# Math 551 2/18/21

- HW #2 due 2/25 @ 9PM
  Th
- Read 3.1 - 3.3 for Tuesday
  Find $x$ such that $\underbrace{f}_{\text{non-linear}}(x) = 0$
- $\boxed{\text{Look over 2.1 - 2.3}}$

# Condition and Stability: Evaluate $f(x)$

Condition number of $f$ at $x$

$$K = \max_{|x-x^*|} \frac{Rel(f(x))}{Rel(x)} \approx \max_{x} \left| \frac{x \cdot f'(x)}{f(x)} \right|$$

$\Rightarrow$ $K = O(1)$, well-conditioned

$K \gg 1$, Ill-conditioned

Ex: $f(x) = \sqrt{x+1} - \sqrt{x}$, uniformly well-conditioned

$$\boxed{K \le \frac{1}{2}}$$

$$\boxed{\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}}$$

$$\left( \frac{\sqrt{x+1} + \sqrt{x}}{\sqrt{x+1} + \sqrt{x}} \right)$$

(1) $F(x) = \sqrt{x+1} - \sqrt{x}$ 4-digit arithmetic

UNSTABLE

$$F(1984) = \sqrt{1985} - \sqrt{1984}$$

Subtractive Cancellation

$$\approx \boxed{0.4455 \times 10^{2}} - \boxed{0.4454 \times 10^{2}}$$

$$= 0.0001 \times 10^{2} = 0.\underbrace{1}_{\text{1-digit}}000 \times 10^{-1}$$

$$\boxed{\text{MATLAB: } \sqrt{1985} - \sqrt{1984} = 0.112239\ldots \times 10^{-1}}$$

(2) $F(x) = \dfrac{1}{\sqrt{x+1} + \sqrt{x}}$

$$F(1984) = \frac{1}{\sqrt{1985} + \sqrt{1984}} \approx \frac{1}{0.4455 \times 10^{2} - 0.4454 \cdot 10^{2}}$$

$$= \frac{1}{0.8909 \times 10^{2}} \approx 0.\underbrace{1122}_{\text{4-digits}} \times 10^{-1}$$

STABLE

Evaluate $F(x)$

Well-conditioned

Ill-conditioned

Algorithm

STABLE

UNSTABLE

Ex: $ax^2+bx+c=0 \Rightarrow x_{+,-} = \dfrac{-b \pm \sqrt{b^2-4ac}}{2a}$

$a = 0.2$

$b = -47.91$ $\xrightarrow{\text{MATLAB}}$ $x_+ = \underbrace{239.4}247\ldots$

$c = 6$ $\qquad x_- = 0.\underline{1}253\ldots$

4-digit arithmetic

$$x_{+,-} \approx \frac{47.91 \pm \sqrt{(-47.91)^2 - 4(0.2)\cdot 6}}{2(0.2)}$$

$$\approx \frac{47.91 \pm \sqrt{2290}}{0.4}$$

$$\approx \frac{47.91 \pm 47.85}{0.4} = \begin{cases} \underbrace{239.4}_{\text{4-digits}} & (+) \\ 0.\underbrace{1}_{\text{1-digit}}500 & (-) \end{cases}$$

Subtractive cancellation

Fix?

$$\frac{-b-\sqrt{b^2-4ac}}{2a}\cdot\frac{-b+\sqrt{b^2-4ac}}{-b+\sqrt{b^2-4ac}} = \boxed{\frac{2c}{-b+\sqrt{b^2-4ac}}}$$

$x_-$

$$x_- \approx \frac{2\cdot 6}{47.91-\sqrt{(-47.91)^2-4\cdot(0.2)\cdot 6}}$$

$$\underbrace{=}_{\text{4-digits}} 0.\underbrace{1253}_{\text{Full 4-digit accuracy!}}$$

# Floating Point Number Systems: $Fl(x)$

$x \in \mathbb{R}$, base $\beta$ (typically 2), $n$

$$Fl(x) = \pm (0.b_1 b_2 b_3 \cdots b_n)_\beta \cdot \beta^e$$

(*) $0 \le b_i \le \beta - 1$

(*) $b_1 \ne 0$

$\left( \begin{array}{ll} \beta = 2 & \{0, 1\} \\ \beta = 10 & \{0, 1, 2, \ldots, 9\} \end{array} \right)$

$\pm$ : sign

$(0.b_1 b_2 \cdots b_n)_\beta$ : mantissa

$e$ : exponent

$(-m \le e \le m)$

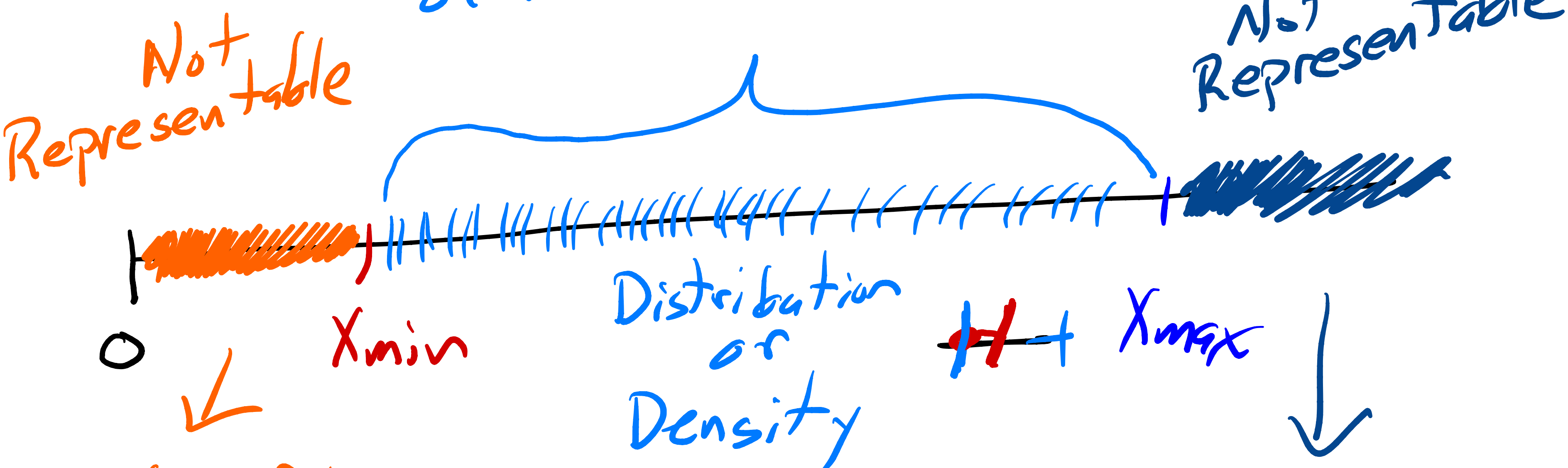
Ex: $n = 4 \quad \beta = 2 \quad m = 3$

$$x_{min} = (0.1000)_2 \times 2^{-3} = 2^{-4} = (0.0625)_{10}$$

$$x_{max} = (0.1111)_2 \times 2^{3} = (111.1)_2 = (7.5)_{10}$$

---

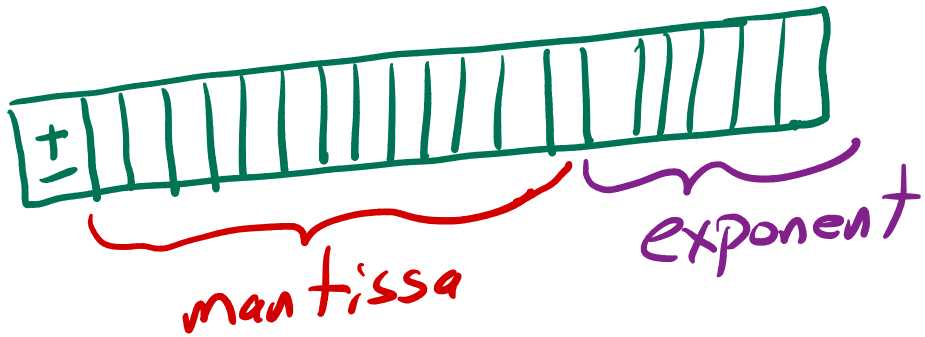
Only a FINITE number of numbers are representable




underflow

overflow

# IEEE 754 Standard    Fl(x)




single-precision (32-bits)    (1, 23, 8)

double-precision (64-bits)    (1, 52, 11)

quadruple-precision (128-bits)    (1, 112, 4)