$(19.625)_{10} = (10011.101)_2 = (1.0011101)_2 \times 2^4$

Machine floating point number

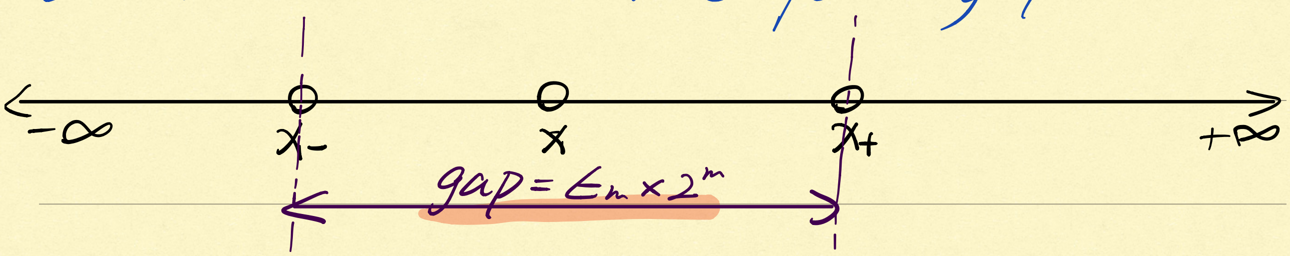
Not all real number can be exactly represented as a machine floating-point number

$$x = \pm 1.\, b_1 b_2 \cdots b_n \underbrace{\cdots}_{b_{n+1} b_{n+2} \cdots} \times 2^m$$

can't represented.

x will be approximated by either x_- or x_+.

the nearest two machine floating point numbers



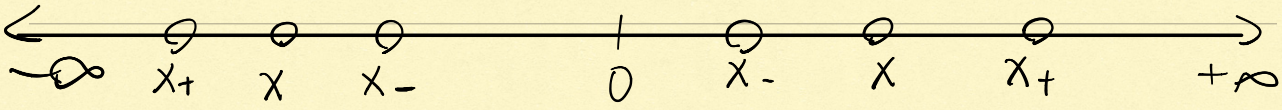
$$x_- = 1.\, b_1 b_2 b_3 \cdots\cdots b_n \times 2^m$$

$$x_+ = x_- + \underbrace{0.00 \cdots 01 \times 2^m}_{2^{-n} \times 2^m \;=\; \epsilon_m \times 2^m}$$

$$|(x_+) - (x_-)| = \epsilon_m \times 2^m$$

Rounding

The process of replacing x by a nearby machine number is called rounding, and the error involved is called roundoff error.



round towards $-\infty$	round towards 0	round towards 0	round towards $+\infty$
round down (floor)	round up (ceil)	round down (floor)	round up (ceil)

①. Round by chopping ↑ $fl(x) = x_+ / x_-$

②. Round to nearst

Rounding errors (roundoff)

Absolute error: $|fl(x) - x| \le \epsilon_m \times 2^m$

Relative error: $\dfrac{|fl(x) - x|}{|x|} \le \dfrac{\epsilon_m \times 2^m}{1.b_1b_2\cdots b_n \times 2^n}$

$= \dfrac{\epsilon_m}{1.b_1b_2\cdots b_n} \le \epsilon_m$

$\Rightarrow \dfrac{|fl(x) - x|}{|x|} \le \epsilon_m.$

IEEE Single precision $\frac{|fl(x)-x|}{|x|} \leq 2^{-23} \approx 1.2 \times 10^{-7}$

10^{-7}

double precision $\frac{|fl(x)-x|}{|x|} \leq 2^{-52} \approx 2.2 \times 10^{-16}$

10^{-16}

Gap: $\delta < \underline{gap}$, $fl(x+\delta) = fl(x)$

$\epsilon_m \times 2^m$

Arithmetic

Possible: $(x+y)+z \neq x+(y+z)$

$z(x+y) \neq zx + zy$

$x + \underset{\text{really small}}{y} = x$

Arithmetic:

①. compute the exact result.

②. round.

$x+y = fl(x+y)$

$x \times y = fl(x \times y)$

Addition and subtraction.

① bring both numbers onto a common exponent.

② Do "grade-school" operation.

③ Round result.

Example: $x = \pm 1.b_1 b_2 b_3 b_4 \times 2^m$

$$a = (1.1011)_2 \times 2^1$$

$$b = (1.1010)_2 \times 2^1$$

$$c = a - b = (0.0001)_2 \times 2^1$$

$$c = (1.????)_2 \times 2^{-3}$$

Lose of precision

machine fills them with its best guess, (not good).

Catastrophic Cancellation

Example: ① 16 decimal digits of accuracy:

$$fl(a) = 3004.45$$

$$fl(b) = 3004.46.$$

$$b-a=\underbrace{0000.\,0}_{\text{lose 5 digits (significands)}}\underbrace{1\text{-------}}_{\text{Still has 11 digits left.}}$$

② 5 decimal accurate digits.

$$f(x)=\sqrt{x^2+1}-1$$

$$f(10^{-3})=\sqrt{10^{-6}+1}-1$$

$$fl(10^{-6}+1)=1 \Rightarrow f(10^{-3})=0.$$

③ $f(x)=\sqrt{x^2+1}-1$

rewrite $\Rightarrow$ $f(x)=\dfrac{x^2}{\sqrt{x^2+1}+1} \Rightarrow f(10^{-3})=\dfrac{10^{-6}}{2}$