1)

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

Standard Norm form:

Max num = $(0.1111) \times 2^4 = 15$

Min num without minus = $(0.1000) \times 2^{-3} = 0.0625$

" " with " $= -(0.1111) \times 2^4 = -15$

IEEE Norm:

Max num = $(0.11111)\ 2^4 = 15.5$

Min num without minus = $(0.10000)\ 2^{-3} = 0.0625$

" " with " $= -(0.11111)\ 2^4 = -15.5$

IEEE deNorm:

Max num = $(1.1111)\ 2^4 = 31$

Min num without minus = $(1.0000)\ 2^{-3} = 0.125$

" " with " $= -(1.1111)\ 2^4 = -31$

b) without minus,

standard form = $2^3 \times 8 = 64$

IEEE = $2^4 \times 8 = 128$

with minus,

standard form = $2^3 \times 8 \times 2 = 128$

IEEE = $2^4 \times 8 \times 2 = 256$

d) $\beta = 2$, $m = 52$, $e = (0, 2047)$

smallest number = $(0.1000\ldots0) \times 2^{0-1023+1+1}$

not taking (0,0)

$= (0.100\ldots\underset{52}{0}) \times 2^{-1021}$

Largest number = $(0.111\ldots1) \times 2^{2047-1023+1-1}$

$= (0.111\ldots\underset{52}{1}) \times 2^{1024}$

e) $\beta = 2$, $e = [0, 2047]$, Bias = 500

$0-500+1+1$

$\therefore$ smallest positive number $= (0.100\ldots\ldots 0_{52}) \times 2^{0-500+1+1}$

$= (0.100\ldots\ldots 0_{52}) \times 2^{-498}$

$\therefore$ Largest " " $= (0.111\ldots\ldots 1_{52}) \times 2^{2047-500+1-1}$

$= (0.111\ldots 1_{52}) \times 2^{1547}$

2) $x = \frac{3}{8} = \frac{2}{8} + \frac{1}{8} = \frac{1}{4} + \frac{1}{8} = (0.011)\ 2^{0}$ $m = 4$

$y = \frac{5}{8} = (0.101)\ 2^{0}$

$fl(x) = (0.011) 2^{0} = \frac{3}{8}$, $fl(y) = (0.101) 2^{0} = \frac{5}{8}$

$x \cdot y = fl(x) \cdot fl(y) = \frac{15}{64} = \frac{1}{64} + \frac{2}{64} + \frac{4}{64} + \frac{8}{64}$

$= \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} = 0.001111$

$= 0.1111 \times 2^{-2}$

$\therefore fl(x \cdot y) = 0.1111 \times 2^{-2}$ $\therefore x \cdot y = fl(x \cdot y)$

No need for rounding. $\therefore$ R. Error $= |x \cdot y - fl(x \cdot y)|$

$= 0$

3)

$x^2 - 60x + 1 = 0$ $\qquad$ $\sqrt{3596} = 59.9666574$

$\therefore x = \dfrac{60 \pm \sqrt{3596}}{2}$

$= \dfrac{60 \pm 59.9667}{2}$

$= 59.9835, \ 0.01665$

as the numbersi are really close, there will be loss of significance while substracting.

$x_1 = 59.9835$

We know,

$x^2 - (x_1 + x_2)x + x_1 x_2 = 0$

$\therefore x_1 x_2 = 1$

$\Rightarrow x_2 = \dfrac{1}{59.9835} = 0.0166713$

$(0.1 d_1 d_2 \ldots)$

4)

$\beta = 2 \ , m = 5 \ , e = [-100, 100]$

a)

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

b) $|x_{min}| = (0.100 \ldots d_m) \, 2^e$

$= 2^{-1} 2^e$

c) non neg numbers total $= 2^5 \times 201 = 6432$

5)

$x^2 - 16x + 3 = 0$

$\Rightarrow x = \dfrac{16 \pm \sqrt{16^2 - 4 \cdot 3}}{2}$

$\sqrt{61} = 7.810249676$

$= 8 \pm \sqrt{61}$

$= 8 \pm 7.810$

$= 15.81, \ 0.19$

when we substroact closen numbers we get loss of significance.
As adding doesn't result in loss of significance

$x_1 = 15.81$

$\therefore x_2 = \dfrac{3}{15.81} = 0.1897$

6/ $B=2,\ m=3,\ e=[-1,2]$

a) $(6.25)_{10} = (110.01)_2 = (0.11001)_2 \times 2^3$

$(6.875)_{10} = (110.111)_2 = (0.110111)_2 \times 2^3$

$fl(6.25)_{10} = (0.1101)_2 \times 2^3$

$fl(6.875)_{10} = (0.1110)_2 \times 2^3$

$fl(6.25)_{10} = 6.5$

$fl(6.875)_{10} = 7$

b) $\delta_1 = |fl(6.25) - 6.25|$

$= |6.5 - 6.25|$

$= 0.5$

$\delta_2 = |fl(6.875) - 6.875|$

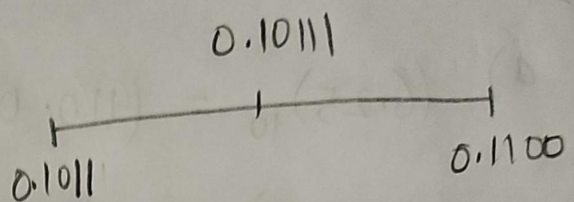
$= |7 - 6.875|$

$= 0.125$

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

$(6.875)_{10} = (1.10111)_2 \times 2^2$

$= (1.1100)_2 \times 2^2$



d) Standard $re = \frac{1}{2}\beta^{1-m} = \frac{1}{2}2^{1-3} = 0.125$

Normal $re = \frac{1}{2}\beta^{-m} = \frac{1}{2}2^{-3} = 0.0625$

deNormal $re = 0.0625$

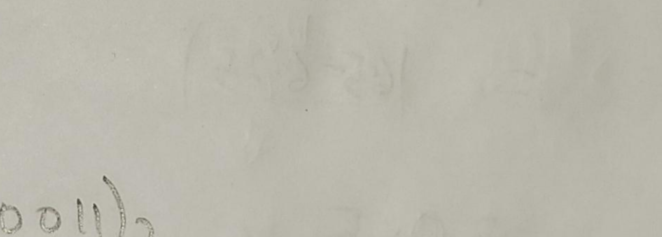
7) a) $(8.235)_{10} = (1000.0011)_2$

$0.235 \times 2 = 0.47$ 0

$0.47 \times 2 = 0.94$ 0

$0.94 \times 2 = 1.88$ 1

$0.88 \times 2 = 1.76$ 1



b) $x = (1.0000011)_2 \times 2^3$

$fl(x) = (1.000010)_2 \times 2^3$

c) $x' = 8.25$

$\therefore RE = |x' - x|$

$= |8.25 - 8.235|$

$= 0.015$

8/ $x^2 - 12x + 5 = 0$

a) $x = \dfrac{12 \pm \sqrt{12^2 - 4.1.5}}{2}$

$\sqrt{31} = 5.5677643$

$= 6 \pm \sqrt{31}$

$= 6 \pm 5.567$

$= 11.57, \; 0.433$

b) $fl(x) = x + \delta_1 x$
$fl(y) = y + \delta_2 y$

$x \pm y = fl(x) \pm fl(y)$
$= x + \delta_1 x \pm y \pm \delta_2 y$
$= x \pm y + \delta_1 x \pm \delta_2 y$
$= (x \pm y)\left(1 + \dfrac{\delta_1 x \pm \delta_2 y}{x \pm y}\right)$

$$\frac{\delta_1 x + \delta_2 y}{x \pm y}$$ is the error part.

when x is closer to y then this value will be high.

c) $x_1 = 11.57$

$\therefore x_2 = \frac{5}{11.57} = 0.4321$