Fixed-point representation.

Example: 8 bits to store a real number.

$$(\underbrace{1\,0\,1\,1\,1}_{\text{integer part}}.\underbrace{0\,1\,1}_{\text{fractional part}})_2$$

$$\underbrace{2^4\ 2^3\ 2^2\ 2^1\ 2^0}_{\text{integer part}}\quad \underbrace{2^{-1}\ 2^{-2}\ 2^{-3}}_{\text{fractional part}}$$

More bits for the integer part → increase range

fractional part → precision

Scientific notation: $x = \pm r \times 10^m$, $r \in [1, 10]$.

floating number in the binary system:

$$x = \pm q \times 2^m$$

q is significand, normally a fractional value in the range $[1.0, 2.0)$

m is the exponent.

leading bit

Numerical Form: $x = \pm q \times 2^m = \pm b_0 \underbrace{.b_1 b_2 \cdots b_n}_{\text{fractional}} \times 2^m$

$b_i \in \{0,1\}$

Exponent range: $m \in [L, U]$.

Precision: $p = n+1$.

$$0.40625 = \overset{2^{-5}}{0.03125}\,x_3 + \overset{2^{-4}}{0.0625}\,x_2 + \overset{2^{-3}}{0.125}\,x_1$$

$x_1 = 3 \qquad x_2 = 0 \qquad x_3 = 1$

$$(0.40625)_{10} = 2^{-2} + 2^{-3} + 2^{-5}$$

$$= (0.01101)_2$$

Normalized floating-point numbers.

$\text{FP number} = \langle\text{sign}\rangle\, 1.\langle\text{fraction field}\rangle \times 2^{(\langle\text{exponent field}\rangle - \text{bias})}$

$$x = \pm 1.b_1 b_2 b_3 \cdots b_n \times 2^m = \pm 1.f \times 2^m.$$

$b_i \in \{0,1\}$.

Exponent range: $m \in [L, U]$

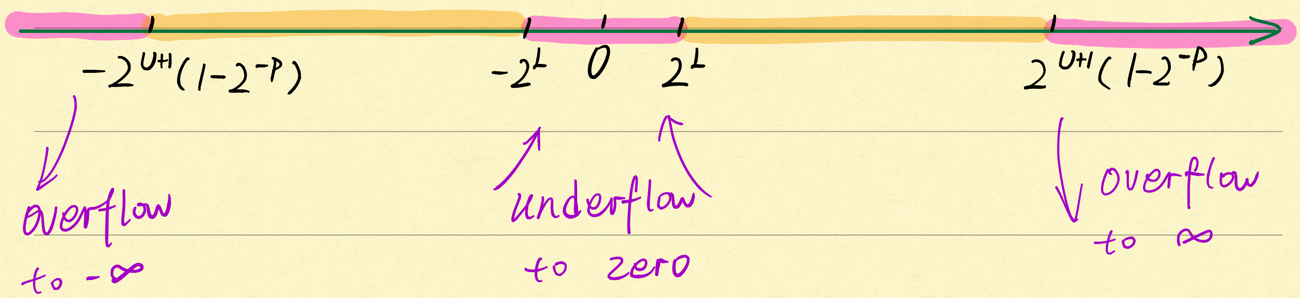
Precision: $p = n+1$.

Smallest positive normalized FP number:

$$1.\underbrace{00\cdots0}_{n} \times 2^{L} = 2^{L} \xrightarrow{\text{depends on}} \text{exponent}$$

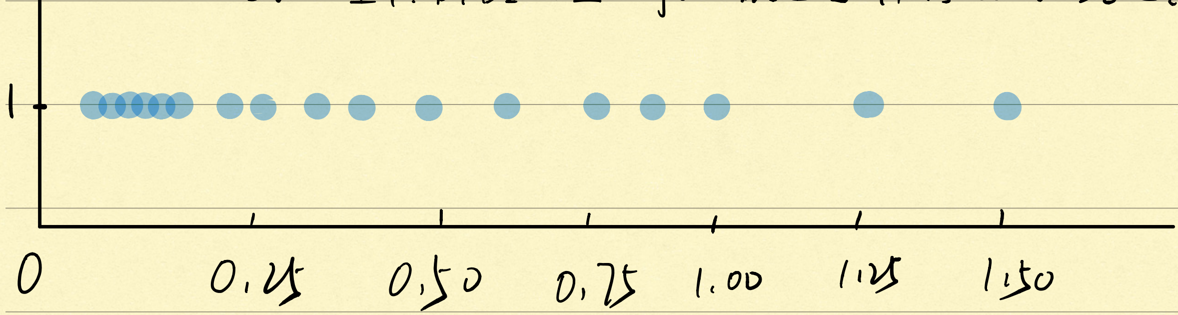
Largest positive normalized FP number:

$$1.\underbrace{111\cdots1}_{n} \times 2^{U} = 2^{U+1}(1-2^{-P}) \quad \underline{\text{depends on}} \quad \text{exponent, precision.}$$



Machine Epsilon: has relative precision. is defined as the distance (gap) between 1 and the next larger floating point number. $\epsilon_m = 1.25 - 1 = 0.25$

$x = \pm 1.b_1b_2 \times 2^{m}$ for $m \in [-4, 4]$ and $b_i \in \{0, 1\}$.



$\longrightarrow$ harder to be represented by x.

$(1.00)_2 = 1.00 \times 2^0 = (1)_{10}$

$(10)_2 = 1.00 \times 2^1 = (2)_{10}$

$(11)_2 = 1.1 \times 2^1 = (3)_{10}$

$\cdots\cdots$

$$\varepsilon_m = 1.\underbrace{00\cdots1}_{n} \times 2^0 - 1.\underbrace{00\cdots0}_{n} \times 2^0$$

$$= 0.\underbrace{00\cdots1}_{n} \times 2^0 = 2^{-n}\cdot 2^0 = 2^{-n}.$$

Subnormal floating point number:
(denormal)

bigger than 0, smaller than 1×2^{min}.

$0.1 \times 2^{min} = 2^{min-1}$

0.01×2^{min} ———

2^{min-n} (smallest).