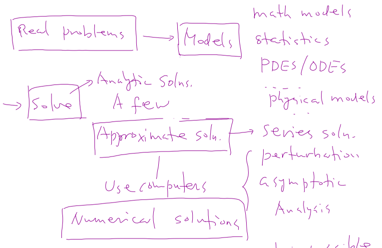
MA402: Computational Mathmatics

Summer session. 1.5 - 2 credits.

Motivations. 1. Get credits towards graduation.

2. Solve practical problems: A flow chart



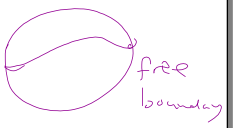
$\int_0^1 e^x \sin x^2 \log(x+1)dx$ Almost impossible to find analytic Soln.

It is easy by computers

Jul 17-1:55 PM

Computational Math can also give theoretical proof or clues.

Hayk Milkayelan. Conjecture.



Numerical solutions → Interpret soln. Graphs, tables etc. → feedback

Improve original models

Problem solving process

Jul 17-2:22 PM

A computer number system.

Does a+b+c always equal c+b+a ?

It may not be always true in a computer number system.

Preparation. Floating point arithmetic

$x \in R$, $\xrightarrow[\text{in a computer}]{\text{represented}}$ $fl(x)$

$fl(x) = \pm . d_1 d_2 \cdots d_n \cdot \beta^s$ — exponential

sign mantissa (fraction) base

$\beta = 2$ binary natural for Computers

$\beta = 8$ octal ✓ for storage purpose

$\beta = 10$ decimal We are used to

$\beta = 16$ hexadecimal ✓

$8 = 2^3$
$16 = 2^4$

1	8	23
sign	exponential	fraction

32 bits single precision

$2^8 = 256$, $2^{256} \approx 10^{38}$

$x > 10^{38}$, over-flow

1034 5612

For the double precision

1	11	52
sign	exponential	fraction

64 bits. 16 digits.

Jul 17-2:27 PM

Device ← More efficient/difficult

Laptop
Iphone
Game

→ CPU → Assembly language

central processing unit

→ High level language

C, C++
Fortran
Pascal
Pathan
KudaA

Applications special languages

Computational Math

Statistics. Matlab
SAS
Splus, R
Maple
Mathematica
Comsl
Games

Each application has its own computer number system.

classify as single 8 / double 16 precision

→ More convinient less efficients.

Jul 17-2:44 PM

Examples of floating numbers.

$-.1011 \cdot 2^5$, binary

$.31416 \times 10^1 \approx \pi$

$\| \quad .031416 \times 10^2$ decimal

Is $fl(x)$ unique ? No

If we set $d_1 \neq 0$, then it is called the normalized floating number.

It is unique unless $x=0$

then $0 = .00 \cdots \beta^0$

Jul 17-2:55 PM

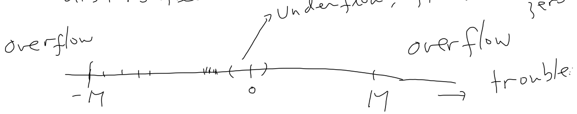
Properties of a computer number system

- Finite: The total numbers are

$2(\beta-1)\beta^{n-1}(S_{max} - S_{min} + 1) + 1$

$d_1 d_2$ exponential 10^{35}

- has maximum/minimum. Not evenly distributed



NaN Not available a is undefined $5/a$

Inf Infinity

Jul 17-3:00 PM

Some limits

single precision. 8 significant digits

0.12046011

$\pm 10^{38}$. Best accuracy we can expect $\underline{\underline{10^{-8}}}$

$3.1416271\underline{872}$
garbage

double precision 16 significant digits
e.g. Matlab

$\pm 10^{138}$ Error $\sim 10^{-16}$

Machine precision

Jul 17-3:06 PM

Error Analysis

π $fl(\pi) = 0.31416 \cdot 10^{1}$

$\pi - fl(\pi) \rightarrow$ rounding.

$$x = \pm .d_1 d_2 \cdots d_n d_{n+1} \cdots \times \beta^s$$

$$fl(x) = \pm \begin{cases} d_1 d_2 \cdots d_n \cdot \beta^s & \text{if } d_{n+1} < \beta/2 \\ d_1 d_2 \cdots (d_n+1) \cdot \beta^s & \text{if } d_{n+1} \geq \beta/2 \end{cases}$$

Error $|x - fl(x)| \leq \frac{1}{2}\beta^{s-n}$

Absolute error

Relative error. if $x \neq 0$

$$\frac{|x - fl(x)|}{|x|} \leq \underline{\frac{1}{2}\beta^{-n}} = \varepsilon$$

machine epsilon precision

$\varepsilon \approx 10^{-8}$ for single presissio

$\varepsilon = 10^{-16}$ for double pressio

Jul 17-3:11 PM

Computer errors are called "round-off errors"

$fl(x) = \underline{x(1+\delta)}$

$= x + \underline{\underline{x\delta}}$

Absolute Error $x\delta$

Relative error δ

$|\delta| \leq \varepsilon$ $= 10^{-8}$ $= 10^{-16}$

✓ $fl(x \cdot y) = xy(1+\delta)$ $|\delta| \leq \varepsilon$

$fl(x/y) = \frac{x}{y}(1+\delta)$ $|\delta| \leq \varepsilon$

$fl(x-y) = (x-y)(1+\delta)$ $|\delta| \leq \frac{|x|+|y|}{|x-y|}\varepsilon$

One rule of thumb

Try to avoid subtraction of two closed numbers.

Ex: $1 - \cos x$ if $|x|$ is very small.

$= 1 - (1 - \frac{x^2}{2}) + O(x^4)$

$= \frac{x^2}{2} + \underline{O(x^4)}$

Jul 17-3:32 PM

Write a Matlab program to solve

$ax^2 + bx + c = 0$

A. $x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$ $x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$

B. $x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$ $x_2 = \frac{c}{x_1}$

C. $x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$ $x_2 = \frac{c}{x_1}$

If $b > 0$. Use C otherwise use B.

Ex: $a = 1$, $x^2 + 2x + e = 0$, $e = 0$, $\begin{cases} x_1 = 0 \\ x_2 = -2 \end{cases}$

$x_1 = \frac{-2 + \sqrt{4 - e}}{2} \approx -1 + \sqrt{1 - \frac{e}{4}}$

$= -1 + (1 - \frac{e}{4})^{1/2} = -1 + 1 - \frac{e}{8} + O(e^2)$

$x_1 = \frac{e}{4}$ $x_2 = \frac{-2 - \sqrt{4 - e}}{2}$

$= -1 - \sqrt{1 - \frac{e}{4}}$

$= -1 - (1 - \frac{e}{4 \cdot 2}) = -2 + \frac{e}{8}$

Catastrophic cancellation

catastrophic

Jul 17-3:43 PM

Better code

$x^2 + bx + c = 0$
$a = 0$

```
function [y1, y2] = quad (b, c)
 If b > 0
   y1 = (-b - sqrt( b*b - 4*c) )/2;
   y2 =  c/y1;
 else
   y1 = ( -b + sqrt(          ))
   y2 = c/y1
 End
```

$x = \frac{-b + \sqrt{b^2 - 4c}}{2}$

Jul 17-4:02 PM