Lecture 1: Error Handling

Abhinav Jha

Indian Institute of Technology, Gandhinagar

9^{th} March 2026

Outline

Why Numerical Methods?

- Existence
- Uniqueness

Why Numerical Methods?

- Existence
- Uniqueness
- What is the solution?

Why Numerical Methods?

- $\sum_{n=1}^{\infty} \frac{1}{n}$
- $\sum_{n=1}^{\infty} \frac{1}{n^2}$

Why Numerical Methods?

- $\sum_{n=1}^{\infty} \frac{1}{n}$
- $\sum_{n=1}^{\infty} \frac{1}{n^2}$
-

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

n	Approximate Sum	Error	Time (s)
10^1	1.54976773	9.52×10^{-2}	0.0000
10^2	1.63498390	9.95×10^{-3}	0.0000
10^3	1.64393457	1.00×10^{-3}	0.0002
10^4	1.64483407	1.00×10^{-4}	0.0024
10^5	1.64492407	1.00×10^{-5}	0.0271
10^6	1.64493307	1.00×10^{-6}	0.2756
10^7	1.64493397	1.00×10^{-7}	2.8460

Table 1: Convergence of the series $\sum_{k=1}^{n} \frac{1}{k^2}$ towards $\pi^2/6$.

Why Numerical Methods?

Numerical Methods

MA203

Numerical Analysis

MA 637

- Binary Number System (Base-2)

  0 and 1

- Decimal Number System (Base-10)

  $619.916 = 6\times10^{2} + 1\times10^{1} + 9\times10^{0} + 9\times10^{-1} + 1\times10^{-2} + 6\times10^{-3}$

Number System

$x = 25_{10}$

Division	Quotient	Remainder
25/2	12	1
12/2	6	0
6/2	3	0
3/2	1	1
1/2	0	1

$25_{10} = 11001_2$

$11001_2 = 1\times2^4 + 1\times2^3 + 0\times2^2 + 0\times2^1 + 1\times2^0$

$= 25_{10}$

$0.125_{10} \longrightarrow$ Binary

Multiplication	Result	Integer Part
0.125×2	0.25	0
0.25×2	0.5	0
0.5×2	1.0	1

$0.125_{10} = 0.001_2$

$25.125_{10} = 11001.001_2$

Scientific Notation

$x_{10} = \pm (r) \times 10^{n}$

r, q_2 = Mantissa

n, m = Exponent

$0.1 \le r < 1$, $n \in \mathbb{Z}$ (Integers)

eg $1301.122 = \underbrace{0.1301122}_{r} \times 10^{4}$ n

$x_2 = \pm q_2 \times 2^{m}$

$0.1_2 \le q_2 < 1.0_2$ and $m \in \mathbb{Z}$

$x_2 = 0 \Rightarrow q_2 = 0$

IEEE Notation

64 bits

$0.1 - - - -$

Sign	Exponent	Mantissa
s	c	f
1 bit	11 bits	52 bits

$2^{-52} = 10^{-d}$ for some $d > 0$

$\Rightarrow \log_{10}(2^{-52}) = -d \log_{10}(10)$

$\Rightarrow d \approx 16$

IEEE Notation

Any number not represented by IEEE and is large leads to overflow
smaller underflow.

Sign	Exponent	Mantissa
s	c	f
1 bit	11 bits	52 bits

$2^{11} = 2048$, -1023 to 1024

$(-1)^s 2^{c-1023}(1+f)$: $0 \le c \le 2047$

Round Off Errors

$y = 0 \cdot d_1 d_2 d_3 \cdots d_k d_{k+1} \cdots \times 10^n$

- **Chopping** : Chop a number to k-digits

$$fe(y) = 0 \cdot d_1 d_2 \cdots d_k \times 10^n$$

- **Rounding** :

$$fe(y) = \begin{cases} 0 \cdot d_1 d_2 \cdots (d_k + 1) \times 10^n & ; \ d_{k+1} \geq 5 \\ 0 \cdot d_1 d_2 \cdots d_k \times 10^n & ; \ \text{else} \end{cases}$$

eg: $\pi = 0 \cdot 3141592 \cdots \times 10^1$

Chopping five-digit $fe(\pi) = 0.31415 \times 10^1$

Rounding $fe(\pi) = 0.31416 \times 10^1$

Measure of Error

p = Known value

p^* = Numerical value

Absolute error : $|p - p^*|$

Relative error : $\dfrac{|p - p^*|}{|p|}$, $p \neq 0$

Arithmetic Errors

$\oplus$ $\ominus$ $\otimes$ $\oslash$

$2+3=5$

$x \oplus y = fe(fe(x) + fe(y))$

$x \otimes y = fe(fe(x) \times fe(y))$

eg: $x = \frac{5}{7}$; $y = \frac{1}{3}$

five-digit chopping: $fe(x) = 0.71428 \times 10^0$

$fe(y) = 0.33333 \times 10^0$

$x \oplus y = fe(fe(x) + fe(y))$

$= fe(0.71428 \times 10^0 + 0.33333 \times 10^0)$

$= fe(1.04761 \times 10^0) = fe(0.104761 \times 10^1)$

$= 0.10476 \times 10^1$

$x+y = \frac{5}{7} + \frac{1}{3} = \frac{22}{21} = 1.0476190\ldots$

$|(x+y) - (x \oplus y)|$

$= 0.19 \times 10^{-4}$

Relative error

$= 0.182 \times 10^{-4}$.

Arithmetic Errors

Example : $x_0^R = 1$

$$x_1^R = \frac{1}{3}$$

$$x_{n+1}^R = \frac{13}{6} x_n^R - \frac{4}{3} x_{n-1}^R \quad : \quad n \geq 1$$

True solution : $x_n^T = \left(\frac{1}{3}\right)^n$

(7) digit chopping in simulations

10

Arithmetic Errors

n	x_n^R	x_n^T	Error
0	1.0000000	1.0000000	0.0000000
1	0.3333333	0.3333333	0.0000000
2	0.1111110	0.1111111	0.0000001
3	0.0370366	0.0370370	0.0000004
4	0.0123439	0.0123457	0.0000018
5	0.0041081	0.0041152	0.0000071
6	0.0013432	0.0013717	0.0000285
7	0.0003431	0.0004572	0.0001141
8	-0.0003042	0.0001524	0.0004566
9	-0.0017757	0.0000508	0.0018265
10	-0.0072891	0.0000169	0.0073060
11	-0.0292185	0.0000056	0.0292241
12	-0.1168947	0.0000019	0.1168966
13	-0.4675857	0.0000006	0.4675863
14	-1.8703451	0.0000002	1.8703453
15	-7.4813812	0.0000001	7.4813813

Table 2: Computed values, exact values, and absolute error

Conditioning !

Taylor Series

$f(x) \in C^{\infty}[a,b]$

$$f(x+h) = f(x) + \frac{h}{1!}f'(x) + \frac{h^2}{2!}f''(x) + \cdots + \frac{h^n}{n!}f^{(n)}(x) + \cdots$$

Truncate it → Some n

Mathematical errors

Root finding Methods

$$x^2 - 1 = 0$$

$$x\sin(x) - e^x\cos(x) + \tan(x) = 0$$