Assignment 1 Solution

Wednesday, February 7, 2024 7:25 PM

1. In the classes, we discussed three forms of floating number representations as shown below, (1) Standard/General Form, (2) Normalized Form, (3) Denormalized Form.
Now, let's take, **$\beta = 2$, $m = 3$ and $-2 \le e \le 4$.** Based on these, answer the following:

(a) (3 marks) What are the **maximum/largest** numbers that can be stored in the system by these three forms defined above (express your answer in decimal values)?
(b) (3 marks) What are the **non-negative minimum/smallest** numbers that can be stored in the system by the three forms defined above (express your answer in decimal values)?
(c) (4 marks) What are the **maximum/largest and minimum/smallest** numbers that can be stored in the system by the three forms defined above if the system has negative support?

1. (a) Standard Form $= (0.111)_2 \times 2^4 = (14)_{10}$

Normalized " $= (1.111)_2 \times 2^4 = (30)_{10}$

Denormalized " $= (0.1111)_2 \times 2^4 = (15)_{10}$

(b) Standard Form $: (0.100)_2 \times 2^{-2} = (0.125)_{10}$

Normalized " $= (1.000)_2 \times 2^{-2} = (0.25)_{10}$

De " " $= (0.1000)_2 \times 2^{-2} = (0.125)_{10}$

(c) If the system has negative support,

	Maximum	Minimum
C-1 :	$(0.111)_2 \times 2^4 = (14)_{10}$	$-(0.111)_2 \times 2^4 = -(14)_{10}$
C-2:	$(1.111)_2 \times 2^4 = (30)_{10}$	$-(1.111)_2 \times 2^4 = -(30)_{10}$
C-3:	$(0.1111)_2 \times 2^4 = (15)_{10}$	$-(0.1111)_2 \times 2^4 = -(15)_{10}$

2. Consider the **real number x =** $(6.235)_{10}$

(a) (3 marks) First convert the decimal number x in binary format at least up to 9 decimal/binary places.
(b) (4 marks) What will be the binary value of x [**Find fl(x)**] if you store it in a system with **m = 5** and **m =6** using the **general/standard** form of Floating point representation.
(c) (3 marks) Now convert back to the decimal form the stored values you obtained in the previous part, and calculate the **rounding error of both numbers**.

2. (a) $x = (6.235)_{10}$

$(6)_{10} = (110)_2$

$(0.235)_{10} = (001111000)_2$

$\therefore (6.235)_{10} = (110.001111000)_2$

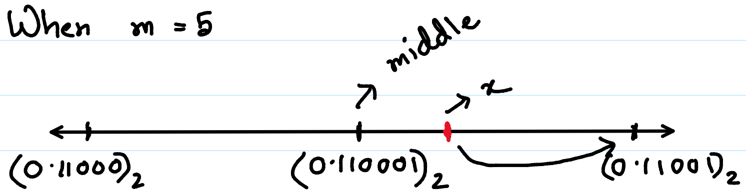
$$
\begin{array}{r|l}
 & 0.235 \\
 & \times 2 \\
\hline
0 & .47 \\
 & \times 2 \\
\hline
0 & .94 \\
 & \times 2 \\
\hline
1 & .88 \\
 & \times 2 \\
\hline
1 & .76 \\
 & \times 2 \\
\hline
1 & .52 \\
 & \times 2 \\
\hline
1 & .04
\end{array}
$$

(b) $(6.235)_{10} = (110.001111000)_2$

(b) $(6.235)_{10} = (110.001111000)_2$
$= (0.\underbrace{11000}_{m=5}111000)_2 \times 2^3$

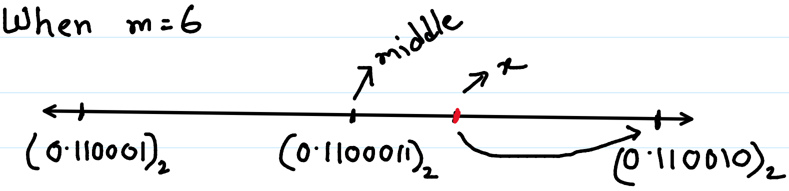
$$\begin{array}{l|l} 1 & .52 \\ & \times 2 \\ \hline 1 & .04 \\ & \times 2 \\ \hline 0 & .08 \\ & \times 2 \\ \hline 0 & .16 \\ & \times 2 \\ \hline 0 & .32 \end{array}$$

When $m = 5$



$\therefore fl(x) = (0.11001)_2 \times 2^3 = (6.25)_{10}$

When $m = 6$



$\therefore fl(x) = (0.110010)_2 \times 2^3 = (6.25)_{10}$

(c) For both $m=5$ and $m=6$, $fl(x) = (6.25)_{10}$

$$\delta = \left|\frac{fl(x) - x}{x}\right| = \left|\frac{6.25 - 6.235}{6.235}\right| = 2.41 \times 10^{-3} \text{ (approx.)}$$

3. Consider the quadratic equation, $2x^2 - 60x + 3 = 0$. Below calculate **up to 6 significant** figures.
(a) (4 marks) Find out where the loss of significance occurs when you calculate the roots?
(b) (3 marks) Show that the roots evaluated in the previous part do not satisfy the fundamental properties of a polynomial.
(c) (3 marks) Evaluate the correct roots such that loss of significance does not occur.

3.(a)
$$2x^2 - 60x + 3 = 0$$
$$\Rightarrow x^2 - 30x + \frac{3}{2} = 0$$
$$\therefore x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Here,
$a = 1$
$b = -30$
$c = 3/2$

Solving,
$$x_1 = \frac{30 + \sqrt{894}}{2} = 15 + 14.9499 \text{ [upto 6 s.f.]}$$
$$= 29.9499$$
$$x_2 = \frac{30 - \sqrt{894}}{2} = 15 - 14.9499$$
$$= 0.0501000 \text{ [upto 6.sf]}$$

Now, $x_1 \times x_2 = 29.9499 \times 0.0501000 = 1.50048$ [upto 6sf]

$\therefore x_1 \times x_2 \neq 1.5$

So, when roots are multiplied, loss of significance occours. as we subtracted two close numbers.

(b) $x_1 + x_2 = 29.9499 + 0.0501000$
$= 30$ $\qquad \therefore x_1 + x_2 = -b/a$

$x_1 \times x_2 = 29.9499 \times 0.0501000$
$= 1.50048$ $\qquad \therefore x_1 \times x_2 \neq c/a$

(c) $x_1 \times x_2 = c/a$

$\Rightarrow 29.9499 \times x_2 = 1.5$

$\Rightarrow x_2 = \dfrac{1.5}{29.9499} = 0.0500836$ [upto 6sf]

Now, $x_1 + x_2 = 29.9499 + 0.0500836 = 29.9999836 \cong 30$

$x_1 \times x_2 = 29.9499 \times 0.0500836 = 1.5$

$\therefore$ Correct root $= 0.0500836$