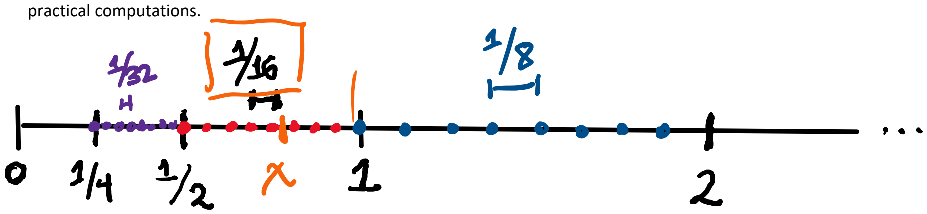
APPM 4600: Floating Point Arithmetic Notes

Wednesday, August 28, 2024 10:33 AM

Class 02: Wed, August 28

Recall: After introducing the course, we described the differences between math with real numbers on pen and paper and in the computer. First big thing to consider: we have limited memory budget per real number. Once we have assigned a number of "bits" (zeroes and ones) for our budget, we discussed how to best build a system to do practical computations.



~~64 bits memory.~~

~~$x = \pm$ ____ . ____~~

~~not practical!~~

$$x = \pm (0.\overset{1}{d_1} d_2 d_3 \overset{4}{d_4})_2 \times 2^e \qquad e \in \{\cancel{-3}, -2, -1, 0, 1, 2, 3, \cancel{4}\}$$

d_1 has to be 1

$$e = \underset{3}{(e_1 e_2 e_3)_2} - \underset{\text{bias}}{3}$$

x true value

fl(x) → closest rounding

- MAX NUMBER?
- MIN NUMBER?
- # of floats for e = 0 :
- TOTAL # of floats :

DEFINITIONS / NOTATION:

fl(x) → rounding x to nearest fl pt #.

abs error → $|x - fl(x)| < 1/32$

rel error → $\dfrac{|x - fl(x)|}{|x|} < \dfrac{1/32}{1/2} = \boxed{1/16}$

Uniform bound for rel error

→ Spacing between # in $[1/2, 1)$ $(e=0)$

"Machine epsilon"

DOUBLE PRECISION (IEEE 754)

DOUBLE PRECISION (IEEE 754)

↳ 64 bits: 1 sign. 52 mantissa. 11 exponent.

$$x = \pm (0.1\, d_2 d_3 \cdots d_{52})_2 \times 2^e$$

$$e = \{0, 1, \ldots, 2047\} - 1023$$

$$= \{\cancel{-1023}, -1022, \ldots, 1023, \cancel{1024}\}$$

0 $\qquad$ $\pm\infty$, NaN

$x_{MAX} \approx 10^{308}$

$x_{MIN} \approx 10^{-308}$

$\varepsilon_{MACH} = 2^{-52} \approx 2 \times 10^{-16}$

$\log_{10}(\varepsilon_{MACH}) \approx -16$

- DOUBLE ⟷ "16 decimal digits (of rel. accuracy".

- 32 bit – "SINGLE PRECISION"

  $\varepsilon_{MACH} \approx 10^{-8}$, 8 decimal digits.

ARITHMETIC OPERATIONS

$x + y$ $\quad$ INPUTS

$x - y$ $\quad$ $fl(x), fl(y)$

$x - y$

xy

x/y

$fl(x), fl(y)$

$\boxed{+}$ — $\underbrace{fl(fl(x) + fl(y))}$

REL ERROR INPUTS ($\leq \varepsilon_{MACH}$)

REL ERROR OUTPUTS?

LOSS OF PRECISION

x, y have the same sign.

$x + y$ ✓

xy ✓*

x/y ✓*

$x - y$ ✗

same sign

same magnitude.

$x = 0.52345$

$y = 0.52343$

5 digits rel acc.

$x - y = 0.00002$

$= 0.2 \times 10^{-4}$