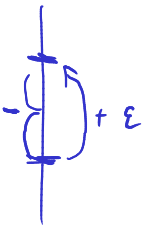
- Exam 1 0
- Quiz deadlines
- HW2



$1 = 1.\underbrace{0\ \cdots\ 0}\cdot 2^0$

$1+\varepsilon = 1.\underbrace{0\ \cdots\ 01}\cdot 2^0$ } ULP "unit in the last place"

$1+\varepsilon_{mach} \neq 1$ (round-to-nearest) } round-to-even

$\hookrightarrow \varepsilon_{mach} = \frac{ULP}{2}$

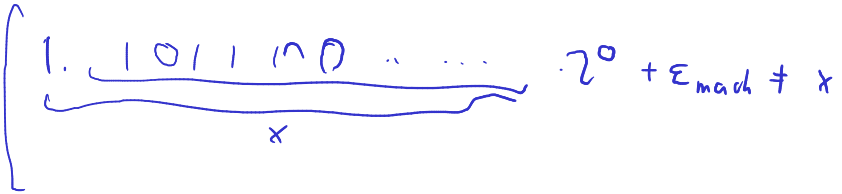
Unit Roundoff

Unit roundoff or *machine precision* or *machine epsilon* or $\varepsilon_{\text{mach}}$ is the smallest number such that

$$\text{float}(1+\varepsilon) > 1.$$

- Technically that makes $\varepsilon_{\text{mach}}$ depend on the rounding rule.

  Assuming round-towards-infinity, in the above system, $\varepsilon_{\text{mach}} = (0.00001)_2$.
- Note the extra zero.
- Another, related, quantity is *ULP*, or *unit in the last place*. ($\varepsilon_{\text{mach}} = 0.5\,\text{ULP}$)


1. 1011 0110 ... 000
x
$\cdot 2^0 + \varepsilon_{mach} \neq x$
1. 1011 0110 ... 000
x
$\cdot 2^{15} + 2^{15} \varepsilon_{mach} \neq x$

FP: Relative Rounding Error

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

$\tilde{x} = x + \varepsilon_{mach} x \neq x$

$x(1 + \varepsilon_{mach}) \neq x$

- Use this as a model for the amount of rounding error introduced.
- Is actually an upper bound:

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

FP: Machine Epsilon

What's that same number for double-precision floating point? (52 stored bits in the significand, 53 total)

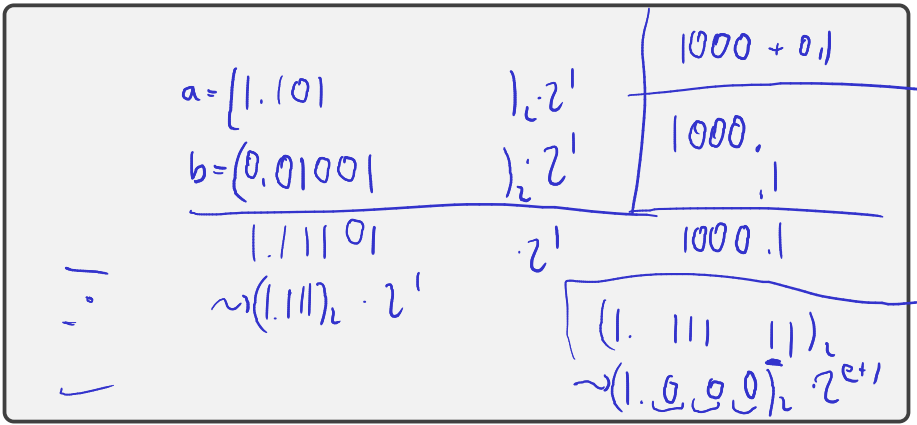
2^{-53} (round-to-nearest, complications with round-to-even)

Demo: Floating Point and the Harmonic Series [cleared]

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 stored bits (four total) in the significand.



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

$a = (1.1011)_2$

$b = (1.1010)_2$

$a - b = (0.0001)_2$

$1.\underline{????} \cdot 2^{-4}$

Demo: Catastrophic Cancellation [cleared]

Supplementary Material

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

Outline

Introduction to Scientific Computing

Systems of Linear Equations
- Theory: Conditioning
- Methods to Solve Systems
- LU: Application and Implementation

Linear Least Squares

Eigenvalue Problems

Nonlinear Equations

Optimization

Interpolation

Numerical Integration and Differentiation

Initial Value Problems for ODEs

Boundary Value Problems for ODEs

Partial Differential Equations and Sparse Linear Algebra

Fast Fourier Transform

Additional Topics

Solving a Linear System

Given:

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

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

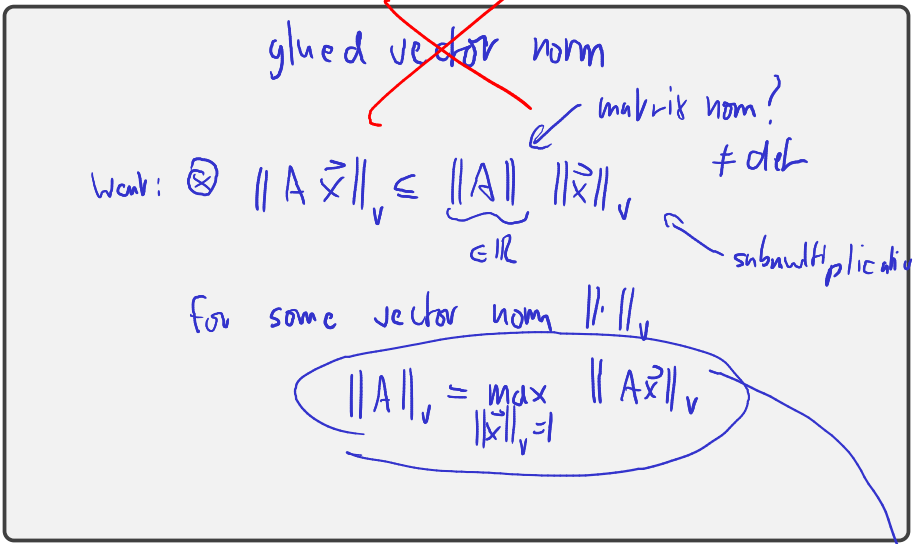
What are we looking for here, and when are we allowed to ask the question?

- n-vector $\vec{x}$ s.t.
- restrict to $m=n$
- solutions may not exist, may not be unique
  - ↳ $\exists!$ for A non singular.

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

Matrix Norms

What norms would we apply to matrices?



Intuition for Matrix Norms

$$\frac{\|Ax\|_v}{\|x\|_v} \leq \|A\|_v \quad (*)$$

Provide some intuition for the matrix norm.

$$\max_{\vec{x}\neq 0} \frac{\|A\vec{x}\|_v}{\|\vec{x}\|_v} = \max_{\vec{x}\neq 0} \left\| \frac{1}{\|x\|_v} Ax \right\| = \max_{\vec{x}\neq 0} \left\| A \cdot \underbrace{\frac{x}{\|x\|_v}}_{\|\cdot\|_v = 1} \right\|_v$$

Identifying Matrix Norms

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

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

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

They agree. (but just these two! as far as we know)

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: