Sep 14, 2020

~~Putting~~ Approximating real numbers on computers $\Rightarrow$ Floating point

Scientific notation: $\pm 5.282 \cdot 10^{4}$ ($\pm$: sign; 5.282: fraction; 4: exponent; 10: base)

$\pm 0.05282 \cdot 10^{6}$

FP is like scientific notation in binary

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

Normalized FP assumes leading 1

$\pm (1.b_1 b_2 \cdots b_f) \cdot 2^{e}$ (significand)

$2^{11} - 1 = 2047$

IEEE FP

64-bit

sign s exponent e fraction f

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

32-bit

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

Bfloat 16

NVIDIA tensor float

Problem 2: representing numbers (64 bits)

$fl(x)$ = 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 $(0.b_1 \cdots b_f) \cdot 2^{-1023}$

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: $|x - fl(x)| / |x|$

ε = "machine epsilon/precision" = ε_{mach}

= max rel. rep. error

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

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

64-bits IEEE FP

Cor: $fl(x) = x(1+\delta)$ for $|\delta| \leq \varepsilon$

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

Problem 2: roundoff error

$fl(x + y)$ = closest FP rep. of $x + y$

$x = fl(x)$ $\quad y = fl(y)$

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

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

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

(ignoring under/over flow)

"$1+\delta$" model

Example: error analysis

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

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

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

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

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

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

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

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

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

$|\delta_2| \le \varepsilon$

$|\delta_3| \le \varepsilon$

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

$= 1 + 2\delta' + O(\varepsilon^2)$ first-order analysis $|\delta'| \le \varepsilon$

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

$$fl(s) = \underbrace{(x_1 + \cdots + x_n)}_{s}(1 + (n-1)\delta) + O(\varepsilon^2) \quad |\delta| \le \varepsilon$$

$$fl(s) = s(1 + (n-1)\delta) \quad |\delta| \le \varepsilon$$

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

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

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

$\|\cdot\|_1 \Rightarrow \|x - \tilde{x}\| \le n^2 \varepsilon \|x\|$

"backwards stable" rel err: $< \varepsilon$