LECTURE

LAST TIME

- ☐ SUMMARY OF GRAM SCHMIDT ALGORITHM.

- ☐ STABILITY OF A FUNCTION $f: \mathbb{C}^n \to \mathbb{C}^m$

  - ERRORS: NOISE, TRUNCATION, ROUNDING

  - CONDITION NUMBER

$$\delta f = f(x+\delta x) - f(x)$$
$$f : f(x).$$

$$\kappa(x) \cdot \lim_{\delta x \to 0} \sup_{\delta x} = \frac{\|\delta f\| / \|f\|}{\|\delta x\| / \|x\|}$$

$\Rightarrow$

RELATIVE OUTPUT ERROR (at point x)

$= \kappa(x)$ RELATIVE INPUT ERROR

- IF f IS DIFFERENTIABLE $\kappa(x) = \dfrac{\|J(x)\| \; \|x\|}{\|f(x)\|}$

- IF f IS LINEAR ($f = Ax$) $\kappa(x) = \dfrac{\|A\| \, \|x\|}{\|Ax\|} \leq \|A\| \, \|A^{-1}\|$

- CONDITION NUMBER OF A SQUARE INVERTIBLE MATRIX:
$\kappa(A) = \|A\| \, \|A^{-1}\|$. FOR THE 2-NORM $\kappa_2(A) = \dfrac{\sigma_1}{\sigma_m}$

$\kappa(A) \geq 1$. (SHOULD BE AN EASY EXERCISE) FOR ANY NORM where $\sigma_{max} = \sigma_1$; $\sigma_{min} = \sigma_m$

THE LARGEST AND SMALLEST SINGULAR VECTORS.

- COROLLARY : $K_2(A) \geq 1$.

TODAY

- FLOATING POINT ARITHMETIC.
- STABILITY TO ROUNDING
- BACKWARD ERROR

[Chapters 13, 14, 15.]

A NOTE OF THE SHARPNESS OF THE CONDITION NUMBER

- CONSIDER $A = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix}$, $0 < \epsilon < 1$.

THEN BY DEFINITION, $K = \frac{1}{\epsilon}$.

Let $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Let $\delta x = \begin{bmatrix} \delta \\ 0 \end{bmatrix}$

THEN $\delta f = \begin{bmatrix} \delta \\ \epsilon \end{bmatrix} - \begin{bmatrix} 0 \\ \epsilon \end{bmatrix} = \begin{bmatrix} \delta \\ 0 \end{bmatrix}$; $f = \begin{bmatrix} 0 \\ \epsilon \end{bmatrix}$

$\|\delta f\|_2 = \delta$; $\|f\|_2 = \epsilon$; $\|x\| = 1$; $\delta x = \delta$.

$\frac{\|\delta f\|}{\|f\|} = \frac{\delta}{}$; $\frac{\|\delta x\|}{\|x\|} = \delta$; $\frac{\|\delta f\|/\|f\|}{\|\delta x\|/\|x\|} = \frac{1}{\epsilon}$.

FLOATING POINT NUMBERS

- Normalized Decimal REPRESENTATION

$$x = \pm S \times 10^{E}$$

$$S = J.d_1 \dots d_{p-1} \qquad d_i : 0 \dots 9 \qquad J : 1, \dots 9 \quad (J \neq 0)$$

e.g. $43.501 = 4.3501 \times 10^{1}$

$-0.0134 = -1.3400 \times 10^{-2}$

- NORMALIZED BINARY REPRESENTATION

$$x = \pm m' \times 2^{E} \rightarrow \text{Exponent}$$

$$m' = 1.b_1 b_2 \dots b_{p-1} \qquad b_i = \{0, 1\}$$

↑ mantissa or fraction or significant

p: precision.

Example: $x = -1.01 \times 2^{0101}$

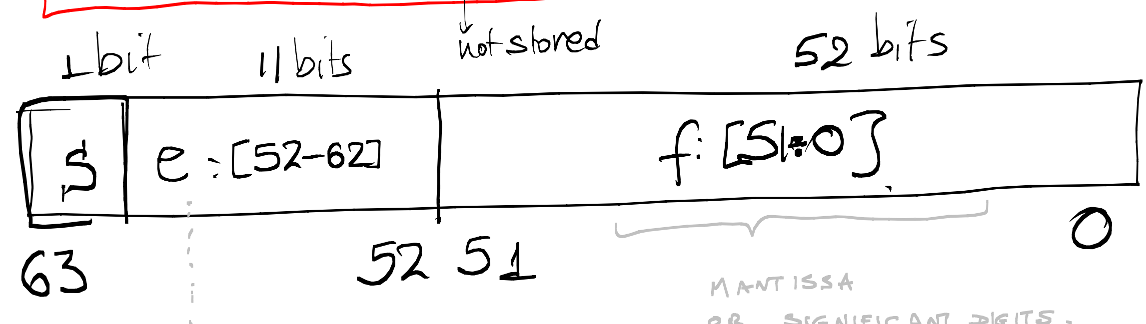
$$= -(1 \times 2^{0} + 0 \times 2^{-1} + 1 \times 2^{-2}) \times 2^{(0 \times 2^{3} + 1 \times 2^{2} + 0 \times 2^{1} + 1 \times 2^{0})}$$

(in base 10) $= -(1 + \frac{1}{2^2})\, 2^{5} = -40$

IEEE-754 FLOATING POINT STANDARD.

- 64-bits to represent a number.
- WE CAN EXACTLY REPRESENT 2^{64} NUMBERS ONLY ($\sim 1.84e19$)
- $$\boxed{x = (-1)^{s} \times \underbrace{1}_{\text{not stored}}.f \times 2^{e-1023}}$$

FORMAT OF AN IEEE-754 NUMBER



sign

NOTE THAT "e" IS NOT THE EXPONENT. THE EXPONENT "E" IS $E = e - 1023$. "e" IS POSITIVE BUT E CAN BE POSITIVE & NEGATIVE

VALID RANGE FOR e

$$0 < e < 2047$$
$$\Downarrow$$
$$-1022 \leq E \leq 1023$$

; Notice that $2^{11} - 1 = 2047$ So not all bits of e ARE ALLOWED TO BE ONES. $\left(\sum_{i=0}^{10} 2^{10} = 2047\right)$

THE "e" = 1...1 IS USED TO CODE SPECIAL CASES. SEE TABLE BELOW.

BIT PATTERN	VALUE
$0 < e < 2047$	$(-1)^{s} \times 2^{e-1023} \times 1.f$
$e = 0$; $f \neq 0$	$(-1)^{s} \times 2^{-1022} \times 0.f$ // Subnormal numbers
$e = 0$; $f = 0$	$(-1)^{s} \times 0.0$ (signed zero)
$s = 0$; $e = 2047$; $f = 0$	INF
$s = -1$; $e = 2047$; $f = 0$	− INF
$e = 2047$; $f \neq 0$	NAN : not a number

powers of two : represented exactly (if in range)

$\pm(1.000\ldots)\,2^{e-1023}$

max number : $\sim 1.8e+308$ (larger numbers $\rightarrow$ overflow)

min number >0 : $\sim 2.2e-308$ (smallest numbers $\rightarrow$ underflow)

min subnormal >0 : $\sim 5e-324$

SPACING OF FLOATING POINT NUMBERS

- Choosing $f=0$ and $e=1023$ we get

$1 \rightarrow +(1.0000\ldots0)\times 2^{1023-1023} = +(1.000)\times 2^{0} = 1$

The next largest number is

$$+(1.000\cdots\underbrace{001}_{1\times 2^{-52}})\,2^{0} = 1+2^{-52}$$

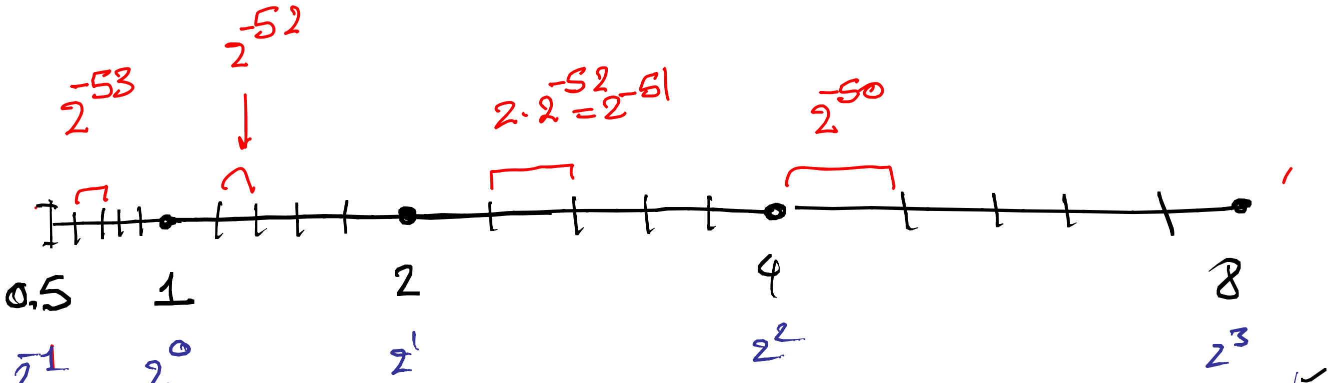
$$= 1+2.20446049e-16$$

$2 \rightarrow (1.000\ldots)\times 2^{1024-1023}$

The next largest number is

$$(1.00\cdots001)\,2^{1} = 2+\underline{2}\cdot 2^{-52}$$

Therefore the gap between floating point numbers in non-constant



POSITIVE FLOATING POINT NUMBERS.

SIDE-NOTE / HISTORICAL REFERENCE).

WM - KAHAN (UC BERKELEY) MAIN

DESIGNER OF IEEE-754.

RECEIVED TURING AWARD IN 1988.

DEFINITIONS.

PRECISION : number of bits for the mantissa

DOUBLE PRECISION : 53 bits $\approx$ 16 decimal digits.

SINGLE PRECISION : 24 bits.

Accuracy : THE NUMBER OF CORRECT DIGITS.

Machine epsilon: ϵ_M : $1+\epsilon_M$ is the next floating point number after 1.

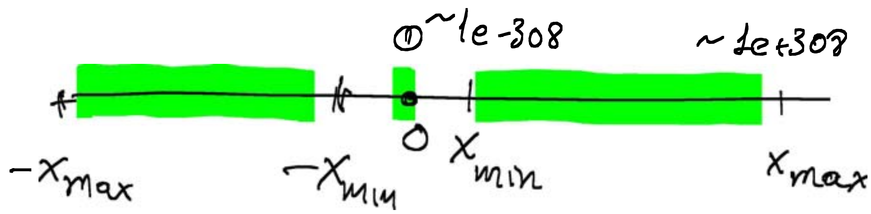
In IEEE-754 $\epsilon_M = 2^{-52} \sim 2.2e-16$

NOTATION

F_0 := SET OF ALL NORMAL FLOATING POINT NUMBERS

F := SET OF ALL FLOATING POINT NUMBERS (NORMAL + SUBNORMAL + (± INF)

- Range (F_0):



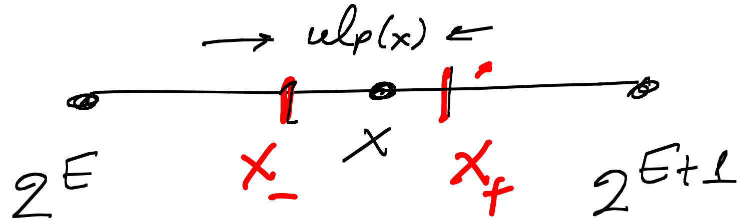
- ulp(x): gap in the power-2 interval that |x| belongs to.

THE EXPONENT IN THE REPRESENTATION OF "x".

$$\Rightarrow ulp(x) = (0.000..001) \times 2^{E} \Rightarrow$$

$$ulp(x) = \epsilon_M 2^{E}$$

- FOR ANY $x \in \mathbb{R}$ that belongs in the range of normal numbers,



$x_+ = x_- + ulp(x)$.

x_-: NEAREST $y \in F_0$: $y < x$; x_+: NEAREST $y \in F_0$: $y > x$

$\tilde{x}$: rounding of $x \in \text{Range}(F_0)$

Depending on the hardware settings

We will also use $[x]$ to indicate rounding.

$\tilde{x} = x_+$

$\tilde{x} = x_-$

$$\tilde{x} = \operatorname*{argmin}_{x_+, x_-} \left(|\tilde{x} - x_+|, |\tilde{x} - x_-| \right)$$

EXERCISE: SHOW THAT

$\forall x \in \text{Range}(F_0)$

$$\boxed{\frac{|x - \tilde{x}|}{|x|} \leq \epsilon_M}$$

$\Rightarrow$ relative rounding error is bounded by the machine epsilon.