BRAC UNIVERSITY

CSE330

Numerical Methods

Assignment 1

Student Information
Name: **Md. Danial Islam** ID: **20101534** Section: **10**

Date: 04 February 2023

Answer to the question no 1

(a) given $\beta = 2$, $m = 4$, $e = [-3, 6]$

general form,

$\Rightarrow \beta^{m-1} \times$ count of exponent

$\Rightarrow 2^{4-1} \times 10$

$\Rightarrow 80$ Possible numbers

normalized form,

$\Rightarrow \beta^{m} \times$ count of exponent

$\Rightarrow 2^{4} \times 10$

$\Rightarrow 160$ Possible numbers

de normalized form,

$\Rightarrow \beta^{m} \times$ count of exponent

$\Rightarrow 2^{4} \times 10$

$= 160$ (Possible numbers)

(b) largest number possible,

$\beta = 2$

$m = 4$

$e = [-3, 6]$

general form,

$$(0.1111)_2 \times \beta^{e_{max}}$$

$$\Rightarrow (0.1111)_2 \times 2^6 \Rightarrow \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}\right) \times 2^6$$

$$\Rightarrow 60$$

normalized form,

$$(1.1111)_2 \times 2^6 \Rightarrow (1 + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4}) \times 2^6$$

$$\Rightarrow 124$$

denormalized form,

$$(0.11111)_2 \times 2^6 \Rightarrow (2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} + 2^{-5}) \times 2^6$$

$$\Rightarrow 62$$

(c) non-negative smallest number

$\beta = 2$

$m = 4$

$e = [-3, 6]$

general form,

$$(0.1000)_2 \times \beta^{e_{min}}$$

$$\Rightarrow (0.1000)_2 \times 2^{-3}$$

$$\Rightarrow 2^{-1} \times 2^{-3} \Rightarrow \frac{1}{16}$$

normalized form,

$$(1.0000)_2 \times \beta^{e_{min}}$$

$$\Rightarrow (1)_2 \times 2^{-3}$$

$$\Rightarrow \frac{1}{8}$$

de-normalized form,

$$(0.10000)_2 \times \beta^{e_{min}}$$

$$\Rightarrow 2^{-1} \times 2^{-3} \Rightarrow \frac{1}{16}$$

(d) using general form,

$$\beta = 2,\ m = 4 \quad e = -1$$

then all the numbers are,

$(0.1000)_2 \times 2^{-1} \Rightarrow \frac{1}{4}$

$(0.1001)_2 \times 2^{-1} \Rightarrow \frac{9}{32}$ $\quad \frac{1}{32}$

$(0.1010)_2 \times 2^{-1} \Rightarrow \frac{5}{16}$ $\quad \frac{1}{32}$

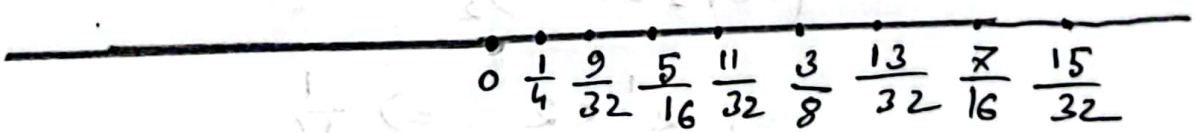
$(0.1011)_2 \times 2^{-1} \Rightarrow \frac{11}{32}$ $\quad \frac{1}{32}$

$(0.1100)_2 \times 2^{-1} \Rightarrow \frac{3}{8}$ $\quad \frac{1}{32}$

$(0.1101)_2 \times 2^{-1} \Rightarrow \frac{13}{32}$ $\quad \frac{1}{32}$

$(0.1110)_2 \times 2^{-1} \Rightarrow \frac{7}{16}$ $\quad \frac{1}{32}$

$(0.1111)_2 \times 2^{-1} \Rightarrow \frac{15}{32}$ $\quad \frac{1}{32}$



here the difference between two numbers are $\frac{1}{32}$ and it's equal for every number of this series. so they are equally spaced.

Answer to the question no 2

(a) given $\beta = 2$, $m = 3$, exp = [0, 15] as 4 bit

as e_{max} & e_{min} reserved then $e = [1, 14]$

∴ new $e_{max} = 14$, $e_{min} = 1$

∴ ~~smallest~~

∴ ~~d~~normalized form,

$$\text{Smallest} = (1.000)_2 \times 2^1$$

$$\text{maximum} = (1.111)_2 \times 2^{14}$$

denormalized form,

$$\text{Smallest} = (0.1000)_2 \times 2^{-3}$$

$$\text{maximum} = (0.1111)_2 \times 2^{14}$$

— o —

(b) Machine epsilon, ϵ_m for normalized form

$$\Rightarrow \frac{1}{2}\beta^{-m} \Rightarrow \frac{1}{2}2^{-3}$$

$$\Rightarrow \frac{1}{16} \text{ (Ans)}$$

(c) For max delta error value for General form

$$\epsilon_m = \frac{1}{2}\beta^{1-m}$$
$$= \frac{1}{2}2^{1-3}$$
$$= \frac{1}{2}\times 2^{-2} = \frac{1}{8} \text{ (Ans)}$$

—o—

Answer to the question no 3

$\beta = 2$, $m = 3$, ~~exp = 4bit~~ $e_{min} = -1$

$e_{max} = 2$

Normalized form,

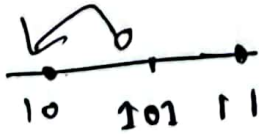
(a) $x = 6.25$

$\therefore (6.25)_{10} \Rightarrow (6)_{10} = 110$

$(.25)_{10} = (0.01)_2$

$(6.25)_{10} = (110.01)_2 \times 2^0 \Rightarrow (1.10001)_2 \times 2^2$

$\Rightarrow (1.10)_2 \times 2^2$



Answer to the question no 3

$\beta = 2$, $m = 3$, $e_{min} = -1$, $e_{max} = 2$

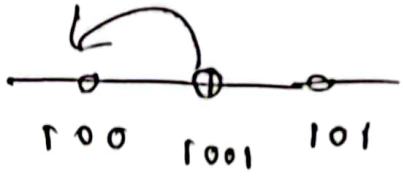
normalized form,

$$x = 6.25 = \frac{25}{4} = \frac{1+8+16}{4} = 2^{-2} + 2^{1} + 2^{2}$$

~~$= (6.01)_2$~~

$$= (110.01)_2$$

$$= (1.1001)_2 \times 2^2$$

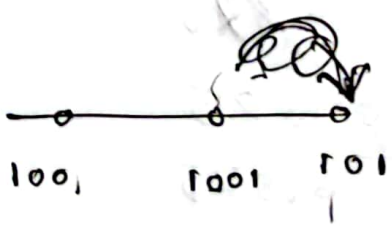


$$= (1.100)_2 \times 2^2 \quad \text{(Ans)}$$

$$x = 6.825 = (110.111)_2 \times 2^0$$

$$= (1.10111)_2 \times 2^2$$

$$= (1.101)_2 \times 2^2 \quad \text{(Ans)}$$



(b) $fl(x) = fl(6.25)$

$= (1.100)_2 \times 2^2$

$= (110.0)_2 \times 2^0$

$= 6$

$\therefore \delta' = \dfrac{|x - fl(x)|}{x} = \dfrac{|6.25 - 6|}{6.25}$

$= 0.04$

$fl(x) = fl(6.875)$

$= (1.101)_2 \times 2^2$

$= (110.1)_2 \times 2^0$

$= 6.5$

$\therefore \delta_2 = \dfrac{|x - fl(x)|}{x} = \dfrac{|6.875 - 6.5|}{6.875}$

$= \dfrac{0.375}{6.875}$

$= 0.0545$

(c) From (b).

~~6.25~~

$(6.25)_{10} = (1.1001)_2 \times 2^2$

$= (110.01)_2 \times 2^0$

and $(6.875)_{10} = (110.111)_2 \times 2^0$

to convert denormalized form the floating point will be in this form

$(0.1 d_1 d_2 d_3 \ldots d_n)_\beta \times \beta^e$

~~6.25~~ $(.1.10.01)_2 \times 2^0$

$\Rightarrow (0.11001)_2 \times 2^3$

& $(110.111)_2 \times 2^0$

$\Rightarrow (0.110111)_2 \times 2^3$

So, if we try to convert it to denormalized form ~~it'll be~~ the exponent will become 3 but an error = 2 bits not possible.

(d)

Machine epsilon, ε_m

Lecture note,

$$\frac{1}{2}\beta^{1-m} = \frac{1}{2} \times 2^{1-3}$$

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

$$= \frac{1}{8} \text{ (Ans)}$$

Normalized and denormalized,

$$\frac{1}{2}\beta^{-m} = \frac{1}{2} \times 2^{-3}$$

$$= \frac{1}{16} \text{ (Ans)}$$