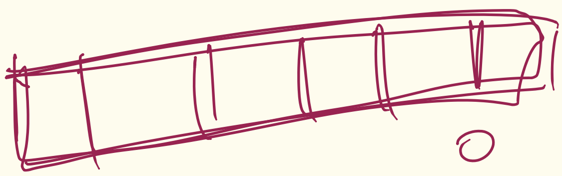
IEEE: 64 bits

1 sign bit;

52 bits for mantissa

11 bits for exp

$$\pm\left(0.1\,d_1 d_2 \dots d_{52}\right)_2 \times 2^{e}$$

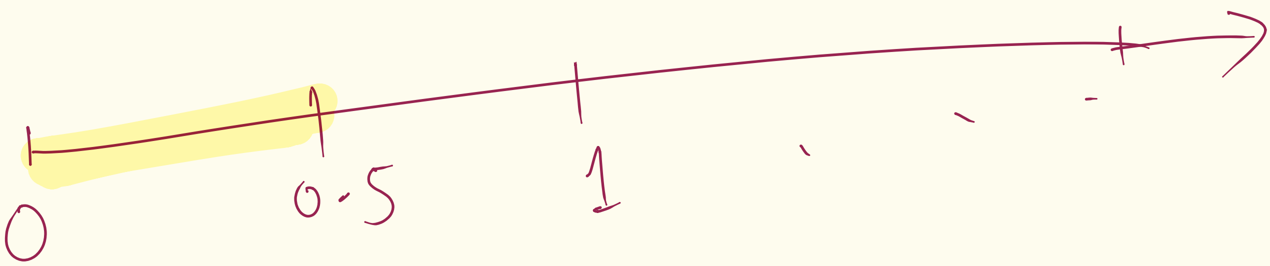


$0 \; . \; . \; . \; 0 = 0$

$1\,1 \; . \; . \; . \; 1 = 2047$

smallest positive: $(0.100\dots0) \times 2^{0}$

$= 0.5$



exp range $[0,\ 2047]$

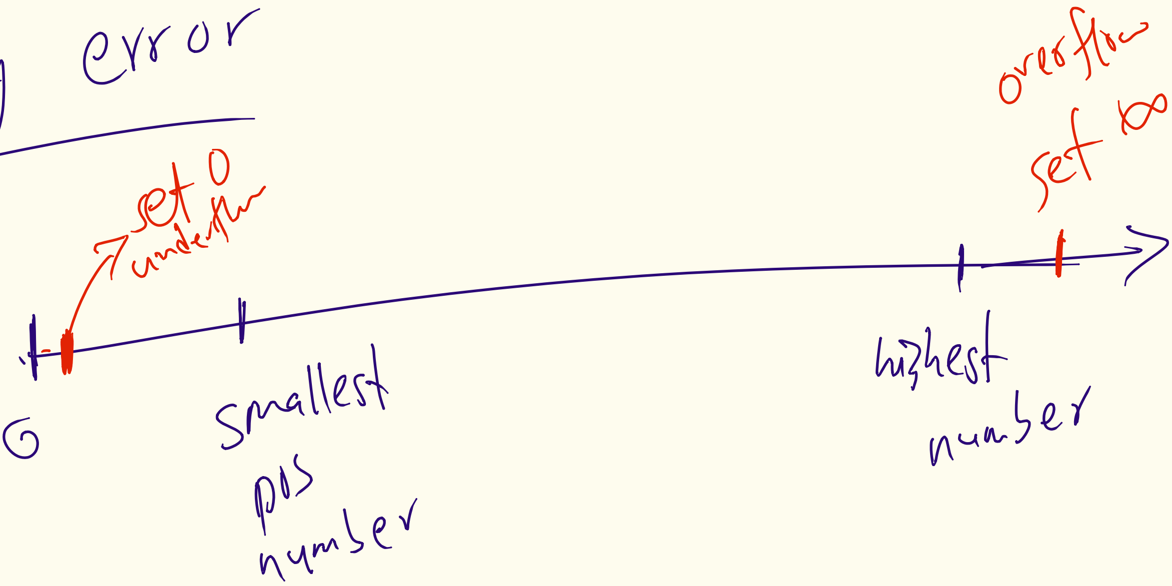
bias $-1022 \rightarrow [-1022,\ +1025]$

$0 \qquad \infty$

min positive value:

$$(0.1\,0 \dots 0) \times 2^{-1021}$$

Rounding error



$$x = \underbrace{3.14159265\,35}_{\text{round to 4 digits}}$$

$$\Downarrow$$

$$3.142 = fl(x)$$

$$|fl(x) - x| = \text{absolute error}$$

$$\text{relative rounding error} = \frac{|fl(x) - x|}{|x|} = \delta$$

example:

x = distance,

$fl(x) - x$ distance

$$\left|\frac{fl(x)-x}{x}\right|.$$ unitless

$$\boxed{|\delta| < 10^{-10}}$$

error is less than 10^{10}

$$\left|\frac{fl(x)-x}{x}\right| < 10^{-10}$$

$$\frac{0.1+0.2}{}\qquad 0.3$$

$(0.1+0.2 - 0.3)$

$= \ldots\ 10^{-17}$

$= 5.55111\ldots e{-17}$

$= 5.55 \times 10^{-17}$

Bounding the relative rounding error:

$$x = (0.d_1 d_2 d_3 \ldots d_m d_{m+1}) \times \beta^e$$

exact

round

$$fl(x) \doteq (0.d_1 d_2 \ldots d_m) \times \beta^e$$

$$x = (0.3141\underline{1}59265) \times 10^1$$

$\downarrow$ round to 4

$$(0.314\underline{2}) \times 10^1$$

$$fl(x) - x = \left((0.d_1 d_2 \ldots d_m) - (0.d_1 d_2 \ldots \ldots)\right) \times \beta^e$$

$$= (0.\underbrace{000\ldots0}_{m \text{ digits}}\boxed{1}) \times \beta^e$$

$$\leq \left(0.000\ldots0\tfrac{\beta}{2}\right)_\beta \times \beta^e$$

$$= \frac{\beta}{2} \times \beta^{-(m+1)} \times \beta^e$$

$$\eta = \frac{1}{2} \times \beta^e \times \beta^{-m}$$

$x = 0.31415$

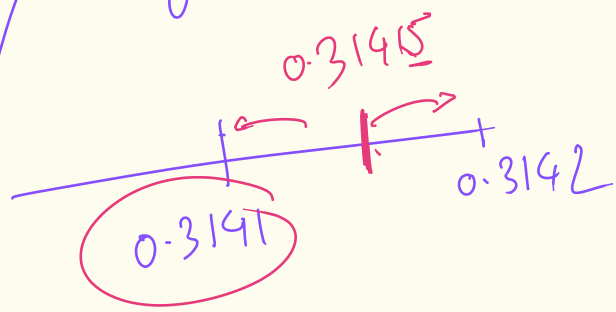
round to 4 digits

0.3142

$\Rightarrow$ error $= 0.0000\underbrace{5}$

always ≤ 5

can this be larger than 5?



base 10 $\longrightarrow$ next digit ≤ 5

" β $\longrightarrow$ ... $\leq \frac{\beta}{2}$

So far,

$$|fl(x) - x| \leq \frac{1}{2}\beta^{e}\cdot\beta^{-m} \quad \text{——— (I)}$$

divide by $|x|$.

$$x = (0.d_1 d_2 d_3 \ \ldots\ ) \times \beta^{e}$$

lowest $x = (0.1\ 0\ 0\ 0\ 0) \times \beta^{e}$

$$= 1 \times \beta^{-1} \times \beta^{e}$$

$$x \geq \beta^{-1} \times \beta^{e} \quad \text{——— (II)}$$

$$\left|\frac{fl(x)-x}{x}\right| \le \frac{\frac{1}{2}\beta^{e}\beta^{-m}}{\beta^{-1}\beta^{e}}$$

$$= \frac{1}{2}\beta^{1-m}$$

maximum rounding error

Machine epsilon $\varepsilon_M = \frac{1}{2}\beta^{1-m}$

IEEE: $\beta = 2,\ m = 52$

$$\varepsilon_M = \frac{1}{2}\cdot 2^{1-52} = 2^{-52}$$

$$= \boxed{2.22\times 10^{-16}}$$

$$\begin{tikzcd}
x \arrow[r] & x(1+\delta_1) \\
y \arrow[r] & y(1+\delta_2)
\end{tikzcd}$$

real

computer

$$\frac{fl(x)-x}{x} = \delta$$

$$fl(x) - x = \delta x$$

$$fl(x) = x + \delta \cancel{x}\, x$$

$$= x(1+\delta)$$

$$x(1+\delta_1) + y(1+\delta_2)$$

$$= x + y + x\delta_1 + y\delta_2$$

relative error: $\dfrac{(x+y+x\delta_1+y\delta_2)-(x+y)}{(x+y)}$

$$= \frac{x\delta_1 + y\delta_2}{x+y}$$

subtraction: $\boxed{\dfrac{x\delta_1 - y\delta_2}{x-y}}$

0.00001
0.00002

loss of significance:
when the relative error $\dfrac{x\delta_1 \pm y\delta_2}{x \pm y}$
becomes larger than ε_M.

$$x^2 - 56x + 1 = 0$$

$$x = 28 \pm \sqrt{783}$$

$28 + \sqrt{783}$

$= 55.9821$

$= 55.98 = 0.5598 \times 10^2$

$28 - \sqrt{783}$

$= 0.0178628$

$= 0.01786$

$= 0.1786 \times 10^{-1}$

$28 + \sqrt{783}$

27.9821

27.98

$28 + 27.98$

$= 55.98$ ✓

~~$28 - \sqrt{783}$~~

~~$= 28 - 27.98$~~

~~$= 0.02$~~

$x_1 \times x_2 = 1$ ✓

$x_2 = \dfrac{1}{x_1}$

$= \dfrac{1}{55.98}$

$= 0.01786$

---

2 digit

$5.9 + 5.5 + 0.4$

$(5.9 + 5.5) + 0.4$

$= 11 + 0.4$

$= 11$

$5.9 + (5.5 + 0.4)$

$= 5.9 + 5.9$

$= 12$

5.01, 5.02 (3 digit)

avg: $\dfrac{5.01 + 5.02}{2} = \dfrac{10.0}{2} = 5.0$

Assignment 1:

Deadline: 19 February

Quiz 1: 19 Febrary

in-person

na17.pdf chapter 1,

= floating point arithmetic note [ARO]

$x = \cancel{3}\ 0.31415$

$\downarrow$

$\bar{x} = 0.3142$

$\bar{x} - x$

$= (0.3142) - (0.31415)$

$= 0.00005$

error $= 0.00006$

$x = 0.31414$

$\downarrow$

round 0.3141

base 10 $\longrightarrow$ 5

" β $\longrightarrow$ $\beta/2$