CSE 330

i) Standard form $x = \pm (0.d_1 d_2 \cdots d_m) \times \beta^e$

$d_1 \neq 0$

ii) Normalized

$x = \pm (0.1 d_1 d_2 \cdots d_m)_\beta \times \beta^e$

iii) Denormalized

$x = \pm (1.d_1 d_2 \cdots d_m)_\beta \times \beta^e$

① $\beta = 2$, $m = 3$ $\quad -2 \leq e \leq 1$ (e_{min}, e_{max})

a) Find minimum and maximum number with and without negative support

i) Standard form → $(0.d_1 d_2 d_3) \times 2^e$

w/o -ve

$Min = (0.100)_2 \times 2^{-2}$

$Max = (0.111)_2 \times 2^1$

$e = \{-2, -1, 0, 1\}$

with -ve

$Min = -(0.111)_2 \times 2^1$

$Max = +(0.111)_2 \times 2^1$

ii) Normalized → $(0.1 d_1 d_2 d_3)_2 \times 2^e$

w/o -ve

$Min = (0.1000)_2 \times 2^{-2}$

Max = $(0.1111)_2 \times 2^{1}$

with -ve

Min = $-(0.1111)_2 \times 2^{1}$

Max = $+(0.1111)_2 \times 2^{1}$

iii) Denormalized → $(1.d_1d_2d_3)_2 \times 2^{e}$

w/o -ve,

Min = $(1.000)_2 \times 2^{-2}$

Max = $(1.111)_2 \times 2^{1}$

w -ve.

Min = $-(1.111)_2 \times 2^{1}$

Max = $(1.111)_2 \times 2^{1}$

b) How many numbers can be represented in each form?

i) Standard → $(0.d_1d_2d_3)_2 \times 2^{e}$ $\{-2, -1, 0, 1\}$

$d_1 \neq 0 \rightarrow d_1 = 1$

d_2 d_3

0/1 0/1

$2 \times 2 = 4$

4×4

$= 2^2 \times 4 = 16$

$(0.100) \times 2^{-2}$

$(0.101) \times 2^{-2}$

$(0.110) \times 2^{-2}$

$(0.111) \times 2^{-2}$

-ve support

16×2

ii) Normalized form

$(0.1d_1d_2d_3)_2 \times 2^{e}$

0/1 0/1 0/1

$2 \times 2 \times 2 = 2^3$

$2^3 \times 4 = 8 \times 4 = 32$

-ve support

(0.1000)
(0.1001)
(0.1010)
(0.1011)
(0.1100)
(0.1101)
(0.1110)
(0.1111)

$\rightarrow$ 32 × 2 = 64

iii) $(1.d_1d_2d_3) \times 2^e$

□ □ □ $= 2^3$

2 × 2 × 2

0/1 0/1 0/1

$\rightarrow$ $2^3 \times 4 = 32$

-ve support 64

c) IEEE format (754) - 64 bit precision

m = 52 bits, e = 11 bits, sign bit = 1 bit

$(0.1d_1d_2 \ldots d_{52}) \times 2^{e-1022}$ Bias

$e_{min} = 0$

$e_{max} = 2^{11} - 1 = 2047$

$\rightarrow (1.d_1d_2 \ldots d_{52}) \times 2^e$

Smallest no $\rightarrow (1.0_10_2 \ldots 0_{52}) \times 2^0 =$ 1

Largest no $\rightarrow (1.1_11_2 \ldots 1_{52}) \times 2^{2047} =$ □

2^{-ve}

Biasing

$(1.d_1 d_2 \ldots\ldots d_{52})_2 \times 2^{e-1023}$ ← Bias

$= (0.1 d_1 d_2 \ldots\ldots d_{52})_2 \times 2^{e-1023} \times 2^1$

$= (0.1 d_1 d_2 \ldots\ldots d_{52})_2 \times 2^{e-1022}$

$e_{min} \to 0 \quad e_{max} = 2047$

Minimum $= (0.1 0_1 0_2 \ldots 0_{52}) \times 2^{-1022}$

Maximum $= (0.1 1_1 1_2 \ldots 1_{52}) \times 2^{1025}$

Theoretical

By convention → $\times 0$ → inf

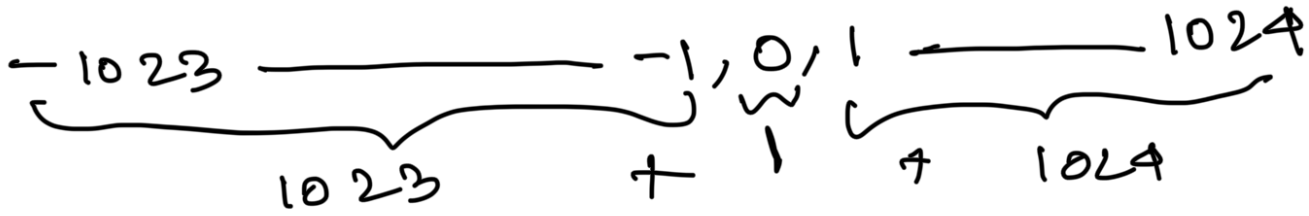
→ Actual

minimum $= (0.1 0_1 0_2 \ldots 0_{52}) \times 2^{-1021}$

maximum $= (0.1 1_1 1_2 \ldots 1_{52}) \times 2^{1024}$

Bias → why 1023?

exponent, e range 0 to 2047, 2048



$= 2048$

$$\text{Bias} = \frac{2^e}{2} - 1$$

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

② $(6.25)_{10}$

Convert into Standard floating point representation where $\beta = 2$, $m = 3$,

$$-1 \leq e \leq 3$$

$(0.d_1d_2d_3) \times 2^e$

$(6.25)_{10} = (110.01)_2$

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

$= (0.110) \times 2^3$

$(0.1d_1d_2d_3)$

$= (0.1100)_2 \times 2^3$

Rounding ?

$0.\underbrace{110}_{m=3}|\boxed{0}1$

i) If $m+1$ th bit $= 0$, just truncate

ii) If m+1th bit = 1

ii-a) 0.110 | 1 0.111 | 1

Round to the nearest even no

0.110

ii-b) 0.110 | 101...

0.111 increase mth bit by 1

Original number $(6.25)_{10}$ After rounding →

$$(0.110)_2 \times 2^3 =$$

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

$$= \left(\frac{1}{2} + \frac{1}{4}\right) \times 2^3 = \frac{3}{4} \times 8 = 6$$

$$R.E = |6.25 - 6| = |fl(x) - x|$$

$$= 0.25$$

(3) If $x = \frac{3}{8}$, $y = \frac{5}{8}$ Find $fl(x \times y)$

where $m = 4$ Find rounding error if any.

$\frac{3}{8}$ $\frac{5}{8}$

$x = \frac{3}{8}$

$= 0.375$

$= (0.011)_2$

$= (0.11)_2 \times 2^{-1}$

$y = \frac{5}{8}$

$= 0.625$

$= (0.101)_2$

$= (0.101)_2 \times 2^0$

$0.375 \times 2 = 0.75$

$0.75 \times 2 = 1.5$

$0.5 \times 2 = 1.0$

$fl(x) = (0.11)_2 \times 2^{-1}$

$fl(y) = (0.101) \times 2^0$

Standard form $(0.d_1 d_2 d_3 d_4) \times 2^e$, $d_1 \neq 0$

$x \times y = fl(x \times y)$

$= fl(x) \times fl(y)$

$= (0.11)_2 \times 2^{-1} \times (0.101)_2 \times 2^0$

$= \frac{3}{4} \times \frac{1}{2} \times \frac{5}{8}$

$= \frac{15}{64} = 0.234375$

$= (0.001111)_2$

$= (0.1111)_2 \times 2^{-2}$

Rounding $= \frac{15}{16} \times 2^{-2}$

$$fl(x*y) = \frac{15}{64}$$

$$R.E = \left| \frac{15}{64} - \frac{15}{64} \right| = \left| fl(x*y) - (x*y) \right|$$

$$= 0$$

Short technique →

Decimal system → $\frac{675}{10 \to 10^1} = 67.5$

$$\frac{675}{100 \to 10^2} = 6.75$$

$$\frac{675}{10^3/1000} \to = 0.675$$

$$\frac{675}{10^4/10000} = 0.0675$$

$$6.75 \times 10^1 = 67.5$$

$$6.75 \times 10^2 = 675$$

$$6.75 \times 10^3 = 675$$

$$\frac{3}{8} = \frac{3}{2^3}$$

$$\frac{(11)_2}{2^3}$$

$$\frac{(11)_2}{2^2} = (0.11)_2$$

$$= (0.011)_2 \to (0.11)_2 \times 2^{-1}$$

5 $(101)_2$

$$\frac{1}{8} = \frac{1}{2^3}$$

$$= (0.101)_2$$

$$\frac{15}{64} = \frac{(1111)_2}{2^6} \quad \frac{(1111)_2}{2^1} = (11.1)_2$$

$$= (0.001111)_2$$

④ $\beta = 2$, $m = 4$, $-100 \leq e \leq 100$

Using IEEE normalized form —

a) Find machine epsilon ε_M

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

$$\frac{\varepsilon_{max}}{\varepsilon_M} = \frac{R.E_{max}}{|x|_{min}} = \frac{1}{2}\beta^{1-m}$$

$$= \frac{1}{2}\beta^{1-4} = \frac{1}{2} \times 2^{-3}$$

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

$$(R.E)_{max} \quad \frac{1}{2}\beta^{-m}\boxed{\beta^e}$$

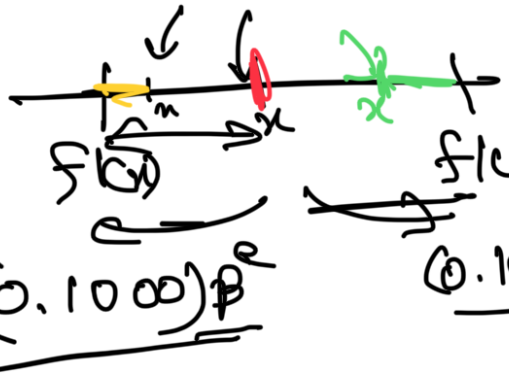
$$|x|_{min} = \frac{\beta^{-1}\boxed{\beta^e}}{1\ \beta^{-m}\ \beta^e} = 2^{-1} \times 2^{-100}$$

$$= 2^{-101}$$

$$\frac{\frac{1}{2}\beta^{-m}}{\beta^{-1}\beta^{R}}$$

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

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



$(0.1000)\beta^{e}$ $(0.100\beta^{e}$

$(0.1001)\beta^{e} - (0.1000)\beta^{e}$

$= (0.0001)\beta^{e}$

$= \beta^{-4}\beta^{e} \rightarrow \frac{1}{2}\beta^{-4}\beta^{e}$

(5) $x^2 - 12x + 5 = 0$

Compute the roots of the eqn keeping four significant figures.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{12 \pm \sqrt{144 - 20}}{2}$$

$$= \frac{12 \pm 2\sqrt{31}}{2}$$

$= 6 + \sqrt{31} = x_1$

$6 - \sqrt{31} = x_2$

1 2 3 4

$x_1 = 11.5677 = 11.57$

$x_2 = 0.43223 = \boxed{0.4322}$

Non-zero 5.f

1 2 3 4

Toy computer →

first calculate $\sqrt{31}$

$= \boxed{5.5677}$

Loss of significance 0.4330 - 0.4322

$x_1 = 6 + 5.567$

$= \underline{11.567} = \underline{11.57}$

$x_2 = \underline{6 - 5.567} = 0.\boxed{4330}$

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

$\alpha\beta = 5$

$\alpha = \underline{11.57}$, $\beta = \frac{5}{\alpha}$

$= \frac{5}{11.57}$

$= 0.43215$

$\boxed{= 0.4322}$