Floating Point Arithmetic

Fixed point Representation:

In our daily lives, we typically use fixed-point representation.

$$X = \pm (d_1 d_2 d_3 \ldots d_{k-1} \cdot d_k \ldots d_n)_\beta$$

Where $d_1 \ldots d_n \in \{0, 1, \ldots, \beta - 1\}$

Example:

$$X = +(11 \cdot 1)_2$$

$$X = -(21 \cdot 5)_{10}$$

Evaluating Fixed point numbers in base 10:

$$(\overset{1}{1}\overset{0}{1} \cdot \overset{-1}{1})_2 = 1 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1}$$

$$= 2 + 1 + \frac{1}{2}$$

$$= 3 + \frac{1}{2}$$

$$= \frac{7}{2}$$

$$= (3 \cdot 5)_{10}$$

Floating Point Representation

Floating point numbers are the subset of all real numbers $F \subset \mathbb{R}$. In computers, floating-point representation allows the system to handle a wide range of numbers, from tiny decimals to huge values, by breaking them down (organized) into a format that the computer can easily store and process.

$$F = \pm \left(0.d_1 d_2 d_3 \cdots\cdots d_n\right)_{\beta} \times \beta^{e}$$

e → exponent

β → Base

$(0.d_1 d_2 d_3 \cdots\cdots d_n)$ → Significand / Fraction / mantissa

where, $\beta, d_i, e \in \mathbb{Z}$ → integers

$$0 \leq d_i \leq \beta - 1$$

$$e_{min} \leq e \leq e_{max}$$

Examples: (Magic of Math)

[We have convert a given number into floating point representation]

* $\pm (110.111)_2 \times 2^0$

$= \pm (11.0111)_2 \times 2^1$

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

$= \pm (0.110111)_2 \times 2^3$

* $(1594.23)_{10} \times 10^0$

$= (159.423)_{10} \times 10^1$

$= (15.9423)_{10} \times 10^2$

$= (1.59423)_{10} \times 10^3$

$= (0.159423)_{10} \times 10^4$

Conventions

Convention 1:

$$\pm (0.d_1 d_2 d_3 \cdots d_m)_\beta \times \beta^e$$

(d_1 ↓ 1)

*** Here $d_1 = 1$ and $d_1 \neq$ any (always) other value ***

d_1 always have to be 1.

Example: Given that, $\beta = 2$, $m = 3$, $e\,[-1, 2]$ (e_{min}, e_{max})

Highest possible Floating point number using convention is $(0.111)_2 \times 2^2$

Normalized Form : (Convention 2)

$$\pm (1.d_1 d_2 d_3 ----- d_m)_\beta \times \beta^e$$

* Here d_1 can be 1 or any other value.

Example: Given that, $\beta = 2$, $m = 3$, e [$\underset{e_{min}}{-1}$, $\underset{e_{max}}{2}$]

Highest possible floating point number using Convention 2/ Normalized Form is $(1.111)_2 \times 2^2$.

Denormalized Form : (Convention 3)

$$\pm (0.1 d_1 d_2 d_3 ---- d_m)_\beta \times \beta^e$$

* Here d_1 can be 1 or any other value.

Example : Given that, $\beta = 2$, $m = 3$, e [-1, 2]

Highest possible floating point number using Convention 3/ denormalized form is $(0.111)_2 \times 2^2$.

Total Combination (How to find total Combination / How many floating numbers can be possible)

1. Given that, $\beta = 2$, $m = 3$, $e = [-1, 2]$

e_{min} e_{max}

Using convention 1 : [Where $d_1 = 1$ (always)]

For $e = -1$

$$(0.1\ d_2\ d_3)_2 \times 2^{-1}$$

d_2	d_3
0	0
0	1
1	0
1	1

[For $e = -1$, Here 4 combinations possible]

$$2^n = 2^2 = 4$$

$e = [-1, 2]$ means "e" can be $-1, 0, 1, 2$.

For $e = 0$

$$(0.1\ d_2\ d_3)_2 \times 2^{-1}$$

d_2	d_3
0	0
0	1
1	0
1	1

[For $e = 0$, Here 4 combinations possible]

Same thing goes for $e = 1$, $e = 2$

4 combinations possible — 4 combinations possible

$\therefore$ total possible combinations $= 4 + 4 + 4 + 4 = 16$.

2. Given that, $\beta = 2$, $m = 3$, $e = [-1, 2]$

* Using normalized form / Convention 2

$$(1.d_1\ d_2\ d_3)_2 \times 2^{-1}$$

d_1	d_2	d_3
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

[For $e = -1$, Here 8 combina possible]

$$2^n = 2^3 = 8$$

Same thing goes for $e = 0$, $e = 1$, $e = 2$

8 combination, 8 combination, 8 combination

∴ Total possible Combinations $= 8 + 8 + 8 + 8$

$= 32$

3. Given that $b = 2$, $m = 3$, $e = [-1, 2]$

* Using normalized Form / Convention 3

$$(0.1\ d_1\ d_2\ d_3)\ 2^{x}\ 2^{-1}$$

d_1	d_2	d_3
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

[For $e = -1$, Here 8 combinati possible]

$\boxed{2^3 = 8}$

Same thing goes for $e = 0$, $e = 1$, $e = 2$

8 combinations 8 combination 8 combinations

$\therefore$ Total possible Combinations $= 8 + 8 + 8 + 8$

$= 4 \times 8$

$= 32$

* Practic Problem

[Find the smallest and largest values]

1. Given that $\beta = 2$, $m = 3$, $e = [-1, 2]$. Find the largest otc maximum value otc Highest value.

[Using convention 1]

Highest value $= +(0 \cdot d_1 d_2 d_3)_2 \times 2^e$

$= +(0 \cdot 1 \; d_2 \; d_3)_2 \times 2^e$

$= +(0 \cdot \underbrace{1}_{d_1=1} \; 1 \; 1)_2 \times 2^{\overbrace{2}^{e_{max}}}$

[As we follow convention 1 here]

$= (7/2)_{10}$

2. Given that $\beta = 2$, $m = 3$, $e = [-1, 2]$, Find the non-negative smallest otc minimum value.

[Using Convention 1]

Smallest value [non-negative] $= +(0.\underbrace{1}_{d_1=1}00)_2 \times 2^{\overbrace{-1}^{e_{min}}} = \frac{1}{4}$

[as we follow convension 1]

3. Give that, $\beta = 2$, $m = 3$, $e = [-1, 2]$

Find the negative minimum otc smallest value.

Negative minimum otc smallest value $=$ − Highest value

$= -(0.111)_2 \times 2^2$.

$= -7/2$

Important

* IF it mentions only the smallest number, you need to include both the smallest non-negative and negative value.

Note

Binary to Decimal

1. $+(0.\overset{-1}{1}\overset{-2}{1}\overset{-3}{1})_2 \times 2^{2}$

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

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

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

$= \frac{4+2+1}{8_2} \times 4$

$= \left(\frac{7}{2}\right)_{10} = 7/2$

2. $(0.\overset{-1}{1}\overset{-2}{0}\overset{-3}{0})_2 \times 2^{-1}$

$= (1\times2^{-1} + 0\times2^{-2} + 0\times2^{-3}) \times \frac{1}{2}$

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

$= \frac{1}{4}$

Practice Problem

Ⓐ $\beta = 2$, $m = 3$, $e = [-1, 2]$. Find the highest value using **Normalized Form.**

$$\pm (1.d_1 d_2 d_3 \ldots d_m)_\beta \times B^e$$

Highest value $= +(1.1\,1\,1)_2 \times 2^e$

$= (1.1\,1\,1)_2 \times 2^2$

$= (1\times 2^0 + 1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3}) \times 2^2$

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

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

$= \frac{15}{8} \times 4$

$= \left(\frac{15}{2}\right)_{10}$

$= (7.5)_{10}$

$B=2$, $m=3$, $e=[-1, 2]$. Find the highest value using **denormalized Form.**

$$\pm (0.1\, d_1\, d_2\, d_3 \ldots d_m)_B \times B^e$$

Highest value $= (0.1\,1\,1\,1)_2 \times 2^2$

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

$$= \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}\right) \times 4$$

$$= \frac{8+4+2+1}{16} \times 4$$

$$= \frac{15}{4}$$

$$= \left(\frac{15}{4}\right)_{10}$$

$$= (3.75)_{10}$$

Floating Point numbers are not equally spaced

Let $b=2$, $m=3$, $e=[-1, 1]$ — For convention 1

* Smallest non-negative number combinations.

For $e=-1$,

$(0.100) \times 2^{-1} = 1/4$

$(0.101) \times 2^{-1} = 5/16$

$(0.110) \times 2^{-1} = 3/8$

$(0.111) \times 2^{-1} = 7/16$

For $e=0$

$(0.100) \times 2^{0} = \frac{1}{2}$

$(0.101) \times 2^{0} = 5/8$

$(0.110) \times 2^{0} = 3/4$

$(0.111) \times 2^{0} = 7/8$

For $e=1$,

$(0.100) \times 2^{1} = 1$

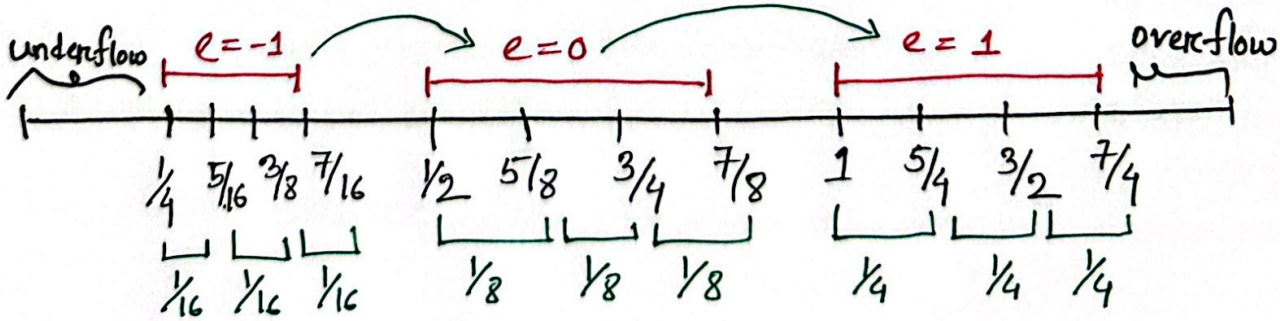
$(0.101) \times 2^{1} = 5/4$

$(0.110) \times 2^{1} = 3/2$

$(0.111) \times 2^{1} = 7/4$

*** important ***

When exponent e is constant, numbers are equally spaced. However, when exponent e changes, the interval or gap between numbers change.



IEEE Standard for double precision (64 bit)

$\beta = 2$

Fraction/mantissa = 52 bits

Exponent = 11 bits

Sign = 1 bits

Using Normalize Form

$$\pm (1.d_1 d_2 d_3 \ldots\ldots d_{52})_2 \times 2^e$$

For e = 11 bits, there will be total $2^{11} = 2048$.

range of e $[0, 2047]$

0 → e_{min}, 2047 → e_{max}

Largest possible number $= + (1.111\ldots\ldots1)_2 \times 2^{2047}$

Smallest possible number (non-negative) $= +(1.00\ldots0)_2 \times 2^0$

$= 1$ [1 is not the enough small number. In this way FP can't store smaller values]

* To get a small number, the exponent value must be small.

* We obtain smaller values through exponent biasing.

Formula :

$$2^{K-1} - 1$$

$$= 2^{11-1} - 1$$

$$= 2^{10} - 1$$

$$= 1024 - 1 = 1023$$

$$\therefore (1.d_1 d_2 - - \cdot d_{52})_2 \times 2^{e-1023}$$

↳ exponent biasing

For denormalized form

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

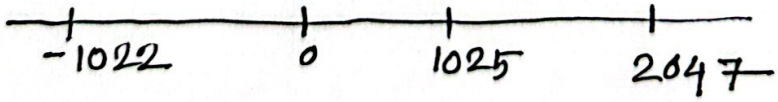
$$e - 1022 \in [-1022, 1025]$$

Largest possible number $= (0.11 - - - \cdot 1)_2 \times 2^{1025}$

$$\approx \infty$$

Smallest possible number $= (0.10 - - - \cdot 0)_2 \times 2^{-1022}$

$$\approx 0$$



* Reduce the range for largest possible number so that we can get the smaller values.

$2^{1025} \rightarrow \pm\infty$ [highest power is used to store ∞]

$2^{-1022} \rightarrow \pm 0$ [used to store 0]

$\therefore$ Highest value $= (0.111\ldots.1)_2 \times 2^{1024}$ [if we put 1025 it will denote ∞]

$\simeq 1.798 \times 10^{308}$

$\therefore$ Smallest value $= (0.100\ldots.0)_2 \times 2^{-1022}$

$\simeq 2.225 \times 10^{-308}$