Lecture-01 Introduction to numerical computation

Lectures: Baojian Zhou
Email: bjzhou@fudan.edu.cn

Books:

- Numerical analysis (3rd), Timothy Sauer
- Matrix Computation (4th), Gene ....

Some notations:

$C(\mathbb{R})$: set of all functions that are continuous on $\mathbb{R}$.

$C'(\mathbb{R})$: set of all f' continuous on $\mathbb{R}$.

$C^n[a,b]$: $f^{(n)}$ exists and continuous.

$\|\cdot\|$: norm

$(a)_{10}$: decimal numbers

$(a)_2$: binary numbers

$fl(a)$: float point of a saved in comp.

problem 1: How to calculate $\sqrt{2}$.

Solution: Babylonian method

Let the numerical value of $\sqrt{2}$ be x.

$\Rightarrow \sqrt{2} \Rightarrow x = \sqrt{2}$

1. $x = \sqrt{2} \Leftrightarrow x^2 = 2 \Rightarrow \frac{x}{2} = \frac{1}{x}$

$\Rightarrow \frac{x}{2} + \frac{x}{2} = \frac{x}{2} + \frac{1}{x} \Rightarrow x = \frac{x}{2} + \frac{1}{x}$

We can guess a value of x., i.e., x_0

Hoping that $\frac{x_0}{2} + \frac{1}{x_0}$ is getting closer to $\sqrt{2}$.

Let $x_1 = \frac{x_0}{2} + \frac{1}{x_0}$, repeat this ...

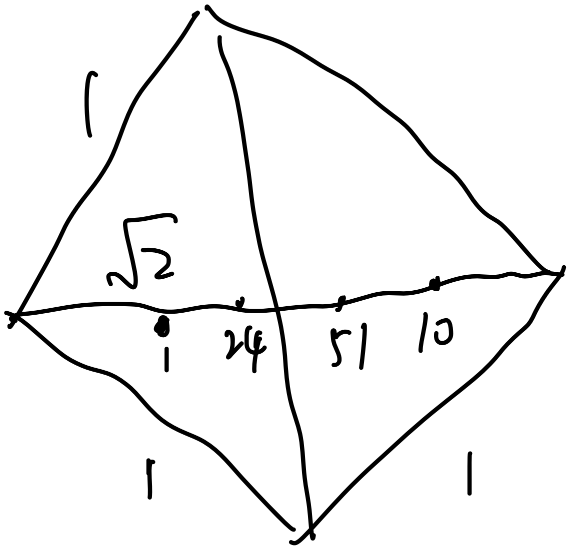
For example,

- $x_0 = 1$, $x_1 = \frac{1}{2} + 1 = 1.5$ x_1 is ~~better~~ than x_0
- $x_2 = \frac{x_1}{2} + \frac{1}{x_1} = \frac{1.5}{2} + \frac{2}{3} \approx 1.4166\cdots$ x_2 ... x_0

$\vdots$

Q1: why does $x_{t+1} = \frac{x_t}{2} + \frac{1}{x_t}$ work?

Illustration of Babylonian method



$$1+\frac{24}{60}+\frac{51}{60^2}+\frac{10}{60^3}$$
$$=\frac{305470}{216000}=1.41421\overline{296}$$

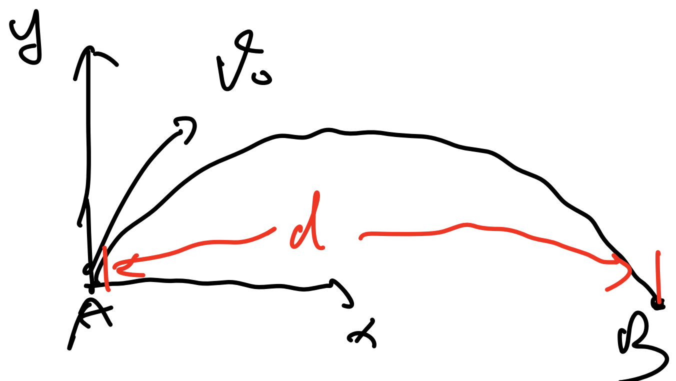
six decimal digits accuracy!

Problem 2: Predict the angle.

Given

- d distance A $\longrightarrow$ B
- v_0 speed of shell
- gravitational acceleration $g=9.8$

Give the right angle so that shell can shell from A to B.



t: total time

$$v_{0x} \cdot t = d.$$

$$v_y = v_{0y} - g\,t_{up}, \quad v_y = 0: \text{ reach the maximum}$$

$$t_{up} = \frac{v_{0y}}{g}. \qquad \text{total time } t = 2t_{up}$$

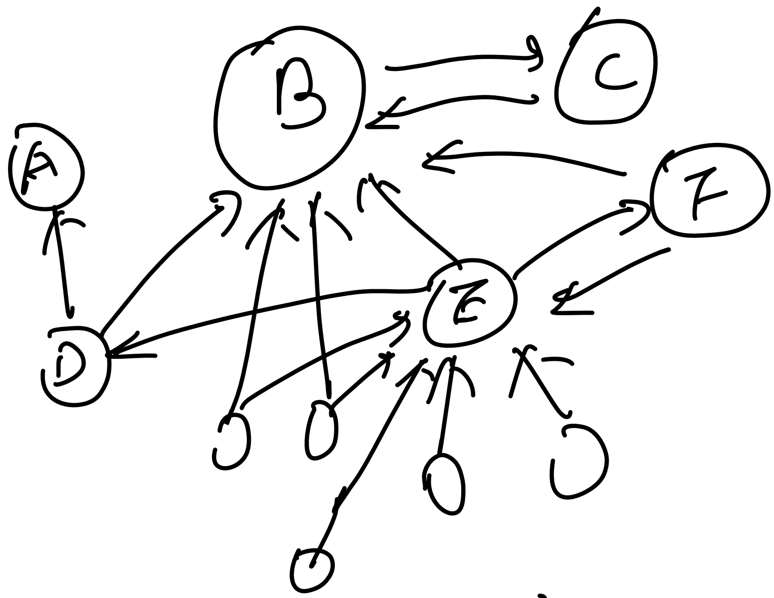
$$v_{0x} = v_0 \cos\theta$$

$$d = 2 \cdot \frac{v_{0y}}{g} \cdot v_{0x} \qquad v_{0y} = v_0 \sin\theta$$

$$\Rightarrow 2v_0^2 \sin\theta \cdot \cos\theta / g - d = 0.$$ How to get θ?

Problem 3: PageRank Problem.

ranking Web pages.



directed graph → adjacency matrix A

degree matrix D.

stochastic matrix $A^T D^{-1}$.

To solve: $\pi = (\alpha A^T D^{-1} + \frac{1-\alpha}{n} E)\pi$.

Machine representation of real numbers

Examples: π, e, $m_e \approx 9.1 \times 10^{-31}$, $c = 2.9 \times 10^{8}$

atoms size $10^{78} \sim 10^{82}$

Computers deal with numbers

→ store into a word.

Naive idea: Fixed-point arithmetic

$$\underbrace{\underbrace{\qquad\qquad}_{\text{Integer part}} . \underbrace{\qquad\qquad}_{\text{Fractional part}}}_{\text{word.}}$$

10^{78} needs $\approx$ 260 bits of integers part

π needs $\simeq$ bits but many digits in fractional part.

So, fixed-point is a bad idea!

Beteer method : float-point rep.

$(x)_{10} = \pm m \times 10^{n}$

sign mantissa power / exponent

尾数

Double precision :

$x=1$, $+1.\underbrace{0\cdots0}_{52} \times 2^{0}$

$x=1+2^{-52}$ $+1.\underbrace{0\cdots0}_{51}1 \times 2^{0}$, next float point

$\Rightarrow$ We call $2^{-52} = \varepsilon_{mach}$.

Rounding — chopping : biased.

$x = \pm 1.(b_1 \cdots b_{52} + 0\cdots 0 b_{53} \cdots) \times 2^{0}$

If $x<0$, then b_{53} remond wil

make $x \to 0$ always.

If $x>0$, then $x \to 0$ always

$|\longrightarrow 0 \longleftarrow|$: chopping.

$x<0$ $\qquad\qquad$ $x>0$

rounding :

$\leftarrow|\rightarrow \quad 0 \quad \leftarrow|\rightarrow$

$x<0$ $\qquad\qquad$ $x>0$.

If $x = \pm 1.\underbrace{\boxed{b_1 \cdots b_{52}}} b_{53} b_{54} \cdots \times 2^p$

$b_{53}=1,\ b_{54}=0,\ b_{55}=0 \ldots$

$x = \pm 1.\boxed{b_1 \cdots b_{52}}\ |\ 00\cdots 0 \times 2^p$

$\begin{cases} \bullet \text{ If } b_{52}=1, \to \text{round up} \to b_{52} \doteq 0 \\ \bullet \text{ If } b_{52}=0, \to \text{round down} \to b_{52}=0 \end{cases}$

$\leftarrow|\rightarrow \quad 0 \quad \leftarrow|\rightarrow$

How to measure the error:

rounding error of x:

$x=9.4$, $fl(9.4)-9.4$: rounding error.

How to measure the error:

absolute error : $|x_c - x|$

relative error : $\dfrac{|x_c - x|}{|x|}$ or $\dfrac{|x_c - x|}{|x_c|}$

If $x \in \mathbb{R}^n$ or $x \in \mathbb{R}^{n\times n}$, use:

$\|x_c - x\|$ or $\|x_c - x\|_{op}$.

what if $x=0$, no worry for saving numbers. Since $x=0$, there is no rounding error : $\frac{0}{0}$: relative error is 0.

Theorem:

$$\frac{|fl(x) - x|}{|x|} \le \underbrace{\frac{1}{2}\varepsilon_{mach}}_{\text{unit round off}}.$$

Verify this:

$|fl(9.4) - 9.4| = 0.2 \times 2^{-49}$

$$\frac{|fl(9.4) - 9.4|}{9.4} = \frac{0.2\times 2^{-49}}{9.4} = \frac{8}{47}\times 2^{-52} \le \frac{1}{2}\varepsilon_{mach}.$$

Proof:

W.L.O.G., Given $x>0$, we want to measure

$$\left|\frac{x-f(x)}{x}\right|.$$ We assume $x = q \times 2^m$

$$q = 1.b_1\cdots b_{52}b_{53}b_{54}\cdots)_2 \times 2^m \quad b_i \in \{0,1\}$$

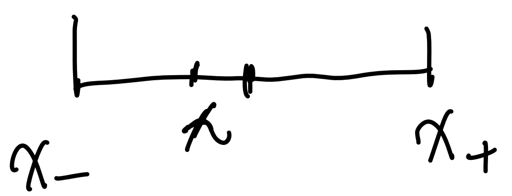
Two cases:

rounding down: $x_- = (1.b_1\cdots b_{52})_2 \times 2^m$.

rounding up: $x_+ = [(1.b_1\cdots b_{52})_2 + 2^{-52}] \times 2^m$.

①. Case 1. $f(x) = x_-$

$$(*) \quad \frac{|x-f(x)|}{|x|} = \left|\frac{x-x_-}{x}\right| \le \frac{1}{2}\frac{|x_+-x_-|}{|x|}$$



$|x-x_-|$ must be no more than half of $|x_+-x_-|$

$$\frac{|x_+-x_-|}{|x|} = \frac{2^{m-52}}{(1.b_1\cdots b_{52})_2 \times 2^m} = \frac{2^{-52}}{(1.b_1\cdots b_{52})_2} \le 2^{-52}$$

$\Rightarrow (*) \le \frac{1}{2} \cdot 2^{-52}.$

③ Case 2; $fl(x) = x_+$

$$(*) \quad \frac{|x - fl(x)|}{|x|} = \frac{|x - x_+|}{|x|} \le \frac{1}{2} \frac{|x_+ - x_-|}{|x|}$$

$$\le \frac{1}{2} \cdot 2^{-52}. \quad \blacksquare$$

Remark: Let $\delta = \frac{fl(x) - x}{x}$, then

①. $fl(x) = x(1+\delta)$ and $|\delta| \le 2^{-53}$.

②. ~~Generis~~ Generization :

$\odot \in \{+, -, \times, \div\}$.

$x = fl(x)$, $y = fl(y)$: machine numbers

$x \odot y$: computed and store

$\rightarrow fl(x \odot y)$.

$fl(x \odot y) = [x \odot y] \cdot (1+\delta)$, $|\delta| \le \varepsilon$.

log-sum-exp trick:

$$\log \pi_i = \log \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}$$

$$\log \pi_i = x_i - \underbrace{\log \sum_{j=1}^{n} \exp(x_j)}$$

$$\text{logsumexp}(x) = \log \sum_{j=1}^{n} \exp(x_j) + b - b$$

$$= b + \log \sum_{j=1}^{n} \exp(x_j - b)$$

$b = \max\{x_i,\ i = 1, 2, \dots n\}$.

This trick can avoid overflow.

Bisection method:

Given: $[a,b]$, f, ε. such that $f(a)\cdot f(b) < 0$

Bisection goes as the following:

for $t = 0, 1, 2, \cdots$

$c = \dfrac{a+b}{2}$

if $f(c) = 0$ return c

if $f(a)\cdot f(c) < 0$ then

$b = c$

else

$a = c$

return c

Error analysis: with $\begin{cases} a_0 = a \\ b_0 = b \end{cases}$

Bisect generats:

$[a_0, b_0], [a_1, b_1], \cdots, [a_n, b_n],$

where

$$a_0 \le a_1 \le a_2 \cdots \le a_n \le b_n \le b_{n-1} \cdots \le b_2 \le b_1 \le b_0$$

and.

$$b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n) \quad (n \geq 0).$$

(Recall every time, it cuts $[a_n, b_n]$ in half.

Recursively,

$$b_n - a_n = 2^{-n}(b_0 - a_0).$$

Thus, $\lim\limits_{n\to\infty} b_n - \lim\limits_{n\to\infty} a_n$

$$= \lim_{n\to\infty} 2^{-n}(b_0 - a_0)$$

$$= 0.$$

If we put $r = \lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} b_n$,

by taking the limit in the inequality

$$f(a_n) \cdot f(b_n) \le 0.$$

$$\Rightarrow \lim_{n\to\infty} f(a_n) \cdot f(b_n) \le 0 \Rightarrow f(r) \cdot f(r) \le 0$$

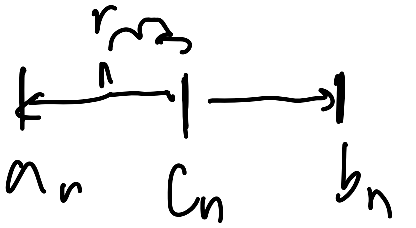
$\Rightarrow f(r) = 0.$

$|a_n - b_n| < \xi$ stoped :

r must be in $[a_n, b_n]$,

$c_n = \frac{a_n + b_n}{2}$ is the estimate of r.

So, $|r - c_n| \leq \frac{1}{2}(b_n - a_n)$



$= 2^{-(n+1)}(b_0 - a_0)$.

Finally, $|r - c_n| \leq 2^{-(n+1)} \cdot (b_0 - a_0)$.

Thm. Error analysis of Bisection :

Let $[a_n, b_n]$ be intervals used in Bisect, then $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$ exist, are equal, and represent a zero of f. Def. $c_n = \frac{a_n + b_n}{2}$ and $r = \lim_{n\to\infty} c_n$, then

$$|r - c_n| \le 2^{-(n+1)}(b_0 - a_0).$$

---

② Time complexity: $O(n)$.

③ Example 3.4 of using bisect.

We know $|r - c_n| \le 2^{-(n+1)}(b_0 - a_0)$

Let $2^{-(n+1)}(b_0 - a_0) \le 0.5 \times 10^{-p}$

$a_0 = 0$

$b_0 = 1$

$p = 6$

$$2^{-n} \le 10^{-6}$$

$$-n \cdot \log 2 \le -6 \log 10$$

$$n \ge \left\lceil \frac{6 \log 10}{\log 2} \right\rceil = 20.$$

④ Exercise: Suppose that the bisection method is started with the interval $[50, 63]$. How many steps should be taken to compute a root with relative accuracy of one part in 10^{-12}?