ex) find $y = \cos x = f(x)$.

For this case, $\delta y \approx -\sin x \, \delta x$. Hence

$$\left|\frac{\delta y}{y}\right| \approx k \left|\frac{\delta x}{x}\right| \qquad k = \frac{x f'(x)}{f} = -x\tan x$$

$$\therefore \left|\frac{\delta y}{y}\right| = |x|\,|\tan x| \left|\frac{\delta x}{x}\right| = k \left|\frac{\delta x}{x}\right|$$

take $x \sim \frac{\pi}{2}$ for example. We see that $f(x)$ is highly sensitive to perturbations in x. //

CONDITION NUMBER OF MATRIX-VECTOR MULTIPLICATION

Take: $x \xrightarrow{Ax} y$ i.e. $y = Ax$

$A \in \mathbb{C}^{m \times n}$

$$k = \sup_{\delta x} \left(\frac{\|A(x+\delta x) - Ax\| \, \|x\|}{\|Ax\| \, \|\delta x\|} \right)$$

$$= \sup_{\delta x} \frac{\|A\delta x\|}{\|\delta x\|} \frac{\|x\|}{\|Ax\|}$$

$$\therefore \quad \kappa = \|A\| \frac{\|x\|}{\|Ax\|}.$$

If A is square & nonsingular

$$\frac{\|x\|}{\|Ax\|} \le \|A^{-1}\|$$

$$\therefore \quad \kappa = \alpha \|A\| \|A^{-1}\|$$

$$\alpha = \frac{\|x\|}{\|Ax\|} \Big/ \|A^{-1}\| \sim 1$$

If $A \in \mathbb{C}^{m \times n}$ use

$$\kappa = \alpha \|A\| \|A^{+}\|$$

//

<u>Thm</u>: Let $A \in \mathbb{C}^{m \times n}$ be nonsingular & $Ax = b$.
The problem of computing b, given x has condition number

$$\kappa = \|A\| \frac{\|x\|}{\|b\|} \le \|A\| \|A^{-1}\|$$

with respect to perturbarbations in x.

The problem of finding x, given b:

$$k = \|A^{-1}\| \frac{\|b\|}{\|x\|} \leq \|A\| \, \|A^{-1}\|$$

CONDITION NUMBER OF A MATRIX

$$k(A) \equiv \|A\| \, \|A^{-1}\|$$

if k is small ~ 1 "well-conditioned"
if k is very large "ill-conditioned"
(k can be ∞)

S'pse $\|\cdot\|_2$ then

$$\|A\|_2 = \sigma_1$$

$$\|A^{-1}\| = \frac{1}{\sigma_m}$$

$$\therefore \quad k(A) = \frac{\sigma_1}{\sigma_m}$$

it "measures" the ellipticity of the hyperellipse.

What happens when $A \to A + \delta A$?

How do we estimate the sensitivity

of outcomes to "errors" in A?

Let's fix b.

$$(A+\delta A)(x+\delta x) = b$$

$$Ax + \delta A x + A\delta x + \delta A \delta x = b$$

$$\delta A x + A \delta x + \underbrace{\delta A \delta x}_{\text{too small} \therefore \text{ remove}} = 0$$

$$\therefore \quad \delta A x = -A\delta x$$

$$\delta x = -A^{-1}(\delta A)x$$

Bound δx:

$$\|\delta x\| \leq \|A^{-1}\| \, \|\delta A\| \, \|x\|$$

Want $\frac{\delta x}{x} \Big/ \frac{\delta A}{A}$:

$$\frac{\frac{\|\delta x\|}{\|x\|}}{\frac{\|\delta A\|}{\|A\|}} \leq \|A^{-1}\| \, \|A\| = k(A)$$

$\therefore$ $k(A)$ captures sensitivity to errors in A

//

FLOATING POINT NUMBERS

Lecture 13 Finite-Precision Computing

Finite Precision Machine (double precision binary machine)

max positive number $\sim 1.79 \cdot 10^{308}$

min positive number $\sim 2.23 \cdot 10^{-308}$

resolution (largest gap) = $2^{-52} \sim 2.22 \cdot 10^{-16}$

representation (binary, octal, hex, decimal, etc)

the IEEE standard is binary (memory representation)

Floating Point Representation: means position of decimal (or binary point) "moves" (stored separately). The gaps between the numbers will scale in proportion to size of the numbers.

$fl(x)$ is the "floating point representation" of the number x

Representation: consists of 4 fields:

- β base
- t precision
- $e \in [L, U]$ exponent
- sign indicator

//

$$x = \pm\left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots \frac{d_t}{\beta^t}\right)\beta^e$$

$$= \pm\underbrace{0.d_1 d_2 \ldots d_t}_{\text{Normalized notation}}\ \beta^e$$

The normalized representation in decimal is $0.\ldots$. The normalized representation in binary is $1.\ldots$

$$0 \le d_i \le \beta - 1 \qquad i = 0, 1, \ldots, t$$

$$e \in [L, U]$$

ex) Normalized representation

100 in $\beta = 10$ $\quad 0.1 \cdot 10^3$

Hexadecimal

$\beta = 16$ $\qquad 0 \le d_i \le 15$

$d_i = 0, 1, 2, \ldots, 9, A, B, C, D, E, F$

$$(23A.D)_{16} = 2 \cdot 16^2 + 3 \cdot 16^1 + 10 \cdot 16^0 + 13 \cdot 16^{-1}$$
$$= 512 + 48 + 10 + 0.8125 = 570.8125$$

<u>Binary</u> $\beta = 2 \qquad 0 \le d_i \le 1$

ex) $(1101.0011)_2 = 10.1101011 \cdot 2^4$

$$= 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 0 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4}$$
$$= 13.1875$$

ex) Convert $\frac{2}{3}$ to binary:

$$\frac{2}{3} = (0.a_1a_2\ldots)_2$$

$\times 2$

$$\frac{4}{3} = (a_1.a_2\ldots)_2 \to a_1 = 1$$

subtract 1 from b.s.

$$\frac{4}{3} - 1 = \frac{1}{3} = (0.a_2\ldots)_2$$

$\times 2$

$$\frac{2}{3} = (a_2.a_3\ldots)_2 \quad a_2 = 0$$

$\times 2$

$$\frac{4}{3} = (a_3.a_4\ldots)_2 \quad a_3 = 1$$

etc

$$\frac{2}{3} = (0.1010\ldots)_2$$

IEEE 754-1985 Standard

	β	t	L	U	wordlength
HP (Half)	2	11	-14	15	16
SP (single)	2	24	-126	127	32
DP (double)	2	53	-1022	1023	64
QP (quad)	2	113	-16383	16384	128

Example SP 32 bit

sign 1 bit | exponent 8 bits | normalized mantissa 23 bits

first entry is an implied 1

$\underbrace{}_{\text{24 bit mantissa}}$

UFL (underflow) β^{L}

OFL (overflow) $\beta^{U+1}(1-\beta^{-t})$

In binary $x = (-1)^{s} q \cdot 2^{m}$

$s=0$ positive $s=1$ negative

$q = (1.f)$

implied (Normalized binary notation)

$m = e - U$

$m = e - 217$ (for 32-bit)

8-bit biased

Range $0 < e < 11111111 = 2^8 - 1 = 255$

the number 0 is reserved for ± 0

" " 255 " " " for NaN

$m = e - 127 \therefore m \in [-126, 127]$.

The smallest positive # $2^{-126} \sim 1.2 \cdot 10^{-38}$

The largest positive # $(2-2^{-23}) \times 2^{127} \approx 3.4 \cdot 10^{38}$

The least significant bit 2^{-23}

$$\text{Eps} = \frac{1}{2}\beta^{1-t} \quad \text{(rounding)}$$

Machine epsilon

$$\text{Eps}_{SP} = \frac{1}{2}2^{-23} = 2^{-24} \approx 6 \cdot 10^{-8}$$

$$\text{Eps}_{DP} = \frac{1}{2}2^{-53} \sim 1 \cdot 10^{-16}$$

The $\frac{1}{2}$ results from using rounding approximation scheme.