Stability $\Rightarrow$ accuracy?

$\Rightarrow$ "backward stable"
measure output perturbation
to the true answer

UFL?

$\Rightarrow$ smallest normal (not subnormal)
number.
Floating Point and Rounding Error

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.

\[(1.0001) \cdot 2^{0}\]

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

\[(1.0001) \cdot 2^{10}\]

- Can we give that number a name?

machine precision \(\varepsilon_{\text{mach}}\)

\[\text{for } (1 + \varepsilon_{\text{mach}}) > 1\]

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

\[(1.0000) \cdot 2^{-53}\]

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

\[2^{-53} \text{ or } 2^{-52} \text{ (dep. on rounding)}\]
\[ f(l(a+b)) = \tilde{x} \quad a+b = x \]

\[ \frac{1}{|x|} \left| \tilde{x} - x \right| = \frac{|x(1+\varepsilon_{\text{mach}}) - x|}{|x|} = \varepsilon_{\text{mach}} \]

Rel. error from FP rounding.
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.

\[
\begin{align*}
  a &= (1.101)_2 \cdot 2^1 \\
  b &= (0.01001)_2 \\
  \hline
  c &= (1.11101)_2 \cdot 2^1
\end{align*}
\]
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 \( b = (1.1010)_2 \cdot 2^0 \).

\[
\begin{align*}
\rightarrow & \quad a = (1.1011)_2 \cdot 2^0 \\
\rightarrow & \quad b = (1.1010)_2 \cdot 2^0 \\
\hline
a - b & = 0.0001 \cdot 2^{-4}
\end{align*}
\]
Demo: Catastrophic Cancellation
2 Systems of Linear Equations
2.1 Theory: Conditioning
Solving a Linear System

Given:

- \(m \times n\) matrix \(A\)
- \(m\)-vector \(b\)

○ 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?

\[ \| A \| : \text{max} \| A \times \| \| \times \| \|_{\| = 1} \]

\[ \| A \| \leq \| A \| \cdot \| x \| \]

"submultiplicativity"

Given \( \| v \| \) - a vector norm

\( \| A \| : \text{max} \| A \times \| \| \times \| \|_{\| = 1} \)  \(-\) matrix norm
If you compute the norm of a \( nx1 \) matrix, you obtain the vector norm of that column vector.

\[
\| \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \|_1 = |x_1| + |x_2| + \cdots + |x_n|
\]

\[
\| A \cdot \left( \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \right) \|_1 = \| \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \|_1
\]

\[
\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}
\]

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

\[
\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{... more subtle.}
\]

\[
\| x \| = 1
\]
\[ \|A\|_1 = \max_{\text{col} j} \sum_{\text{row} i} |A_{ij}| \]

\[ \|A\|_{\infty} = \max_{\text{row} i} \sum_{\text{col} j} |A_{ij}| \]
Demo: Matrix norms

In-class activity: Matrix norms