Welcome to Ma221! Lecture 5, Sep 1

(Almost) all you need to know about FP:

$$fl(a \underbrace{op}_{+,-,*,/} b) = (a \text{ op } b)(1+\delta) \quad |\delta| \le \varepsilon$$

$$fl(\sqrt{a}) = \sqrt{a}(1+\delta)$$

Some more details

- Important to understand or write reliable code
  - class projects available!
- Lots of recent HW developments
- Analyzing code reliability hard
  - lots of tools being developed
- See posted notes

(1) Exception Handling

IEEE Standard has rules for

Underflow: Tiny/Big = 0, or "subnormal": special numbers with smallest exponent, with leading zeros

Ex $\underset{\text{not } 1}{\underline{0}}.00101 \cdot 2^{\text{min\_exp}}$

if no subnormal numbers, what happens to

if $(x \neq y)$ $z = \frac{10^{-10}}{x-y}$?

can divide by zero

Overflow:

$\pm 1/0 = \pm \text{Inf}$ Inf = Infinity

Rules: Big + Big = Inf

$3 - \text{Inf} = -\text{Inf}$ etc.

Invalid:

$0/0$ = NaN = "Not a Number"

Rules: Inf − Inf = NaN $\sqrt{-1}$ = NaN

3 + NaN = NaN etc

Flags available to checks if an exception occured

Impact of Exceptions on Software:

Compute $s = \|x\|_2 = \sqrt{\sum x_i^2}$

What could go wrong with

$s = 0$, for $i = 1:n$, $s = s + x_i^2$, $s = \sqrt{s}$

Overflow or underflow could cause wrong answers even if "exact" s is OK

careful version of $\|x\|_2$ in BLAS =
Basic Linear Algebra Subroutines

Worst case examples
Crash of Ariane 5
Robotic car crash

Current work to make BLAS + LAPACK
more reliable - class projects

Better Error Analysis for Underflow
$$fl(a \text{ op } b) = (a \text{ op } b)(1+\delta) + \eta$$
$$|\eta| \le \text{tiny number}$$

How to write reliable code without
slowing down (too much)

Run "reckless" code, fast but
ignores possible exceptions

Check flag (or output) for exceptions

In rare case of exceptions,
redo slowly/carefully

Other topics:
High Precision (eg HW Q1.18)

Exploiting Low Precision:

but want "usual" accuracy

do most work in faster low precision,

a little in higher precision

Reproducibility, despite nonassociativity:

$$fl((1-1)+10^{-20}) = 10^{-20}$$

$$fl(1+(-1+10^{-20})) = 0$$

Norms, SVD, condition numbers

How to understand accuracy:

Backward error analysis (scalar case)

want $f(x)$

get $alg(x) = f(x+\delta) \approx f(x) + f'(x)\delta$

error bound

$$\frac{alg(x) - f(x)}{f(x)} \cong \frac{x \cdot f'(x)}{f(x)} \cdot \frac{\delta}{x}$$

$$\underbrace{\left|\frac{alg(x) - f(x)}{f(x)}\right|}_{\text{relative error in output}} \lessapprox \underbrace{\left|\frac{x f'(x)}{f(x)}\right|}_{\text{condition number}} \cdot \underbrace{\left|\frac{\delta}{x}\right|}_{\text{relative error in input}}$$

if cond very large, $f(x)$ very small, i.e.

x near root of f

root ~ $\hat{x} = x - \frac{f(x)}{f'(x)}$ Newton

$$\frac{x-\hat{x}}{x} = \frac{f(x)}{x f'(x)} = \frac{1}{\text{condition \#}}$$

Same approach for $Ax=b$, $Ax=\lambda x$ etc

Get $(A+\Delta)\hat{x} = b$ where

Δ "small" compared to A

What does small mean?

Need vector and matrix norms

want $x = f(A,b)$ get $\hat{x} = \text{alg}(A,b)$

$= f(A+\Delta, b)$

if Δ "small" enough for Taylor exp.

error ~ $J_f(A)\cdot\Delta$, J = Jacobian

want to bound $|J_f(A)|\cdot|\Delta|$

need matrix norms

Matrix and Vector Norms

Def: Let B be linear space ($\mathbb{R}^n$ or $\mathbb{C}^n$)

It is normed if there is

$\|\circ\| : B \to \mathbb{R}$ s.t.

(1) $\|x\| \geq 0$ and $\|x\| = 0$ iff $x = 0$

"positive definite"

(2) $\|c\cdot x\| = |c|\cdot\|x\|$ "homogeneous"

(3) $\|x+y\| \leq \|x\| + \|y\|$ "triangle inequality"

Examples: p-norm $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ $p \geq 1$

Euclidean Norm $\|x\|_2$ = 2-norm

$= \sqrt{\sum x_i^2} = \sqrt{x^T x}$

∞-norm $\|x\|_\infty = \max_i |x_i|$

C-norm $= \|Cx\|$ where $\|\cdot\|$ any norm, C any full column rank matrix (HW Q1.5)

Lemma (1.4): All norms equivalent: given any $\|\cdot\|_a$ and $\|\cdot\|_b$ there are positive constants α, β

$\alpha \|x\|_a \leq \|x\|_b \leq \beta \cdot \|x\|_a$

(proof: Compactness)

Lemma is an excuse to use easiest norm in any proof (HW Q 1.14)

Matrix Norms: on $A^{m \times n}$

Def: Matrix norm: vector norm on m·n vector

(1) $\|A\| \geq 0$, $\|A\| = 0$ iff $A = 0$

(2) $\|cA\| = |c| \cdot \|A\|$

(3) $\|A+B\| \leq \|A\| + \|B\|$

Ex: max norm $\max_{ij} |A_{ij}|$

Frobenius Norm $= \|A\|_F = \sqrt{\sum_{ij} |A_{ij}|^2}$

Def: Operator Norm: given any vector norm

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

Lemma (1.6) An operator norm is a matrix norm (HW Q1.15)

Lemma (1.7) if $\|A\|$ is an operator norm then $\exists x$ such that $\|x\| = 1$ and $\|Ax\| = \|A\|$

proof: $\|A\| = \max_{x \neq 0} \|Ax\| / \|x\|$

$$= \max_{x \neq 0} \left\| A \frac{x}{\|x\|} \right\|$$

$$= \max_{\|y\|=1} \|Ay\|$$

y attaining maximum exists since $\|Ay\|$ continuous function on closed bound set $y : \|y\| = 1$