Assignment 01

Abdullah Al Mazid Zomader

ID: 24241189

Sec: 26

BRAC University

CSE330: Numerical Methods

Mr. Towshik Anam Taj

July 01, 2025

Ans to the ques. no-01

(a) Given, $\beta = 2$, $m = 4$, $e_{min} = -5$ and $e_{max} = 2$

$\therefore$ In General form,

$$x = (0.d_1 d_2 d_3 d_4)_\beta \times \beta^e \quad \text{where, } d_1 \neq 0$$

$$= (0.1111)_2 \times 2^2 = (2^{-1}+2^{-2}+2^{-3}+2^{-4}) \times 2^2 = 3.75$$

$\therefore$ In Normalized form,

$$x = (0.1\, d_1 d_2 d_3 d_4)_\beta \times \beta^e$$

$$= (0.11111)_2 \times 2^e = (2^{-1}+2^{-2}+2^{-3}+2^{-4}+2^{-5}) \times 2^2 = 3.875$$

$\therefore$ In Denormalized form,

$$x = (1.d_1 d_2 d_3 d_4)_\beta \times \beta^e$$

$$= (1.1111)_2 \times 2^2 = (2^0+2^{-1}+2^{-2}+2^{-3}+2^{-4}) \times 2^2 = 7.875$$

(b)

In Generalized form:

$$(0.d_1 d_2 d_3 d_4)_\beta \times \beta^e \quad \text{where } d_1 \neq 0$$

$$= (0.1000)_2 \times 2^{-5} = 2^{-1} \times 2^{-5} = \frac{1}{2^6} = 0.015625$$

In Normalized form:

$$(0.1 d_1 d_2 d_3 d_4)_\beta \times \beta^e$$

$$= (0.10000) \times 2^{-5} = 2^{-1} \times 2^{-5} = 0.015625$$

In Denormalized form:

$$(1.d_1 d_2 d_3 d_4)_\beta \times \beta^e$$

$$= (1.0000)_2 \times \beta^e = 2^0 \times 2^{-5} = 1 \times 2^{-5} = 0.03125$$

(c)

$\pm (0.1 d_1 d_2 d_3 d_4)_\beta \times \beta^e \rightarrow$ Normalized form

thus, $2^m = 2^4 = 16$ values/combination for 4 bits for one exponent $\rightarrow -5$:

$$(0.10000)_2 \times 2^{-5} = 2^{-1} \times 2^{-5} = 0.5 \times 2^{-5} = 0.015625$$

$$(0.10001)_2 \times 2^{-5} = (2^{-1} + 2^{-5}) \times 2^{-5} = 0.53125 \times 2^{-5} = 0.0166015625$$

$$(0.10010)_2 \times 2^{-5} = (2^{-1} + 2^{-4}) \times 2^{-5} = 0.5625 \times 2^{-5} = 0.017578125$$

$$(0.10011)_2 \times 2^{-5} = (2^{-1} + 2^{-4} + 2^{-5}) + 2^{-5} = 0.59375 \times 2^{-5} = 0.0185546875$$

$(0.10100)_2 \times 2^{-5} = 0.625 \times 2^{-5} = 0.01953125$

$(0.10101)_2 \times 2^{-5} = 0.65625 \times 2^{-5} = 0.0205078125$

$(0.10110)_2 \times 2^{-5} = 0.6875 \times 2^{-5} = 0.021484375$

$(0.10111)_2 \times 2^{-5} = 0.71875 \times 2^{-5} = 0.0224609375$

$(0.11000)_2 \times 2^{-5} = 0.75 \times 2^{-5} = 0.0234375$

$(0.11001)_2 \times 2^{-5} = 0.78125 \times 2^{-5} = 0.0244140625$

$(0.11010)_2 \times 2^{-5} = 0.8125 \times 2^{-5} = 0.025390625$

$(0.11011)_2 \times 2^{-5} = 0.84375 \times 2^{-5} = 0.0263671875$

$(0.11100)_2 \times 2^{-5} = 0.875 \times 2^{-5} = 0.02734375$

$(0.11101)_2 \times 2^{-5} = 0.90625 \times 2^{-5} = 0.0283203125$

$(0.11110)_2 \times 2^{-5} = 0.9375 \times 2^{-5} = 0.0292968875$

$(0.11111)_2 \times 2^{-5} = 0.96875 \times 2^{-5} = 0.0302734375$

$\therefore$ Unique value per exponent, $n = 16$

$\therefore$ no. of exponent, $m = 7$.

$\therefore$ Total ($\pm$) number $= 16 \times 7 = 112$

$\therefore$ 112 positive numbers

$\therefore$ 112 negative numbers

224 numbers can be represented.

(d) For normalized form, $\pm(0.1\,d_1 d_2 d_3 d_4)_\beta \times \beta^e$

$\therefore$ The numbers for $e=-1$:

$(0.10000)_2 \times 2^{-1} = 0.25$

$(0.10001)_2 \times 2^{-1} = 0.265625$

$(0.10010)_2 \times 2^{-1} = 0.28125$

$(0.10011)_2 \times 2^{-1} = 0.296875$

$(0.10100)_2 \times 2^{-1} = 0.3125$

$(0.10101)_2 \times 2^{-1} = 0.328125$

$(0.10110)_2 \times 2^{-1} = 0.34375$

$(0.10111)_2 \times 2^{-1} = 0.359375$

$(0.11000)_2 \times 2^{-1} = 0.375$

$(0.11001)_2 \times 2^{-1} = 0.390625$

$(0.11010)_2 \times 2^{-1} = 0.40625$

$(0.11011)_2 \times 2^{-1} = 0.421875$

$(0.11100)_2 \times 2^{-1} = 0.4375$

$(0.11101)_2 \times 2^{-1} = 0.453125$

$(0.11110)_2 \times 2^{-1} = 0.46875$

$(0.11111)_2 \times 2^{-1} = 0.484375$

$\therefore$ common difference $= 0.265625 - 0.25 = 1/64$

The numbers for $e=0$:

$(0.10000)_2 \times 2^{0} = 0.5$

$(0.10001)_2 \times 2^{0} = 0.53125$

$(0.10010)_2 \times 2^{0} = 0.5625$

$(0.10011)_2 \times 2^{0} = 0.59375$

$(0.10100)_2 \times 2^{0} = 0.625$

$(0.10101)_2 \times 2^{0} = 0.65625$

$(0.10110)_2 \times 2^{0} = 0.6875$

$(0.10111)_2 \times 2^{0} = 0.71875$

$(0.11000)_2 \times 2^{0} = 0.75$

$(0.11001)_2 \times 2^{0} = 0.78125$

$(0.11010)_2 \times 2^{0} = 0.8125$

$(0.11011)_2 \times 2^{0} = 0.84375$

$(0.11100)_2 \times 2^{0} = 0.875$

$(0.11101)_2 \times 2^{0} = 0.90625$

$(0.11110)_2 \times 2^{0} = 0.9375$

$(0.11111)_2 \times 2^{0} = 0.96875$

$\therefore$ commmon difference $= 0.53125 - 0.5 = 1/32$

$0.28125 - 0.265625$

$= 1/64$

$\vdots$

$0.5 - 0.484375$

$= 1/64$

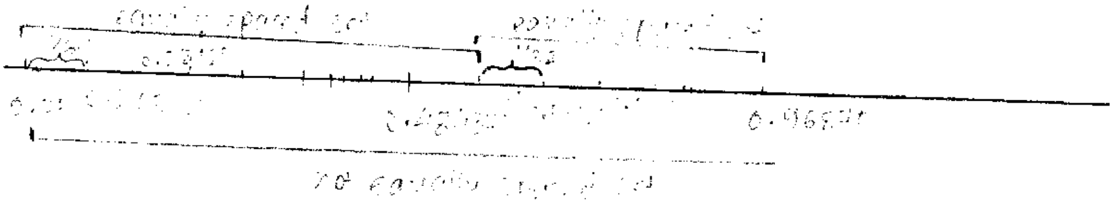
$0.5625 - 0.53125$

$= \frac{1}{32}$

$\vdots$

$0.96875 - 0.9375$

$= \frac{1}{32}$



Elements per set = 16

Number of equally spaced set for $e=-1$ and $e=0$ is 2

Number of equally spaced set for whole set of exponents is 8

Equally spaced set for $e=-1$ is {0.25, 0.265625, 0.28125, 0.296875, 0.3125, 0.328125, 0.34375, 0.359375, 0.375, 0.390625, 0.40625, 0.421875, 0.4375, 0.453125, 0.46875, 0.484375}

Equally spaced set for $e=0$ is {0.5, 0.53125, 0.5625, 0.59375, 0.625, 0.65625, 0.6875, 0.71875, 0.75, 0.78125, 0.8125, 0.84375, 0.875, 0.90625, 0.9375, 0.96875}.

Ans to the ques no-02

(a) Given, $\beta = 2$

$m = 7$

$e_{min} = -4$

$e_{max} = 8$

General: $(0 \cdot d_1 d_2 d_3 d_4 d_5 d_6 d_7) \times 2^{-4}$ where $d_1 \neq 0$

$= (0.1000000) \times 2^{-4}$

$= 2^{-1} \times 2^{-4} = 2^{-5} = 0.03125$

Denormalized: $(1 \cdot d_1 d_2 d_3 d_4 d_5 d_6 d_7) \times 2^{-4}$

$= (1.0000000) \times 2^{-4}$

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

$= 1 \times 2^{-4} = 0.0625$

(b) Actual value $= |x|_m = \beta^e \times \beta^{-1}$

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

$\therefore \delta_{max} = \dfrac{|fl(x) - x|}{|x|} = \dfrac{\frac{1}{2} \times \beta^{-m} \times \beta^e}{\beta^e \times \beta^{-1}} = \frac{1}{2} \times \beta^{(1-m)}$

$\therefore$ General form of Machine Epsilon, $\delta_{max} = \frac{1}{2} \times \beta^{(1-m)}$

$= \frac{1}{2} \times 2^{(1-7)}$

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

$= 2^{-7} = \frac{1}{128}$

∴ For Normalized form:

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

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

∴ Machine Epsilon in Normalized form; $\delta_{max} = \dfrac{\frac{1}{2} \times \beta^{-m} \times \beta^{-1} \times \beta^{e}}{\beta^{-1} \times \beta^{e}}$

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

$$= \frac{1}{2} \times 2^{-7}$$

$$= 2^{-8} \text{ (Ans)}$$

(c)

For Denormalized Form,

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

and $|x|_{min} = \beta^{e}$

∴ Machine Epsilon/Maximum scale invariant Error, $\delta_{max} = \dfrac{\frac{1}{2} \times \beta^{-m} \times \beta^{e}}{\beta^{e}}$

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

Here, the formula have a properties of mantissa precision, not the exponent range. Changing e_{min} affects how small a number can be represented, but not the precision with which any number is represented. So, the maximum scale intervention error do not change.

Ans to the ques no-03

Given,

$5x^2 - 70x + 4 = 0$

$\therefore a = 5,\ b = -70,\ c = 4$

(a)

Using quadratic equation $= \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

$$= \frac{-(-70) \pm \sqrt{70^2 - 4.5.4}}{2.5}$$

$$= \frac{70}{10} \pm \frac{2\sqrt{1205}}{10} = \frac{70}{10} \pm \frac{\sqrt{1205}}{5}$$

$$= 7 \pm \frac{\sqrt{1205}}{5}$$

$\therefore \alpha = 7 + \dfrac{\sqrt{1205}}{5} = 7 + 6.9426\ldots$

$= 13.942$ [significant figure, sf = 5]

$\therefore \beta = 7 - \dfrac{\sqrt{1205}}{5} = 7 - 6.9426 = 0.0574$ (Ans)

(b) From (a),

$\dfrac{\sqrt{1205}}{5} = 6.942621983$

and,

with significant figure = 5, $\dfrac{\sqrt{1205}}{5} = 6.9426$

$\therefore$ loss of significance $= 6.942621983 - 6.9426$

$= 0.000021983$ (Ans)

(c) $x^2 + (\alpha+\beta)x + \alpha\beta = 0$

Given,

$$5x^2 - 70x + 4 = 0$$

$\therefore a = 5, b = -70, c = 4$

$\therefore \alpha+\beta = -b/a \Rightarrow \alpha+\beta = -70/5 = 14$

and, $\alpha\beta = c/a \Rightarrow \beta = \frac{c}{a} \times \frac{1}{\alpha}$

$$= \frac{4}{5} \times \frac{1}{13.942}$$

From (b) we get,

$\alpha = 13.942$

$$= 0.0\underbrace{57378}_{SF=5}10$$

$$= 0.057378 \text{ (Ans)}$$