Math 151A

TA: Bumsu Kim

Today…

- Floating Point Arithmetic
- Bisection Method
- Fixed Point Methods

Floating Point Arithmetic

- What to avoid and why?
- Avoid "Loss of significance"
  - $x = 0.123456789$ $\quad fl(x) = 0.123457$
  - $y = 0.123455486$ $\quad fl(y) = 0.123456$
  - $x - y = 0.000001303$ $\quad fl(x) - fl(y) = 0.000001$
  - 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

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

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

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.

UCLA

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 $\frac{x^2}{\sqrt{x^2+1}+1}=0.5\times10^{-8}$
  - Reduce the number of arithmetic operations
    - Nested multiplication
    - $a_0+a_1x+a_2x^2+a_3x^3+\cdots+a_nx^n=a_0+x(a_1+x(a_2+x\big(a_3+\cdots+x(a_{n-1}+a_nx)\big)\cdots)$

Roots of a Nonlinear Equation

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

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

- We will essentially find an estimate of the solution

UCLA

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}

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

Bisection Method

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

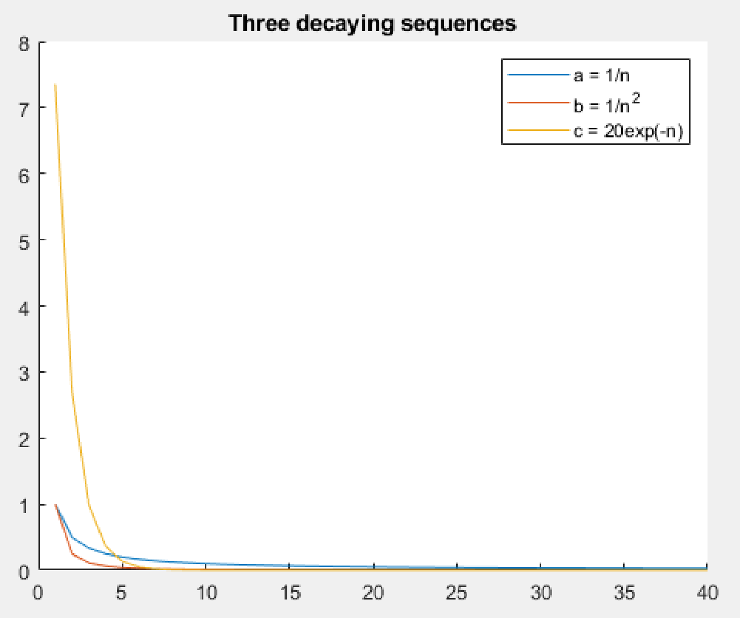
- If $f(a) < 0$ and $f(b) > 0$ then there exists $c \in (a, b)$ s.t. $f(c) = 0$.
  - (of course, when f is continuous $\Leftrightarrow$ $f \in C[a, b]$)

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

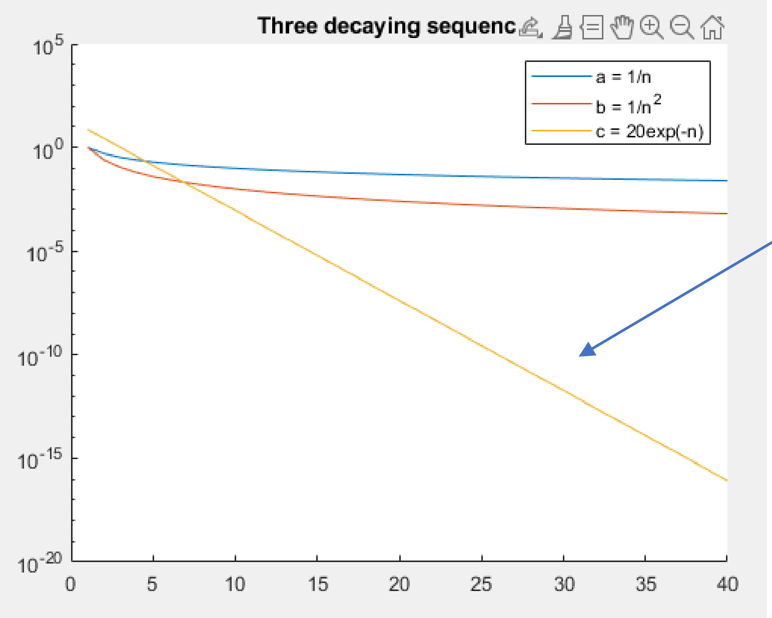
- Bisection method converges "linearly"
  - The current error bound is bounded by a *linear function* of the previous error bound
  - Indeed, we have $e_{k+1} \leq \frac{1}{2} e_k$ for the bisection method, where $|x_k - x_*| \leq e_k$

Order of Convergence

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

c_n is preferred!

[Week2_Convergence_Rate.m]

Order of Convergence

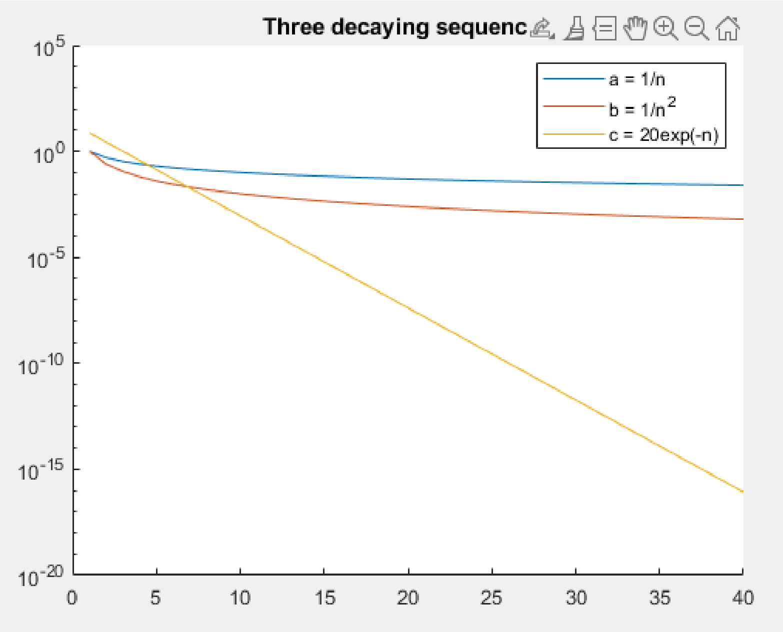
- Thus, we can define the "rate of convergence" by

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

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

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



UCLA

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 $f(h) = 1 + O(h^2)$

- Taylor's Theorem is very useful

Order of Convergence

- Exercises
  - $e^h = 1 + O(h)$ and $\log(1+h) = O(h)$
  - If f is even and analytic at 0 (roughly speaking, f is “smooth”), then $f(h) = 1 + O(h^2)$
  - 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

UCLA

Bisection Method

- Implementation: [Week2_Bisection_method.m]

Bisection Method

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

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))

Theorem 2.3

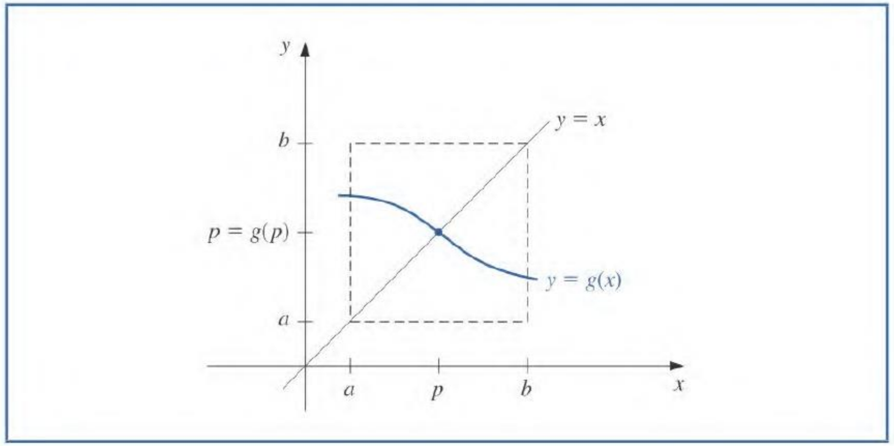
(i) If $g \in C[a, b]$ and $g(x) \in [a, b]$ for all $x \in [a, b]$, then g has at least one fixed point in $[a, b]$.

(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, $|g'(x)| \leq c < 1$ 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 it has a unique fixed point in $[-1, 1]$

UCLA

Exercise on the Fixed Point Theorem

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

$$\cos\alpha = \alpha,$$

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.*

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

Rate of Convergence, revisited

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