Machine precision is the distance between 1 and the next fp number. In this case,

$1 = (0 \mid E=0 \mid 1.0\ldots0)$ (23 bits)

$1 + 2^{-23} = (0 \mid E=0 \mid 1.0\ldots01)$ $\leftarrow 2^{-23}$

$\hookrightarrow \approx 1.2\times10^{-7}$

Usually calculations are done in double precision:

sign: 1 bit

exponent: 11

mantissa: 52

machine precision: $eps = 2^{-52} \approx 2.2\times10^{-16}$

How to calculate machine precision? Demo

Floating point representations were standardized by the IEEE in 1980s

↳ Inst. of Electrical & Electronics Engin.

Consistent Rules

- Representation (# bits)
- Correct / consistent rounding
- sensible treatment of exceptions: $\frac{1}{0}$

Hidden bit rep.: 0, NaN, $\pm\infty$ (over/underflow)

single precision: 32 bits ($\epsilon \sim 10^{-7}$)

double precision: 64 bits ($\epsilon \sim 10^{16}$)

extended precision: 80 bits ($\epsilon \sim 10^{-20}$) (no hidden bits)

Rounding

Example: Decimal notation: $\frac{1}{10} = 0.1$

Binary notation: $\frac{1}{10} = (1.100\overline{1100})_2 \times 2^{-4}$

How is this rounded?

$\frac{1}{3} - \frac{1}{4} = \frac{1}{12} * \frac{1}{16}$

$\frac{1}{12} - \frac{1}{16} = \frac{4}{48} - \frac{3}{48}$

$= \frac{1}{48} - \frac{1}{64}$

4 options: Up

Down

to 0

nearest.

$1/10 = 1.100\overline{1100} \times 2^{-4}$

$= 1.100\ 1100\ 1100\ 1100 \cdots 1100\ 1$ (down, toward 0)

$= 1.100\ 1100\ 1100\ 1100 \cdots 1101\ 0$ (up, (nearest))

↓ default

Absolute rounding error = $|\text{round}(x) - x|$

Relative rounding error = $\dfrac{|\text{round}(x) - x|}{x}$

Most important rule; IEEE says $\text{Round}(a+b) = a \oplus b$

↑ exact addition; ↑ computer addition

$= (a+b)(1+\delta)$

↑ $\leq \epsilon$ machine precision

- DEMO to calculate ϵ_{MACH} in MATLAB.

- IEEE rules make sure that arithmetic is done to precision <u>at least</u> that of machine precision:

<u>Ex:</u> $a \oplus b = \text{round}(a+b)$

$= (a+b)(1+\delta)$

$\oplus$: addition in floating point

$(a+b)$: true exact result

$(1+\delta)$: $|\delta| < \epsilon$.

This is a statement of <u>relative accuracy</u>:

$$\Rightarrow \left| \frac{\text{round}(a+b) - (a+b)}{(a+b)} \right| = |\delta| < \epsilon$$

<u>Ex:</u> $|\text{round}(a+b) - (a+b)| = |a+b||\delta|$

$\leq (|a|+|b|)\epsilon$

$\leq 2 \cdot \max(|a|, |b|)\epsilon$.

<u>absolute accuracy</u> $\longrightarrow$