Math 151A

TA: Bumsu Kim

Today…

- Floating Point Arithmetic

- Bisection Method

- Fixed Point Methods

Floating Point Arithmetic

✓ $1, \frac{1}{2}, \frac{3}{2}$

$\pi = 3.141592\cdots$

$= \underbrace{0.3141592\cdots}_{\text{Mantissa}} \times 10^{1}$ ← exp.

$\approx 0.314159 \times 10^{1}$

- What to avoid and why?
- Avoid "Loss of significance"

0.000000211

  - $x = 0.123456789$ ← $fl(x) = 0.123457$
  - $y = 0.123455486$ $fl(y) = 0.123456$
  - $x - y = 0.000001303$ $fl(x) - fl(y) = 0.000001$
  - Relative error is large (≈7.7%)

  - Avoid subtraction of two nearly equal numbers ("subtractive cancelation")
    - e.g. $\sqrt{x^2 + 1} - 1$ when x is small

  - Reduce the number of arithmetic operations

UCLA

Avoid Subtractive Cancelation

1) Avoid substraction of 2 nearly equal numbers.

Why? It causes cancelation of significant digits. Given 2 nearly equal numbers $x > y$ of k-digit representation:

$$fl(x) = 0.d_1 d_2 \ldots d_p \alpha_{p+1} \ldots \alpha_k \times 10^n$$

and

$$fl(y) = 0.d_1 d_2 \ldots d_p \beta_{p+1} \ldots \beta_k \times 10^n$$

Then

$$fl(fl(x) - fl(y)) = 0.\underbrace{\sigma_{p+1}\sigma_{p+2} \ldots \sigma_k}_{(k-p)} \times 10^{n-p}$$

$$= \overbrace{0.\underbrace{\sigma_{p+1}\sigma_{p+2} \cdots \sigma_k}_{(k-p)} \boxed{\underbrace{\gamma_1 \cdots \gamma_p}_{0 \cdots 0}}}^{k\text{-digits}} \times 10^{n-p}$$

padded

where

$$0.\sigma_{p+1}\sigma_{p+2} \ldots \sigma_k = \alpha_{p+1} \ldots \alpha_k - \beta_{p+1} \ldots \beta_k$$

Note that we have at most $k-p$ significant digits i.e. we lost p significant digits! In most machines, $x-y$ will be assigned k-significant digits with last p digits either 0 or randomly assigned.

Floating Point Arithmetic

- Examples
  - e.g. $\sqrt{x^2+1}-1$ when x is small
    - Use $\frac{x^2}{\sqrt{x^2+1}+1}$ instead
    - With 9 significant digits, $\sqrt{x^2+1}-1 = 10^{-8}$ and $\frac{x^2}{\sqrt{x^2+1}+1} = 0.5\times 10^{-8}$

$x = 10^{-4}$

- Reduce the number of arithmetic operations
  - Nested multiplication
  - $\underbrace{a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n}_{\frac{n(n+1)}{2}\text{ mult. \& } n \text{ additions}} = a_0 + \boxed{x}(a_1 + x(a_2 + x(a_3 + \cdots + x(a_{n-1} + a_nx))\cdots)$

$1\otimes$ $2\otimes$ $3\otimes$ $n\otimes$

a_{n-2} $\times$ $\oplus$ $\otimes$

n mult.
n add.

UCLA

Roots of a Nonlinear Equation

- We want to find a zero/solution/root of a function $f(x): \mathbb{R} \to \mathbb{R}$

- Can't find the *exact* solution due to the finite precision

- We will essentially find an estimate of the solution

Bisection Method

- We will find an estimate of the solution x^*

- A good type of estimation is to say: $x^* \in [\alpha, \beta]$ for two close numbers α and β

- For instance, if we can guarantee $x^* \in [0.99999, 1.00001]$, the absolute error is at most 10^{-5}

  $|1 - x^*| \leq 10^{-5}$

- The **bisection method** provides this type of estimate

Bisection Method

- Bisection method is based on the *Intermediate Value Theorem*

- If $f(a) < 0$ and $f(b) > 0$ then there exists $c \in (a, b)$ s.t. $f(c) = 0$.
  - (of course, when f is continuous $\Leftrightarrow f \in C[a, b]$)

- Then we just reduce the size of the interval by half at each iteration
  - The accuracy is doubled at every time

- Bisection method converges "linearly"
  - The current error bound is bounded by a *linear function* of the previous error bound
  - Indeed, we have $e_{k+1} \leq \frac{1}{2} e_k$ for the bisection method, where $|x_k - x_*| \leq e_k$

$\rightarrow$ exponential decay of $e_k \leq 2^{-k} \cdot (\text{Const.})$

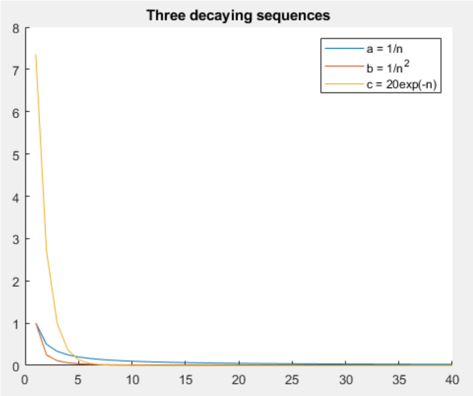
UCLA

Order of Convergence

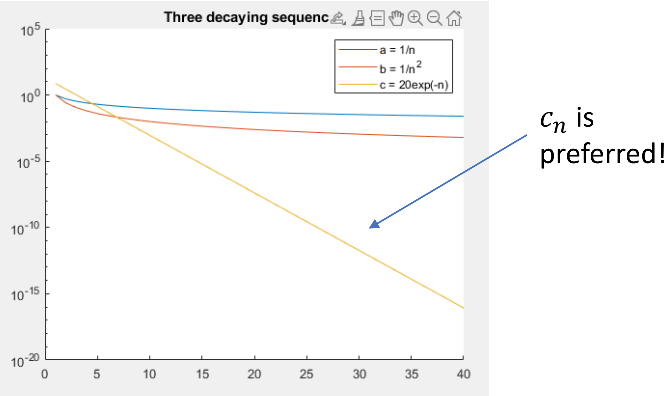
- Consider three sequences $a_n = \frac{1}{n}$, $b_n = \frac{1}{n^2}$, and $c_n = 20e^{-n}$

  $a_n \leq b_n$

- At first c_n is much larger than the other two, however, it *decays* much faster



a_n, b_n, and c_n



y-axis in log scale

[Week2_Convergence_Rate.m]

UCLA

Order of Convergence

- Thus, we can define the "rate of convergence" by

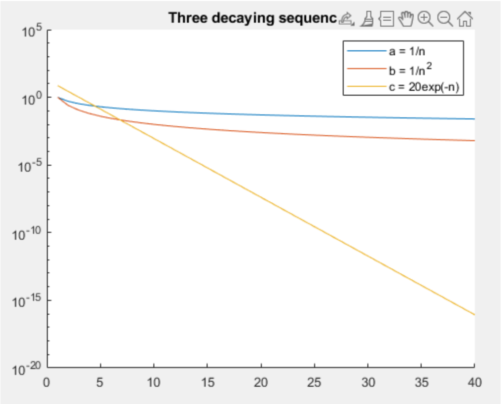
Definition 1.18 Suppose $\{\beta_n\}_{n=1}^{\infty}$ is a sequence known to converge to zero and $\{\alpha_n\}_{n=1}^{\infty}$ converges to a number α. If a positive constant K exists with

$\exists N_0 > 0$ s.t. ← holds for any $n > N_0$

$$|\alpha_n - \alpha| \le K|\beta_n|, \quad \text{for large } n,$$

then we say that $\{\alpha_n\}_{n=1}^{\infty}$ converges to α with **rate, or order, of convergence** $O(\beta_n)$. (This expression is read "big oh of β_n".) It is indicated by writing $\alpha_n = \alpha + O(\beta_n)$. ■

- Notice the condition "for large n"
- e.g. $c_n = 20e^{-n}$ is $O(a_n)$ and also $O(b_n)$
- But better to say $c_n = O(e^{-n})$ because it is more accurate



$c_n = O(a_n)$, $c_n = O(b_n)$

$c_n = O(1)$

Order of Convergence

- For functions, we can come up with a similar definition:

Definition 1.19 Suppose that $\lim_{h\to 0} G(h) = 0$ and $\lim_{h\to 0} F(h) = L$. If a positive constant K exists with

$$|F(h) - L| \le K|G(h)|, \quad \text{for sufficiently small } h,$$

then we write $F(h) = L + O(G(h))$. ■

- In fact, where h goes doesn't matter (i.e. can define the order of convergence for $h \to 3$ or $h \to -\infty$.)

- e.g.
  - $\sin(h) = O(h)$ near 0 (i.e. as $h \to 0$)
  - $e^h = 1 + O(h)$ and $\log(1+h) = O(h)$
  - If f is even and analytic at 0 (roughly speaking, f is "smooth"), then $f(h) = 1 + O(h^2)$

$\cos(h) = 1 + O(h^2)$

- Taylor's Theorem is very useful

Order of Convergence

- Exercises $\boxed{h \to 0}$
  - $e^h = 1 + O(h)$ and $\log(1+h) = O(h)$

    near 0

    $$\begin{cases} e^h = 1 + h + \frac{h^2}{2} + \cdots \le 1 + 2h & \forall h \le 0.1 \\ |\log(1+h)| = \left|h - \frac{h^2}{2} + \frac{h^3}{3} - \cdots \right| \le 2h & \forall h \le 0.1 \end{cases}$$

  - If f is even and analytic at 0 (roughly speaking, f is "smooth"), then $f(h) = 1 + O(h^2)$

    f is even, $f(-x) = f(x)$

    the first order terms in the Taylor expansion of $f = 0$
    (odd)

    $$f(h) = f(0) + \underbrace{\frac{h^2}{2} f''(0) + \frac{h^4}{4!} f^{(4)}(0) + \cdots}_{O(h^2)}$$

  - The sequence $x_0 = 0.75$ and $x_n = \left(\frac{e^{x_{n-1}}}{3}\right)^{1/2}$ ← Will be revisited after going over the Fixed Point Methods

Bisection Method

- Implementation: [Week2_Bisection_method.m]

Bisection Method

- Error Analysis – Number of operations to achieve ϵ-accuracy

(See textbook)

Bisection Method

- Pros and Cons:
  - Guaranteed convergence
  - But the convergence is slow
  - Must find two points with different signs first
- There are other methods that provide faster convergence with/without convergence guarantee

Fixed Point Methods

- Banach Fixed Point Theorem

Banach Fixed Point Theorem. Let (X, d) be a non-empty complete metric space with a contraction mapping $T : X \to X$. Then T admits a unique fixed-point x^* in X (i.e. $T(x^*) = x^*$). Furthermore, x^* can be found as follows: start with an arbitrary element $x_0 \in X$ and define a sequence $(x_n)_{n\in\mathbb{N}}$ by $x_n = T(x_{n-1})$ for $n \geq 1$. Then $\lim_{n\to\infty} x_n = x^*$.

- Real number version (part (ii))

$g([a,b]) \subseteq [a,b] \Rightarrow g$ has at least one F.P.

Theorem 2.3

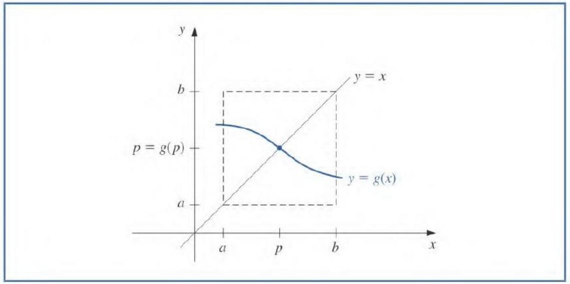
(i) If $g \in C[a, b]$ and $g(x) \in [a, b]$ for all $x \in [a, b]$, then g has at least one fixed point in $[a, b]$.

(ii) If, in addition, $g'(x)$ exists on (a, b) and a positive constant $k < 1$ exists with

$$|g'(x)| \leq k, \quad \text{for all } x \in (a, b),$$

then there is exactly one fixed point in $[a, b]$. (See Figure 2.3.)

Figure 2.3



Fixed Point Methods

- To apply the fixed point theorem, you need to show
  - First, g is continuous on $[a, b]$. (If it's obvious, at least mention that it is continuous)
  - Secondly, $g([a, b]) \subseteq [a, b]$
  - Finally, to show the uniqueness, $|g'(x)| \le c < 1$ for all $x \in (a, b)$

- (If any one of the above is missing you'll lose points)

- For strictly increasing/decreasing functions, it's easy to check the first two conditions
  - e.g. $g(x) = e^{x/2} - 1$ on $[-1, 1]$
    - Note that g does ~~not~~ satisfy the 3rd condition, ~~but~~ and it has a unique fixed point in $[-1, 1]$

$g(-1) = e^{-1/2} - 1 > -1$, $g(1) = e^{1/2} - 1 < 1$

$\Updownarrow$

$\sqrt{e} < 2$

$\Updownarrow$

$e < 4$

Exercise: Fixed Point Theorem

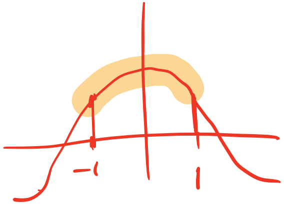
Exercise 2.6 *Consider fixed-point iteration to compute the solution of*

$$\cos\alpha = \alpha, \quad \text{on } (-\infty, \infty)$$

using $g(x) = \cos x$. *Prove that this converges for any starting guess. Compute a few iterations to see what the approximate value of* α *is.*

Let $x_0 \in \mathbb{R}$, $\quad x_1 = \cos(x_0) \in [-1, 1]$

$\Rightarrow$ Equivalent to a F.P.I. of $\cos(x)$ on $[-1,1]$



$\max\limits_{[-1,1]} \cos(x) = 1$

$\min\limits_{[-1,1]} \cos(x) = \cos(1) > 0$

$\Rightarrow$ $\cos(x)$ has at least one F.P. on $[-1,1]$

$\Rightarrow$ next step, $\quad x_2 \in [\cos(1), 1] \Rightarrow |g'(x)| = |\sin(x)| \leq \sin(1) < 1$ on $[\cos(1), 1]$

Thus, $\exists!$ F.P. α in $[\cos(1), 1]$.

Exercise: Rate of Convergence, revisited

- Does the sequence $x_0 = 0.75$ and $x_n = \left(\frac{e^{x_{n-1}}}{3}\right)^{1/2}$ converges?
  - If so, find the rate of convergence

Sol x_n converges to $x_* \in (0,1)$ linearly! (exponentially fast convergence)

pf) $g(x) = \left(\frac{e^x}{3}\right)^{1/2} = \frac{e^{x/2}}{\sqrt{3}}$, $g(0) = \frac{1}{\sqrt{3}} > 0$, $g(1) = \frac{\sqrt{e}}{\sqrt{3}} < 1$ since $e < 3$

and g is increasing. $\Rightarrow$ g has a fixed point on $[0,1]$

$0 < g'(x) = \frac{1}{2}\frac{e^{x/2}}{\sqrt{3}} < \frac{\sqrt{e}}{2\sqrt{3}} < \frac{1}{2}$ on $[0,1]$, the F.P. is unique!!

let x_* be the F.P. $\Rightarrow$ $g(x_*) = x_*$, i.e., $\frac{e^{x_*/2}}{\sqrt{3}} = x_*$

Now, Taylor expansion about x_*

$\Rightarrow$ $g(x_*+h) = g(x_*) + h g'(x_*) + \frac{h^2}{2} g''(\xi(h) + x_*)$

arb. small as long as n large

$|x_{n+1} - x_*| = |g(x_n) - x_*| = (x_n - x_*) g'(x_*) + \frac{(x_n - x_*)^2}{2} \underbrace{g''(\xi(x_n))}_{\leq M} \leq (x_n - x_*)\left(g'(x_*) + \varepsilon\right)$ $\square$

UCLA

Exercise: Fixed Point Iteration

6. The following four methods are proposed to compute $7^{1/5}$. Rank them in order, based on their apparent speed of convergence, assuming $p_0 = 1$.

a. $p_n = p_{n-1}\left(1 + \dfrac{7 - p_{n-1}^5}{p_{n-1}^2}\right)^3$

b. $p_n = p_{n-1} - \dfrac{p_{n-1}^5 - 7}{p_{n-1}^2}$

c. $p_n = p_{n-1} - \dfrac{p_{n-1}^5 - 7}{5p_{n-1}^4}$

d. $p_n = p_{n-1} - \dfrac{p_{n-1}^5 - 7}{12}$

- Represent each method in the form of "fixed point iteration"

- Find the apparent speed(rate) of convergence

(See GitHub Repo)

UCLA

Exercise: Fixed Point Iteration

24. **a.** Show that if A is any positive number, then the sequence defined by

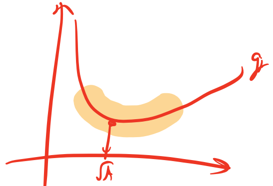
$$x_n = \frac{1}{2}x_{n-1} + \frac{A}{2x_{n-1}}, \quad \text{for } n \geq 1,$$

converges to $\sqrt{A}$ whenever $x_0 > 0$.

b. What happens if $x_0 < 0$?

$A = 2.$

$$\left(x_n = \frac{1}{2}x_{n-1} + \frac{1}{x_{n-1}} \qquad x_n \to \sqrt{2} \right)$$



a. pf) Let $x_0 > 0$. Then $x_1 = \frac{1}{2}x_0 + \frac{A}{2x_0} \geq \sqrt{x_0 \cdot \frac{A}{x_0}} = \sqrt{A}$ (AM $\geq$ GM)

Now, let $g(x) = \frac{1}{2}x + \frac{A}{2x}$, $g(x) \geq \sqrt{A}$ $\forall x \geq 0$ as shown above.

Also note that $g'(x) = \frac{1}{2} - \frac{A}{2x^2}$ and g has a min at $x = \sqrt{A}$, $g' > 0$ on $(\sqrt{A}, \infty)$ and $g' < 0$ on $(0, \sqrt{A})$.

Thus $\max(g([\sqrt{\tfrac{A}{2}}, \sqrt{2A}])) = \max(g(\sqrt{\tfrac{A}{2}}), g(\sqrt{2A})) = \frac{3}{2\sqrt{2}}\sqrt{A} < \sqrt{2A}$

i.e., $g([\sqrt{\tfrac{A}{2}}, \sqrt{2A}]) \subseteq [\sqrt{\tfrac{A}{2}}, \sqrt{2A}]$ and g has a fixed pt in $[\sqrt{\tfrac{A}{2}}, \sqrt{2A}]$.

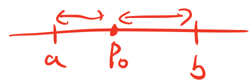
Finally, $\frac{1}{2} \geq g'(x) \geq \frac{1}{2} - \frac{A}{2 \cdot \frac{1}{2}A} = -\frac{1}{2}$ $\forall x \in [\sqrt{\tfrac{A}{2}}, \sqrt{2A}]$, and thus $|g'(x)| \leq \frac{1}{2} < 1$.

$\Rightarrow$ The F.P. of g is unique and equal to $\sqrt{A}$. □

b. $x_n \to -\sqrt{A}$, because $-x_n = \frac{1}{2}(-x_{n-1}) + \frac{A}{2(-x_{n-1})}$. □

iterations :

Cor 2.5: $|g'(x)| \le k < 1$



$$\Rightarrow |p_n - p_*| \le k^n \cdot \underbrace{\max(p_0 - a,\ b - p_0)}$$

to minimize this, choose $p_0 = \frac{b+a}{2}$

$= \frac{b-a}{2}$

e.g. $g'(x) = \frac{1}{2} e^{x/2}$, $x \in [\overset{a}{0}, \overset{b}{1}]$

$|g'(x)| \le g'(1) = \frac{1}{2}\sqrt{e} = k < 1$

$\downarrow$

$$|p_n - p_*| \le \frac{e^{n/2}}{2^n} \cdot \frac{b-a}{2} \le 0.00001$$

WANT: $\left(\frac{\sqrt{e}}{2}\right)^n \cdot \frac{1}{2} \le 10^{-5}$

Take log both sides

$$n \underbrace{\log(\sqrt{e}/2)}_{\text{negative}} \le \underbrace{\log(2 \times 10^{-5})}_{\text{negative}}$$

$$n \ge \frac{\log(2 \times 10^{-5})}{\log(\sqrt{e}/2)}$$