Sep 16, 2020

Last time: floating point "1+δ" model

$fl(x \odot y) = (x \odot y)(1+\delta) \quad |\delta| \le \varepsilon \le \begin{cases} 10^{-16} & \text{64-bit} \\ 10^{-7} & \text{32-bit} \end{cases}$

$\odot \in \{+, -, *, /\}$

$$\frac{|fl(x \odot y) - x \odot y|}{|x \odot y|} = |\delta| \le \varepsilon$$

Given two FP numbers x, y, $fl_\odot(x, y)$ is accurate

Alg. $\tilde{f}$ for f is accurate if

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = O(\varepsilon)$$

What if ill-conditioned?

$$\frac{\|\delta f(x)\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|}$$ could be large

Example: subtraction with FP reps

$\hat{x} = fl(x) = x(1+\delta_1)$ $\quad |\delta_1| \le \varepsilon$

$\hat{y} = fl(x) = y(1+\delta_2)$ $\quad |\delta_2| \le \varepsilon$

$$fl(\hat{x}-\hat{y}) = (\hat{x}-\hat{y})(1+\delta_3)$$

$$= x(1+\delta_1)(1+\delta_3) - y(1+\delta_2)(1+\delta_3)$$

$$= x(1+2\delta_4) - y(1+2\delta_5)$$

$(|\delta_4|, \ |\delta_5| \le \varepsilon + O(\varepsilon^2) = O(\varepsilon))$

$$= x - y + x2\delta_4 - y2\delta_5$$

$$\frac{|fl(\hat{x}-\hat{y}) - (x-y)|}{|x-y|} = \frac{|x2\delta_4 - y2\delta_5|}{|x-y|}$$

$x, y = O(\frac{1}{\sqrt{\varepsilon}})$, $x-y = O(1)$

$O(\sqrt{\varepsilon})$ $\quad O(\sqrt{\varepsilon})$

$= O(\sqrt{\varepsilon})$ (not accurate)

Problem comes from FP representation error

$fl(\hat{x} - \hat{y}) = \hat{x} - \hat{y}$ if $1/2 < \hat{x}/\hat{y} < 2$ (HW2)

Still, we had: $fl(\hat{x} - \hat{y}) = x(1+2\delta_4) - y(1+2\delta_5) = \tilde{x} - \tilde{y}$

$$\frac{|\tilde{x} - x|}{|x|} = O(\varepsilon) \qquad \frac{|\tilde{y} - y|}{|y|} = O(\varepsilon)$$

This is called backward stability

Alg $\tilde{f}$ for f is backward stable if,

$\forall$ input x: $\tilde{f}(x) = f(\tilde{x})$ $\quad \frac{\|x - \tilde{x}\|}{\|x\|} = O(\varepsilon)$

$\tilde{f}$ always solves a nearby problem exactly

Slightly weaker: plain stability

$$\frac{\|\tilde{f}(x) - f(\tilde{x})\|}{\|f(\tilde{x})\|} = O(\varepsilon), \qquad \frac{\|x - \tilde{x}\|}{\|x\|} = O(\varepsilon)$$

Last time: sums

(f) $s = \sum_{i=1}^{n} x_i$

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

$\hat{s}_k$ = k^{th} partial sum in alg

returns $\hat{s}_n$

$\hat{s}_0 = 0, \quad \hat{s}_1 = x_1, \quad \hat{s}_k = (\hat{s}_{k-1} + x_k)(1+\delta_k)$ $\quad |\delta_k| \le \varepsilon$

$\hat{s}_2 = (x_1 + x_2)(1+\delta_2)$

$\hat{s}_3 = (\hat{s}_2 + x_3)(1+\delta_3) = (x_1 + x_2)(1+\delta_2)(1+\delta_3) + x_3(1+\delta_3)$

$\vdots$

$$\hat{s}_n = (x_1 + x_2)\prod_{j=2}^{n}(1+\delta_j) + x_3\prod_{j=3}^{n}(1+\delta_j) + \dots + x_n(1+\delta_n)$$

$$= x_1(1+(n-1)\delta_1') + x_2(1+(n-1)\delta_2') + x_3(1+(n-2)\delta_3') + \dots + x_n(1+\delta_n')$$

$\tilde{x}_1$ $\qquad$ $\tilde{x}_2$

$$= f(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$$

$\|x - \tilde{x}\|_\infty \le \|x\|_\infty n\delta \quad |\delta| < \varepsilon$

backward stable

$\|x - \tilde{x}\|_\infty / \|x\|_\infty = O(\varepsilon)$ (dimension in big O)

$f(A, b) \longrightarrow z = A^{-1}b \quad (Az = b)$

$\star \quad \sup_{\|\delta A\|} \frac{\|\delta z\|}{\|z\|} \Big/ \frac{\|\delta A\|}{\|A\|} = \|A\| \cdot \|A^{-1}\| = \kappa(A)$

Suppose we have a backward stable alg $\tilde{f}$ for $z = A^{-1}b$

$\tilde{f}_b(A) = f_b(\tilde{A})$ for $\frac{\|\tilde{A} - A\|}{\|A\|} = O(\varepsilon)$ (δA)

$\|\tilde{f}_b(A) - f_b(A)\| / \|f_b(A)\|$

$= \|f_b(\tilde{A}) - f_b(A)\| / \|f_b(A)\|$

$= \|f_b(A + \delta A) - f_b(A)\| / \|f_b(A)\|$

(δz over the numerator, z over the denominator)

$\star \quad \leq \kappa(A) \cdot \|\delta A\| / \|A\| = O(\kappa(A)\,\varepsilon)$

$\kappa(A)$: indep. of alg

ε: backward stability

Can we always have backward stable algos?

SVD: $A \xrightarrow{f} (U, \Sigma, V^T)$

$U^TU = I, \quad \Sigma = \begin{pmatrix} \ddots & \\ & \end{pmatrix}, \quad V^TV = I$

carry out some arithmetic ...

$\hat{U}\hat{\Sigma}\hat{V}^T \stackrel{?}{=} \text{svd}(\tilde{A}) \qquad \frac{\|\tilde{A}-A\|}{\|A\|} = O(\varepsilon)$

Need: $\hat{U}^T\hat{U} = I \qquad \hat{U} = U + E \qquad (U+E)^T(U+E) = \underbrace{U^TU}_{I} + E^TU + U^TE + E^TE$

$\|E\| = O(\varepsilon)$

Alternative: forward error analysis?

① problem ② work thru flops ③ show that solution is close

Solving $Ax = b$ depends on $\kappa(A)$