Overview

- Error / cond. nr.
- FP

Example 2
- Interp
- MC
- Errors / cond.
- FP.

HW3 due

HW4 timeline

$$\text{Abs} = |\overbrace{x_0 - \tilde{x}}^{\Delta x}| = |\Delta x|$$

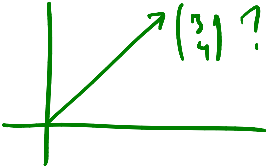
$$\text{Rel.} = \frac{\text{Abs.}}{|x_0|} = \frac{|\Delta x|}{|x_0|}$$

Measuring Error

Why is $\cancel{|\tilde{x}| - |x_0|}$ a bad measure of the error?

$|\tilde{x} - x_0|$ ← Yes.

If $\tilde{x}$ and x_0 are vectors, how do we measure the error?



Need equivalent to the abs. value:

"norm" $\left\|\begin{pmatrix}3\\4\end{pmatrix}\right\|_2 = \sqrt[2]{3^2+4^2}$

$\|\vec{x}_0 - \tilde{\vec{x}}\|_2$

Sources of Error

What are the main sources of error in numerical computation?

- Truncation error
- Rounding error

---

abs. error $= |x_0 - \tilde{x}| =$ "$\underline{\text{trunc.}}$ + $\underline{\text{rounding}}$"

$\tilde{x}$: both

$\boxed{\text{abs. error} = 10^{-5}}$

Scen. 1 $\quad x_0 = 10^{-6}$

Scen. 2 $\quad x_0 = 10^{0}$

1.000 0|0

1.0000|1

5 accurate digits

rel. error $= \dfrac{\text{abs. error}}{1} = 10^{-5}$

Digits and Rounding

Establish a relationship between '*accurate digits*' and rounding error.

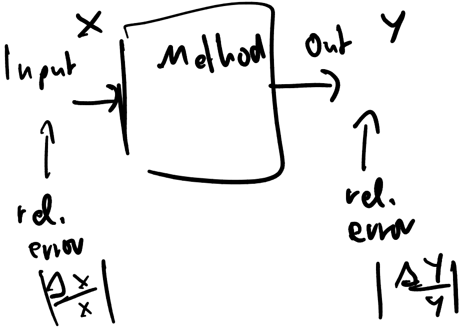
$3.14159\ldots$

finite precision $\rightarrow 3.14$ rounded to 3 digits

rel. error $\dfrac{|3.14159 - 3.14|}{|3.14159|} \approx 10^{-3}$

Condition Numbers

Methods f take input x and produce output $y = f(x)$.
Input has (relative) error $|\Delta x| \,/\, |x|$.
Output has (relative) error $|\Delta y| \,/\, |y|$.
Q: Did the method make the relative error bigger? If so, by how much?



$f(x) \to$ output

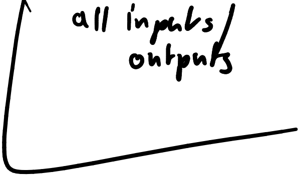
input

$$\text{Cond.} = \frac{\text{rel. error out}}{\text{rel. error in}} = \frac{|\Delta y| / |y|}{|\Delta x| / |x|}$$

most interested in: upper bound of CN for

$10^{-3} \longrightarrow 10^{-1}$

For this example:
$CN = 10^2$



"Good CN": small

"Bad CN": big (error gets bigger)

$$\text{Abs. CN} = \frac{|\text{Abs. error in out}|}{|\text{Abs. error in in}|} = \frac{|\Delta y|}{|\Delta x|}$$

nth-Order Accuracy

Often, *truncation error* is controlled by a parameter h.

Examples:

- distance from expansion center in Taylor expansions
- length of the interval in interpolation

A numerical method is called 'nth-order accurate' if its truncation error $E(h)$ obeys

$$E(h) = O(h^n).$$

Outline

Wanted: Real Numbers... in a computer

Computers can represent *integers*, using bits:

$$23 = 1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = (\underline{10111})_2$$

16   8   4   2   1

How would we represent fractions, e.g. $\underline{23.625}$?

$23 = 1 \cdot 2^4 + \cdots + 1 \cdot 2^0$

$23.625 = 1 \cdot 2^4 + \cdots + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + \underline{1} \cdot 2^{-3}$

23.5

$2^{-3} = .125$

$23.625 = (10111.101)_2$

Fixed-Point Numbers

32

> Suppose we use units of 64 bits, with 32 bits for exponents $\geqslant 0$ and 32 bits for exponents < 0. What numbers can we represent?

32

0 0 0 0 0 0 0 1 1

2^{31} 2^{0} 2^{-1} 2^{-32}

Biggest: $(111111\ldots1.111\ldots)_2$

> How many 'digits' of relative accuracy (think relative rounding error) are available for the smallest vs. the largest number?

$2^{-31} \leftarrow 2^{-32} \rightarrow$ smaller 0

bigger

$(111.111)_2 \rightarrow (1111.1110)_2$

Smallest: $(0.\ldots001)_2 = 1\cdot 2^{-32}$

actual result: 2^{-32}

computed result: 2^{-31}

$\hookrightarrow$ rel. error = $\dfrac{|2^{-32} - 2^{-31}|}{2^{-32}} = \dfrac{2^{-32}}{2^{-32}}$

In fixed point:

uneven relative error.

big results: small rel. error

small results: big rel. error

Floating Point numbers

Convert $13 = (1101)_2$ into floating point representation.

$$13 = \underline{(1.101)_2} \cdot \underline{2^{\underbrace{3}}}$$

$$= (1101)_2 = (110.1)_2 \cdot 2 = (11.01)_2 \cdot 2^2$$

What pieces do you need to store an FP number?

$$\underline{1.1100\,11 \,-}\; 2^{32}$$

$$(1.11000\,111)_2 \cdot 2^{-32}$$

In-class activity: Floating Point

Unrepresentable numbers?

Can you think of a somewhat central number that we cannot represent as

$$x = (1.\_\_\_\_\_\_\_\_\_)_2 \cdot 2^{-p}?$$