Numerical Analysis : [MA214]
Lecture 1

Instructor: Prof. Tony J. Puthenpurakal

Department of Mathematics
Indian Institute of Technology Bombay

tputhen@math.iitb.ac.in

Textbooks:

1. Elementary Numerical Analysis by S. D. Conte and C. deBoor.
2. Numerical Analysis by Richard L. Burden and J. D. Faires.

Instructions:

- 80% Attendance Required.
- 2 Quizes (10% each) : 20%.
- Midsem : 30%.
- EndSem : 50%
- Passing approximately 40% of Max. score.

Note: Last time max score was above 90%. So passing marks $\approx$ 36%

Calculators are a must for this course. You need it in tutorials, quizzes, mid-sem, and end-sem.

All calculations will be done to 4 sig. digit accuracy. (SCI-4 in Casio Calculators)

Note: Programmable calculators, tablet P.C's are not allowed in exam.

Note: All angles will be in radians. All logarithms will be to base e.

What is Numerical Analysis?

I will explain it through examples

1. $\int_a^b f(x)dx$ exists for example when $f : [a, b] \to \mathbb{R}$ is continuous.

However in most cases it is impossible to compute it.

Examples:-

1. $I_1 = \int_0^1 sin(x^2)dx$
2. $I_2 = \int_0^1 e^{-x^2}dx$
3. $I_3 = \int_0^{\frac{1}{4}} \frac{1}{\sqrt{1-x^3}}dx$

Not only do we have to approximate the integrals, we also have to do it with pre-assigned accuracy. For example, approximate I_1 upto 10^{-6}, *i.e* $|I_1 - approx| \leq 10^{-6}$.

Suppose you are given a function $f : [0, 2] \to \mathbb{R}$ at some values $x_0, x_1, \ldots, x_n$.

Problem: approximate $f(t)$ at $t \in [0, 2] - \{x_1, x_2, \ldots, x_n\}$

Graphically:

Goal: To find a curve passing through these points mentioned in above figure.

Methods

There are two standard methods to do this job.

1. Lagrange Interpolation
2. Piecewise Methods
   (a) Piecewise Linear
   (b) Cubic Spine Interpolation

Initial Value Differential Equations

$$\frac{dy}{dx} = f(x, y), \text{ where } y(x_0) = y_0. \tag{1}$$

In general not possible to find y exactly.

Example:

$$\frac{dy}{dx} = sin(x + y^2), \text{ where } y(0) = 1$$

However for many applications approximate value is enough.

Suppose in the previous example you have to approximate $y(1)$.

All methods will first approximate in between points and then find $Y(1)$.

for example $y(0.1)$ is approximated first. Then using this $y(0.2)$ is approximated.
$y(0.3)$ is approximated using $y(0.2)$.
$y(0.4)$ is approximated using $y(0.3)$.
$\vdots$
$y(1)$ is approximated using $y(0.9)$.

This creates an additional issue and that is of error propagation.

Let A be a $n \times n$ matrix.

Recall $\lambda \in \mathbb{R}$ is said to be an eigenvalue of matrix A if there exists $\overline{x}(\neq 0)$ such that

$$A\overline{x} = \lambda\overline{x}$$

Here, $\overline{x}$ is called eigen-vector corresponding to λ.

Question: How do you find eigenvalues and eigenvectors ?

In applications the size of the matrix is large.

$n \geq 10,000$ is common

$n \geq 1$ million for significant % of cases.

So usual method of finding

$$p(t) = |tI - A|$$

and then finding roots of $p(t)$ is not feasible.

In practice, matrix A will have a dominant eigenvalue $i.e$, there exists λ_0 s.t.

$$|\lambda_0| \geq |\lambda_i|$$

for all other eigenvalues λ_i and it is enough for applications to find λ_0.

Floating Point Arithmetic

n-digit Floating Point number in base β has the form

$$n = \pm(.d_1 d_2 \ldots d_n)_\beta \beta^e$$

$$(.d_1 d_2 \ldots d_n)_\beta \rightarrow \textit{mantissa}$$

$$e \rightarrow \textit{exponent}$$

For most computers, $\beta = 2$

For Calculators $\beta = 10$

$x =$ real number, $fl(x) =$ floating point representation of x.

Example: $\sqrt{0.5} = 7.071E{-1}$ in 4 significant digits.

Suppose x^* is an approximation to x. Then

$$|x - x^*| = \text{ Absolute error}$$

$$\frac{|x - x^*|}{|x|} := \text{ relative error, (provided } x \neq 0)$$

Problem: We do not know x. So how to find relative error?

$$\begin{aligned} \alpha &= \frac{x - x^*}{x^*} \\ \frac{x - x^*}{x} &= \frac{\alpha}{1+\alpha} \approx \alpha, \text{ if } \alpha \text{ is small} \end{aligned}$$

Definition:- x^* is said to be approximate x to t significant digits if

$$\left|\frac{x - x^*}{x}\right| \leq 5 \times 10^{-t}, \text{ we are assuming } x \neq 0$$

Loss of Significant digits

It is not true that if

$$x \sim x^* \text{ and } y \sim y^* \text{ in 4 sig digits}$$

$$\Longrightarrow$$

$$\begin{aligned} x + y &\sim x^* + y^* && \text{in 4 sig digits} \\ x - y &\sim x^* - y^* && \text{in 4 sig digits} \\ xy &\sim x^* y^* && \text{in 4 sig digits} \\ \frac{x}{y} &\sim \frac{x^*}{y^*} && \text{in 4 sig digits} \end{aligned}$$

Things which create loss of significant digits

1. Subtraction of nearly equal quantities
2. division by number which is close to zero

Example 1

$$f(x) = 1 - cosx$$
$$cos(0.01) = 1 \text{ in 4 sig digits}$$
$$f(0.01) = 1 - 1 = 0$$

Actual answer is $5E{-}5$, ($i.e.$ 5×10^{-5})

Loss of sig digits arises since $cos(0.01) = 1$ in 4 sig digits

To avoid this

$$\begin{aligned} f(x) &= 1 - cosx \\ &= \frac{1 - cos^2x}{1 + cosx} \\ &= \frac{sin^2x}{1 + cosx} \\ f(0.01) &= \frac{1E{-}4}{1+1} = 5E{-}5 \end{aligned}$$

Example 2

$$x^2 + 111.11x + 1.2121 = 0$$

$$\begin{aligned} b^2 &= 1.235E4 \\ b^2 - 4ac &= 1.234E4 \\ \sqrt{b^2 - 4ac} &= 1.111E2 = b \text{ in 4 sig digits} \end{aligned}$$

therefore

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = 0$$

Again loss of sig digits occur because we are subtracting nearly equal quantities.

$$\begin{aligned} x_1 &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} \times \frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}} \\ &= \frac{-2c}{b + \sqrt{b^2 - 4ac}} = \frac{1.212}{222.1} \\ &= 1.091E{-2} \text{ correct upto 4 sig digits} \end{aligned}$$

Example 3

$$f(x) = \frac{x - sinx}{tanx}\ \text{, so}\ f(0.01) = \frac{0.01 - 1E^{-2}}{1E^{-2}} = 0$$

$$\text{Rewrite } f(x) = \frac{x - sinx}{tanx} \times \frac{x + sinx}{x + sinx}$$

$$= \frac{x^2 - sin^2x}{tanx(x + sinx)}$$

$$f(0.01) = \frac{1(E{-}4) - 1E(-4)}{1(E{-}2)(0.01 + 1(E{-}2))} = 0$$

Actual value $f(0.01) \approx 1.667E{-}5$ 4 sig digits. So one has to use Taylor series

$$sinx \approx x - \frac{x^3}{6} \text{ and } tanx \approx x - \frac{x^3}{3}$$

$$f(x) \approx \frac{x^3/6}{x - x^3/3} \approx \frac{x^2/6}{1 - x^2/3}$$

$$f(0.01) = 1.667E{-}5$$

Error Propagation

Once an error is committed it contaminates subsequent results

Error propagation is studied in terms of two related concepts:

1. condition
2. instability

1. condition:

$$\begin{aligned}
\text{Condition} \quad &\leftrightarrow \quad \text{sensitivity of } f(x) \text{ to changes in } x \\
&= \quad max \left\{ \frac{\frac{|f(x)-f(x^*)|}{|f(x)|}}{\frac{|x-x^*|}{|x|}} : |x - x^*| \text{ small} \right\} \\
&\approx \quad \left| \frac{f'(x)x}{f(x)} \right|
\end{aligned}$$

The larger the condition, the more ill-conditioned the function is said to be.

Examples

Example 1 : $f(x) = \sqrt{x}$, $f'(x) = \frac{1}{\sqrt{2x}}$

$$\left|\frac{f'(x)x}{f(x)}\right| = \left|\frac{\frac{1}{\sqrt{2x}}x}{\sqrt{x}}\right| = \frac{1}{2}$$

So taking square-root is well conditioned, since it actually reduces the relative error.

Example 2 : $f(x) = \frac{10}{1-x^2}$

$$\left|\frac{f'(x)x}{f(x)}\right| = \left|\frac{2x^2}{1-x^2}\right| \text{ , large when } |x| \text{ is close to 1}$$

What to do:

$$\text{put } x = 1 - y \text{ , then } y \sim 0 \text{ if } x \sim 1$$

$$f(x) = \frac{10}{1-(1-y)^2} = \frac{10}{2y - y^2} \approx \frac{10}{2y} = \frac{5}{y} \text{ , since } y^2 \sim 0$$

$g(y) = \frac{5}{y}$

$$\left|\frac{g'(y)y}{g(y)}\right| = 1$$

Example of instability

$$\begin{aligned}
\frac{du_1}{dt} &= 9u_1 + 24u_2 + 5cost - \frac{1}{3}sint \\
\frac{du_2}{dt} &= -24u_1 - 51u_2 - 9cost + \frac{1}{3}sint \\
u_1(0) &= \frac{4}{3} \text{ and } u_2(0) = \frac{2}{3}
\end{aligned}$$

exact solution

$$\begin{aligned}
u_1(t) &= 2e^{-3t} - e^{-39t} + \frac{1}{3}cost \\
u_2(t) &= -e^{-3t} + 2e^{-39t} - \frac{1}{3}cost
\end{aligned}$$

RK method of order 4 with step size 0.1

$\tilde{u}_1$	approximate to u_1
$\tilde{u}_2$	approximate to u_2

t	$u_1(t)$	$\tilde{u}_1(t)$
0.1	1.793061	-2.645169
0.2	1.423901	-18.45158
$\vdots$	$\vdots$	$\vdots$
0.9	0.3416143	-695332
1.0	0.2796748	-3099671

t	$u_2(t)$	$\tilde{u}_2(t)$
0.1	-1.032001	7.844527
0.2	-0.8746809	38.87631
$\vdots$	$\vdots$	$\vdots$
0.9	-0.2744088	1390664
1.0	-0.2298877	6199352

Instability

Instability=sensitivity of the numerical process for the calculation of $f(x)$ from x to the inevitable rounding error committed in a calculator or a computer.

Example:- $f(x) = \sqrt{x+1} - \sqrt{x}$

$$condition = \left|\frac{f'(x)x}{f(x)}\right| = \frac{1}{2}\frac{x}{\sqrt{x+1}\sqrt{x}} \approx \frac{1}{2}$$

So conditioning is good

$$f(12345) = 1.111E2 - 1.111E2 = 0.0$$

Actual value $= 4.5E{-3}$. So we analyze what goes wrong

$$\begin{aligned}
x_0 &= 12345 \\
x_1 &= f_1(x_0) = x_0 + 1 = 12346 \\
x_2 &= f_2(x_1) = \sqrt{x_1} \\
x_3 &= f_3(x_0) = \sqrt{x_0} \\
f_4(t) &= x_2 - t
\end{aligned}$$

Condition of f_4 is

$$\left|\frac{f_4'(t)t}{f_4(t)}\right| = \left|\frac{t}{x_2 - t}\right|$$

f_4 is well conditioned except when $t = x_2$

In our example $x_2 - x_3 \approx 0.005$, $x_3 - t \approx 111.11$

So condition of $f_4 \approx 22,222 \approx 40,000$ times condition of f.

What to do

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

$$\begin{aligned} f(12345) &= \frac{1}{1.111E2 + 1.111E2} \\ &\approx 4.502E{-3} \text{ correct up to 3 sig digits} \end{aligned}$$

Note: It is possible to estimate the effect of instability by considering the rounding errors once at a time

Theorem (Intermediate-value theorem for continuous function's)

Let $f : [a, b] \longrightarrow \mathbb{R}$ be a continuous function. $x_1, x_2 \in [a, b]$ and say $f(x_1) < \alpha < f(x_2)$. Then there exists $c \in [a, b]$ such that $f(c) = \alpha$.

Corollary

Let $f : [a, b] \longrightarrow \mathbb{R}$ be a continuous function. Let $x_1, x_2, \cdots, x_n \in [a, b]$ and let $g_1, g_2, \cdots, g_n$ be real numbers all of one sign. Then

$$\sum_{i=1}^{n} f(x_i) g_i = f(\xi) \sum_{i=1}^{n} g_i, \text{ for some } \xi \in [a, b]$$

Proof (sketch)

We may assume $g_i \geq 0$. Without loss of generality assume

$$f(x_1) = min\{f(x_i) : i = 1, 2, \cdots, n\}$$
$$f(x_n) = max\{f(x_i) : i = 1, 2, \cdots, n\}$$

Then

$$f(x_1)\sum_{i=1}^{n} g_i \leq \sum_{i=1}^{n} f(x_i)g_i \leq f(x_n)\sum_{i=1}^{n} g_i$$

$$h(x) = f(x)\sum_{i=1}^{n} g_i \text{ is continuous}$$

$$h(x_1) \leq \sum_{i=1}^{n} f(x_i)g_i \leq h(x_n)$$

So by the intermediate value Theorem there exists ξ such that

$$h(\xi) = \sum_{i=1}^{n} f(x_i)g_i$$

Similarly One has the following:-

Corollary

Let $g : [a, b] \longrightarrow \mathbb{R}$ *be non-negative (or non-positive) integrable function. Let* $f : [a, b] \longrightarrow \mathbb{R}$ *be continuous. Then*

$$\int_a^b f(x)g(x)dx = f(\xi)\int_a^b g(x)dx \text{ for some } \xi \in [a, b]$$

Remark:- The assumption $g(x)$ is of one sign is essential.

Example:- $f(x) = g(x) = x,\ x \in [-1, 1]$

$$\int_{-1}^{1} f(x)g(x)dx = \int_{-1}^{1} x^2 dx = \frac{2}{3}$$

However $\int_{-1}^{1} g(x)dx = 0$

Existance of Maximum and Minimum of continuous function

Theorem

Let $f : [a, b] \longrightarrow \mathbb{R}$ be continuous. Then there exists $\alpha, \beta \in [a, b]$ such that

$$f(\alpha) \leq f(x) \leq f(\beta), \forall x \in [a, b]$$

Let us recall:-

Theorem (Rolle's Theorem)

Let $f : [a, b] \longrightarrow \mathbb{R}$ be a continuous function and assume $f : (a, b) \longrightarrow \mathbb{R}$ is differentiable. If $f(a) = f(b)$, then there exists $\xi \in (a, b)$ such that $f'(\xi) = 0$

Rolle's Theorem implies the famous mean value Theorem

Theorem (M.V.T.)

Let $f : [a, b] \longrightarrow \mathbb{R}$ be a continuous function and assume $f : (a, b) \longrightarrow \mathbb{R}$ is differentiable. Then

$$\frac{f(b) - f(a)}{b - a} = f'(\xi) \text{ for some } \xi \in (a, b)$$

Proof.

Apply Rolle's Theorem to

$$F(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a)$$

$\square$

Taylor's formula with integral remainder

Theorem (Taylor's formula with integral remainder)

If $f(x)$ has $(n+1)$ continuous derivatives on $[a, b]$ and c is some point in $[a, b]$. Then for all $x \in [a, b]$

$$\begin{aligned} f(x) &= f(c) + f'(c)(x-c) + \frac{f''(c)}{2}(x-c)^2 + \cdots \\ &\quad \cdots + \frac{f^{(n)}(c)}{n!}(x-c)^n + R_{n+1}(x) \\ \text{where } R_{n+1}(x) &= \frac{1}{n!}\int_c^x (x-s)^n f^{n+1}(s)ds \end{aligned}$$

Let $p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$, $a_n \neq 0$, $n \geq 1$ and all $a_i \in \mathbb{C}$.

It is easy to see that $p(x)$ has atmost n roots.

However the following is non-trivial to prove

Theorem (FTA)

$p(x)$ has a root, i.e., there exists $\xi \in \mathbb{C}$ such that $p(\xi) = 0$

Note $\frac{1}{n^p} \longrightarrow 0$ for any $p \geq 0$

Intuitively, $\frac{1}{n} \longrightarrow 0$ more slowly than $\frac{1}{n^2}$ and so on.

Definition

Let $\{\alpha_n\}_{n\geq 1}$ and $\{\beta_n\}_{n\geq 1}$ be two sequences. We say α_n is *big − oh* of β_n and we write

$$\alpha_n = O(\beta_n)$$

If $|\alpha_n| \leq k|\beta_n|$, for some constant k and for all $n >> 0$

Examples:-

$\{\frac{1}{n}\}$, $\{\frac{100}{n}\}$, $\{\frac{10}{n} - \frac{40}{n^2} + e^{-n}\}$, $\{\frac{1}{n^p}\}$ (here $p \geq 1$) are all $O(\frac{1}{n})$.

Definition

Let $\{\alpha_n\}_{n\geq 1}$ and $\{\beta_n\}_{n\geq 1}$ be two sequences. We say α_n is *little − oh* of β_n and we write it as $\alpha_n = o(\beta_n)$, if

$$\lim_{n\to\infty} \frac{\alpha_n}{\beta_n} = 0$$

Example:

1. $\{\frac{1}{n^2}\}$
2. $\{\frac{1}{n\log n}\}$

both are $o(\frac{1}{n})$

Important Remark:-

A convergence order rate of $\frac{1}{n}$ is much too slow to be useful in calculations.

Example

$$\frac{\pi}{4} = \sum_{i=0}^{\infty} \frac{(-1)^i}{2i+1} = 1 - \sum_{j=1}^{\infty} \frac{2}{16j^2 - 1}$$

Set $\alpha_n = 1 - \sum\limits_{j=1}^{n} \frac{2}{16j^2-1}$.

The sequence $\{\alpha_n\}$ is monotonically decreasing. Moreover

$$0 \leq \alpha_n - \frac{\pi}{4} \leq \frac{1}{4n+3} \text{ , } n = 1, 2, \cdots$$

to calculate $\frac{\pi}{4}$ correctly to within 10^{-6} we would need $10^6 \leq 4n + 3$ or roughly $n = 250,000$ calculations.

However round of errors in calculation $\alpha_{250,000}$ will be usually greater that 10^{-6}. Here

$$\begin{aligned} \alpha_n &= \frac{\pi}{4} + O\left(\frac{1}{n}\right) \\ \alpha_n &\neq \frac{\pi}{4} + o\left(\frac{1}{n}\right) \end{aligned}$$