Lecture 18

⊛ Lecture 16 posted soon ⊛

Really Small #'s w/ Floats!

Do we need any tricks?? ✓

Float 128 / Float 256 ✓

- Richardson + K. Last step in Video

- 0, ∞ have relative Error.

→ · Floating point
Relative Error.

$$1.5 \times 10^{-16}$$

$$1.\underbrace{\boxed{}}_{\text{16 digits}}$$

⊛ Compute small differenc

$S_1 = M + e$ $S_2 = M$

then $S_1 - S_2$ is a bad c

always gives

→

$\times 10^{-16}$

$= 7$

e = mass of Electron

M = Some mass you have

way to compute e.

If $M = 1g$

then $M + e \simeq M$.

⊕

FP Ca

↙

ML

⊕ Want Everything as fast as possible ⊕

\$ \$ \$

BFloat16

7-bit Mantissa
8-bit Exponent
1-bit Sign.

mps

Applied Math

⊕ would like
Float128/256 ⊕

↓ ↓

$$\underbrace{\text{Float64x2} \quad \text{Float64x4}}_{\text{Simulate 256-bit math.}}$$

⊕ don't Expand Exponent
range! ⊕

$$\sqrt{\sum_{i=1}^{n} x_i^2} \;.$$

↓

$$\boxed{x_{max} \cdot \sqrt{\sum_{i=1}^{n} \left(\frac{x_i}{x_{max}}\right)^2}} \;.$$

MultiFloats.jl

$x \in \mathbb{R}$ assume it's in range!

$$fl(x) = x(1+\varepsilon) \qquad |\varepsilon| \le u$$

if $x \ge$ float max, then

$$fl(x) = \pm\infty$$

$$fl(0) = 0(1+\overset{0.}{\varepsilon})$$

$$x \oplus y = (x+y)(1+\varepsilon)$$

Richardson

$\rho(I$

$e^{(k+1)} =$

$x^{(k+1)} = x^{(k)}$

$-x^{(*)} = (-x^{(*}$

$\longrightarrow$

$-\omega A) \le (1 - \frac{1}{\kappa(A)})$

$$(I - \omega A) e^{(k)}$$

$$+ \omega(\underbrace{b - Ax^{(k)}}_{\text{Resdu}})$$

$Ax^* = b$

$$(I - \omega A) x^{(k)} + \omega b - x^*$$

$$\underbrace{(I - \omega A) x^{(k)} \underbrace{+ \omega A x^* - x^*}_{- (I - \omega A) x^*}}$$

$\checkmark$

$$(I - \omega A)(x^{(k)} - x^*)$$

$$\Rightarrow e^{(k+1)} = (I - \omega A) e^{(k)}$$

$e^{(k+1)}$

Assume A is sym., so

$$\rho(A) = \|A\|_2 \Rightarrow$$

$$\|e^{(k+1)}\| \le \rho(A) \cdot \|e^{(0)}$$

$$\le \left(1 - \frac{1}{\kappa(A)}\right) \|e^{(0)}\|$$