MA 322: Scientific Computing

Department of Mathematics
Indian Institute of Technology Guwahati

January 5, 2023

CHAPTER -1: COURSE RELATED MATTERS

About the course

MA322 SCIENTIFIC COMPUTING [3-0-2-8]

Prerequisites: Nil

Errors; Numerical methods for solving scalar nonlinear equations; Interpolation and approximations, spline interpolations; Numerical integration based on interpolation, quadrature methods, Gaussian quadrature; Initial value problems for ordinary differential equations - Euler method, Runge-Kutta methods, multi-step methods, predictor-corrector method, stability and convergence analysis; Finite difference schemes for partial differential equations - explicit and implicit schemes; Consistency, stability and convergence; Stability analysis (matrix method and von Neumann method), Lax equivalence theorem; Finite difference schemes for initial and boundary value problems (FTCS, backward Euler and Crank-Nicolson schemes, ADI methods, Lax Wendroff method, upwind scheme).

Texts:

1. D. Kincaid and W. Cheney, Numerical Analysis: Mathematics of Scientific Computing, 3rd Ed., AMS, 2002.
2. G. D. Smith, Numerical Solutions of Partial Differential Equations, 3rd Ed., Calrendorn Press, 1985.

References:

1. K. E. Atkinson, An Introduction to Numerical Analysis, Wiley, 1989.
2. S. D. Conte and C. de Boor, Elementary Numerical Analysis - An Algorithmic Approach, McGraw-Hill, 1981.
3. R. Mitchell and S. D. F. Griffiths, The Finite Difference Methods in Partial Differential Equations, Wiley, 1980.
4. Richard L. Burden and J. Douglas Faires, Numerical analysis, Brooks/Cole, 2001.

- Lecture: **C1** (Tue, Wed, Thu: 15:00-15:55); Venue: 5102.
- Lab: **ML-2** (Tue: 09:45-11:40); Venue: Mathematics Department Lab (E).

About the course

MA322 SCIENTIFIC COMPUTING [3-0-2-8]

Prerequisites: Nil

Errors; Numerical methods for solving scalar nonlinear equations; Interpolation and approximations, spline interpolations; Numerical integration based on interpolation, quadrature methods, Gaussian quadrature; Initial value problems for ordinary differential equations - Euler method, Runge-Kutta methods, multi-step methods, predictor-corrector method, stability and convergence analysis; Finite difference schemes for partial differential equations - explicit and implicit schemes; Consistency, stability and convergence; Stability analysis (matrix method and von Neumann method), Lax equivalence theorem; Finite difference schemes for initial and boundary value problems (FTCS, backward Euler and Crank-Nicolson schemes, ADI methods, Lax Wendroff method, upwind scheme).

Texts:

1. D. Kincaid and W. Cheney, Numerical Analysis: Mathematics of Scientific Computing, 3rd Ed., AMS, 2002.
2. G. D. Smith, Numerical Solutions of Partial Differential Equations, 3rd Ed., Calrendorn Press, 1985.

References:

1. K. E. Atkinson, An Introduction to Numerical Analysis, Wiley, 1989.
2. S. D. Conte and C. de Boor, Elementary Numerical Analysis - An Algorithmic Approach, McGraw-Hill, 1981.
3. R. Mitchell and S. D. F. Griffiths, The Finite Difference Methods in Partial Differential Equations, Wiley, 1980.
4. Richard L. Burden and J. Douglas Faires, Numerical analysis, Brooks/Cole, 2001.

- Lecture: **C1** (Tue, Wed, Thu: 15:00-15:55); Venue: 5102.
- Lab: **ML-2** (Tue: 09:45-11:40); Venue: Mathematics Department Lab (E).

Course Policy

- Quizes/Assignments: 20%.
- Mid-sem exam: 20%.
- End-sem exam: 35%.
- Lab tests: 25%.
- Attendance in the lectures is **mandatory**. You will not be allowed to appear in the exam if your **attendance < 75%.**
- You **must attend** the LAB sessions **without fail** to gain maximum from the lab and it will play a crucial role in your **GRADE** in this course.

Course Policy

- Quizes/Assignments: 20%.
- Mid-sem exam: 20%.
- End-sem exam: 35%.
- Lab tests: 25%.
- Attendance in the lectures is **mandatory**. You will not be allowed to appear in the exam if your **attendance $<$ 75%.**
- You **must attend** the LAB sessions **without fail** to gain maximum from the lab and it will play a crucial role in your **GRADE** in this course.

About the instructor and TA(s)

- Instructor: SATYAJIT PRAMANIK
- TA(s): Mr. PUSPENDU JANA, TBD

Where you can 'get hold of' the instructor!

- Physically: E1-305, Department of Mathematics
- Electronically: satyajitp [AT] iitg [DOT] ac [DOT] in
- Office Hours: MONDAY **14:00-15:00** with **prior appointment**

About the instructor and TA(s)

- Instructor: SATYAJIT PRAMANIK
- TA(s): Mr. PUSPENDU JANA, TBD

Where you can 'get hold of' the instructor!

- Physically: E1-305, Department of Mathematics
- Electronically: satyajitp [AT] iitg [DOT] ac [DOT] in
- Office Hours: MONDAY **14:00-15:00** with **prior appointment**

CHAPTER 0: PRELIMINARIES

Recall (from MA101, MA102 and ...)

Theorem (Intermediate Value Theorem)

On an interval $[a, b]$, a continuous function assumes all values between $f(a)$ and $f(b)$.

Theorem (Taylor's Theorem with Lagrange Remainder)

If $f \in C^n[a, b]$ and if $f^{(n+1)}$ exists on the open interval (a, b), then for any points c and x in the closed interval $[a, b]$,

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + E_n(x),$$

where, for some point ξ between c and x, the error term is

$$E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x-c)^{n+1}.$$

Recall (from MA101, MA102 and ...)

Theorem (Intermediate Value Theorem)

On an interval $[a, b]$, a continuous function assumes all values between $f(a)$ and $f(b)$.

Theorem (Taylor's Theorem with Lagrange Remainder)

If $f \in C^n[a, b]$ and if $f^{(n+1)}$ exists on the open interval (a, b), then for any points c and x in the closed interval $[a, b]$,

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + E_n(x),$$

where, for some point ξ between c and x, the error term is

$$E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x-c)^{n+1}.$$

Recall (from MA101, MA102 and ...)

Corollary (Maclaurin series)

An important special case arises when $c = 0$. In this case, Taylor theorem gives

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(0) x^k + E_n(x),$$

where, for some point ξ between c and x, the error term is

$$E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) x^{n+1}.$$

Recall (from MA101, MA102 and ...)

Theorem (Mean-Value Theorem)

If f is in $C[a, b]$ and if f' exists on the open interval (a, b), then for x and c in the closed interval $[a, b]$,

$$f(x) = f(c) + f'(\xi)(x - c),$$

where ξ is between c and x.

We will use this theorem to approximate $f'(x)$.

Theorem (Rolle's Theorem)

If f is continuous on $[a, b]$ and if f' exists on the open interval (a, b), and if $f(a) = f(b)$, then $f'(\xi) = 0$ for some ξ in the open interval (a, b).

Recall (from MA101, MA102 and ...)

Theorem (Mean-Value Theorem)

If f is in $C[a,b]$ and if f' exists on the open interval (a,b), then for x and c in the closed interval $[a,b]$,

$$f(x) = f(c) + f'(\xi)(x - c),$$

where ξ is between c and x.

We will use this theorem to approximate $f'(x)$.

Theorem (Rolle's Theorem)

If f is continuous on $[a,b]$ and if f' exists on the open interval (a,b), and if $f(a) = f(b)$, then $f'(\xi) = 0$ for some ξ in the open interval (a,b).

Recall (from MA101, MA102 and ...)

Theorem (Taylor's Theorem with Integral Remainder)

If $f \in C^{n+1}[a, b]$, then for any points c and x in the closed interval $[a, b]$,

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + R_n(x),$$

where

$$R_n(x) = \frac{1}{n!} \int_c^x f^{(n+1)}(t)(x-t)^n \mathrm{d}t.$$

Recall (from MA101, MA102 and ...)

Theorem (Alternative form of Taylor's Theorem)

If $f \in C^{n+1}[a, b]$, then for any points x and $x + h$ in the closed interval $[a, b]$,

$$f(x+h) = \sum_{k=0}^{n} \frac{h^k}{k!} f^{(k)}(x) + E_n(h),$$

where

$$E_n(h) = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(\xi),$$

in which the point ξ lies between x and $x + h$.

Recall (from MA101, MA102 and ...)

Theorem (Taylor's Theorem in Two Variables)

Let $f \in C^{n+1}([a,b],[c,d])$. If (x,y) and $(x+h,y+k)$ are points in the rectangle $[a,b]\times[c,d] \subseteq \mathbb{R}^2$, then

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

where

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

in which θ lies between 0 and 1.

Recall (from MA101, MA102 and ...)

Theorem (Mean-Value Theorem for Integrals)

Let u and v be continuous real-valued functions on an interval $[a, b]$, and suppose that $v \geq 0$. Then there exists a point ξ in $[a, b]$ such that

$$\int_a^b u(x)v(x)\mathrm{d}x = u(\xi)\int_a^b v(x)\mathrm{d}x.$$

Definition (Order of convergence)

Let $\{x_n\}$ be a sequence of real numbers tending to a limit x^*. If there positive constants C and α, and an integer N such that

$$|x_{n+1} - x^*| \leq C|x_n - x^*|^\alpha \qquad (n \geq N)$$

we say that the rate of convergence is of order α at least.

Recall (from MA101, MA102 and ...)

Theorem (Mean-Value Theorem for Integrals)

Let u and v be continuous real-valued functions on an interval $[a, b]$, and suppose that $v \geq 0$. Then there exists a point ξ in $[a, b]$ such that

$$\int_a^b u(x)v(x)\mathrm{d}x = u(\xi)\int_a^b v(x)\mathrm{d}x.$$

Definition (Order of convergence)

Let $\{x_n\}$ be a sequence of real numbers tending to a limit x^*. If there positive constants C and α, and an integer N such that

$$|x_{n+1} - x^*| \leq C|x_n - x^*|^\alpha \qquad (n \geq N)$$

we say that the rate of convergence is of order α at least.

Order notations

Definition (Big O)

Let $\{x_n\}$ and $\{\alpha_n\}$ be two sequences. We write

$$x_n = O(\alpha_n)$$

if there are constants C and $n_0 \in \mathbb{N}$ such that $|x_n| \leq C|\alpha_n|$ when $n \geq n_0$. Here, we say that x_n is **BIG "Oh"** of α_n.

Definition (Little o)

Let $\{x_n\}$ and $\{\alpha_n\}$ be two sequences. We write

$$x_n = o(\alpha_n)$$

if, intuitively, $\lim_{n\to\infty}(x_n/\alpha_n) = 0$. Here, we say that x_n is **little "oh"** of α_n.

Order notations

Definition (Big O)

Let $\{x_n\}$ and $\{\alpha_n\}$ be two sequences. We write

$$x_n = O(\alpha_n)$$

if there are constants C and $n_0 \in \mathbb{N}$ such that $|x_n| \leq C|\alpha_n|$ when $n \geq n_0$. Here, we say that x_n is **BIG "Oh"** of α_n.

Definition (Little o)

Let $\{x_n\}$ and $\{\alpha_n\}$ be two sequences. We write

$$x_n = o(\alpha_n)$$

if, intuitively, $\lim_{n\to\infty}(x_n/\alpha_n) = 0$. Here, we say that x_n is **little "oh"** of α_n.

CHAPTER 1: ERRORS

Floating-point numbers

- Most computers have an *integer mode* and a **floating-point mode** for representing numbers.
- A nonzero number x in a computer using base $\beta \in \mathbb{N}$ is stored essentially in the form

$$x = \sigma \cdot (.a_1 a_2 \cdots a_t)_\beta \cdot \beta^e,$$

where $0 \leq a_i \leq \beta - 1$, $\sigma = \pm 1$ is called the sign, $e \in \mathbb{Z}$ is called the exponent, and $(.a_1 a_2 \cdots a_t)_\beta$ is called the mantissa of the floating-point number x. The number β is also called the *radix*, and the point preceding a_1 is called the *radix point*. The integer t gives the number of base β digits in the representation.
- For $a_1 \neq 0$, we call the representation the *normalized floating-point represntation*.

Floating-point numbers

- Computers are not able to operate using real numbers expressed with more than a fixed number of digits. The word length of the computer places a restriction on the precision with which real numbers can be represented.
- Even a simple number like $1/10$ cannot be stored exactly in any binary machine.
- It requires an infinite binary expression:

$$\frac{1}{10} = (0.0\ 0011\ 0011\ 0011\ 0011 \cdots)_2$$

- If we read 0.1 into a 32-bit computer and then print it out to 40 decimal places, we obtain the following result:

$$0.10000\ 00014\ 90116\ 11938\ 47656\ 25000\ 00000\ 00000$$

Floating-point numbers

- Computers are not able to operate using real numbers expressed with more than a fixed number of digits. The word length of the computer places a restriction on the precision with which real numbers can be represented.
- Even a simple number like $1/10$ cannot be stored exactly in any binary machine.
- It requires an infinite binary expression:

$$\frac{1}{10} = (0.0\ 0011\ 0011\ 0011\ 0011 \cdots)_2$$

- If we read 0.1 into a 32-bit computer and then print it out to 40 decimal places, we obtain the following result:

$$0.10000\ 00014\ 90116\ 11938\ 47656\ 25000\ 00000\ 00000$$

Floating-point numbers

- Computers are not able to operate using real numbers expressed with more than a fixed number of digits. The word length of the computer places a restriction on the precision with which real numbers can be represented.
- Even a simple number like $1/10$ cannot be stored exactly in any binary machine.
- It requires an infinite binary expression:

$$\frac{1}{10} = (0.0\,0011\,0011\,0011\,0011\cdots)_2$$

- If we read 0.1 into a 32-bit computer and then print it out to 40 decimal places, we obtain the following result:

$$0.10000\ 00014\ 90116\ 11938\ 47656\ 25000\ 00000\ 00000$$

Floating-point numbers

$$
\begin{aligned}
& \frac{1}{2^4}+\frac{1}{2^5}+\frac{1}{2^8}+\frac{1}{2^9}+\frac{1}{2^{12}}+\frac{1}{2^{13}}+\frac{1}{2^{16}}+\frac{1}{2^{17}}+\cdots \\
= \; & \frac{2^{13}+2^{12}+2^9+2^8+2^5+2^4+2+1}{2^{17}}+\cdots \\
= \; & \frac{8192+4096+512+256+32+16+2+1}{131072}+\cdots \\
= \; & \frac{13107}{131072}+\cdots
\end{aligned}
$$

- We shall be careful/aware of *roundoff errors* — they may contaminate computer calculations.
- We shall also be careful about *a loss of significance*, which may arise when two nearly equal numbers are subtracted.

Floating-point numbers

$$
\begin{aligned}
& \frac{1}{2^4}+\frac{1}{2^5}+\frac{1}{2^8}+\frac{1}{2^9}+\frac{1}{2^{12}}+\frac{1}{2^{13}}+\frac{1}{2^{16}}+\frac{1}{2^{17}}+\cdots \\
= & \frac{2^{13}+2^{12}+2^9+2^8+2^5+2^4+2+1}{2^{17}}+\cdots \\
= & \frac{8192+4096+512+256+32+16+2+1}{131072}+\cdots \\
= & \frac{13107}{131072}+\cdots
\end{aligned}
$$

- We shall be careful/aware of *roundoff errors* — they may contaminate computer calculations.
- We shall also be careful about *a loss of significance*, which may arise when two nearly equal numbers are subtracted.

Floating-point numbers

$$
\begin{aligned}
& \frac{1}{2^4}+\frac{1}{2^5}+\frac{1}{2^8}+\frac{1}{2^9}+\frac{1}{2^{12}}+\frac{1}{2^{13}}+\frac{1}{2^{16}}+\frac{1}{2^{17}}+\cdots \\
= & \frac{2^{13}+2^{12}+2^9+2^8+2^5+2^4+2+1}{2^{17}}+\cdots \\
= & \frac{8192+4096+512+256+32+16+2+1}{131072}+\cdots \\
= & \frac{13107}{131072}+\cdots
\end{aligned}
$$

- We shall be careful/aware of *roundoff errors* — they may contaminate computer calculations.
- We shall also be careful about *a loss of significance*, which may arise when two nearly equal numbers are subtracted.

Rounding vs. chopping/truncating

- If x is rounded so that $\tilde{x}$ is the n-digit approximation to it, then

$$|x - \tilde{x}| \leq \frac{1}{2} \times 10^{-n} \qquad \text{(verify!)}.$$

- If x is chopped/truncated so that $\hat{x}$ is the n-digit approximation to it, then

$$|x - \hat{x}| \leq 10^{-n} \qquad \text{(trivial!)}.$$

Rounding vs. chopping/truncating

- If x is rounded so that $\tilde{x}$ is the n-digit approximation to it, then

$$|x - \tilde{x}| \leq \frac{1}{2} \times 10^{-n} \qquad \text{(verify!)}.$$

- If x is chopped/truncated so that $\hat{x}$ is the n-digit approximation to it, then

$$|x - \hat{x}| \leq 10^{-n} \qquad \text{(trivial!)}.$$

Absolute and Relative Errors: Loss of Significance

Definition (Absolute and relative errors)

When a real number x is approximated by another number x^*, the error is $x - x^*$. The **absolute error** is

$$|x - x^*|$$

and the relative error is

$$\left|\frac{x - x^*}{x}\right|$$

Theorem (Theorem on Loss of Precision)

If x and y are positive normalized floating-point binary machine numbers such that $x > y$ and

$$2^{-q} \leq 1 - \frac{y}{x} \leq 2^{-p}$$

then at most q and at least p significant binary bits are lost in the subtraction $x - y$.

Absolute and Relative Errors: Loss of Significance

Definition (Absolute and relative errors)

When a real number x is approximated by another number x^*, the error is $x - x^*$. The **absolute error** is

$$|x - x^*|$$

and the relative error is

$$\left|\frac{x - x^*}{x}\right|$$

Theorem (Theorem on Loss of Precision)

If x and y are positive normalized floating-point binary machine numbers such that $x > y$ and

$$2^{-q} \leq 1 - \frac{y}{x} \leq 2^{-p}$$

then at most q and at least p significant binary bits are lost in the subtraction $x - y$.