Normalized floating-point numbers

$$x = \pm q \times 2^{m} = \pm 1.b_1 b_2 b_3 \dots b_n \times 2^{m} = \pm 1.f \times 2^{m}$$

- **Exponent range**: $[L, U]$
- **Precision**: $\mathrm{p} = n + 1$
- **Smallest positive normalized FP number:**

$$\textbf{UFL} = 2^{L}$$

- **Largest positive normalized FP number:**

$$\textbf{OFL} = 2^{U+1}(1 - 2^{-p})$$

Floating-point number representation

Numerical form:

$$x = \pm 1.f \times 2^{m}$$

Representation in memory:

$$x = \underbrace{\boxed{\;s\;}}_{\text{sign}}\underbrace{\boxed{\qquad c \qquad}}_{\text{exponent}}\underbrace{\boxed{\qquad\qquad f \qquad\qquad}}_{\text{significand}}$$

$$x = (-1)^{s}\, 1.f \times 2^{c-shift} \qquad m = c - shift$$

Single Precision

$c \Rightarrow 8$ bits $\to (000\ldots00)_2 \to (0)_{10}$; $(111\ldots11)_2 \to (255)_{10}$

$1 \leq c \leq 254$

$m = c - \text{shift}$

$\text{shift} = 127$

$$\boxed{-126 \leq m \leq 127}$$

L U

$$\boxed{\epsilon_m = 2^{-23}}$$

$f \Rightarrow 23$ bits

$p = 24$

UFL : $2^{-126} \approx 10^{-38}$

OFL : $2^{128}(1 - 2^{-24}) \approx 10^{38}$

Double Precision

$c = 11$ bits $\to (000\ldots00)_2 \to (0)_{10}$; $(111\ldots11)_2 \to (2047)_{10}$

$1 \leq c \leq 2046$

$\text{shift} = 1023$

$$\boxed{-1022 \leq m \leq 1023}$$

L U

$f = 52$ bits

$p = n+1 = 53$

UFL : $2^{-1022} \approx 10^{-308}$

OFL : $2^{1024}(1 - 2^{-53}) \approx 10^{308}$

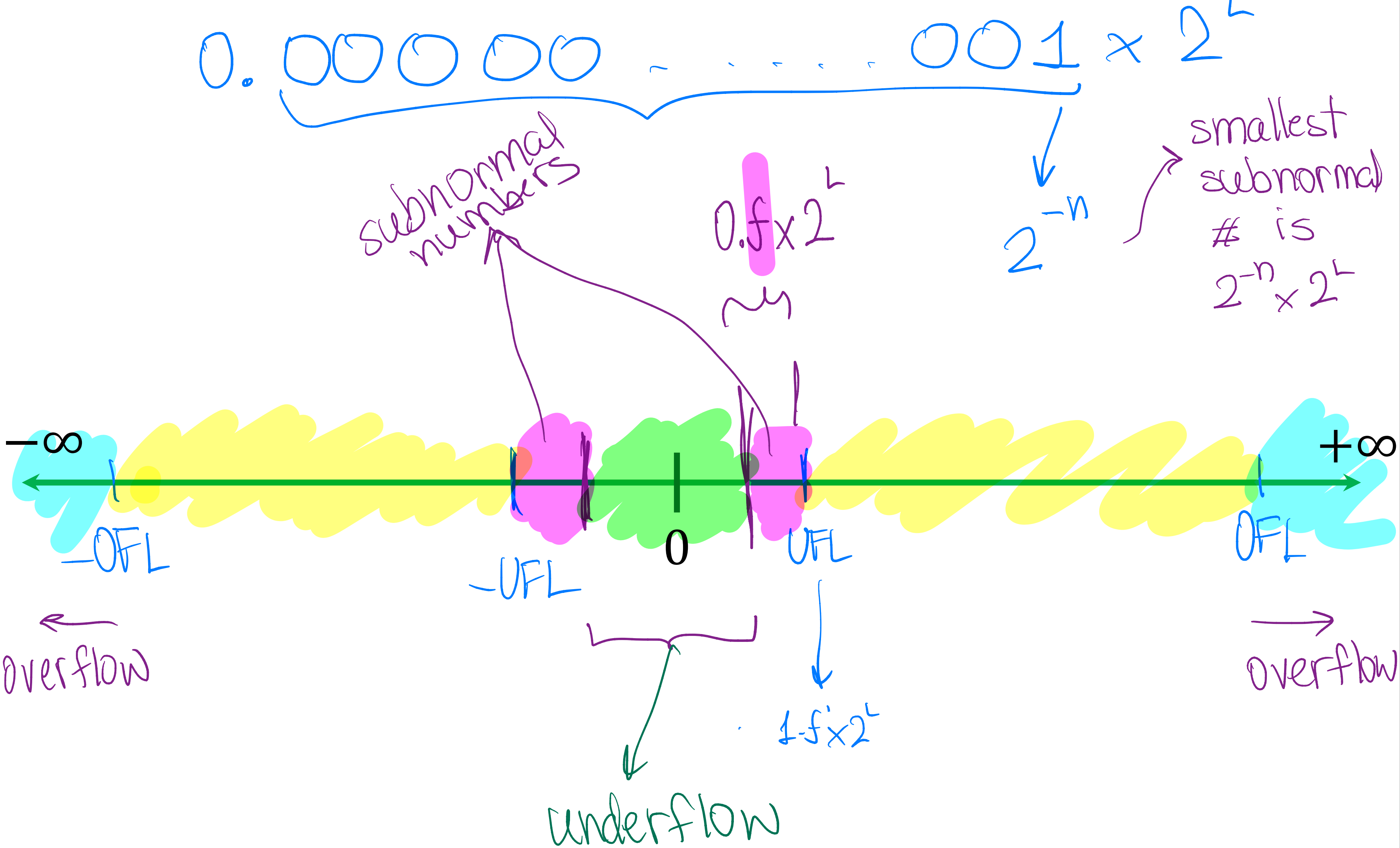
$$\boxed{\epsilon_m = 2^{-52}}$$

Special Cases

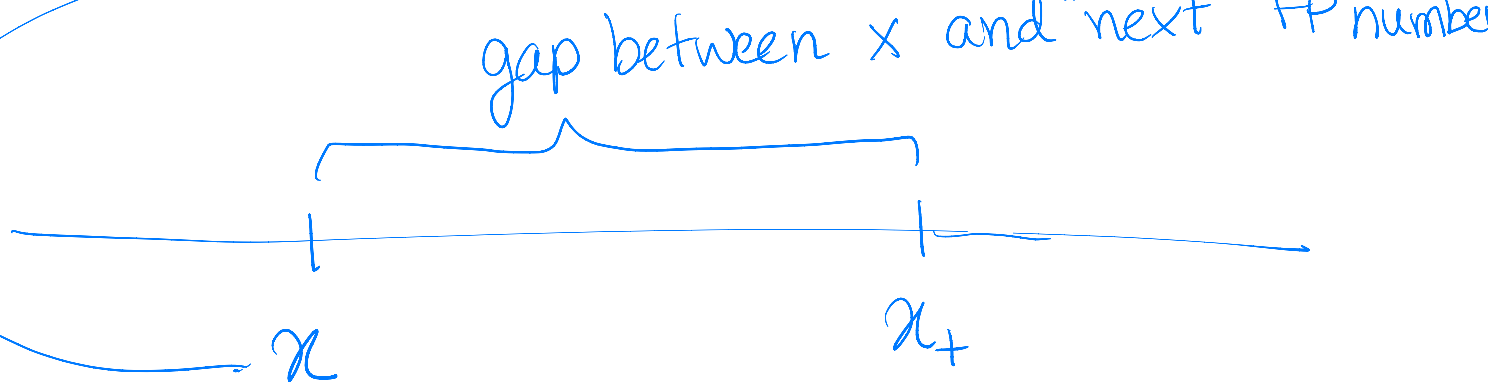
$$x = \pm 1.f \times 2^m$$

c	f	
All zeros	All zeros	± 0
All zeros	$f \neq 0$	Subnormal $x = \pm 0.f \times 2^L$
All ones	All zeros	$\pm \infty$
All ones	$f \neq 0$	NaN

Normalized floating point number scale



$$|x| = 1.f \times 2^m$$



$$\text{gap} = |x - x_+| = \epsilon_m \times 2^m = e_a$$

Example: $2^4 \times 2^{-52} = 2^{-48}$ $(x = 2^4)$