CSE330: Numerical Methods

Assignment 1

Prepared by:
Saad Bin Sohan
BRAC University

Email: sohan.academics@gmail.com
GitHub: https://github.com/saad-bin-sohan

Instructions for preparing the solution script:

- Write your name, ID#, and Section number clearly in the very front page.
- Write all answers sequentially.
- Start answering a question (not the pat of the question) from the top of a new page.
- Write legibly and in orderly fashion maintaining all mathematical norms and rules. Prepare a single solution file.
- Start working right away. There is no late submission form. If you miss the deadline, you need to use the make-up assignment to cover up the marks.

---

1. In the classes, we discussed three forms of floating number representations as shown below,

$$\begin{aligned} \text{Lecture Note Form} \quad &: \quad F = \pm(0.d_1d_2d_3\cdots d_m)_\beta\, \beta^e \ , \qquad (1)\\ \text{Normalized Form} \quad &: \quad F = \pm(1.d_1d_2d_3\cdots d_m)_\beta\, \beta^e \ , \qquad (2)\\ \text{Denormalized Form} \quad &: \quad F = \pm(0.1d_1d_2d_3\cdots d_m)_\beta\, \beta^e \ ,, \qquad (3)\end{aligned}$$

where $d_i, \beta, e \in \mathbb{Z}$, $0 \leq d_i \leq \beta - 1$ and $e_{\min} \leq e \leq e_{\max}$. Now, let's take, $\beta = 2$, $m = 4$ and $-4 \leq e \leq 2$. Based on these, answer the following:

   (a) (3 marks) What are the maximum numbers that can be stored in the system by the three forms defined above?

   (b) (3 marks) What are the non-negative minimum numbers that can be stored in the system by the three forms defined above?

   (c) (4 marks) Using Eq.(1), find all the decimal numbers for $e = -3$, plot them on a real line, and show if the number line is equally spaced or not.

2. Let $\beta = 2$, $m = 5$, $e_{\min} = -2$ and $e_{\max} = 5$. Answer the following questions:

   (a) (4 marks) Compute the minimum of $|x|$ for normalized and denormalized form.

   (b) (4 marks) Compute the Machine Epsilon value for the normalized and denormalized form.

   (c) (2 marks) Compute the maximum delta value for the form given in Eq.(2).

3. Let $\beta = 2$, $m = 3$, $e_{\min} = -2$ and $e_{\max} = 2$. Answer the following questions:

   (a) (4 marks) Find the floating point representation of the numbers $(2.23)_{10}$ and $(2.2018)_{10}$ in the Normalized form.

   (b) (2 marks) Compute the rounding errors for Part (a).

   (c) (4 marks) Can the numbers $(2.23)_{10}$ and $(2.2018)_{10}$ be represented in denormalized form? If so, find the floating-point representations. If not, then concisely explain why?

Ans to the que-1

1(a)

given, $\beta = 2$ $\qquad -4 \leq e \leq 2$

$m = 4$

In convention-1 or the lecture note form,

the maximum number,

$$F_{max} = +(0.1111)_2 \times 2^2$$

$\Rightarrow F_{max} = (1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} +$

$1 \times 2^{-4}) \times 4$

$= 3.75$ (Ans)

in normal form,

the maximum number,

$F_{max} = +(1.1111)_2 \times 2^2$

$\Rightarrow F_{max} = (1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} +$

$1 \times 2^{-3} + 1 \times 2^{-4}) \times 2^2$

$= 7.75$ (Ans)

in denormalised form:

the maximum number,

$$F_{max} = +(0.11111)_2 \times 2^2$$

$$= (1\times2^{-1} + 1\times2^{-2} + 1\times2^{-3} + 1\times2^{-4} + 1\times2^{-5}) \times 2^2$$

$$= 3.875$$ (Ans)

1(b)

in lecture note form,

the non-negative minimum number,

$$F_{min} = +(0.1000)_2 \times 2^{-4}$$

$$= (1\times2^{-1}) \times 2^{-4}$$

$$= 0.63125$$ (Ans)

in normalized form,

the minimum non negative number,

$$F_{min} = +(1.0000)_2 \times 2^{-4}$$

$$= 1 \times 2^0 \times 2^{-4}$$

$$= 0.0625$$ (Ans)

in denormalised form,

the minimum number,

$$F_{min} = +(0.10000)_2 \times 2^{-4}$$

$$= 1 \times 2^{-1} \times 2^{-4}$$

$$= 0.03125$$ (Ans)

1(c)

in lecture note form,

$$F = \pm (0.d_1 d_2 \dots d_m)_\beta \times \beta^e$$

$\beta = 2$
$e = -3$
$m = 4$
$d_1 = 1$

for $e = -3$ the numbers will be

$$F = \pm (0.1 d_2 d_3 d_4)_2 \times 2^{-3}$$

the numbers are

$$\pm (0.1000)_2 \times 2^{-3} = \pm 0.0625 = \pm \frac{1}{16}$$

$$\pm (0.1001)_2 \times 2^{-3} = \pm 0.0703125 = \pm \frac{9}{128}$$

$$\pm (0.1010)_2 \times 2^{-3} = \pm 0.078125 = \pm \frac{5}{64}$$

$$\pm (0.1011)_2 \times 2^{-3} = \pm 0.0859375 = \pm \frac{11}{128}$$

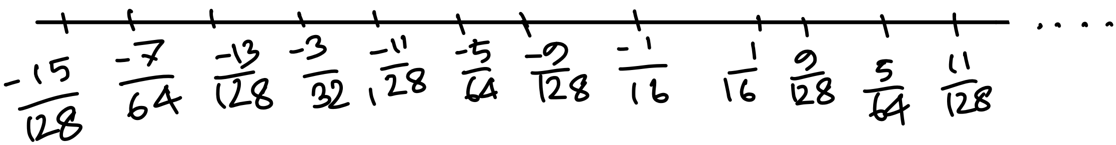
$$\pm (0.1100)_2 \times 2^{-3} = \pm 0.09375 = \pm \frac{3}{32}$$

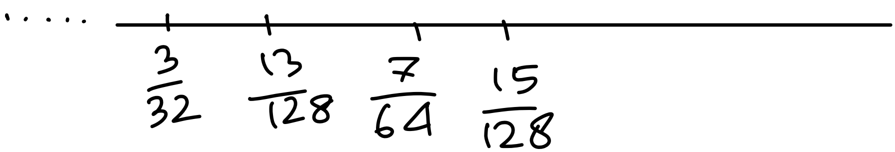
$\pm (0.1101)_2 \times 2^{-3} = \pm\ 0.1015625 = \pm \frac{13}{128}$

$\pm (0.1110)_2 \times 2^{-3} = \pm 0.109375 = \pm \frac{7}{64}$

$\pm (0.1111)_2 \times 2^{-3} = 0.1171875 = \pm \frac{15}{128}$

plotting them on a number line, we get,





in the numbers we see,

$diff_1 = \frac{9}{128} - \frac{1}{16} = \frac{1}{128} = 7.8125 \times 10^{-3}$

$diff_2 = \frac{5}{64} - \frac{9}{128} = \frac{1}{128} = 7.8125 \times 10^{-3}$

so we see the numbers are equally spaced except for $-\frac{1}{16}$ and $\frac{1}{16}$. they have a distance of $\frac{1}{16} - (-\frac{1}{16}) = \frac{1}{8}$.

Ans to the que no-2

2(a)

given, $\beta = 2$, $m = 5$, $-2 \leq e \leq 5$

in normalized form:

$$\min, x = (1.00000)_2 \times \beta^e$$

$$= 2^0 \times \beta^e$$

$$= \beta^0 \times \beta^e$$

as a value we can get $\min, x = 2^0 \times 2^{-2}$

$$= 0.25$$

in denormalised form.

$$\min, x = (0.100000)_\beta \times \beta^e$$

$$= 2^{-1} \times \beta^e$$

$$\min, x = \beta^{-1} \times \beta^e$$

as a value we can get, $\min, x = 2^{-1} \times 2^{-2}$

$$= 0.125$$

2(b)

machine epsilon, $\epsilon_m = \dfrac{|fl(x) - x|}{|x|}$

in normalized form:

lets take two number $x_1 = (1.00000)_\beta \times \beta^e$

and $x_2 = (1.00001)_\beta \times \beta^e$

so, $\left|fl(x) - x\right| = \frac{1}{2}\left[(1.00001)_\beta \beta^e - (1.00000)_\beta \beta^e\right]$

$$= \frac{1}{2}\beta^e \left[1.00001 - 1.00000\right]$$

$$= \frac{1}{2}\beta^e (0.00001)$$

$$= \frac{1}{2}\beta^e \times 2^{-5}$$

$$= \frac{1}{2}\beta^e \times \beta^{-5}$$

from 2(a), min, $x = \beta^e$

so in normalised form,

$$\epsilon_m = \frac{|fl(x) - x|}{|x|} = \frac{\frac{1}{2}\beta^e \times \beta^{-m}}{\beta^e}$$

$$\epsilon_m = \frac{1}{2}\beta^{-m} = \frac{1}{2}\times(2)^{-5} = 0.015625$$

in denormalised form:

$$\epsilon_m = \frac{|fl(x) - x|}{|x|}$$

lets take two values of x,

$$x_1 = (0.100000)_2 \times \beta^e$$

$$x_2 = (0.100001)_2 \times \beta^e$$

for ϵ_m, $|fl(x) - x|$,

$$= \frac{1}{2}\left[(0.100001)_2\,\beta^e - (0.100000)_2\,\beta^e\right]$$

$$= \frac{1}{2}\beta^e\left[0.000001\right]$$

$$= \frac{1}{2}\beta^e \times 2^{-6} = \frac{1}{2}\beta^e \times \beta^{-m-1}$$

from 2(a) min, $x = \beta^{-1} \times \beta^e$

so, $\epsilon_m = \dfrac{|fl(x) - x|}{|x|} = \dfrac{\frac{1}{2}\beta^e \times \beta^{-m-1}}{\beta^{-1} \times \beta^e}$

$$\epsilon_m = \frac{1}{2}\beta^{-m} \quad \text{(Ans)}$$

$$= \frac{1}{2} \times \beta^{-5} = 0.015625$$

2(c)

for normalized form, from 2(b),

maximum delta, $\epsilon_m = \dfrac{1}{2}\beta^{-m}$

so $\epsilon_m = \dfrac{1}{2} \times 2^{-5}$

$$= \frac{1}{64} = 0.015625$$

Ans to the que-3

3(a)

$x_1 = (2.23)_{10}$

$= (10.00111\ldots\ldots)_2 \times 2^0$

$= (10.00111)_2 \times 2^0$

for normalized form,

$x_1 = (10.00111)_2 \times 2^0$

$= (1.000111)_2 \times 2^1$

but since $m = 3$ the rounded value will be $fl(x_1) = (1.001)_2 \times 2^1$

now, $x_2 = (2.2018)_{10}$

$= (10.0011\ldots)_2 \times 2^0$

$= (10.0011)_2 \times 2^0$

for normalized form,

$x_2 = (1.00011)_2 \times 2^1$

$fl(x_2) = (1.001)_2 \times 2^1$

3(b)

For $x_1 = (2.23)_{10} = (1.00011)_2 \times 2^1$

rounding error, $\delta = |fl(x) - (x)|$

$$\delta_1 = |(1.001)_2 \times 2^1 - (1.000111)_2 \times 2^1|$$

$$= |2.25 - 2.21875|$$

$$= \frac{1}{32} = 0.03125$$

for x_2,

$$\delta_2 = |(1.001)_2 \times 2^1 - (1.00011)_2 \times 2^1|$$

$$= |2.25 - 2.1875|$$

$$= \frac{1}{16} = 0.0625$$

3(c)

$$x_1 = (2.23)_{10} = (1.00011)_2 \times 2^1$$

$$= (0.100011)_2 \times 2^2$$

in denormalized form, under $m=3$

$$x_1 = (0.1001)_2 \times 2^2$$

for $x_2 = (2.2018)_{10}$

$$= (10.0011)_2 \times 2^0$$

$$= (0.100011)_2 \times 2^2$$

denormalized form $= (0.1001)_2 \times 2^2$

so both the number can be represented in denorm form