Math 174: Numerical Methods

Topics: Numerical methods for solving linear & non-linear equations and for approximation.

Instructor: Prof Reagan Saab (rsaab@ucsd.edu)
Office hours: Thurs. 5:00-7:00 pm (or by appointment)

TA: TBD

Course webpage: math.ucsd.edu/~rsaab → teaching → click on course
check regularly for HW & announcements

Homework: 5 assignments (essentially bi-weekly)
20% of grade

Midterms: 2 exams (20% each) weeks 4 & 8
Final: (40% of grade)

Alternative grading scheme (HW: 20%
Higher midterm 20%, Final 60%)
Lecture 1: Review of Calculus

Limits: \( \lim_{x \to c} f(x) = L \iff \forall \epsilon > 0, \exists \delta > 0 \)
\[ \text{st. } |f(x) - L| < \epsilon \text{ when } 0 < |x - c| < \delta \]

If there is no such \( L \), then the limit does not exist at \( c \).

Example: \( \lim_{x \to 2} x^2 = 4 \)

Proof (informal): Want \( |x^2 - 4| < \epsilon \) whenever \( 0 < |x - 2| < \delta \).

But then \( |x^2 - 4| = |x - 2||x + 2| < \delta \frac{4 + \delta}{\delta} \).

Thus \( \delta + 4\delta = \epsilon \) so \( \delta = -2 + \sqrt{4 + \epsilon} \).

Formal Proof: \( \forall \epsilon > 0 \) let \( \delta = -2 + \sqrt{4 + \epsilon} \).

If \( |x - 2| < \delta \), then
\[ |x^2 - 4| = |x - 2||x + 2| < \delta(\delta + 4) = (-2 + \sqrt{4 + \epsilon})(2 + \sqrt{4 + \epsilon}) = \epsilon. \]
Example 2: \( f(x) = \frac{|x|}{x} = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \end{cases} \) (undefined at 0)

\[
\lim_{x \to 0} \frac{|x|}{x} = ?
\]

Suppose \( \lim_{x \to 0} \frac{|x|}{x} = L \) then \( \forall \varepsilon > 0, \exists \delta > 0 \) s.t. \( |\frac{|x|}{x} - L| < \varepsilon \)

when \( |x| < \delta \)

Take \( \varepsilon = 1 \)

then \( |\frac{|x|}{x} - L| < 1 \) when \( |x| < \delta \)

but choosing \( x = \frac{\delta}{2} : |1 - \frac{\delta}{2}| < 1 \) \( \iff \frac{\delta}{2} < 1 \)

choosing \( x = -\frac{\delta}{2} : |-1 + \frac{\delta}{2}| < 1 \) \( \iff -\frac{\delta}{2} < 1 \)

\[\text{Contradiction}\]

so no such \( L \) exists!

Continuity: \( f \) is cont. at \( c \) if \( \lim_{x \to c} f(x) = f(c) \)

E.g. \( f(x) = x^2 \) is cont. at \( x = 2 \)

\( f(x) = \frac{|x|}{x} \) is cont. at \( x = 2 \) (why?)

\( f(x) = \frac{|x|}{x} \) is not cont. at \( x = 0 \) (why?)

Intermediate Value Theorem: Let \( [a, b] \) be an interval & let \( f \) be a cont. \( f \) on \( [a, b] \). Then for every \( y \in [f(a), f(b)] \) (or \( [f(b), f(a)] \) ), there exists \( x^* \in [a, b] \) s.t. \( f(x^*) = y \).

In words: \( f \) assumes all values between \( f(a) \) & \( f(b) \).
Derivative: \[ f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c} \]

\[ = \lim_{h \to 0} \frac{f(c+h) - f(c)}{h} \]

**If \( f'(c) \) exists, we say \( f \) is differentiable at \( c \).**

**If \( f \) is differentiable at \( c \Rightarrow f \) is continuous at \( c \).**

**Proof:**

\[ \lim_{x \to c} f(x) - f(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c} \cdot (x - c) = f'(c) \cdot \lim_{x \to c} (x - c) \]

\[ = 0 \]

\[ \Rightarrow \lim_{x \to c} f(x) = f(c) \Rightarrow \text{cont.} \]

**Warning:** Not all continuous \( f \) are differentiable!

**Notation:**
- \( C(\mathbb{R}) \) set of all continuous functions on \( \mathbb{R} \)
- \( C'(\mathbb{R}) \) set of all \( f' \)s on \( \mathbb{R} \) with \( f' \) continuous

\( C'(\mathbb{R}) \subset C(\mathbb{R}) \)

- \( C^n(\mathbb{R}) \) set of \( f \)s on \( \mathbb{R} \) with \( f^{(n)} \) continuous
- \( C^\infty(\mathbb{R}) \) set of \( f \)s with all derivatives continuous

\( C^\infty(\mathbb{R}) \subset C^2(\mathbb{R}) \subset C'(\mathbb{R}) \subset C(\mathbb{R}) \)

**E.g.:** \( \sin(x), \cos(x), e^x \in C^\infty(\mathbb{R}) \)
\( C^n([a,b]) = \text{set of } f \text{ on } [a,b] \text{ for which } f^{(n)} \text{ exists and } f^{(n)} \text{ is cont.} \)

**Taylor's theorem**

**Version 1 (Lagrange remainder)**

If \( f \in C^n([a,b]) \) & \( f^{(n+1)} \) exists on \((a,b)\) then \( \forall x, c \in [a,b] \), there is a \( \xi \) between \( x \) \& \( c \) s.t.

\[
L(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1}
\]

\( E_n(x) \): error term

**Example:** \( e^x = \sum_{k=0}^{\infty} \frac{1}{k!} \cdot 1 \cdot x^k = \sum_{k=0}^{\infty} \frac{x^k}{k!}, \quad x < \infty \)

**Exercise:** Taylor series for \( \sin(x) \), \( \cos(x) \), \( \ln(x) \). \( x \in (0,\infty) \)

**Example:** What is the approx. error in computing \( e^3 \) with \( n \)-terms of the Taylor series around 0 (i.e. \( c = 0 \))
By Taylor's theorem with $c=0$

\[e^x = \sum_{k=0}^{m-1} \frac{x^k}{k!} + \frac{e^c x^m}{m!} \cdot \frac{1}{E_{m-1}(x)}\]

\[E_{m-1}(3) = \frac{e^c 3^m}{m!} < \frac{e^3 3^m}{m!}\]

So $E_0(3) < e^3$

$E_1(3) < e^3 \cdot 3$

$E_2(3) < e^3 \cdot 3^2 / 2$

$E_3(3) < e^3 \cdot 3^3 / 3!$

Mean Value Theorem: Let $f \in C[a,b]$ and suppose $f'$ exists on $(a,b)$ then $\forall x, c \in [a,b]$

\[f(x) = f(c) + f'(c)(x-c)\] For some $\xi$ between $x$ and $c$

All statement: There exists $\xi$ such that $f'(\xi) = \frac{f(x) - f(c)}{x-c}$
Alternate Form of Taylor's Theorem:

If \( f \in C^{n+1}([a,b]) \) & \( x, x+h \in [a,b] \)

then \( f(x+h) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!} h^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} h^{n+1} \frac{1}{E_n(h)} \)

Taylor's Theorem in 2-variables:

Let \( f \in C^{n+1}([a,b] \times [c,d]) \) & let \((x,y)\) and \((x+h, y+k)\) both be in \([a,b] \times [c,d] \subseteq \mathbb{R}^2\) then

\[ f(x+h, y+k) = \sum_{i=0}^{n} \frac{1}{i!} \left( h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y} \right)^i f(x,y) \]

\[ + \frac{1}{(n+1)!} \left( h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y} \right)^{n+1} f(x+oh, y+ok) \]

where \( E_n(h,k) = \frac{1}{(n+1)!} \left( h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y} \right)^{n+1} f(x+oh, y+ok) \)

where \( \xi \in [0,1] \).
what is \( (h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y})^n \)?

Think of it as an operator which acts on functions to produce other functions.

Specifically

\[
(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y})^0 f(x,y) = f(x,y)
\]

\[
(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y}) f(x,y) = (h \frac{\partial f}{\partial x} + k \frac{\partial f}{\partial y})(x,y)
= hf_x + kf_y
\]

\[
(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y})^2 f(x,y) = h^2 \frac{\partial^2 f}{\partial x^2} + k^2 \frac{\partial^2 f}{\partial y^2} + 2hkh \frac{\partial^2 f}{\partial x \partial y}
\]

\[
\ldots
\]

So

\[
L(x+h,y+h) = f(x,y) + hf_x(x,y) + kf_y(x,y) + \left( h^2f_{xx} + 2hkf_{xy} + k^2f_{yy} \right) + \ldots
\]

\[
\uparrow \text{ polynomial in } h,k
\]

**Example:** Write the 2nd order Taylor series of \( f(x,y) = e^{xy} \)

ie, find \( f(x+h,y+k) \)

**Solu:**

we'll need \( f(x,y) \)

\[
L_x = \frac{\partial f}{\partial x} = ye^{xy} \quad L_y = xe^{xy}
\]

\[
L_{xx} = y^2e^{xy}, \quad L_{xy} = xe^{xy}, \quad L_{yy} = x^2e^{xy}
\]

\[
f(x+h,y+k) = f(x,y) + hf_x(x,y) + kf_y(x,y) + \left( h^2f_{xx} + 2hkf_{xy} + k^2f_{yy} \right) + \ldots
\]

\[
= e^{xy} + hxe^{xy} + ke^{xy} + h^2y^2e^{xy} + 2hkxe^{xy} + k^2x^2e^{xy} + \ldots
\]
So \( h(x+h, y+k) = E_2(h, k) + e^{x+h} + (kye^{x+h} + kx e^{x+h}) \)

\[ + \frac{1}{2!}(h^2 y e^{x+h} + h^2 x e^{x+h} + 2hk (e^{x+h} + xye^{x+h})) \]

Example for the same \( h \) we write \( E_1(h, k) \)

\[ E_1(h, k) = \frac{1}{2} [oh^2 y e^{x+h} + (oh)^2 x e^{x+h} + 2ohk (e^{x+h} + xye^{x+h})] \]

where \( x = x + oh \) & \( y = y + oh \) for some \( \sigma \in [0, 1] \)

---

**Orders of Convergence**

For a sequence \( x_n \), we write

\[ \lim_{n \to \infty} x_n = L \]

if \( \forall \epsilon > 0, \exists \delta \in \mathbb{N} \) s.t. \( |x_n - L| < \epsilon \)

whenever \( n > \delta \)

**Example:** \( \lim_{n \to \infty} \frac{n+1}{n} = 1 \) dec. \( |\frac{n+1}{n} - 1| = \left| \frac{1}{n} \right| < \epsilon \)

\[ \forall n > \Omega(\epsilon) : = \frac{1}{\epsilon} . \]

Orders of convergence: Suppose \( \lim_{n \to \infty} x_n = L \), then we say...
the order of conv. of $x_n$ to $L$ is

- **Linear** if $|x_{n+1} - L| \leq C|x_n - L| \quad \forall n \geq N$

- **Superlinear** if $|x_{n+1} - L| \leq \varepsilon_n |x_n - L| \quad \forall n \geq N$
  \[\text{sequence with } \varepsilon_n \to 0\]

- **Quadratic** if $|x_{n+1} - L| \leq C|x_n - L|^2 \quad \forall n \geq N$
  \[\text{const. (not necessarily } < 1)\]

- **Of order $\alpha$** if $|x_{n+1} - L| \leq C|x_n - L|^{\alpha} \quad \forall n \geq N$

**Big "Oh" & Little "oh":**

- $x_n = O(d_n)$ if $\exists C > 0 \& N \in \mathbb{N}$ s.t. $|x_n| \leq Cd_n \quad \forall n \geq N$

- $x_n = o(d_n)$ if $\exists \varepsilon_n \to 0 \text{ s.t. } |x_n| \leq \varepsilon_n |d_n| \quad \forall n \geq N$

  \[\text{(Intuitively } \lim_{n \to \infty} \frac{x_n}{d_n} = 0) \text{ or } (d_n \text{ dominates } x_n \text{ asymptotically)}\]

**Examples:**

\[e^x - \sum_{k=0}^{n-1} \frac{x^k}{k!} = O\left(\frac{1}{n!}\right) \quad \text{ (why?)}\]

\[\text{Fast convergence}\]

\[\ln 2 - \sum_{k=1}^{n-1} \frac{(-1)^{k-1}}{k} = O\left(\frac{1}{n}\right) \quad \text{ (why?)}\]

\[\text{Slow convergence}\]
Similar notation for functions (See book!)

- \( f(x) = O(g(x)) \) if \( \exists c, r \text{ s.t. } f(x) \leq C g(x) \forall x \geq r \)
- \( f(x) = o(g(x)) \) if \( \lim_{x \to \infty} \frac{f(x)}{g(x)} = 0 \)

Mean Value Theorem for Integrals

Let \( u \) and \( v \in C([a, b]) \) and suppose \( v \geq 0 \). Then, \( \exists \xi \in [a, b] \) with

\[
\int_{a}^{b} v(x) u(x) \, dx = u(\xi) \int_{a}^{b} v(x) \, dx
\]

Proof: exercise
Implicit Function Theorem:

Let $G$ be a function of two real variables that is continuously differentiable in a neighborhood $\mathcal{O}$ of $(x_0, y_0)$. If $G(x_0, y_0) = 0$ and $\frac{\partial G}{\partial y} \neq 0$ at $(x_0, y_0)$, then there exists a $\delta > 0$ and $F$ continuously differentiable for $|x-x_0| < \delta$ such that $f(x) = y_0$ and $G(x, f(x)) = 0$ for all $x : |x-x_0| < \delta$. 

\[ G(x) : f(x) = y_0, \quad G(x, f(x)) = 0 \]
Example: \( G(x, y) = x^2 + 2y^3 - y^3 \)

Note: \( G(x, y) \) is continuously differentiable

\[ \frac{\partial G}{\partial y} = 16y^2 - 3y^2 \]

- \( G(0, 0) = 0 \) but \( \frac{\partial G}{\partial y} \bigg|_{(0,0)} = 0 \) so IFT does not apply
- \( G(-1, -1) = 0 \) & \( \frac{\partial G}{\partial y} \bigg|_{(-1, -1)} = 13 \) so IFT applies!

So IFT: \( f(-1) = -1 \) & \( G(x, h(x)) = 0 \) in the neighborhood of \( x_0 = -1 \).

Example: Let \( y \) be defined implicitly by the equation

\[ x^3 - y^3 + 4x^2 + y^4 - 24 = 0 \]

Find \( \frac{dy}{dx} \bigg|_{(2,1)} \)

Recall that \( y \) is a function of \( x \) by IFT.

Differentiate: \( 3x^2 - 3y^2 \frac{dy}{dx} + 8x + 4y^3 \frac{dy}{dx} = 0 \)

\[ \Rightarrow \frac{dy}{dx} = \frac{(7y^2 - 4y^3)}{(3x^2 + 8x)} \]

\[ \Rightarrow \frac{dy}{dx} \bigg|_{(2,1)} = \frac{1}{3} \cdot 28. \]
Solutions of non-linear equations

Goal: Want to find roots of equations (zeros of functions)

i.e.,

\[ \text{Find } x \text{ so that } f(x) = 0 \]
\[ x \in \mathbb{R} \quad \text{or} \quad x \in \mathbb{R}^n \]
\[ \mathbb{R} \rightarrow \mathbb{R} \quad \text{or} \quad \mathbb{R}^n \rightarrow \mathbb{R}^n \]

Next few lectures: Bisection method, Newton's method, Secant method

Section 3.1: Bisection Method

Based on simple observations

- If \( f(a) > 0 \) & \( f(b) < 0 \) (or vice versa) & if \( f \) is continuous on \([a, b]\) then (by the Intermediate value theorem) there is some point in \([a, b]\) s.t. \( f(x) = 0 \).
  - Say \( x \)

- If \( c = \frac{a + b}{2} \) (midpt of the interval) then \( x \) is either in \([a, c]\) or \((c, b]\). So we check: if \( f(c) f(a) < 0 \) then \( \exists x \in [a, c] \) with \( f(x) = 0 \)
otherwise \(f(c)f(b)<0\) \& \(\exists x \in (c,b)\) with \(f(x)=0\)  
(\(or\) \(f(c)=0\))

- We can keep going this way until \(|f(c)|<\varepsilon\) for some tolerance \(\varepsilon\).

**Example:**

Use **bisection** method to find a root of \(e^x=x^2\).

**Solve:** Let \(f(x)=e^x-x^2\) \& note that

- \(f(-1)=\frac{1}{e}-1<0\) \& \(f(0)=e^0-1>0\)
  
  so we can take \([a_1,b_1]=[-1,0]\)

  \(c_1=\frac{-1+0}{2}=-\frac{1}{2}\), \(f(c_1)=e^{-\frac{1}{2}}-\frac{1}{4}=\frac{1}{\sqrt{e}}-\frac{1}{4}>0\)
  
  \(\Rightarrow f(a_1)f(c_1)<0\)

- \(a_2=a_1\), \(b_2=c_1\)

  \(c_2=\frac{-1+(-\frac{1}{2})}{2}=-\frac{3}{4}\), \(f(c_2)=e^{-\frac{3}{4}}-\frac{9}{16}<0\)

  \(\Rightarrow f(b_2)f(c_2)<0\)

- \(a_3=c_2\), \(b_3=b_2=-\frac{1}{2}\) \(\Rightarrow c_3=\frac{-\frac{3}{4}-\frac{1}{2}}{2}=-\frac{5}{8}\)

  \(f(-\frac{5}{8})=e^{-\frac{5}{8}}-\frac{\frac{15}{64}>0}\) \(so\) \(c_4=...\)

**Obviously, computers are better suited for this**
Bisection algorithm (pseudocode)

Define the interval

Input: $a, b, M, S, E$ — stop if $f(c) < S$

Max number of iterations

Initialization: $u = f(a)$

$v = f(b)$

$e = b - a$

If $\text{sign}(u) = \text{sign}(v)$ then stop

For $k = 1$ to $M$ do

$e = e/2$

$c = a + e$

$w = f(c)$

(output the error)

$(a + e = a + b - a = b - a)$

(compute the $f$ value at midpt)

Output $k, c, w, e$

If $|e| < S$ or $|w| < E$ then stop

If $\text{sign}(w) \neq \text{sign}(u)$ then

$b = c$

$v = w$

(update the interval & the $f$ value)

Else

$a = c$

$u = w$

EndIf

End Do
Error analysis:

**Theorem:** If \([a_0, b_0], [a_1, b_1], \ldots, [a_n, b_n], \ldots\) denote the intervals that arise in the bisection method then

\[
\lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n = r \quad \text{s.t.} \quad f(r) = 0
\]

Moreover, \( r = \lim_{n \to \infty} c_n \) where \( c_n = \frac{a_n + b_n}{2} \)

and \( |r - c_n| \leq 2^{-(n+1)}(b_0 - a_0) \)

**Proof:** Note that either \( a_1 = a_0 \) & \( b_1 \leq b_0 \)

or \( a_1 \geq a_0 \) & \( b_1 = b_0 \)

& either way \( b_1 - a_1 = \frac{b_0 - a_0}{2} \)

\( a_0 \leq a_1 \) & \( b_0 \geq b_1 \)

continuing this way

\[
\begin{align*}
& a_0 \leq a_1 \leq \ldots \leq b_0 \\
& b_0 \geq b_1 \geq \ldots \geq a_0 \\
& b_n - a_n = \frac{1}{2}(b_{n-1} - a_{n-1})
\end{align*}
\]

So \((a_n)\) is a non-decreasing seq. with upper bound \(b_0\)

\( \Rightarrow \) it converges

Similarly \((b_n)\) converges
Moreover \( b_n - a_n = \frac{1}{2} (b_{n-1} - a_{n-1}) = \frac{1}{2} \cdot \frac{1}{2} (b_{n-2} - a_{n-2}) = \ldots = \frac{1}{2^n} (b_0 - a_0) \)

\[ \Rightarrow \lim_{n \to \infty} b_n - \lim_{n \to \infty} a_n = \lim_{n \to \infty} 2^{-n} (b_0 - a_0) = 0 \]

But \( f(a_n) f(b_n) \leq 0 \) so taking limits
\[ (f(r))^2 \leq 0 \Rightarrow f(r) = 0 \] (\( r \) is a root)

Finally
\[ |r - c_n| \leq \frac{|a_n - a_{n-1}|}{2} \leq 2^{-n+1} (b_0 - a_0) \]

\( r \) is in \([a_n, b_n]\) & \( c_n \) is the midpoint.

Example: If \( a_0 = 10 \) & \( b_0 = 20 \)

How many steps of the bisection method ensure that the relative accuracy is \( 10^{-7} \).

Relative accuracy \( \Rightarrow \frac{|r - c_n|}{|r|} \leq 10^{-7} \)

but \( r > a_0 > 10 \) so it's enough to have
\[ |r - c_n| \leq 10^{-8} \]

we know that \( |r - c_n| \leq 2^{-n+1} (20 - 10) \)

So we want \( 10 \cdot 2^{-(n+1)} \leq 10^{-8} \)

\[ -(n+1) \leq \log_2 10^{-8} \Rightarrow n \geq (\log_2 10^8) - 1 \] so \( n \geq 28 \)