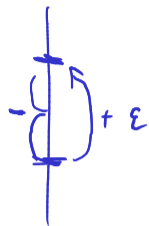


- Exponent 0
- Quiet denorms
- HVZ



$$1 = 1. \underbrace{0 \dots 0}_{\text{ULP}} \cdot 2^0$$

$$1 + \epsilon^? = 1. \underbrace{0 \dots 01}_{\text{ULP}} \cdot 2^0$$

ULP  
"unit in the last place"  
rounds to even

$1 + \epsilon_{\text{mach}} \neq 1$  (round-to-nearest)

$$\hookrightarrow \epsilon_{\text{mach}} = \frac{\text{ULP}}{2}$$

## Unit Roundoff



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

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

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

Assuming round-towards-~~infinity~~, in the above system,

$$\epsilon_{\text{mach}} = (0.00001)_2.$$

nearest

4 stored bits in the significand

- ▶ Note the extra zero.
- ▶ Another, related, quantity is *ULP*, or *unit in the last place*.

$$(\epsilon_{\text{mach}} = 0.5 \text{ ULP})$$

relative to rtn

$$\left[ \underbrace{1. \underbrace{1011100 \dots}_x}_{\times} \cdot 2^0 + \epsilon_{\text{mach}} \neq x \right]$$

$$\underbrace{1. \underbrace{1011100 \dots}_x}_{\times} \cdot 2^{15} + 2^{15} \epsilon_{\text{mach}} \neq x$$

## FP: Relative Rounding Error

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

$$\tilde{x} = x + \epsilon_{\text{mach}} x \neq x$$

$$x(1 + \epsilon_{\text{mach}}) \neq x$$

- Use this as a model for the amount of rounding error introduced.

- Is actually an upper bound:

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

## 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.

The diagram illustrates the addition of two floating-point numbers in binary. It is divided into two main sections by a vertical line.

**Left Section:**

- Number  $a = (1.101)_2 \cdot 2^1$  is written with its significand  $1.101$  and exponent  $2^1$ .
- Number  $b = (0.01001)_2 \cdot 2^1$  is written with its significand  $0.01001$  and exponent  $2^1$ .
- A horizontal line separates the numbers from their sum.
- The sum is  $1.11101$ .
- The sum is rounded to  $(1.111)_2 \cdot 2^1$ .

**Right Section:**

- The original number  $b = (1.001)_2 \cdot 2^{-1}$  is written with its significand  $1.001$  and exponent  $2^{-1}$ .
- A horizontal line separates it from its shifted version.
- The shifted number is  $0.1001$ .
- The sum of the shifted numbers is  $1.000.1$ .
- The sum is rounded to  $(1.11111)_2 \cdot 2^0$ .
- The final result is  $(1.0000)_2 \cdot 2^{+1}$ .

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

$$\begin{array}{r} a = (1.1011)_2 \\ b = (1.1010)_2 \\ \hline a - b = (0.0001)_2 \\ \qquad \qquad \qquad 1. \underline{\quad ? ? ? ? \quad} \cdot 2^{-4} \end{array}$$

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  $\mathbf{b}$

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

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?

~~glued vector norm~~

Want:  $\otimes \quad \|A \vec{x}\|_v \leq \underbrace{\|A\|}_{\in \mathbb{R}} \|\vec{x}\|_v$

matrix norm?  
≠ def

submultiplicativity

for some vector norm  $\|\cdot\|_v$

$$\|A\|_v = \max_{\|\vec{x}\|_v=1} \|A\vec{x}\|_v$$

## Intuition for Matrix Norms

$$\frac{\|Ax\|_v}{\|x\|_v} \leq \|A\|_v$$

Provide some intuition for the matrix norm.

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

## Identifying Matrix Norms

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

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

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  $\gamma$ .
- ▶ Obeys triangle inequality  $\|A + B\| \leq \|A\| + \|B\|$

But also some more properties that stem from our definition:

