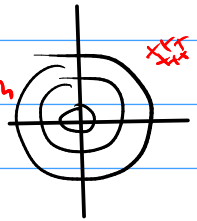
Stability $\Rightarrow$ accuracy ?

→ "backward stable"
measure output perturbation
to the true answer



UFL?

→ smallest normal (not subnormal)
number.

Floating Point and Rounding Error

$(1.0001)_2 \cdot 2^{p<0}$

What is the relative error produced by working with floating point numbers?

- What is smallest floating point number > 1? Assume 4 bits in the significand.

$(1.0001)_2 \cdot 2^{p<0}$

- What's the smallest FP number > 1024 in that same system?

$(1.0001)_2 \cdot 2^{10}$

- Can we give that number a name?

machine precision ε_{mach} → the smallest number so that $fl(1+\varepsilon_{mach}) > 1$

- What does this say about the relative error incurred in floating point calculations?

$(1.0001)_2 \approx fl((1.0001)_2)$

- What's that same number for double-precision floating point? (52 bits in the significand)

2^{-53} or 2^{-52} (dep. on rounding)

$$fl(a+b) = \tilde{x} \qquad a+b = x$$

$$\uparrow$$

$$= x \cdot (1 + \varepsilon_{mach})$$

Rel. error from FP rounding

$$\frac{|\tilde{x} - x|}{|x|} = \frac{|x(1+\varepsilon_{mach}) - x|}{|x|} = \varepsilon_{mach}$$

Demo: Floating Point and the Harmonic Series

Implementing Arithmetic

- How is floating point addition implemented?
  Consider adding $a = (1.101)_2 \cdot 2^1$ and $b = (1.001)_2 \cdot 2^{-1}$ in a system with three bits in the significand.

$$
\begin{array}{lll}
a = & (1.101)_2 \cdot & 2^1 \\
+ & \\
b = & 0.01001)_2 & 2^1 \\
\hline
 & 1.11101 & \cdot 2^1
\end{array}
$$

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 $b = (1.1010)_2 \cdot 2^0$.

→ $a = (1.1011)_2 ? \cdot 2^0$

→ $b = (1.1010)_2 ? \cdot 2^0$

$a - b = 0.0001 - ?$

→ $= 1.? \cdot 2^{-4}$

Demo: Catastrophic Cancellation

2 Systems of Linear Equations

2.1 Theory: Conditioning

Solving a Linear System

Given:

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

$Ax = b$

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

  - If $b \in \text{span}(\text{columns}(A))$ → even if the answer is not unique
  - For unique answer, need A to be invertible.

Matrix Norms

- What norms would we apply to matrices?

Need:

$$\|Ax\| \leq \|A\| \cdot \|x\|$$

"submultiplicativity"

Given $\|v\|$ ← a vector norm

$$\|A\| := \max_{\|x\|=1} \|Ax\|$$

← matrix norm

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

$n \times 1 \rightarrow \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}}_{A}$

$$\|A \cdot (\underbrace{\pm 1}_{\uparrow})\| = \left\| \pm \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right\|_1 = \left\| \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right\|_1$$

If you compute the norm of a nx1 matrix, you obtain the vector norm of that column vector.

$1 \times n$

$$(\alpha_1 \cdots \alpha_n) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

... more subtle.

$\|x\| = 1$

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

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

Demo: Matrix norms

In-class activity: Matrix norms