Overview

- Error / cond. num.
- δp

Example 2
- Interp
- MC
- Errors / cond.
- FP.

HW3 due
HW4 timeline

Abs = \left| x_0 - x \right| = \left| \Delta x \right|

Rel. = \frac{Abs}{\left| x_0 \right|} = \frac{\left| \Delta x \right|}{\left| x_0 \right|}
Measuring Error

Why is $|\tilde{x}| - |x_0|$ a bad measure of the error?

If $\tilde{x}$ and $x_0$ are vectors, how do we measure the error?

Need equivalent to the abs value:

$\| \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \|_2 = \sqrt{y_1^2 + y_2^2}$
Sources of Error

What are the main sources of error in numerical computation?

- **Truncation error**
- **Rounding error**

\[
\text{abs. error} = |x_0 - \tilde{x}| = \text{trunc.} + \text{rounding}
\]

\[
\text{abs. error} = 10^{-5} \quad \text{both}
\]

**Scenario 1**
\[
x_0 = 10^{-6}
\]

**Scenario 2**
\[
x_0 = 10^0
\]

\[
\text{rel. error} = \frac{\text{abs. error}}{1} = 10^{-5}
\]
Establish a relationship between ‘accurate digits’ and rounding error.

\[
\text{finite precision} \quad \Rightarrow 3.14159 \quad \Rightarrow \text{rounded to 3 digits}
\]

\[
\text{relative error} \Rightarrow \frac{|3.14159 - 3.14|}{|3.141591|} \approx 10^{-3}
\]
Methods $f$ take input $x$ and produce output $y = f(x)$. Input has (relative) error $|\Delta x|/|x|$. Output has (relative) error $|\Delta y|/|y|$. Q: Did the method make the relative error bigger? If so, by how much?

Condition Numbers

\[
\text{Cond.} = \frac{\text{rel. error out}}{\text{rel. error in}} = \frac{|\Delta y|/|y|}{|\Delta x|/|x|}
\]

most interested in: upper bound of CN for
\[ 10^{-3} \rightarrow 10^{-1} \]

For this example:
\[ CN = 10^2 \]

- Good \( CN \): small
- Bad \( CN \): big (error gets bigger)

\[ \text{Abs. } CN = \left| \frac{\text{Abs. error in out}}{\text{Abs. error in in}} \right| = \frac{|\Delta y|}{|\Delta x|} \]
nth-Order Accuracy

Often, truncation error is controlled by a parameter \( h \).

Examples:
- distance from expansion center in Taylor expansions
- length of the interval in interpolation

A numerical method is called ‘nth-order accurate’ if its truncation error \( E(h) \) obeys

\[
E(h) = O(h^n).
\]
Outline

Python, Numpy, and Matplotlib
Making Models with Polynomials
Making Models with Monte Carlo
Error, Accuracy and Convergence

Floating Point
Modeling the World with Arrays
  The World in a Vector
  What can Matrices Do?
  Graphs
  Sparsity
Norms and Errors
The ‘Undo’ Button for Linear Operations: LU
LU: Applications
  Linear Algebra Applications
  Interpolation
Repeating Linear Operations: Eigenvalues and Steady States
Eigenvalues: Applications
Approximate Undo: SVD and Least Squares
SVD: Applications
  Solving Funny-Shaped Linear Systems
  Data Fitting
  Norms and Condition Numbers
  Low-Rank Approximation
Iteration and Convergence
Solving One Equation
Solving Many Equations
Finding the Best: Optimization in 1D
Optimization in $n$ Dimensions
Wanted: Real Numbers... in a computer

Computers can represent integers, using bits:

\[ 23 = 1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = (10111)_2 \]

How would we represent fractions, e.g. 23.625?

\[
23 = 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0
\]

\[
23.625 = 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3}
\]

\[ 2^{-3} = 0.125 \]

\[ 23.625 = (10111.101)_2 \]
Fixed-Point Numbers

Suppose we use units of 64 bits, with 32 bits for exponents $\geq 0$ and 32 bits for exponents $< 0$. What numbers can we represent?

\[\text{Biggest: } (1, 1111 \ldots 1, 111111)_{2}\]

How many ‘digits’ of relative accuracy (think relative rounding error) are available for the smallest vs. the largest number?

\[\text{Smallest: } (0 \ldots 001)_{2} = 1 \cdot 2^{-32}\]
Actual result: $2^{-32}$

Computed result: $2^{-33}$

$$\text{rel. error} = \frac{2^{-32} - 2^{-33}}{2^{-32}} = \frac{2^{-32}}{2^{-32}} = 2^{-1}$$

In fixed point:

uneven relative error.

big results: small rel. error
small results: big rel. error
Floating Point numbers

Convert $13 = (1101)_2$ into floating point representation.

$$13 = \frac{(1.101)_2 \cdot 2^3}{2} = (1101)_2 = (110.1)_2 \cdot 2 = (11.01)_2 \cdot 2^2$$

What pieces do you need to store an FP number?

$$\frac{1.110011_2}{2^{32}} - 2^{-32}$$
In-class activity: Floating Point
Unrepresentable numbers?

Can you think of a somewhat central number that we cannot represent as

\[ x = (1.\underline{\text{-------}})_2 \cdot 2^{-p}? \]