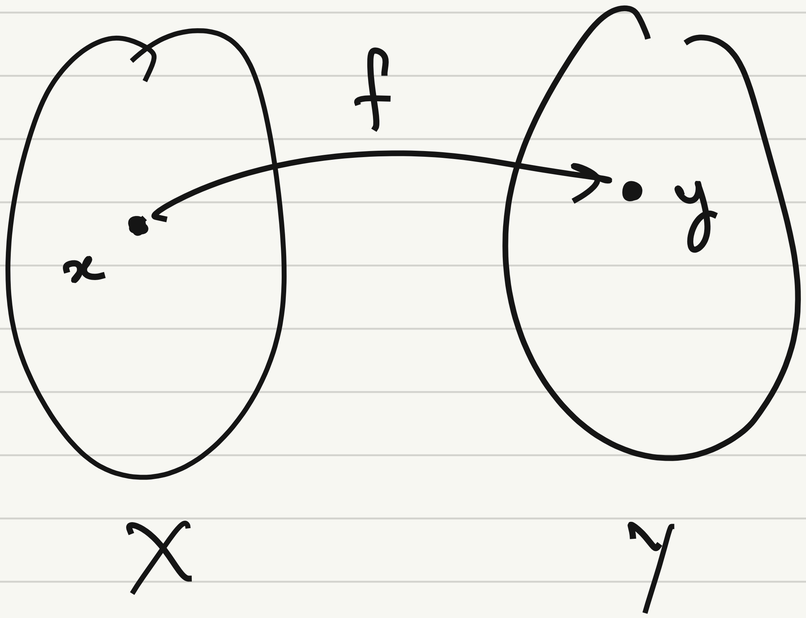
Floating-Point Numbers, Accuracy, Stability

Review



Condition number : $\sup_{\delta x} \dfrac{\text{error in } f(x)}{\text{error in } x}$

$$\hat{\kappa} = \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|}$$

$$\kappa = \sup_{\delta x} \frac{\|\delta f\|/\|f\|}{\|\delta x\|/\|x\|} = \frac{\|x\|}{\|f\|}\hat{\kappa}$$

$$f(x_1, x_2) = f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = x_1 - x_2 \quad : \quad \hat{\kappa} = \sup_{\delta x} \frac{2\delta}{\delta} = 2 \qquad \text{in } \|\cdot\|_\infty$$

$$\kappa = \frac{\|x\|}{\|f\|} \cdot 2 = \frac{\max(|x_1|, |x_2|)}{|x_1 - x_2|} \cdot 2$$

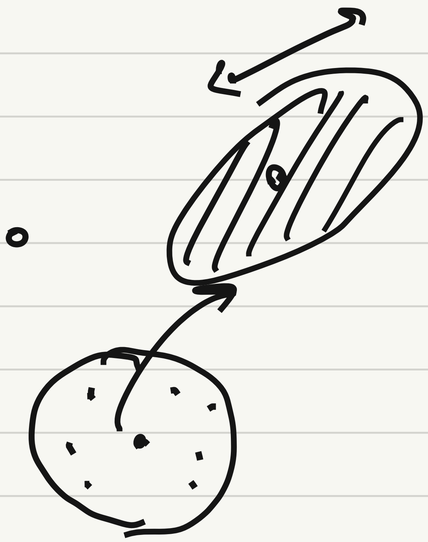
Roots of polynomials : $f(a_0, a_1, \dots, a_n) = (z_1, z_2, \dots, z_n)$

$$p(z) = \prod_{i=1}^{20} (x - i) = (x-1)(x-2)\cdots(x-20)$$

$$= a_{20}x^{20} + a_{19}x^{19} + \cdots + a_1 x^1 + a_0$$

$$K \approx 5 \times 10^{13}$$

$$\frac{\|\delta x\|}{\|x\|} = 10^{-10} \longrightarrow \frac{\|\delta f\|}{\|f\|} \leq 5 \times 10^{3}$$



$$A = \begin{bmatrix} 3 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad f(x) = Ax = \begin{bmatrix} 3x_1 \\ 0.5x_2 \end{bmatrix}$$

Show That $\kappa(x) = \dfrac{3\|x\|}{\|Ax\|}$

whether $\|x\| = \max(|x_1|, |x_2|)$

or $\|x\| = \sqrt{x_1^2 + x_2^2}$

Generalize to $A = \begin{bmatrix} a_{11} & & \\ & a_{22} & \\ & & \ddots \\ & & & a_{nn} \end{bmatrix}$

Floating-point arithmetic

IEEE 754:

Single precision: 32 bits (c/c++/Java: float)

Double precision: 64 bits (c/c++/Java: double. Python: float)

Floating point ≈ Scientific notation

$2022 = \underbrace{2.022}_{} \times 10^3$

floating point number : $\pm M \times \beta^{E}$ or 0

β : base or radix (here 2)

$M = d.ddd\ldots d$ or $0.ddd\ldots d$

M : mantissa

p digits

p : precision

E : exponent

double precision ($\beta = 2$)

	±	E	M	
64 bits	1	11	52	p = 53
	1	8	23	p = 24

single

$1.dd\ldots d$

p digits

$2^{-128} - 2^{128}$

$10^{-38} \sim 10^{38}$

In our course: Idealized floating point #s

$\beta \in \mathbb{N}$, fixed p

$M = d.dd\ldots d$ (p digits), $E \in \mathbb{Z}$

F = set of all floating-point numbers

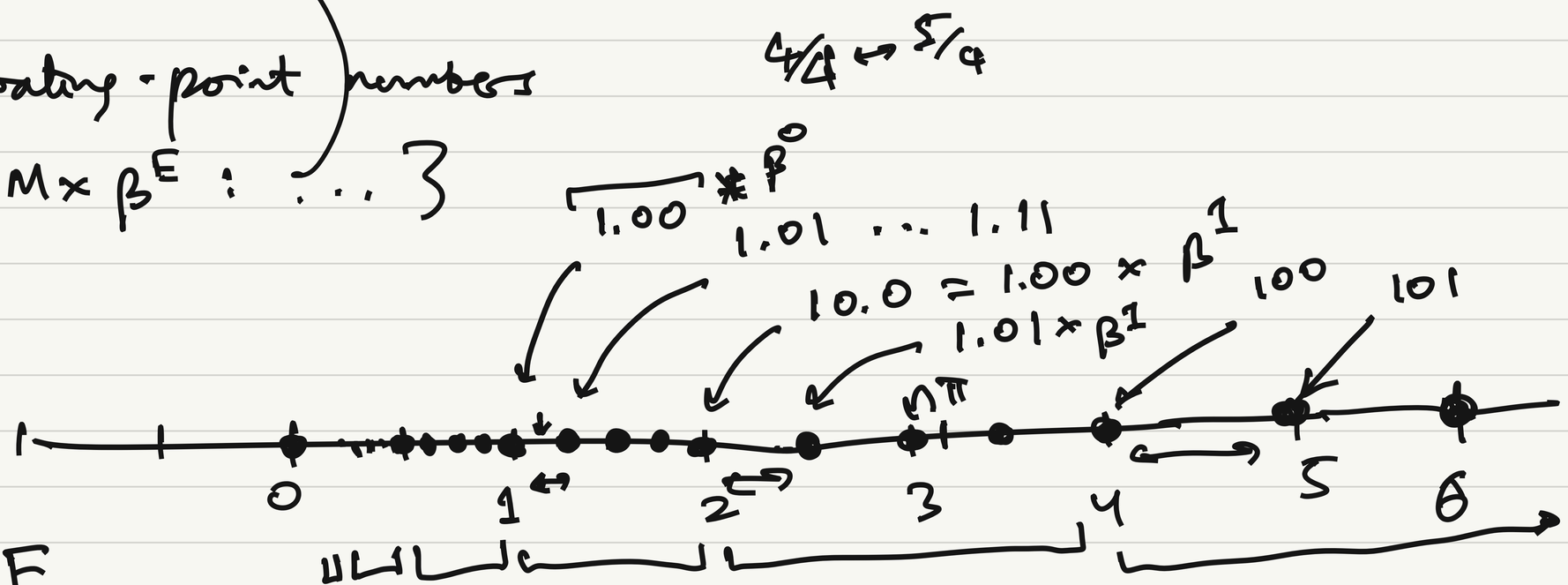
$= \{0\} \cup \{\pm M \times \beta^E : \ldots\}$

$\beta = 2, \quad p = 3$

$x \in \mathbb{R}$

$fl(x) \qquad fl: \mathbb{R} \to F$

Goldberg, "What Every CS should know about floats"



What is the largest relative error between x and $fl(x)$?

largest rel. error is between 1 and next float $= 1 + \beta^{(1-p)}$

β, p

$\underbrace{d.ddd\ldots d}_{p \text{ digits}}$

$x = \frac{1}{2}(1 + (\ldots))$

$1.00\ldots0$
$1.00\ldots1$

max error $= \frac{1}{2}\beta^{1-p} =: \varepsilon_{machine}$

machine epsilon → Single precision $\varepsilon_m = 2^{-24} \approx 6 \cdot 10^{-8}$

Double: $\varepsilon_m = 2^{-53} \approx 10^{-16}$

for any $x \in \mathbb{R}$, $|\underbrace{fl(x) - x}| \leq \varepsilon_m |x|$

$\Leftrightarrow$ $\underbrace{fl(x) = x(1+\varepsilon) \text{ for some } \varepsilon \text{ with } |\varepsilon| \leq \varepsilon_m}_{\text{for all } x \in \mathbb{R}}$ ← (1)

$x, y \in F$, $\quad x+y,\ x-y,\ x \cdot y,\ x/y \notin F$

$$x \circledast y = fl(x * y)$$

$$\boxed{x \circledast y = (x * y)(1+\varepsilon) \quad \text{for some } |\varepsilon| \leq \varepsilon_m} \longrightarrow (2)$$

for all $x, y \in F$

ε_m : measure of <u>worst-case relative error</u> introduced by floating point system

$x, y, z, \in \mathbb{R}$, compute $x - yz$

$\tilde{x} = fl(x) = x(1+\varepsilon_1)$ for some $|\varepsilon_1| \leq \varepsilon_m$

$\tilde{y} = fl(y) = y(1+\varepsilon_2)$

$\tilde{z} = fl(z) = z(1+\varepsilon_3)$

$\tilde{y} \otimes \tilde{z} = (\tilde{y} \cdot \tilde{z})(1+\varepsilon_4) = yz(1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_4)$

$\tilde{x} \ominus (\tilde{y} \otimes \tilde{z}) = (\tilde{x} - \tilde{y} \otimes \tilde{z})(1+\varepsilon_5) = \cdots$

$= \cdots = x - yz + \cdots$

$yz(1 + \underbrace{\varepsilon_2 + \varepsilon_3 + \varepsilon_4}_{|\varepsilon_2+\varepsilon_3+\varepsilon_4| \leq 3\varepsilon_m} + \underbrace{\varepsilon_2\varepsilon_3 + \cdots}_{= yz + \underbrace{yz\varepsilon_6}_{|\varepsilon_6| \leq 3\varepsilon_m}})$

① $fl(x) = x(1+\varepsilon)$ for some $|\varepsilon| < \varepsilon_m$

② $x \circledast y = (x * y)(1+\varepsilon)$ for some $|\varepsilon| \leq \varepsilon_m$

$0.1 + 0.1 + 0.1 \neq 0.3$

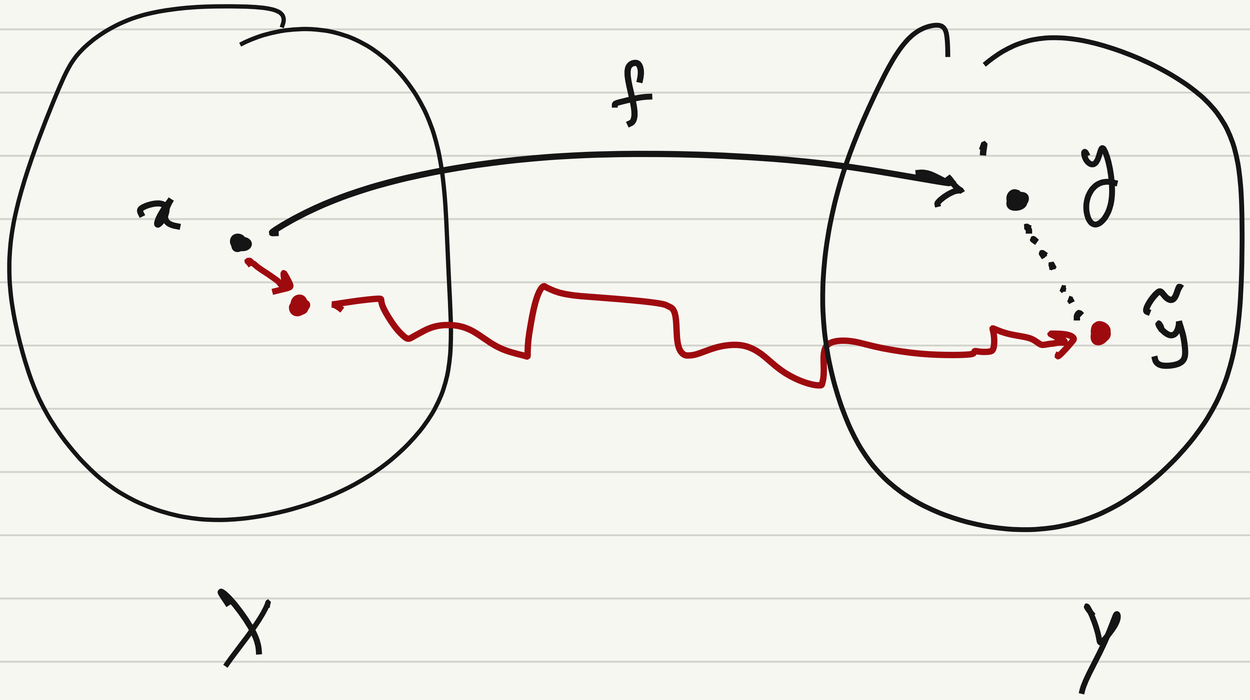
$a \oplus b = b \oplus a$ but $a \oplus (b \oplus c) \neq (a \oplus b) \oplus c$

$$\begin{array}{r} 1.2345\overbrace{0000} \\ - \underline{1.2321\ldots} \\ 0.0024 = \overbrace{2.4000} \times 10^{-3} \end{array}$$

$p(x) = (x-10)^6$ ←

$= x^6 - 6 \times 10 x^5 + 15 \times 10^2 x^4 - \ldots + 10^6$ ←

Accuracy and stability of algorithms



$$\tilde{f}(x) = \tilde{y}$$

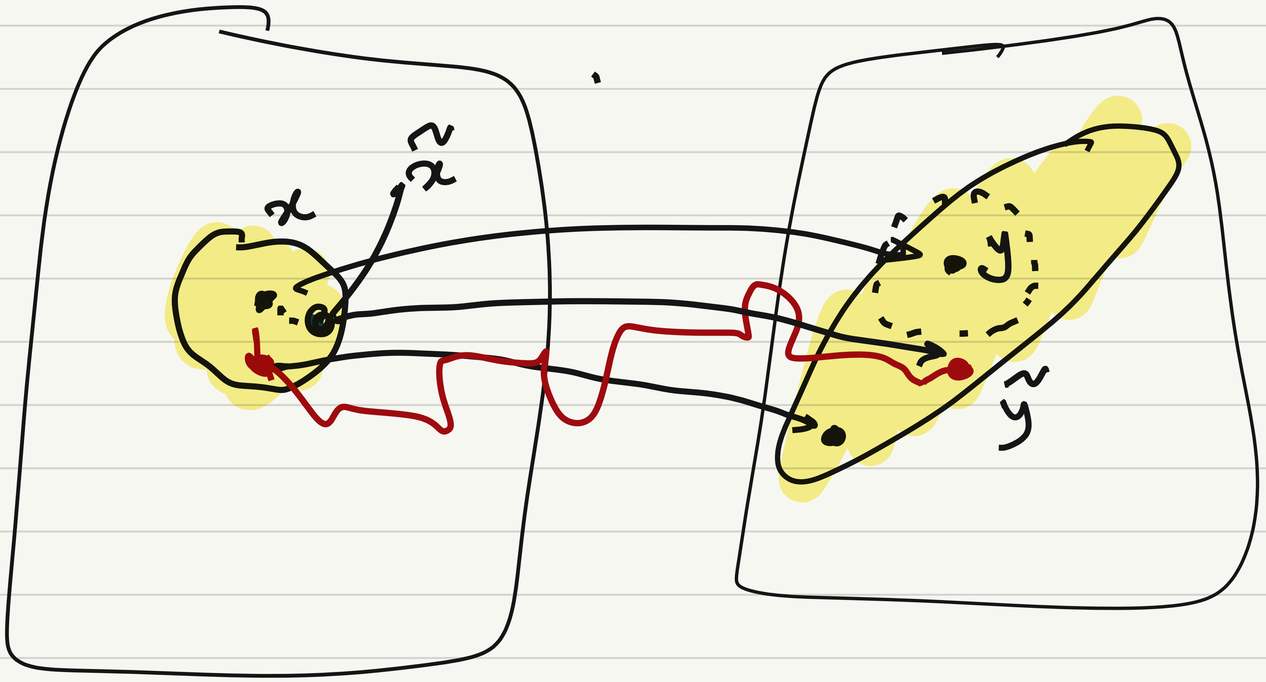
Abs. forward error $= \|\tilde{y} - y\|$

Rel. " " $= \dfrac{\|\tilde{y} - y\|}{\|y\|}$

Alg. is **accurate** if

rel. f.e. is "small" for all inputs

r.f.e. $= O(\varepsilon_m)$



Computed $\tilde{y} = \hat{f}(x)$

Suppose there exists $\tilde{x}$

s.t. $f(\tilde{x}) = \tilde{y}$

Abs. backward error $= \|\tilde{x} - x\|$

Rel. b. " $= \dfrac{\|\tilde{x} - x\|}{\|x\|}$

Alg. is **backward stable** if r.b.e. is small for all x

r.b.e. $= O(\varepsilon_m)$

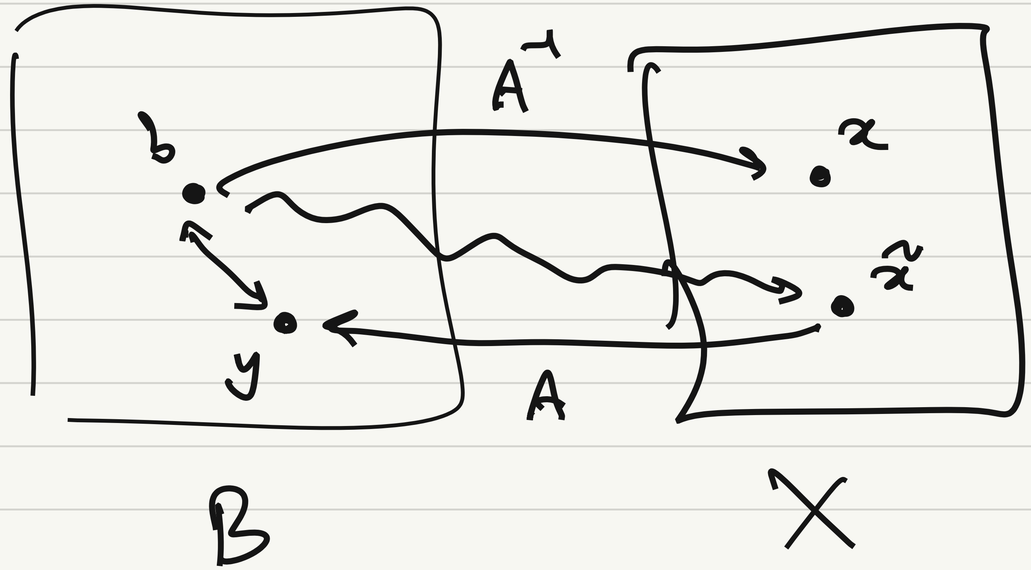
Alg. gives <u>exactly</u> the right <u>answer</u> to <u>nearly</u> the right <u>question</u>

Solve $Ax = b$.

$f(A, b) = A^{-1}b$

$\tilde{f}(A, b) = \tilde{x}$

$A\tilde{x} \to y$ $\quad\quad$ $\|b - A\tilde{x}\|$: b.e.



error

$= O(\varepsilon_m)$ means $\exists$ const. C s.t. for all input x,

$$\text{error} \leq C\varepsilon_m \quad \text{as} \quad \varepsilon_m \to 0$$