

Fixed point Representation

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

$$\text{where } d_1, \dots, d_n \in \{0, 1, \dots, \beta-1\}$$

Fixed points or floating points are how numbers are stored/represented in a computer

Example

$$X = +(10.1)_2$$

$$X = -(123.12)_{10}$$

Evaluating fixed point numbers in base 10:

$$\begin{aligned}(10.1)_2 &\rightarrow 1 \times 2^1 + 0 \times 2^0 + 1 \times 2^{-1} \\ &= 2 + 0 + \frac{1}{2} \\ &= (2.5)_{10}\end{aligned}$$

Floating point Representation:

$$F = \{ \pm (\underbrace{0.d_1 d_2 \dots d_n}_{\text{fraction/mantissa}}) \beta^e \}$$

$e \rightarrow \text{exponent}$
 $\beta \rightarrow \text{base}$

$$\text{Where } \beta, d_i, e \in \mathbb{Z} \rightarrow \text{integers}$$

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

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

Evaluating floating point numbers in base 10

Examples

$$\begin{aligned}123.45 \\&= 12.345 \times 10^1 \\&= 1.2345 \times 10^2 \\&= 0.12345 \times 10^3\end{aligned}$$

$$\begin{aligned}1001.11 \\&= \underbrace{0.100111}_{\text{Fraction}} \times \underbrace{2^4}_{\text{Base}} \quad \text{4 — exponent}\end{aligned}$$

Conventions :

$$\textcircled{1} \pm (0.\underbrace{d_1}_{\downarrow} d_2 \dots d_m)_\beta \beta^e$$

$d_1 = 1$ always.

Example :

$$\begin{aligned}\beta &= 2 & e_{\min} &= -1 \\m &= 3 & e_{\max} &= 2\end{aligned}$$

$$\rightarrow \text{highest possible FP number} = (0.111)_2 \times 2^2$$

② Normalized Form:

$$\pm (1.\underbrace{d_1}_{\downarrow} d_2 \dots d_m)_\beta \beta^e$$

$\left. \begin{matrix} d_1 = 1 \\ d_1 \neq 1 \end{matrix} \right\} \text{both correct}$

Example :

$$\begin{aligned}\beta &= 2 & e_{\min} &= -1 \\m &= 3 & e_{\max} &= 2\end{aligned}$$

$$\rightarrow \text{highest possible FP number} = (1.111)_2 \times 2^2$$

Denormalized Form

$$\pm (0.1 d_1 d_2 \dots d_m)_\beta \beta^e$$

Example

$$\beta = 2 \quad e_{\min} = -1$$

$$m = 3 \quad e_{\max} = 2$$

$$(0.111)_2 \times 2^2$$

$$\beta = 2 \quad e_{\min} = -1 \quad \text{Convention 1}$$

$$m = 3 \quad e_{\max} = 2$$

Ignoring the \pm sign

Find the smallest and largest non-negative number

$$\begin{array}{ccc} d_1 & d_2 & d_3 \\ 0.1 & \square & \square \end{array}$$

Fixed

$$\begin{array}{cc} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{array}$$

4 possible #

$$\begin{array}{c} 0 \\ \downarrow \\ -1 \\ 0 \\ 1 \\ 2 \end{array}$$

4 possible #

\therefore Total possible # that can be represented = $4 \times 4 = 16$

Smallest #

$$\begin{aligned} & (0.100)_2 \times 2^{-1} \\ & = (1 \times 2^{-1}) \times 2^{-1} \\ & = \frac{1}{4} \end{aligned}$$

Largest #

$$\begin{aligned} & (0.111)_2 \times 2^2 \\ & = (1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3}) \times 2^2 \\ & = \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} \right) \times 2^2 \\ & = \frac{7}{2} \end{aligned}$$

Considering sign bit

Smallest possible # = $-\frac{7}{2}$

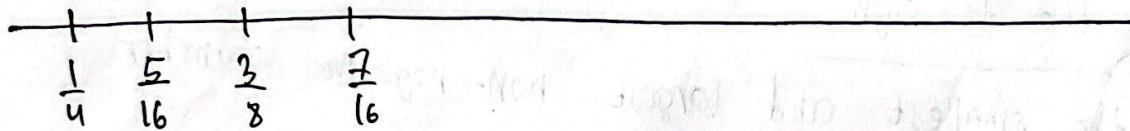
$e = -1$

1st smallest non-negative number = $(0.100) \times 2^{-1} = \frac{1}{4}$

2nd " " " " = $(0.101) \times 2^{-1} = \left(\frac{1}{4} + \frac{1}{8}\right) \times 2^{-1} = \frac{5}{16}$

3rd " " " " = $(0.110) \times 2^{-1} = \frac{3}{8}$

4th " " " " = $(0.111) \times 2^{-1} = \frac{7}{16}$



Equally spaced

$$\frac{5}{16} - \frac{1}{4} = \frac{3}{8} - \frac{5}{16} = \frac{7}{16} - \frac{3}{8} = \frac{1}{16} \quad [\text{exponent constant} = \text{equally spaced}]$$

$e = 0$

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

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

⋮

$e = 1$

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

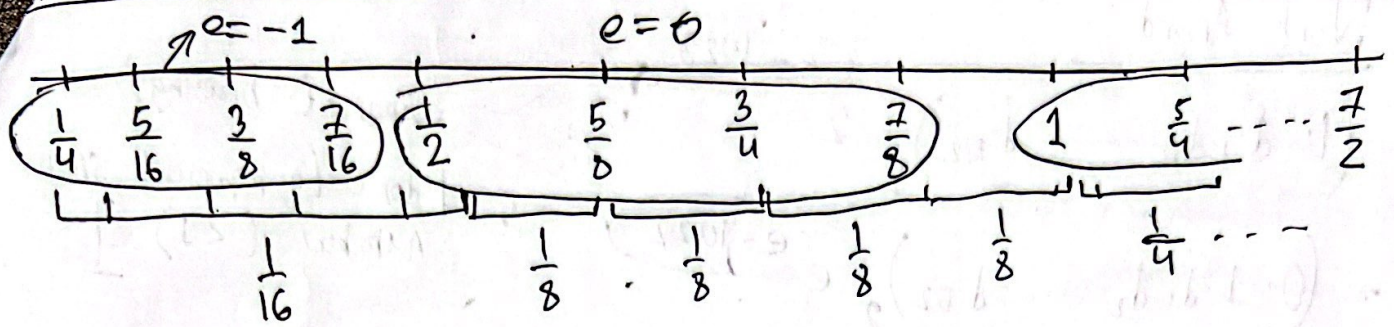
⋮

$e = 2$

$$(0.100) \times 2^2 = 2$$

$$(0.111) \times 2^2 = \frac{7}{2}$$

-negative number line:



Important

If convention 1/ is used, we don't have 0 in our number system. Bcz (0.1000...) ^{denormalized form}

we have 1 fixed here.

But 0 can be approximated using, for example, $(0.1000...) 2^{-50} \approx 0$

IEEE standard for double precision

$\beta = 2$ 52 - bits for fraction/mantissa
 11 - bits for exponent
 1 - bit for sign

Starting with Normalized Form:

$$(1. d_1 d_2 \dots d_{52}) \times 2^e$$

min value of $e = 0$ [all 11 bits 0]

max value of $e = 2^{11} - 1 = 2047$ [all 11 bits 1]

$$e_{\min} = 0$$

$$e_{\max} = 2047$$

largest possible ^{non-neg} num = $(1.11\dots1)_2 \times 2^{2047}$

smallest " " " = $(1.0\dots0)_2 \times 2^0 = 1$ (not that small)

cannot express 0.001 for example.

Work Around

$$(1 \cdot d_1 d_2 \dots d_{52})_2 \cdot 2^{e-1023}$$
$$= (0 \cdot 1 d_1 d_2 \dots d_{52})_2 \cdot 2^{e-1022}$$

exponent biasing.
[done to represent small numbers (< 1)]

previously $e \in [0, 2047]$

Now, with exponent biasing $[-1022, 1025]$

$$\left. \begin{aligned} \text{Now, highest possible num} &= (0 \cdot 1 \underbrace{11 \dots 1}_{52 \text{ 1s}}) \times 2^{1025} \approx \infty \\ \text{smallest possible number} &= (0 \cdot 1 0 \dots 0) \times 2^{-1022} \approx 0 \end{aligned} \right\}$$

In IEEE standard, 2 bits from exponent is reserved for ∞ and 0.

Highest possible exponent is used to store infinity

$$2^{1025} \rightarrow \infty$$

smallest possible exponent is used to store 0

$$2^{-1022} \rightarrow 0$$

$$\text{Now, highest possible num, except inf} = (0 \cdot 1 11 \dots 1)_2 \times 2^{1024}$$
$$\approx 1.798 \times 10^{308}$$

$$\text{lowest " " , except 0} = (0 \cdot 1)_2 \times 2^{-1021}$$
$$\approx 2.225 \times 10^{-308}$$