SFU
muraki

FRIDAY

- QUIZ TODAY - quiz sheets at back of room.
- MATLAB WARM-UPs - go through worksheets carefully due MON.

LAST DAY

- running Matlab's GE → "backslash" $\vec{x} = [A] \backslash \vec{b}$
- the GOOD, the BAD & the UGLY. → 1st computing report.
- finite-precision arithmetic leads to errors
  - we want to find $\vec{x}$ so that $[\ ]\vec{x} = \vec{b}$
  - so we use GE: $\vec{x}_{GE} = [A]\ \vec{b}$
  - if we then check $\vec{x}_{GE}$: $[A]\vec{x} - \vec{b} \equiv$ error $= 0$
  - expect these errors to as matrix size N (arithmetic)
  - demo #1 → for typical "GE" matrices ERROR GROWTH seems

    obs: $\text{rms}\left([A]\vec{x} - \vec{b}\right) = O\left(10^{-} - 10^{-}\right)$, $N = (10)$
  - demo #2 → there are small matrices ($N < 10$) for which the unusually

SFU
muraki

MATLAB FLOATING-POINT (from   )

sign → ± ← mantissa

$$\pm (d_1 . d_2 d_3 \dots d_N) * 10^p$$ ← exponent

- largest magnitude $\pm$ realmax $\sim 1.79 \times 10^{+308}$

  violating this is

- smallest magnitude $\pm$ realmin $\sim 2.22 \times 10^{-308}$

  violating this is

- estimate of # $10^{72} - 10^{87}$

UNAVOIDABLE ERRORS IN COMPUTER ARITHMETIC

- addition of floating point

$$a + b = (+D_a * 10^{P_a}) + (+D_b * 10^{P_b})$$

  - need to add   of   (say $P_a > P_b$)
  - example, say $P_b = P_a - 5$

	a.aaa	aaaa	aaaa	aaaa			$* 10$
+	.	bbb	bbbb	bbbb	bbbb	b	$* 10$

5 for $10^{P_b - P_a}$ last digits used

SFU
muraki

- can simulate smaller N on calculator
  - add $\left(\frac{1}{3} * 10^5\right) + \left(\frac{1}{7} * 10^3\right)$

```
>> (1/3)*1e5
ans =
      3.333333333333333e+04
```

keeping only 5 digits (IEEE std uses rounding*)

$$\begin{array}{rl} & 3.3333 \times 10^4 \\ + & \quad\ 3 \times 10 \\ \hline & 3.3476 \times 10^4 \end{array}$$

(*rounding occurs in binary.)

↳ if carry,

- multiplication has less of a last digit problem

$$\begin{array}{rl} & a.aaa \ldots \ldots aaaa \times 10^{P_a} \\ \times & 6.666 \ldots \ldots 6666 \times 10^{P_b} \\ \hline & c.ccc \ldots \ldots ccc\boxed{c} \times 10^{P_a+P_b} \end{array}$$

if carry ⟶ then

possible or

(similar for division)

- week 03 quiz - artificial truncation error

PRACTICE & BRING CALCULATOR

SFU
muraki

- subtraction is interesting for nearly equal numbers

$$\begin{array}{rl} & A.AAa\ aaaa\ \ldots\ aaaa \times 10^{Pa} \\ - & A.AA6\ 6666\ \ldots\ 6666 \times 10^{Pa} \\ \hline & 0.00c\ cccc\ \ldots\ cccc \times 10^{Pa} \\ & c.ccc\ cccc\ \ldots\ c \times 10^{Pa} \end{array}$$

↳ "padded" zeros

ANSWER contains only sig digits ( )

- in sequential calculations (a-b)*d → only sig digits (even though - digits)

- in sequential calculations, need to round-off at step

- NOTE: can "lose" lost digits by addition to a magnitude #

- example.

errorneous

$$\begin{array}{rl} & a.aaaa\ aaaa\ aaaa\ xxxx \times 10^{p} \\ - & 6.666\ 6666\ 666 \times 10^{p} \\ \hline & 6.666c\ cccc\ cccc\ ccc \end{array}$$

irrelevant.

SFU
muraki

ABSOLUTE vs RELATIVE ERRORS (p17)

- for x^* an approximated value for x (exact)

$$\text{absolute error} \equiv |\ -\ |$$

$$\text{relative error} \equiv \frac{|\ -\ |}{|\ |}$$

- what we really mean by finite-precision truncation error is really error of 10^{-16}

```
>> factorial(19)

ans =

   1.216451004088320e+17
```

- in last day's GE residual experiment, it might have been better to use error relative to rms($\vec{\ }$).

That is, $\text{relative residual error} = \text{rms}([A]\vec{x}-\vec{b}) \Big/ \text{rms}(\vec{\ })$

(but the experiment was designed to have rms($\vec{\ }$) ≈ 1

SIMPLE

STRATEGIES for REDUCING FINITE-PRECISION ERROR

- recall the calculus limit $\lim_{\theta \to 0} \frac{1-\cos\theta}{\theta} = -$

SFU
muraki

```
>> calcLim = @(th) (1-cos(th))./(th.^2)

calcLim =

  function_handle with value:

    @(th)(1-cos(th))./(th.^2)
```

1) there is an analytical "fix" for this –

$$\frac{1-\cos\theta}{\theta^2} = \frac{1}{2}\left(\frac{\theta/2}{\theta/2}\right)^2$$

2) reducing # of operations?

· how many operations to evaluate polynomial

$$p(x) = \sum_{k=0}^{N} c_k \cdot x^k$$

multiplications per term

additions

$$(\; + \; + \; + \; \ldots + \;) = \; + \; \underline{(\;\;)}$$

· NESTED EVALUATION (p24)

$$p(x) = (\cdots((c_N \cdot \; + c_{N-1}) \; + c_{N-2}) \; + \; \ldots \; + c_0)$$

additions & multiplications

SFU
muraki

3) respecting magnitudes.

$S = (1 \ldots) e^{17} + 1$

- SORTED SUMMATION (, but)
  - add , magnitude quantities
  - MatHab's sort command & abs.