- Forum invites
- one question/concern per post
- FP
- numerical linalg
  - ↳ norms/errors/cond.
  - ↳ solving

2^{-51} .010

2^{-52} .001

1 .000

"1 accurate digit" → 0.1

"2 accurate digits" → 0.01

"

0.0003147777

5.0003147777

Rounding Modes

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

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

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

Demo: Density of Floating Point Numbers [cleared]

Demo: Floating Point vs Program Logic [cleared]

Smallest Numbers Above...

- What is smallest FP number > 1? Assume 4 bits in the significand.

$1.\underline{0001}$

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

$1.0001 \times 2^{10} \longleftrightarrow 1.0000 \times 2^{10}$

Can we give that number a name?

Unit Roundoff

1.1001

1.1010

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

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

- Assuming round-to-nearest, 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*. ($\varepsilon_{\text{mach}} = 0.5\,\text{ULP}$)

FP: Relative Rounding Error

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

$$\left|\frac{x-\tilde{x}}{x}\right| = \left|\frac{x - x(1+\varepsilon)}{x}\right|$$

$$= |\varepsilon| \le \varepsilon_{mach}.$$

$$|\tilde{x}| = |x \cdot (1+\varepsilon)|$$

FP: Machine Epsilon

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

$$\varepsilon_{mach} = 0.5 \cdot 2^{-52}$$
$$= 2^{-53} \sim 10^{-16}$$

Demo: Floating Point and the Harmonic Series [cleared]

In-Class Activity: Floating Point

In-class activity: Floating Point

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.

$$a = (1.101)_2 \cdot 2^1$$
$$b = (0.01001) \cdot 2^1$$
$$a+b = (1.11001) \cdot 2^1$$

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

$$a = (1.1011)_2$$

$$b = (1.1010)_2$$

$$a - b = 0.0001 ???$$

Demo: Catastrophic Cancellation [cleared]

Supplementary Material

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