Lecture 1

Material covered: Up to Chapter 2 in the slides

Symbolic vs Numeric Computing

Symbolic: algebraic manipulation
- exact
- slow
- (takes functions ! returns functions)

Numerical: Approximate (if we think of real numbers)
- fast
- (takes numbers ! returns functions)

> Even with symbolic we usually need a number at the end

Packages

Symbolic	Numeric
SymPy	Julia
Mathematica	NumPy
Sage	Eigen
Magma	BLAS
	LAPACK

Overview

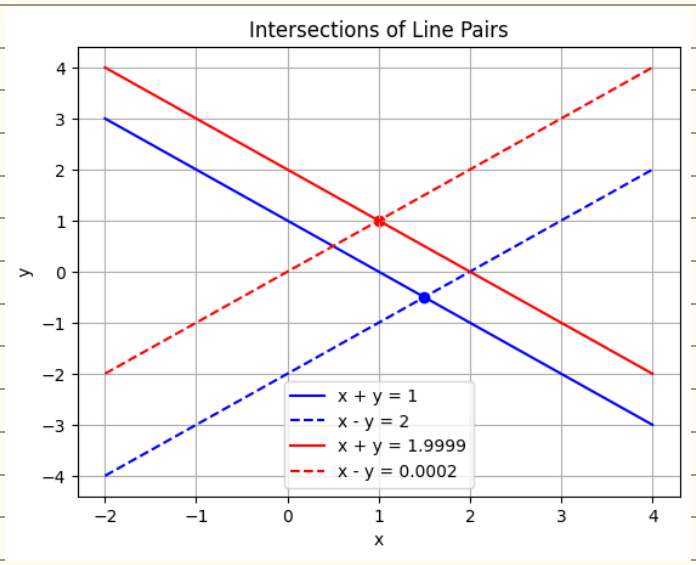
* Mathematical formulation
  ↓
  Algorithm (mostly concerned with this)
  ↓
  Implementation

* What is a "good" numerical algorithm?
  - stable (small change to input leads to small changes in output)
  - robust (not many assumptions)
  - fast to compute
  - easy to implement

* Sensitive / ill-conditioned problem
  - no method can work (so we usually change the problem)

Ex 1 $x+y=2$
$x-y=2$

$x+y=1.9999$
$x-y=0.0002$



Stable

$x=y=1$ $x=1.0005$
$y=0.99985$

Ex 2

$x + 0.99y = 1.99$

$0.99x + 0.98y = 1.98$

$x + 0.99y = 1.9899$

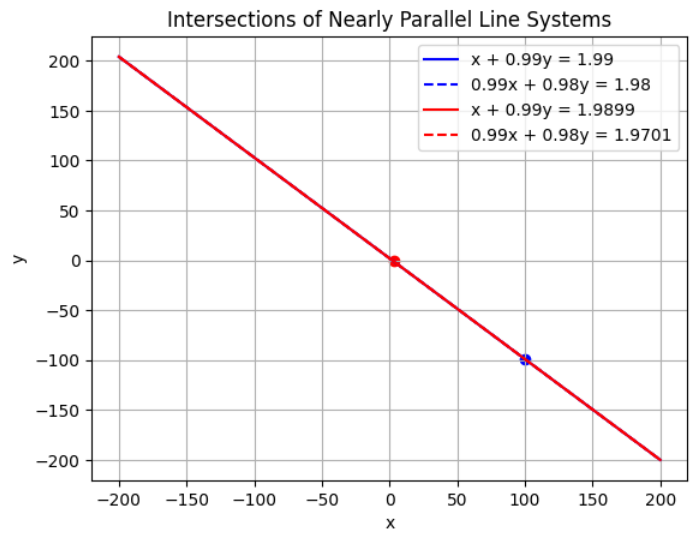
$0.99x + 0.98y = 1.9701$

$x = y = 1$

$x = 2.97$

$y = -0.99$

Unstable



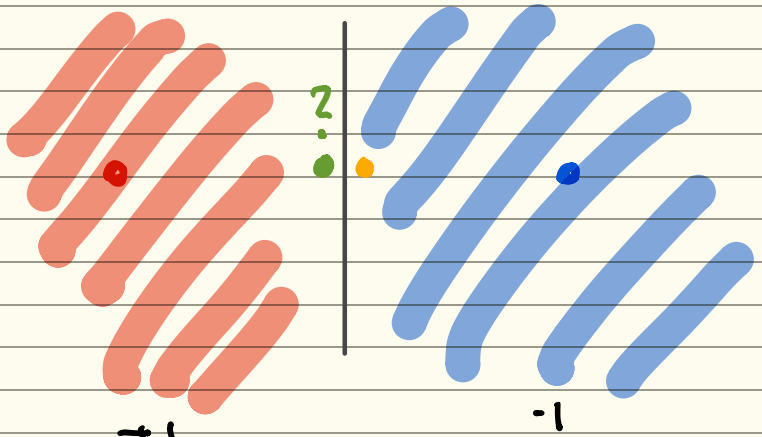
(There are 2 lines)

Ex 3

To which point are we closer to

red → +1

blue → −1



very close together ← { ● → −1, ● → +1

Representation of numbers

$$x = \pm \; 0.d_1 d_2 \ldots d_m \times \beta^e$$

↑ sign bit

β = base

e = exponent

$d_1 \ldots d_m$ - mantissa

If $\beta = 2$ (binary) $d_1 = 1$ (almost always)

when is d_1 not 1?

- Overflow vs underflow (see slides)

Alternative representation

$\frac{p}{q}$ where p & q are integers

Precision $\pm \frac{n}{2}$ (with n bits)

half for p !

half for q

* floating point is more flexible

Rounding Error

Closest number to $x \in \mathbb{R}$

$$fl(x) = \begin{cases} x_- & \text{largest representable number } x \geq x_- \\ x_+ & \text{smallest representable number } x \leq x_+ \end{cases}$$

$$\rightarrow \boxed{x_- \leq x \leq x_+}$$

Error

Assume $fl(x) = x_-$

$$x = (0.1 b_2 \ldots b_{m+1})_2 \times 2^e$$

$$x_- = (0.1 b_2 \ldots b_m)_2 \times 2^e$$

$$x_+ = x_- + 2^{-m} = \left((0.1 b_2 \ldots b_m)_2 + 2^{-m}\right) \times 2^e$$

Absolute error:

$$\max(x - x_-) = \frac{x_+ - x_-}{2} = 2^e \left(\frac{2^{-m}}{2}\right) = 2^{e-m-1}$$

Relative error

$$\frac{x - x_-}{x} \leq \frac{2^{e-m-1}}{\underbrace{2^{e-1}}} \leq 2^{-m}$$

x is certainly bigger than this, if $b_{m+1} = 0$ $(2^e / 2)$

Assume

$$u = 2^{-m}$$

$$x_- = x_- + x - x \geq -ux + x = x(1-u)$$

$$\Rightarrow x_+ \leq x(1+u)$$

$$\boxed{fl(x) \leq x(1+\delta) \quad |\delta| \leq u}$$

Computing with representations

Let $\odot$ be an operation

$$\odot \in \{+, -, \cdot, :\}$$

Operations in representations are not **associative**!

Example: addition

$$(a+b)+c \neq a+(b+c)$$

Cause: Adding two represented numbers does not result in a represented number

Claim: Add small numbers first (then large ones)

$$(a+b)+c = fl(fl(a+b)+c)$$
$$= fl((a+b)(1+\delta_1)+c)$$
$$= ((a+b)(1+\delta_1)+c)(1+\delta_2)$$
$$= ((a+b+c)+(a+b)(1+\delta_1))(1+\delta_2)$$
$$= (a+b+c)\left(1+\frac{a+b}{a+b+c}\delta_1\right)(1+\delta_2)$$
$$= (a+b+c)\left(1+\frac{a+b}{a+b+c}\delta_1+\delta_2+\frac{a+b}{a+b+c}\delta_1\delta_2\right)$$
$$\approx (a+b+c)\left(1+\frac{a+b}{a+b+c}\delta_1+\delta_2\right)$$

$\delta_1\delta_2 \ll \delta_1, \delta_2$

Likewise

$$a+(b+c) = (a+b+c)\left(1+\frac{b+c}{a+b+c}\delta_3+\delta_4\right)$$

if $a<b<c \Rightarrow$ $\boxed{\frac{a+b}{a+b+c} < \frac{b+c}{a+b+c}}$

Addition & Subtraction Error

Subtracting similar numbers

$a = 0.\overbrace{x\ldots x1}^{m}\ \overbrace{ssss}^{\text{loss to precision}} \cdot 10^e$

$b = 0.y\ldots y0\ tttt \cdot 10^e$ → difference may be lost

$a-b = 0.0\ldots 01$ error: 2^{-m+e}

Addition is ok but multiple addition errors accumulate (so many steps → problematic)

Example

Solving $x^2 + 2ax + b = 0 \quad a > 0 \quad b < a^2$

$$x_2 = \frac{-2a + \sqrt{4a^2 - 4b}}{2} = -a + \sqrt{a^2 - b}$$

Steps

$s_1 := a^2$

$s_2 := s_1 - b$

$s_3 := \sqrt{s_2}$

$s_4 := s_3 - a$

if $a^2 >> b$ s_4 can have a large error

(Possible)
Solution:

$$x_2 = -a + \sqrt{a^2 - b} \cdot \underbrace{\frac{a + \sqrt{a^2 - b}}{a + \sqrt{a^2 - b}}}_{1}$$

$$= \frac{-b}{a + \sqrt{a^2 - b}}$$

More steps $\Rightarrow$ avoid subtraction of similar terms

A common theme

$$f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right)$$

Bad if $\sqrt{x+1} \approx \sqrt{x}$

$$f(x) = f(x)\,\frac{\sqrt{x+1}+\sqrt{x}}{\sqrt{x+1}+\sqrt{x}} = \frac{x}{\sqrt{x+1}+\sqrt{x}}$$

Summation

$$\frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \cdots + \frac{1}{n(n+1)}$$

has a smaller error

than

$$\frac{1}{n(n+1)} + \cdots + \frac{1}{2\cdot 3} + \frac{1}{1\cdot 2}$$

Integration

$$I_n = \int_0^1 x^n e^{-x}\,dx$$

① $I_n = -\frac{1}{e} + nI_{n-1}$ $\quad I_0 = 1 - \frac{1}{e}$

② Backward: $I_{n-1} = \frac{1}{n} I_n + \frac{1}{ne}$

② is better than ① $\Rightarrow$ chose any value for I_n (n large) $\Rightarrow$ error is divided each step by n (becomes small very quickly)

General Analysis - addition / subtraction

$x, y \in \mathbb{R} \qquad \bar{p} \approx p = x + y$

$$\bar{p} = fl(fl(x) + fl(y)) = fl(x(1+\delta_1) + y(1+\delta_2))$$

$$= (x(1+\delta_1) + y(1+\delta_2))(1+\delta_3)$$

$$= x(1+\delta_1)(1+\delta_3) + y(1+\delta_2)(1+\delta_3)$$

$$= x + y + x(\delta_1 + \delta_3 + \delta_1\delta_3) + y(\delta_2 + \delta_3 + \delta_2\delta_3)$$

Relative error

$$\frac{|\bar{p} - p|}{|p|} = \frac{|x(\delta_1 + \delta_3 + \delta_1\delta_3) + y(\delta_2 + \delta_3 + \delta_2\delta_3)|}{|x+y|}$$

if $|x+y| \approx 0 \quad \frac{|\bar{p}-p|}{|p|} \gg 0$

Depends on size of $|x+y|$

Multiplication & Division

$$p = x \cdot y$$

$$\bar{p} = fl(fl(x) \cdot fl(y)) = fl(x(1+\delta_1) \cdot y(1+\delta_2))$$

$$= x \cdot y\,(1+\delta_1)(1+\delta_2)(1+\delta_3)$$

$$\approx x+y\,(1+\delta_1+\delta_2+\delta_3+\ldots)$$

product terms δ_i

$$|\delta_i| \leq u$$

Relative Error

$$\frac{\bar{p}-p}{p} = \frac{|x \cdot y|\,(\delta_1+\delta_2+\delta_3+O(u^2))}{|x \cdot y|}$$

$$= (\delta_1+\delta_2+\delta_3+O(u^2))$$

Product independent of size of $x \cdot y$