

Sep 14, 2020

Putting ^{real} numbers on computers \Rightarrow Floating point
Approximating

Scientific notation: \pm (5) (282) \cdot (10)⁴ exponent
base $\pm 0.05282 \cdot 10^6$

FP is like scientific notation in binary

$$1.011 \cdot 2^{-3} = (1 + \frac{1}{4} + \frac{1}{8}) \cdot 2^{-3} = 0.171875$$

Normalized FP assumes leading 1

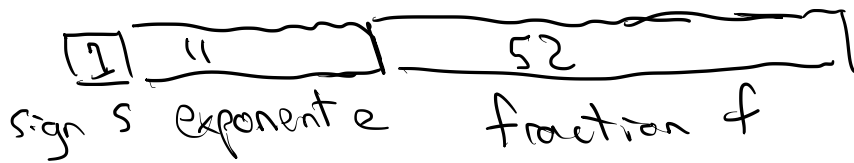
$$\pm (1.b_1b_2 \dots b_f) \cdot 2^e$$

significant

$$2^n - 1 = 2047$$

IEEE FP

64-bit



$$(-1)^s \cdot (1+f) \cdot 2^{e-1023}$$

32-bit



$$(-1)^s \cdot (1+f) \cdot 2^{e-127}$$

Bfloat 16



NVIDIA tensor float



Problem 2: representing numbers (64 bits)

$f(x) \hat{=}$ closest floating point rep. of $x \in \mathbb{R}$

x too large? $(1 + 1 - 2^{-52}) \cdot 2^{2^{11}-1-1023} \approx 2^{1024} \approx 10^{308}$

x too small? $(1 + 0) \cdot 2^{0-1023} = 2^{-1023} \approx 10^{-308}$

too large \Rightarrow Inf

too small \Rightarrow subnormal $(0.b_1 \dots b_f) \cdot 2^{-1023}$

x inexact $1/3, 1/10, x$ nonsensical $0.0 / 0.0$

overflow, underflow, invalid, divide by 0

\Rightarrow all exceptions

At a given exponent, 52 bits fraction
more space b/w larger numbers
less smaller
relative gaps the same

relative representation error: $|x - f(x)| / |x|$

$\epsilon =$ "machine epsilon/precision" = ϵ_{mach}

$=$ max rel. rep. error

$= \frac{1}{2} \text{dist}(1.0, \text{nextfloat}(1.0))$

64-bits
IEEE FP

$$= \frac{1}{2} |1 + 2^{-52} - 1| = 2^{-53} \approx 10^{-16}$$

Cor: $f(x) = x(1 + \delta)$ for $|\delta| \leq \epsilon$

(Julia, MATLAB: $\text{eps}(1) \Rightarrow 2^{-52}$)

Problem 2: roundoff error

$$f(x + y) = \text{closest FP rep. of } x + y$$

\uparrow \uparrow
 $x = f(x)$ $y = f(y)$

Model: for $\odot \in \{+, -, *, /\}$

$$f(x \odot y) = (x \odot y)(1 + \delta), \quad |\delta| \leq \epsilon$$

$$\frac{|f(x \odot y) - x \odot y|}{|x \odot y|} = |\delta| \leq \epsilon$$

(ignoring under/over flow)

" $1 + \delta$ " model

Example: error analysis

Math: $s = \sum_{i=1}^n x_i$



```
s = 0
for i = 1:n
    s = s + x_i
return s
```

$s_k = \sum_{i=1}^k x_i$

$s_1 = s_0 + x_1 = f_1(s_0 + x_1) = f_1(x_1) = x_1$

$s_2 = s_1 + x_2 = f_2(x_1 + x_2) = (x_1 + x_2)(1 + \delta_2)$

$s_3 = s_2 + x_3 = f_3(s_2 + x_3)$

$|\delta_2| \leq \epsilon$
 $|\delta_3| \leq \epsilon$

$= f_3((x_1 + x_2)(1 + \delta_2) + x_3)$

$= ((x_1 + x_2)(1 + \delta_2) + x_3)(1 + \delta_3)$

$= (x_1 + x_2)(1 + \delta_2)(1 + \delta_3) + x_3(1 + \delta_3)$

$1 + \delta_2 + \delta_3 + \delta_2\delta_3$

$= 1 + 2\delta + O(\epsilon^2)$

first-order analysis

$|\delta'| \leq \epsilon$

$s_3 = (x_1 + x_2 + x_3)(1 + 2\delta)$

$|\delta| \leq \epsilon$

$$f(s) = \underbrace{(x_1 + \dots + x_n)}_s (1 + (n-1)\delta) + O(\epsilon^2) \quad |\delta| \leq \epsilon$$

$$f(s) = s(1 + (n-1)\delta) \quad |\delta| \leq \epsilon$$

$$f(x) = \text{sum}(x), \quad \tilde{f}(x) = \text{floating point sum}$$

$$\tilde{F}(x) = f(\tilde{x}) \quad \tilde{x} = x + e, \quad e = \begin{pmatrix} (1 + (n-1)\delta) \\ \vdots \\ (1 + (n-1)\delta) \end{pmatrix}$$

$$\frac{\|x - \tilde{x}\|}{\|x\|} = \frac{\|e\|}{\|x\|} = \frac{(n-1)|\delta| \left\| \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|}{\|x\|}$$

$$\|\cdot\|_2 \Rightarrow \|x - \tilde{x}\| \leq n^2 \epsilon \|x\|$$

↳ "backwards stable" rel err: $C\epsilon$