Goal:

- Implications of FP inner working

- matrix norms

- matrix cond. nr.

HW2: due tomorrow

- linear systems
- lst sq
- eigenvalue
- nonlinear
- fi

Problems with FP Addition

What happens if you subtract two numbers of very similar magnitude?
As an example, consider $a = (1.1011)_2 \cdot 2^0$ and $b = (1.1010)_2 \cdot 2^0$.

rel.
rounding error $\le \varepsilon_m$ $a = (\cancel{1.1011})_2 \cdot 2^0$
rel.
rounding error $\le \varepsilon_m$ $b = (\cancel{1.101}0)_2 \cdot 2^0$

?!
$1.0000 \cdot 2^{-4}$

rel.
rd. error $\le 2^4 \cdot \varepsilon_{mach}$

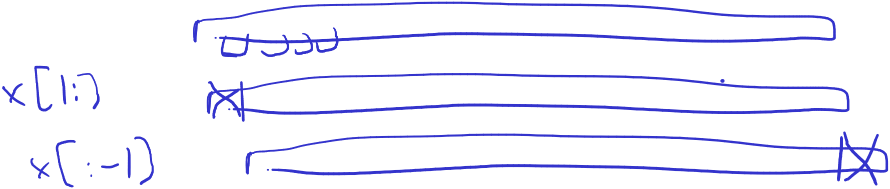
Demo: Catastrophic Cancellation [cleared]

Rel. error vs "digits"

accurate significant digits

$|\text{rel. err.}| \leq M$

$\iff$ "The result has $-\log_{10} M$ a.s.d"



Demos:

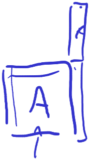
- Density
- Harmonic series
- FP vs Program Logic

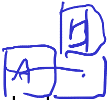
Supplementary Material

- Josh Haberman, Floating Point Demystified, Part 1
- David Goldberg, What every computer programmer should know about floating point
- Evan Wallace, Float Toy
- Julia Evans, Examples of Floating Point Problems, 2022

Outline

Solving a Linear System





Given:

- $m \times n$ matrix A
- m-vector $\boldsymbol{b}$

What are we looking for here, and when are we allowed to ask the question?

Want $\vec{x}$ from $A\vec{x} = \vec{b}$

- lin. comb. of cols to get $\vec{b}$
- $m = n$
- A not singular: there exists a unique sol.

if it is: no sol, or ∞ many

Next: Want to talk about conditioning of this operation. Need to measure distances of matrices.

$\|x - \hat{x}\|$

Solving a Linear System

Given:

- $m \times n$ matrix A
- m-vector $\boldsymbol{b}$

What are we looking for here, and when are we allowed to ask the question?

Want: n-vector $\boldsymbol{x}$ so that $A\boldsymbol{x} = \boldsymbol{b}$.

- Linear combination of columns of A to yield $\boldsymbol{b}$.
- Restrict to square case ($m = n$) for now.
- Even with that: solution may not exist, or may not be unique.

Unique solution exists iff A is *nonsingular*.

Next: Want to talk about conditioning of this operation. Need to measure distances of matrices.

Linearity of matrices?

$$A(\alpha \vec{x}) = \alpha (A\vec{x})$$

$$A(\vec{x}+\vec{y}) = A\vec{x} + A\vec{y}$$

$$\|A\vec{x}\| \qquad \|x\|$$

Matrix Norms

What norms would we apply to matrices?

$\|Ax\|$

$\|A\|$

$$\underbrace{\|A\vec{x}\|}_{\text{vec}} \le \underbrace{\|A\|}_{?} \underbrace{\|x\|}_{\text{vec}}$$

"mat norm" ⇐ defining

Assume $\vec{x} \ne 0$:

$$\max_{\vec{x}\ne 0} \frac{\|A\vec{x}\|}{\|\vec{x}\|} = \|A\|$$

$$= \max_{x\ne 0} \|Ax\| \cdot \frac{1}{\|x\|} = \max_{x \ne 0} \Big\|A \overbrace{\underbrace{\frac{x}{\|x\|}}_{\left\|\frac{x}{\|x\|}\right\| = 1}}^{\vec{y}}\Big\| = \max_{\|y\|=1} \|Ay\|$$

Intuition for Matrix Norms

Provide some intuition for the matrix norm.

Identifying Matrix Norms

$\|x\|_1 = \sum |x_i| \qquad \|\vec{x}\|_\infty = \max |x_i|$

What is $\|A\|_1$? $\|A\|_\infty$?

$$\left.\begin{aligned} \|A\|_1 &= \max_{\text{col } j} \sum_{\text{row } i} |A_{ij}| \\ \|A\|_\infty &= \max_{\text{row } i} \sum_{\text{col } j} |A_{ij}| \end{aligned}\right\} \text{by invisible proof}$$

How do matrix and vector norms relate for $n \times 1$ matrices?

$$\|A\|_1 = \cancel{\max_{\text{col } j}} \sum_{\text{row } i} |A_{ij}|$$
$$\|A\|_\infty = \max_{\text{row } i} \cancel{\sum_{\text{col } j}} |A_{ij}|$$

Demo: Matrix norms [cleared]

Properties of Matrix Norms

Matrix norms inherit the vector norm properties:

- $\|A\| > 0 \Leftrightarrow A \neq 0$.
- $\|\gamma A\| = |\gamma| \, \|A\|$ for all scalars γ.
- Obeys triangle inequality $\|A + B\| \leq \|A\| + \|B\|$

But also some more properties that stem from our definition: