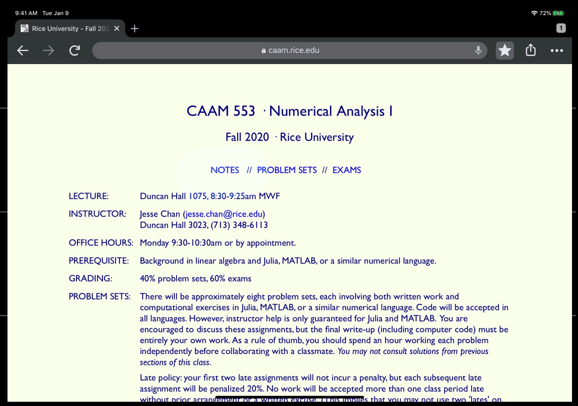
Class overview:



www.caam.rice.edu/~caam553/

- homework, exams, & notes will be posted here.
- Canvas for homework submission & gradebook.

- What should you learn?
    - Analysis of algorithms for problems in continuous (vs discrete) mathematics.
    - Tools for research: analysis & implementation of algorithms
    - Qualifying exam preparation
        - ↳ Make exams like quals.
        - ↳ How is practice

Floating point number systems

- What you should learn from this lecture:
  - how real numbers are represented on computers
  - when "catastrophic" rounding errors occur

- Naive ideas for representations
  - $3.14159 \ldots\ldots \times 10^{e}$

    $\Rightarrow$ store digits and a scale factor
  - What are some issues w/ this

    $\Rightarrow$ fixed precision numbers

$$a \times b = 560\cancel{88}$$

$$123 \quad 456 \qquad 5.60 \times 10^4$$

$$\uparrow\uparrow\uparrow \quad \uparrow\uparrow\uparrow$$

$$= 1.23 \times 10^2 \quad 4.56 \times 10^2 \quad = 3 \text{ digits} + \text{exponent}$$

"Floating" vs "fixed" point

- aims for relative vs absolute precision

$$x = \underset{\substack{\uparrow \\ \text{sign}}}{\underline{\pm}} \Big(\underset{\substack{\downarrow \\ \text{coefficients}}}{d_0} \beta^{0} + d_1 \beta^{-1} + \cdots + d_{\underbrace{p-1}} \beta^{-(p-1)} \Big) \beta^{e}$$

base$^{\text{power}}$ $\qquad$ p = precision

- everything is an integer
  - β = base / radix
  - p = precision
  - e = exponent (usually with $e_{min} \le e \le e_{max}$)

Example: suppose $\beta = 2$, $p = 3$

$$x = \pm \left(d_0 2^{0} + d_1 2^{-1} + d_2 2^{-2} \right) 2^{e}$$

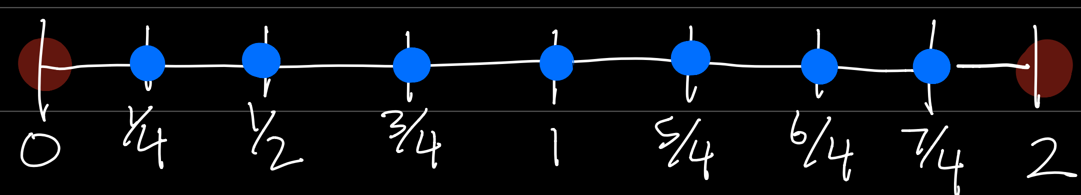
$$= \pm \left(d_0 + d_1 \tfrac{1}{2} + d_2 \tfrac{1}{4} \right) 2^{e}$$

Suppose $e = 0$: what numbers can we represent?

$$x = \pm \left(d_0 + d_1 \tfrac{1}{2} + d_2 \tfrac{1}{4} \right) 2^{e}$$

$$e=0\begin{cases} x=\frac{1}{4} \Rightarrow d=(d_0,d_1,d_2)=(0,0,1) \\ x=1/2 \Rightarrow d=(0,1,0) \\ x=3/4 \Rightarrow d=(0,1,1) \\ x=1 \Rightarrow d=(1,0,0) \\ x=5/4 \qquad d=(1,0,1) \\ x=6/4 \qquad d=(1,1,0) \\ x=7/4 \qquad d=(1,1,1) \end{cases}$$

$x \in (0,2) \Rightarrow 0,2$ not representable.



Suppose now $e=1$: what numbers are representable?

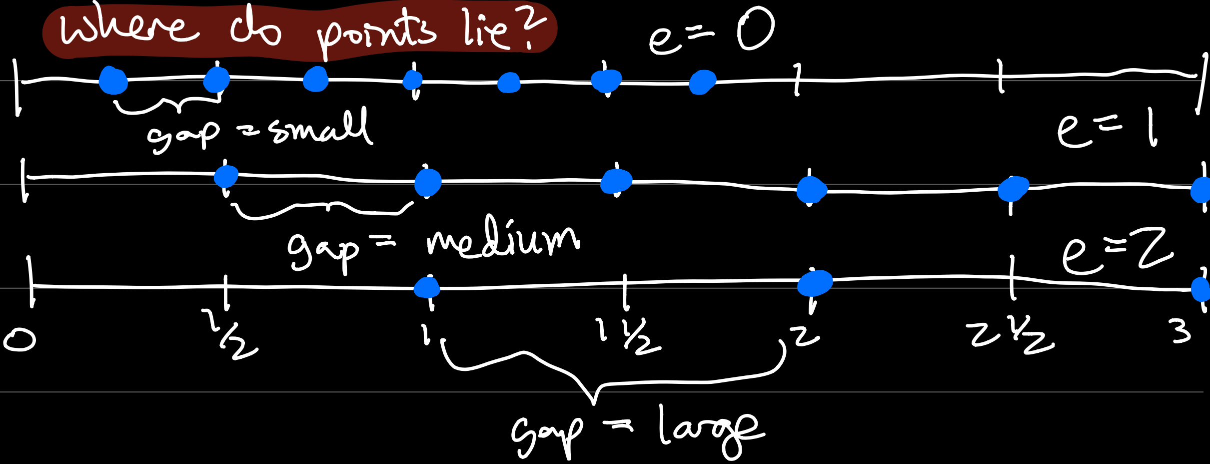
$$x = \pm\left(d_0 + d_1\tfrac{1}{2} + d_2\tfrac{1}{4}\right)2$$

$\Rightarrow x = \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, 5/4, 6/4, 7/4\} \times 2$

$\Rightarrow x = \{\tfrac{1}{2}, 1, 6/4, 2, 2\tfrac{1}{2}, 3, 3\tfrac{1}{2}\}$ for $e=1$

$x = \{1, 2, 3, 4, 5, 6, 7\}$ for $e=2$

where do points lie?



Observations:

- representations of numbers is not unique. Can fix by assuming $d_0 = 1$ (referred to as normalization)
- Main idea: spacing between numbers wider for larger numbers. precision depends on e.
- radix point $\Rightarrow$ moves or "floats"

$\underbrace{1\,2\,3}_{\text{stored separately}}.\underbrace{4\,5\,6}$

stored separately in fixed pt. format

IEEE - Inst. of Electrical + Electronic Engineers

Floating point standards

- half : $p = 11$ (16 bit)

$$[\underbrace{\text{sign}}_{1\text{ bit}} \mid \underbrace{\text{exponent}}_{5\text{ bits}} \mid \underbrace{\text{significand}}_{10\text{ bits}}]$$

- single : $p = 24$ (32 bit)

$$[\underbrace{\text{sign}}_{1\text{ bit}} \mid \underbrace{\underbrace{\text{exponent}}_{8\text{ bits}}}_{\text{range}} \mid \underbrace{\underbrace{\text{significand}}_{23\text{ bits}}}_{\text{precision}}]$$

- double : $p = 53$ (64 bit)

$$[\underbrace{\text{sign}}_{1\text{ bit}} \mid \underbrace{\text{exponent}}_{11\text{ bits}} \mid \underbrace{\text{significand}}_{52\text{ bits}}]$$

- quad: $p = 113$ (128 bit)

<u>Note</u>: cost of operations do not scale linearly wrt. number of bits!

Machine Epsilon

- Binary $\Rightarrow$ $\beta = 2$, double prec.

For $e = 0$, $x \in [0, 2]$ and

$$x = d_0 2^0 + d_1 2^{-1} + \dots \underbrace{d_{52} 2^{-52}}$$

$$\varepsilon_{mach} = \text{smallest rel. diff b/w 2 floating pt numbers}$$

Rounding in floating point systems

- Floating point systems designed to ensure that the relative floating point representation error is bounded.
- Let $fl(x)$ denote the floating point representation of x.

$$\Rightarrow \quad \boxed{\frac{|x - fl(x)|}{|x|} \leq \varepsilon_{mach}}$$

Pf: Assume $x \in \mathbb{R}$ + within representable range of floating point system

$\Rightarrow$ x is b/w 2 floating pt numbers

$a \leq x \leq b$, $|b-a| \leq \varepsilon_{mach}$

$fl(x) = x(1+\delta) \Rightarrow \frac{|x - fl(x)|}{|x|} \leq \varepsilon_{mach}$ $\left(|\delta| \leq \varepsilon_{mach}\right)$

- Floating point arithmetic is not exact, & requires truncating information.

Let $\mathbb{F}$ = set of all floating point numbers

Let $x, y \in \mathbb{F}$.

$$\Rightarrow x = \pm\left(x_0\beta^0 + x_1\beta^1 + \dots x_{p-1}\beta^{-(p-1)}\right)\beta^{e_x}$$

$$+ \quad y = \pm\left(y_0\beta^0 + y_1\beta^1 + \dots y_{p-1}\beta^{-(p-1)}\right)\beta^{e_y}$$

Assume $e_y = e_x - 1$

$$\beta^{e_y} = \beta^{e_x}\beta^{-1}$$

$$y = \pm\left(y_0\beta^{-1} + y_1\beta^{-2} + \dots y_{p-1}\beta^{-(p-2)}\right)\beta^{e_x}$$

$$x+y = \pm\left(x_0\beta^0 + (x_1+y_0)\beta^{-1} + \dots\right)\beta^{e_x}$$

Main issue: $\mathbb{F}$ (set of floating point numbers) is not closed, e.g.,

$\Rightarrow$ for $x, y \in \mathbb{F}$, $x+y \notin \mathbb{F}$

Solution: design operations s.t.

$$fl(x+y) = (x+y)(1+\delta)$$
$$fl(x-y) = (x-y)(1+\delta)$$
$$fl(x*y) = (x*y)(1+\delta)$$
$$fl(x/y) = (x/y)(1+\delta)$$
$$\text{w/ } |\delta| \leq \varepsilon_{mach}$$

- Can still have issues ....

Rounding and catastrophic cancellation

Let $x, y \in \mathbb{R} \Rightarrow$ consider $fl(x-y)$

Let $\hat{x} = (1+\delta_x)x$
$\hat{y} = (1+\delta_y)y \Longrightarrow$

$$fl(x-y) = fl(\hat{x}-\hat{y})$$
$$= (x - y + \delta_x x - \delta_y y)$$

$|\delta| \leq \varepsilon_{mach}$

$$\frac{|fl(\hat{x}-\hat{y}) - (x-y)|}{|x-y|} = \frac{|\delta_x x - \delta_y y|}{|x-y|}$$

$$\leq \frac{|\delta_x||x| + |\delta_y||y|}{|x-y|} \leq$$

$$\leq \underbrace{\max\{\delta_x, \delta_y\}}_{\leq \varepsilon_{mach}} \underbrace{\frac{|x| + |y|}{|x-y|}}$$

Error can be large if $x \approx y$ or if $x, y \gg 1$!!