$m = c - \text{shift}.$

Numerical Form: $x = (-1)^s \cdot 1.f \times 2^m$

Representation in memory: $x = \underbrace{\pm}_{\text{sign}} \underbrace{m}_{\text{exponent}} \underbrace{f}_{\text{significand}}$

$\underset{\text{unsigned int}}{c} = \underset{\text{signed int.}}{m} + \text{shift}$

IEEE-754 Single precision (32 bits):

$s = \{0, 1\}$

$x = \underbrace{s}_{1 \text{ bit}} \ \underbrace{c}_{8 \text{ bits}} \ \underbrace{f}_{23 \text{ bits}} \Rightarrow p = 24.$

$(0\,0\,0\,0\,0\,0\,0\,0)_2 = (0)_{10}$ $\qquad c \in [0, 255].$

$(1\,1\,1\,1\,1\,1\,1\,1)_2 = (255)_{10}$

reserve 0, 255 for special cases.

$\Rightarrow c \in [1, 254]. \Rightarrow 1 \leq m + \text{shift} \leq 254.$

~~set~~ shift $= 127 \longrightarrow -126 \leq m \leq 127.$

$m \in [-126, 127].$

Example: $x = -67.125$

$(67.125)_{10} = (1000011.001)_2 = (1.000011001)_2 \times 2^6$

$s = 0, \ f = \underbrace{0000110010000 \cdots 0}_{23 \text{ bits}}$

$m = 6, \ c = m + \text{shift} = 6 + 127 = (133)_{10}$

$$c = \underbrace{(10000101)_2}_{8 \text{ bits}}$$

$$\underbrace{0}_{1 \text{ bit}} \; \underbrace{10000101}_{8 \text{ bits}} \; \underbrace{00001100100000\cdots0}_{23 \text{ bits}}$$

Machine Epsilon ($\epsilon_m$): $\epsilon_m = 2^{-23}$

UFL

Smallest positive normalized FP number: $2^{-126} \approx 10^{-38}$

Largest positive normalized FP number: $2^{127+1} \cdot (1-2^{-24})$

OFL

$$= 2^{128} - 2^{104}$$

$$\approx 10^{38}$$

IEEE-754 Double precision (64 bits):

$x = \underbrace{\pm}_{1 \text{ bit}} \; \underbrace{c}_{11 \text{ bits}} \; \underbrace{f}_{52 \text{ bits}} \Rightarrow p = 53.$

Set shift $= 1023 \Rightarrow c = m + 1023.$

$(00000000000)_2 = (0)_{10}$

$(11111111111)_2 = (2047)_{10}$

$c \in [1, 2046].$

$\Rightarrow m \in [-1022, 1023].$

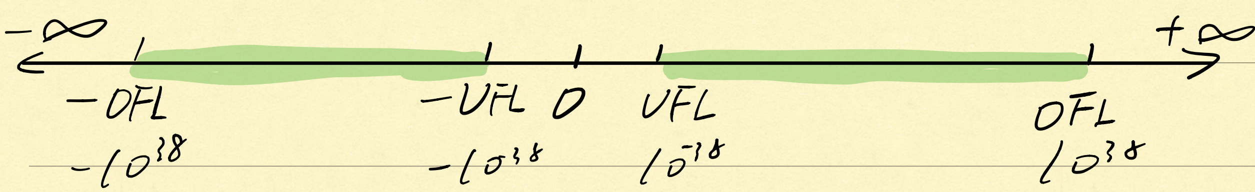
$\epsilon_m = 2^{-52} \approx 2.2\times10^{-16}$

$UFL = 2^L = 2^{-1022} \approx 2.2\times10^{-308}$

$OFL = 2^{1024}(1-2^{-53}) \approx 1.8\times10^{308}$

## Special Cases

Single precision




① represent $\underline{0}$ : $x = s\ \underbrace{00\cdots000}_{8/11\text{ bits}}\ \underbrace{000\cdots0000}_{23/52\text{ bits.}}$

② represent $\underline{+\infty}$ $(s=0)$ and $\underline{-\infty}$ $(s=1)$

$x = s\ \underbrace{111\cdots111}_{8/11\text{ bits}}\ \underbrace{000\cdots0000}_{23/52\text{ bits.}}$

③ represent $\underline{NaN}$

$x = s\ \underbrace{111\cdots111}_{8/11\text{ bits}}\ \underbrace{\text{anything} \neq 00\ldots00}_{23/52\text{ bits.}}$

④ represent subnormal numbers

$$x = s \underbrace{000\cdots000}_{8/11 \text{ bits}} \underbrace{\text{anything} \neq 00\cdots00}_{23/52 \text{ bits}}$$

Smaller than UFL.

$0.f \times 2^L$

## Subnormal (or denormalized) numbers.

$$x = (-1)^s \times 0.f \times 2^L.$$

IEEE-754 Single precision (32-bits)

$c = (00000000)_2 = (0)_{10}$

$m = -126$

Smallest positive subnormal FP number:

$$0.000\cdots1 \times 2^{-126} = 2^{-23} \times 2^{-126} = 2^{-149} \approx 1.4 \times 10^{-45}$$

IEEE-754 double precision (64-bits)

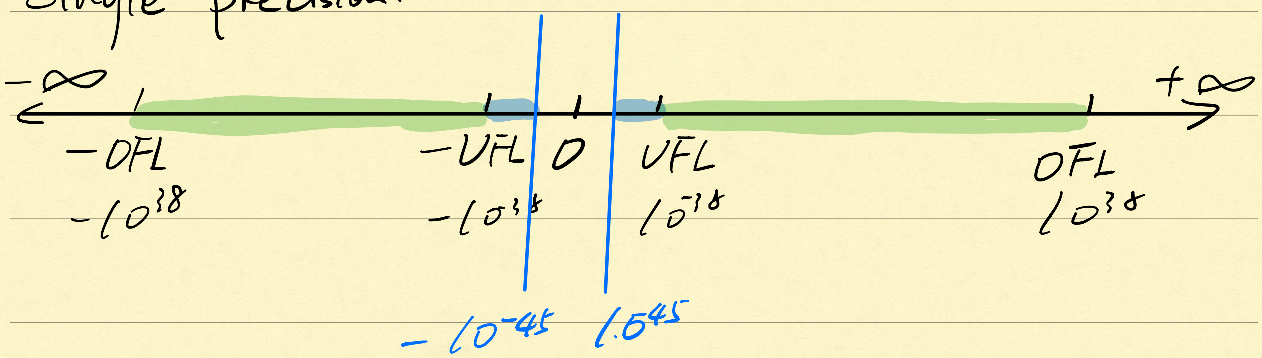
$c = (00000000000)_2 = (0)_{10}$

$m = -1022$

Smallest positive subnormal FP number:

$$2^{-52} \times 2^{-1022} = 2^{-1074} \approx 10^{-324}$$

Single precision.




$$\underbrace{1.000\cdots0}_{24} \times 2^{-126} \qquad p = 24$$

$$0.\underbrace{111\cdots1}_{23} \times 2^{-126} \qquad p = 23$$

$$0.00\cdots0\underbrace{1010}_{4} \times 2^{-126} \qquad p = 4$$

Overflow: $x: |x| > OFL \to x = \pm OFL$

Gradual Underflow: $x: |x| < UFL \to x = 0, \pm UFL$

subnormal number