September 10, 2024

Announcements

- Exam 1
- HW2

Goals

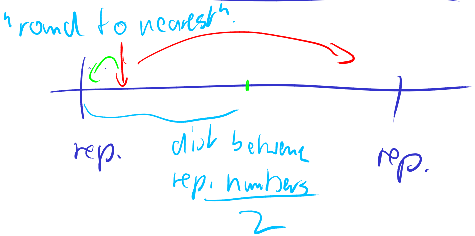
- Floating P
- Lin. Systems Th.

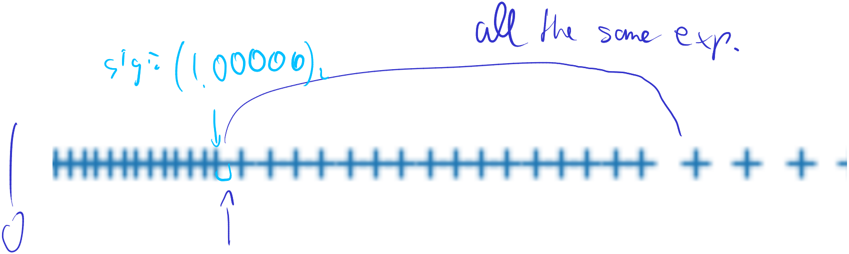
Review

$$\underbrace{1.\overbrace{\qquad\qquad\qquad}^{\text{fraction}}}_{\text{significand}} \cdot 2^{exp}$$

subnormals: based on sp. exp.

- no implicit leading one




all the same exp.
sig: $(1.00000)_2$
0

Rounding Modes

How is rounding performed? (Imagine trying to represent π.)

$$(\underbrace{1.1101010}_{\text{representable}}11)_2$$

$1.1101010\,10$

0

$0\ 0$

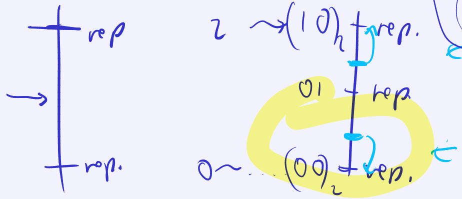
$(1.11010\,1\,1)_2$

What is done in case of a tie? $0.5 = (0.1)_2$ ("Nearest"?)

$(1.1101011\,10)$

$(1.1101100)_2$

round-to-nearest-even.



Demo: Density of Floating Point Numbers [cleared]

Rounding Modes

How is rounding performed? (Imagine trying to represent π.)

$$\left(\underbrace{1.1101010}_{\text{representable}}11\right)_2$$

What is done in case of a tie? $0.5 = (0.1)_2$ ("Nearest"?)

Up or down? It turns out that picking the same direction every time introduces *bias*. Trick: *round-to-even*.

$$0.5 \to 0, \qquad 1.5 \to 2$$

Demo: Density of Floating Point Numbers [cleared]

Smallest Numbers Above…

$1.\underline{\ \ }\,\underline{\ \ }\,\underline{\ \ }\,\underline{\ \ }$

- What is smallest FP number > 1? Assume 4 stored bits (5 total) in the significand.

$$1.\underline{0}\,\underline{0}\,\underline{0}\,\underline{1}$$

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

$$(1.000)_2 \cdot 2^{10}$$

Can we give that number a name?

Unit Roundoff

Unit roundoff or *machine precision* or *machine epsilon* or $\varepsilon_{\text{mach}}$ is...

smallest number $\varepsilon > 0$ so that

$$fl(1+\varepsilon) \neq 1$$

For round-to-even:

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

FP: Relative Rounding Error

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

$$\hat{x} = x + x\epsilon_{much} = x(1+\epsilon_{mach})$$

$\uparrow$ smallest number $> x$

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

$$\frac{\Big(x_{sig}\cdot\cancel{2^{x_{exp}}} + \hat{x}_{sig}\cdot\cancel{2^{x_{exp}}}}{x_{sig}\cdot\cancel{2^{x_{exp}}}}$$

FP: Machine Epsilon

What's machine epsilon for double-precision floating point with round-to-nearest? (52 stored bits in the significand, 53 total)

Demo: Floating point and the Harmonic Series [cleared]

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

$$\to a = (1.1011)_2 \ldots$$

$$\to b = (1.1010)_2$$

$$a - b = (0.0001)_2$$

$$(1.\underset{0000}{????})_2 \cdot 2^{-4}$$

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
- Julia Evans, Examples of Floating Point Problems, 2022

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