= 0.9879 \times 10^{-2}$

T.E. $= 0.9879 \times 10^{-2}$

(iii) 0.0034563216789 (last 4 digits)

$= 0.34563216789 \times 10^{-2}$

R.E. $= 0.3456322 \times 10^{-2}$

T.E. $= 0.3456321 \times 10^{-2}$

(iv) 1.5678912 (last 3 digits)

$= 0.15678912 \times 10^{1}$

R.E $= 0.15679 \times 10^{1}$

T. E $= 0.15678 \times 10^{1}$

- **Absolute Error : (E_A)**

Absolute Error is defined as the positive difference between true value and approximated value.

- If x is actual value (true value) and x_a is approximated value then Absolute Error is calculated by,

E_A = | Actual Value – Approximated Value |

$\therefore E_A = |x - x_a|$

Example : Let x = 0.00458529. Find the absolute error if x is truncated to 3 decimal digits.

Solution:

Given x = 0.00458529

$\therefore x = 0.458529 \times 10^{-2}$

$\therefore x_a = 0.458 \times 10^{-2}$ (after truncating to the 3 decimal places)

$\therefore$ Absolute Error = $| x - x_a | = | 0.458529 \times 10^{-2} - 0.458 \times 10^{-2} |$

$\therefore E_A = | (0.458529 - 0.458) \times 10^{-2} |$

$\therefore E_A = 0.000529 \times 10^{-2} = 0.529 \times 10^{-2 + (-3)}$

$\therefore E_A = 0.529 \times 10^{-5}$

Example : Let x = 0.00458529. Find the value of absolute error if x is rounded off to three decimal digits.

Given x = 0.00458529

$\therefore x = 0.458529 \times 10^{-2}$

$\therefore x_a = 0.459 \times 10^{-2}$ (after rounding off to the three decimal places)

Therefore Absolute Error (E_A) $= |x - x_a| = | 0.458529 \times 10^{-2} - 0.459 \times 10^{-2} | =$

$| -0.000471 \times 10^{-2} |$

$\therefore E_A = 0.000471 \times 10^{-2}$

$\therefore E_A = 0.471 \times 10^{-5}$

- **Relative Error : (E_R)**

Relative Error is the ratio of the error to the actual value of a variable. If x is the actual value and x_a is the approximated value then relative error is given by,

$E_R = \frac{x - x_a}{x}$

Example : Let x = 0.00599821. Find the relative error if x is truncated to 3 decimal digits.

Solution:

Here x = 0.00599821

$\therefore x = 0.599821 \times 10^{-2}$

$\therefore x_a = 0.599 \times 10^{-2}$

Now, Relative Error = $\frac{x - x_a}{x}$

$$= \frac{0.599821 \times 10^{-2} - 0.599 \times 10^{-2}}{0.599821 \times 10^{-2}}$$

$$= \frac{0.000821 \times 10^{-2}}{0.599821 \times 10^{-2}}$$

$= \frac{0.821\times10^{-5}}{0.599821\times10^{-2}}$

$= 1.36874\times 10^{-5-(-2)}$

$= 1.36874\times 10^{-3}$

$= 0.136874\times10^{-3+1}$

$= 0.136874\times10^{-2}$

Example : Let x = 0.00597621. Find the relative error if x is round off to three decimal digits.

$x = 0.597621 \times 10^{-2}$

$x_a = 0.598 \times 10^{-2}$

$E_R = \frac{x-x_a}{x}$

$= -0.63418119 \times 10^{-3}$

Example : Let x = 0.005998. Find the relative error if x is round off to three decimal digits.

(Do it yourself)

❖ Normalized Floating Point Form :

While noting a decimal number in the standard 'floating point form', the digit on the left of the decimal point should be zero, and the digit to the right of the decimal point should be non zero. Also, note that the notation En denotes multiplication with 10 to the power n

Example : abc E5 = abc $\times 10^5$.

For the number abc E5, the part 'abc' is called the mantissa and '5' is called the exponent

Examples : Convert the following values in the Normalized Floating Point Form.

(1) 0.0312 E3

Answer : 0.3120 E2

(2) 0.009723 E9

Answer : 0.9723 E7

(3) 1.2375 E5

Answer: 0.1237 E6

(4) 0.145744 E18

Answer : 0.1457 E18

(5) 10.371 E7

Answer : 0.1037 E9

(6) 4.4440 E2

Answer : 0.4444 E3

(7) 9.9008 E104

Answer : 0.9900 E105

(8) 0.00294 E(-17)

Answer: 0.2940 E(-19)

(9) 0. 45703 E(-21)

Answer: 0.4570 E(-21)

(10) 12 .314 E(-24)

Answer: 0.1231 E(-22)

- **Addition of Floating Point numbers:**

Condition : The exponents must be equal (select the larger of the two).

Note : Express the addition in the standard floating point form.

Examples : Add the following floating point numbers.

(1) 0.3254 E5 and 0.5464 E5

0.3254 E5 + 0.5464 E5

= 0.8718 E5

(2) 0.3254 E2 and 0.5462 E5

0.3254 E2 = 0.0003 E5 (E2 to E5)

0.3254 E2 + 0. 5462 E5

= 0.0003 E5 + 0. 5462 E5

= 0.5465 E5

(3) 0.5467 E5 and 0.7253 E3

0.7253 E3 = 0.0072 E5 (E3 to E5)

0.5467 E5 + 0.7253 E3

= 0.5467 E5 + 0.0072 E5

= 0.5539 E5

(4) 0.7254 E2 and 0.5467 E2

0. 7254 E2 + 0. 5467 E2

= 1.272 E2

= 0.1272 E3

(5) 0.07 E(-1) and 0.66 E(-3)

0.66 E(-3) = 0.0066 E(-1)

0.0700 E(-1) + 0.66 E(-3)

0.0700 E(-1) + 0.0066 E(-1)

= 0.0766 E(-1)

= 0.7660 E(-2)

- **Subtraction of Floating Point numbers :**

Condition : The exponents must be equal (select the larger of the two).

Note : Express the subtraction in the standard floating point form.

Examples : Subtract the following floating point numbers.

(1) Subtract 0.7254 E5 from 0.7288 E5

0.7288 E5 – 0.7254 E5

= 0.0034 E5

= 0.3400 E3

(2) Subtract 0.7253 E2 from 0.5467 E5

Now 0.7253 E2 = 0.0007 E5

∴ 0.5467 E5 – 0.7253 E2

= 0.5467 E5 – 0.0007 E5

= 0.5460 E5

(3) Subtract 0.7254 E(-99) from 0.7288 E(-99)

0.7288 E(-99) – 0.7254 E(-99)

= 0.0034 E(-99)

= 0.3400 E(-101)

(4) Subtract 0.5423 E(-1) from 0.6298 E2

Now 0.5423 E(-1) = 0.0005 E2

∴ 0.6298 E2 – 0.5423 E(-1)

= 0.6298 E2 – 0.0005 E2

= 0.6293 E2

(5) Subtract 0.2834 E(-99) from 0.5492 E(-97)

Now 0.2834 E(-99) = 0.0028 E(-97)

∴ 0.5492 E(-97) – 0.2834 E(-99)

= 0.5492 E(-97) – 0.0028 E(-97) = 0.5464 E(-97)

- **Multiplication of Floating Point numbers :**

Note : Mantissa parts are to be multiplied and the exponents are to be added. The product is to be expressed in standard floating point format.

Examples : Multiply the following floating point numbers.

(1) 0.6543 E5 and 0. 2253 E3

0.6543 E5 × 0. 2253 E3

= (0.6543 × 0.2253) E(5 + 3)

= 0.1474 E8

(2) 0.1234 E5 by 0.1111 E13

0.1234 E5 × 0.1111 E13

= (0.1234 × 0.1111) E(5 + 13)

= 0.0137 E18

= 0.1370 E17

(3) 0.1234 E(-75) by 0.1111 E37

0.1234 E(-75) × 0.1111 E37

= (0.1234 × 0.1111) E(-75 + 37)

= 0.0137 E(-38)

= 0.1370 E(-39)

(4) 0.1235 E20 by 0.1298 E(-11)

0.1235 E20 × 0. 1298 E(-11)

= (0.1235 × 0.1298) E(20 – 11)

= 0.0160 E9

= 0.1600 E8

- **Division of Floating Point Numbers :**

Note : The Mantissa part is to be divided and the exponent of second number is to be subtracted from the exponent of the first number. The quotient is written in standard form.

Examples : Divide the following floating point numbers.

(1) Divide 0.8888 E5 by 0.2000 E3

0.8888 E5 ÷ 0.2000 E3

= (0.8888 ÷ 0.2000) E(5 – 3)

= 4.4444 E2

= 0.4444 E3

(2) Divide 0.9998 E5 by 0.1000 E(-99)

0.9998 E ÷ 0.1000 E(-99)

= (0. 9998 ÷ 0.1000) E(5 – (-99))

= 0.9998 E 105

- **M.C.Q.'s :**

1)Which of the following is correct normalized floating point form of 0.009723 E9?

a) 0.9723 E7

b)0.9723 E8

c) 0.9723 E11

d) None of these

2) Which of the following is correct normalized floating point form of 1.2375 E5?

a) 0.1238 E6

b) 0.1237 E6

c) 0.12375 E4

d) None of these

3) Which of the following is correct form of 0.12347891 after truncating last 4 digits?

a) 0.1235

b) 0.7891

c) 0.1291

d) 0.1234

4) 0.1234 E75 × 0.1111 E37 = _____

a) 0.1370 E75

= 0.1234 * 0.1111 E(75+37)
= 0.1371 E112

0.1371 E112

b) 0.1370 E111

c) 0.1370 E34

d) None of these

5) Which of the following is correct formula for relative error?

a) $E_r = (x - x_a) / x$

b) $E_r = | x - x_a |$

c) $E_r = (x + x_a) / x$

d) $E_r = (x \cdot x_a) / x$

6) Which of the following is correct formula for Absolute Error?

a) $E_a = x - x_a$

b) $E_a = | x + x_a |$

c) $E_a = | x - x_a |$

d) $E_a = x / x_a$

7) 0.5467 E2 + 0.7254 E2 = _____

a) 0.1272 E3

= 0.5467 + 0.7254 E2
= 1.2721 E2
= 0.1272 E(2+1)
= 0.1272 E3

b) 0.1272 E2

c) 0.1272 E4

d) 0.1272 E1

8) 0.5467 E5 – 0.7253 E2 = _____

a) 0.5460 E0

b) 0.5460 E5

c) 0.5460 E3

d) 0.5460 E7

9) A positive difference between actual value and approximate value is called ____.

a) Relative Error

b) Error

c) Absolute Error

d) None of these