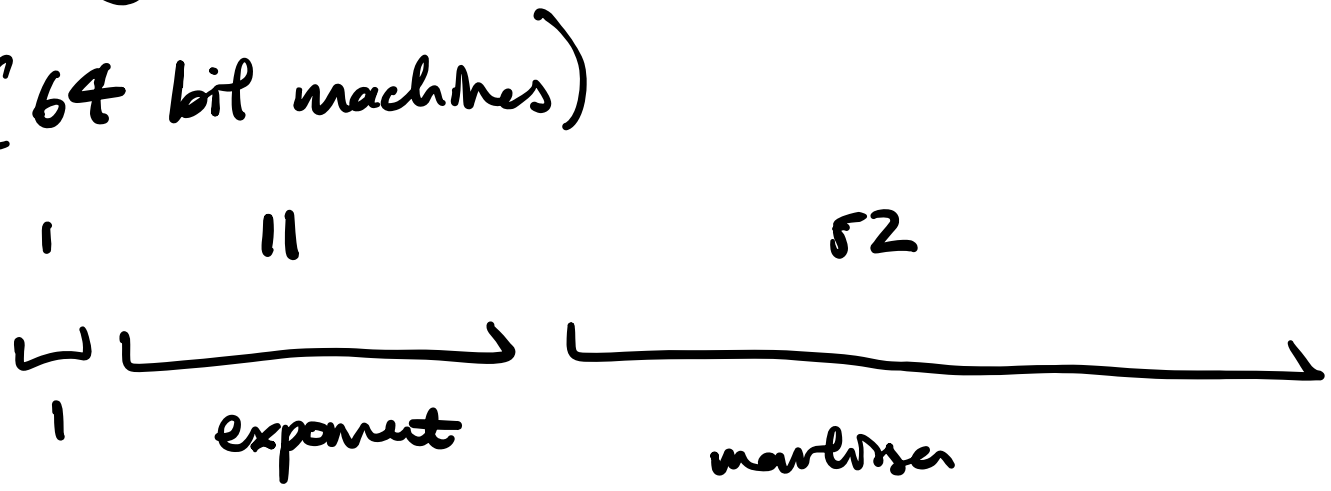
Floating Point Numbers

(64 bit machines)



$$\Rightarrow (-1)^{sign} \times 1.\underbrace{\ldots\ldots\ldots}_{\text{mantissa}} \times 2^{\text{exponent}-1023}$$

---

example 1.75 as Float 64

sign = 0

mantissa = $0.75 = 2^{-1} + 2^{-2}$

$\Rightarrow$ 0.11 (in binary)

exponent = 1023 = 01111111111 (in binary)

$\Rightarrow$ $1.75 = 0\ 01111111111\ 1100\ldots0$ (52 zeros)

as a Float 64

Def. Machine Precision ε

$\rightarrow$ The smallest representable floating point number s.t.

$$1+\varepsilon \neq 1$$

In FP: $1 = 0\ 01111111111\ 00\ldots00$

$+1$ ↓

$1+\varepsilon = 0\ 01111111111\ 00\ldots01$

$\Rightarrow$ $\varepsilon = 2^{-52} \approx 10^{-16}$

FP Operations

$$fl(x \; op \; y) = (x \; op \; y)(1+\delta)$$

$op \in \{+, \times, -, \div\}$

$|\delta| < \varepsilon$

eg: Nested operations

$\otimes$ is fp multiplication

$$(a_1 \otimes a_2) \oplus a_3$$

$$= (a_1 \times a_2)(1+\delta_1) \oplus a_3 \quad , |\delta_1| < \varepsilon$$

$$= \left[(a_1 \times a_2)(1+\delta_1) + a_3 \right](1+\delta_2), |\delta_2| < \varepsilon$$

$$= (a_1 \times a_2)(1+\delta_1)(1+\delta_2) + a_3(1+\delta_2)$$

$$= (a_1 \times a_2)(1+\delta_1+\delta_2+\delta_1\delta_2) + a_3(1+\delta_2)$$

$\sim$

$\sim \varepsilon^2$, ignore

$= \delta_3$, $|\delta_3| < 2\varepsilon$

$$= (a_1 \times a_2)(1+\delta_3) + a_3(1+\delta_2)$$

- Two operations, get error of $|\delta_3| < 2\varepsilon$, makes sense

Big O Notation and Order of Convergence

Def. If $f(h) \to q$ as $h \to 0$, and

$$|f(h) - q| \leq K h^p \text{ for some } K \neq 0,$$

then $f(h) \to q$ with order of convergence p.

(Equivalently, converges as $O(h^p)$)

eg. $\cos(2h) + 2h \sin(h)$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

$$\Rightarrow \cos(2h) = 1 - \frac{(2h)^2}{2!} + \frac{(2h)^4}{4!} - \cdots$$

$$2h\sin(h) = 2h\left(h - \frac{h^3}{3!} + \frac{h^5}{5!} - \cdots\right)$$

$$\Rightarrow \cos(2h) + 2h\sin(h)$$

$$= 1 - \frac{(2h)^2}{2!} + 2h^2$$

$$+ \frac{(2h)^4}{4!} - \frac{2h^4}{3!} + \cdots$$

$$= 1 + \left(\frac{2^4}{4!} - \frac{2}{3!}\right)h^4 + \cdots$$

$$\Rightarrow \cos(2h) + 2h\sin(h) \longrightarrow 1$$

with rate $O(h^4)$