CS2262: Numerical Methods

January 16, 2024

Dr. James Ghawaly

Introduction

- **History**:
  - **Education**: University of Tennessee Knoxville (PhD, MS, BS)
  - **Oak Ridge National Laboratory**: Staff Data Scientist
  - **LSU**: Assistant Professor in CSE, previously Senior Research Scientist
- **Research**: AI/ML, applications to national security, cybersecurity
- **Teaching**:
  - HNRS3025: Large Language Models for Real World Applications
  - CS2262: Numerical Methods
  - More to come!

Oak Ridge National Laboratory

What is Numerical Methods?

- **Numerical methods** are algorithms designed to solve mathematical problems by approximating numerical solutions, especially when exact analytical solutions are infeasible or impractical.
- Computers are very good at this – lots of iterative algorithms based on simple calculations.
- Challenges in Computerized Numerical Methods:
  - CPUs only have hardware for addition and multiplication.
  - Numbers with inherently limited **precision**
  - **Discrete** rather that continuous computation
- Because of these challenges, computerized numerical methods have **errors**.

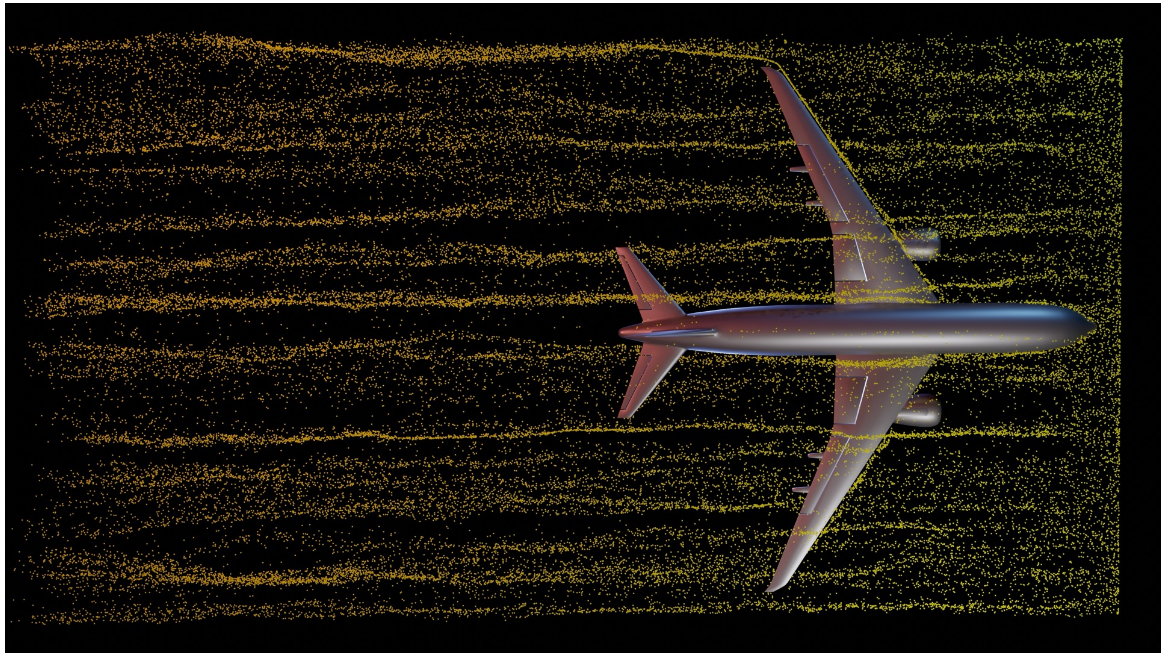
What will be Covered?

1. Computer Arithmetic and Errors
2. Taylor Approximations
3. Root Finding
4. Interpolation
5. Numerical Differentiation & Integration
6. Linear Equations
7. Numerical Linear Algebra
8. Differential Equations and Applications

Applications of Numerical Methods

Numerous Applications!

- Chemical Systems
- Particle Physics
- Seismology
- Video Games
- Climate Modeling
- Financial Modeling
- Weather Prediction
- Population Dynamics
- Biological Systems
- Materials Science
- Machine Learning
- Fluid Flow
- Neuroscience
- Structural Engineering



Syllabus

- Syllabus can be accessed through jamesghawaly.org
- Link here: https://jamesghawaly.org/files/CSC2262_Ghawaly_syllabus.pdf

Computer Arithmetic

Numerical Representation

- **Decimal** : base-10
- **Base** or **radix** of 10: indicates that there are 10 unique digits for representing numbers: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
- What does the number 547.65_{10} actually mean?

$$5\ 4\ 7\ .\ 6\ 5$$

100's, 10's, 1's, $\frac{1}{10}$'s, $\frac{1}{100}$'s

$$547_{10} = 7 \cdot 10^0 + 4 \cdot 10^1 + 5 \cdot 10^2$$

$$0.65_{10} = 6 \cdot 10^{-1} + 5 \cdot 10^{-2}$$

Numerical Representation & Conversion

- **Binary** : base-2
- **Base** or **radix** of 2: indicates that there are 2 unique digits for representing numbers: 0, 1
- What does the number 1011.001_2 actually mean?

$$1\ 0\ 1\ 1\ .\ 0\ 0\ 1$$

8's, 4's, 2's, 1's, $\frac{1}{2}$'s, $\frac{1}{4}$'s, $\frac{1}{8}$'s

$$1011_2 = 1 \cdot {2_{10}}^0 + 1 \cdot {2_{10}}^1 + 0 \cdot {2_{10}}^2 + 1 \cdot {2_{10}}^3$$
$$= 11_{10}$$

$$0.001_2 = 0 \cdot {2_{10}}^{-1} + 0 \cdot {2_{10}}^{-2} + 1 \cdot {2_{10}}^{-3} = \frac{1}{8_{10}}$$

Numerical Representation & Conversion

- **Hexadecimal** : base-16
- **Base** or **radix** of 16: indicates that there are 16 unique digits for representing numbers: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F
- What does the number $\mathrm{F32.C}_{16}$ actually mean?

$$\mathrm{F}\ 3\ 2\ .\ \mathrm{C}$$

256's 16's 1's $\frac{1}{16}$'s

$$\mathrm{F32}_{16} = 2 \cdot 16^0 + 3 \cdot 16^1 + 15 \cdot 16^2 = 3890_{10}$$

$$0.\mathrm{C}_{16} = 0 \cdot 16^{-1} = \frac{1}{16_{10}}$$

General Numerical Conversion

- Let's say we have a decimal integer x containing n digits that we wish to convert to base-r
- We can express x as follows
  $x = a_n \cdot r^n + a_{n-1} \cdot r^{n-1} + a_{n-2} \cdot r^{n-2} + \cdots + +a_1 \cdot r^1 + a_0 \cdot r^0$
- We wish to determine the coefficients $a_n, a_{n-1}, \ldots , a_0$ for $0 \leq a < r$
- We can solve this by repeatedly dividing x by r and recording the quotient and the remainder.
  - The remainders become the coefficients, starting with LSB
  - After each division, the integer part of the quotient becomes the new dividend
  - Continue until quotient is 0

Decimal to Base-2 Examples

Convert 463_{10} to base-2

2	463	
2	231	1
2	115	1
2	57	1
2	28	1
2	14	0
2	7	0
2	3	1
2	1	1
	0	1

MSB

111001111_2

256 128 64 32 16 8 4 2 1

Convert 463_{10} to base-2

2	463	
2	231	1
2	115	1
2	57	1
2	28	1
2	14	0
2	7	0
2	3	1
2	1	1
	0	1

← MSB

$463_{10} = 111001111_2$

Decimal to Base-2 Examples

Convert 463_{10} to base-2

$2\overline{)463}$	
$2\overline{)231}$	1
$2\overline{)115}$	1
$2\overline{)57}$	1
$2\overline{)28}$	1
$2\overline{)14}$	0
$2\overline{)7}$	0
$2\overline{)3}$	1
$2\overline{)1}$	1
0	1 ← MSB

$463_{10} = 111001111_2$

Convert 256_{10} to base-2

$2\overline{)256}$	
$2\overline{)128}$	0
$2\overline{)64}$	0
$2\overline{)32}$	0
$2\overline{)16}$	0
$2\overline{)8}$	0
$2\overline{)4}$	0
$2\overline{)2}$	0
$2\overline{)1}$	0
0	1

$256_{10} = 100000000$

Decimal to Base-16 Examples

Convert 16243_{10} to base-16

16	$\overline{)16243}$	
16	$\overline{)1015}$	3
16	$\overline{)63}$	7
16	$\overline{)3}$	F
	0	3

$16243_{10} = 3F73_{16}$

$\Updownarrow$

0x3F73

Convert 63406_{10} to base-16

16	$\overline{)63406}$	
16	$\overline{)3962}$	E
16	$\overline{)247}$	A
16	$\overline{)15}$	7
	0	F

$63406_{10} = F7AE_{16}$

Decimal Fraction to Any Base

- Let's say the decimal x has a fractional part z containing n digits that we wish to convert to base-r
- We can express z as follows
$$z = a_n \cdot r^{-n} + a_{n-1} \cdot r^{-(n-1)} + a_{n-2} \cdot r^{-(n-2)} + \cdots + +a_1 \cdot r^{-1}$$
- We want to solve for the coefficients $a_1, a_2, \ldots, a_n$, which will be the digits representing the number in base-r
- We can solve this by repeatedly multiplying z by r.
  - the fractional part of $r \cdot z$ becomes the next value to be multiplied by r
  - the integer part becomes the coefficient
  - Continue until the fractional part is 0

Decimal Fraction to Base-2 Example

756.375_{10} to base-2

• split into integer and fractional part

Integer Part

756_{10} to base-2

$2\,\overline{)756}$
$2\,\overline{)378}$ 0
$2\,\overline{)189}$ 0
$2\,\overline{)94}$ 1
$2\,\overline{)47}$ 0
$2\,\overline{)23}$ 1
$2\,\overline{)11}$ 1
$2\,\overline{)5}$ 1
$2\,\overline{)2}$ 1
$2\,\overline{)1}$ 0
0 1

$756_{10} = 1011110100_2$

Fractional Part

0.375_{10} to base-2

$(2)(0.375) = 0.75 \rightarrow 0$
$(2)(0.75) = 1.5 \rightarrow 1$
$(2)(0.5) = 1.0 \rightarrow 1$

$0.375_{10} = 0.011_2$

$756.375 = 1011110100.0.011_2$

Decimal Fraction to Base-16 Example

8974.109619140625_{10} to base-16

8974_{10} to base-16

$16 \overline{)8974}$	
$16 \overline{)560}$	E
$16 \overline{)35}$	0
$16 \overline{)2}$	3
0	2

$8974_{10} = 230E_{16}$

0.109619140625_{10} to base-16

$$(16)(0.109619140625) = 1.75390625 \rightarrow 1$$

$$(16)(0.75390625) = 12.0625 \rightarrow C$$

$$(16)(0.0625) = 1.0 \rightarrow 1$$

$$0.1096191410625_{10} = 0.1C1_{16}$$

$$\longrightarrow 230E.1C1_{16}$$

NOTE: Many calculators will have roundoff errors when doing this calculation!!

Repeating Decimals

- Different numerical bases have fractional numbers that cannot be represented in a finite number of digits.
- For example: base-10, $^1/_3 = 0.333333\ldots_{10} = 0.\overline{3}_{10}$
- However, in base-3, $^1/_3 = 0.1_3$
- So how would you convert $0.1010101010\ldots_2 = 0.\overline{10}_2$ to decimal?
- Remember, a number with base-y containing n digits can be converted to decimal by summing n powers of y and multiplying each by the corresponding digit.
  - But this would be an infinite series!

Repeating Decimals

- Geometric Series to the rescue!
- From the geometric series, we have the following:

$$\sum_{i=0}^{n} r^i = \frac{1 - r^{n+1}}{1 - r}, \qquad r \neq 1$$

$$\sum_{i=1}^{n} r^i = \frac{r - r^{n+1}}{1 - r}, \qquad r \neq 1$$

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1 - r}, \qquad |r| < 1$$

Repeating Decimal: Example with Geometric Series

Convert $0.\overline{10}_2$ to base-10

$$0.1010101\ldots = 1\cdot 2^{-1} + \cancel{0\cdot 2^{-2}} + 1\cdot 2^{-3} + \cancel{0\cdot 2^{-4}} + 1\cdot 2^{-5} + \cancel{0\cdot 2^{-6}} + \ldots$$

$$= 2^{-1} + 2^{-3} + 2^{-5} + \ldots$$

$$= 2^{-1}\left(2^{0} + 2^{-2} + 2^{-4}\right) \leftarrow \text{this is an infinite series with } r = 2^{-2}$$

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r} \Rightarrow$$

$$= 2^{-1}\cdot\frac{1}{1-2^{-2}} = \frac{1}{2}\cdot\frac{1}{1-\frac{1}{4}} = \frac{1}{2}\cdot\frac{4}{3} = \frac{4}{6} = \frac{2}{3}$$

Repeating Decimal: Example with Geometric Series

Convert $0.\overline{1011}_2$ to decimal

$$0.1011\,1011\,1011\,1011 = 2^{-1}+2^{-3}+2^{-4}+2^{-5}+2^{-7}+2^{-8}+2^{-9}+2^{-11}+2^{-12}+2^{-13}+2^{-15}+2^{-16}$$

$$= 2^{-1}\left(2^{0}+2^{-2}+2^{-3}+2^{-4}+2^{-6}+2^{-7}+2^{-8}+2^{-10}+2^{-11}+2^{-12}+2^{-14}+2^{-15}\right)$$

$$= 2^{-1}\left(\underbrace{2^{0}+2^{-2}+2^{-4}+2^{-6}+2^{-8}+2^{-10}+2^{-12}+2^{-14}}\right)+2^{-1}\left(\underbrace{2^{-3}+2^{-7}+2^{-11}+2^{-15}}\right)$$

Infinite Geometric Series with $r=2^{-2}$

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r} \Rightarrow (2^{-1})\left(\frac{1}{1-2^{-2}}\right) = \frac{2}{3}$$

$$= 2^{-1}\left[2^{-3}\left(\underbrace{2^{0}+2^{-4}+2^{-8}+2^{-12}}\right)\right]$$

Infinite Geometric Series with $r = 2^{-4}$

$$= 2^{-1}\left[2^{-3}\left(\frac{1}{1-2^{-4}}\right)\right] = \frac{1}{2}\cdot\frac{1}{8}\cdot\frac{1}{1-\frac{1}{16}} = \frac{1}{16}\cdot\frac{16}{15} = \frac{16}{240} = \frac{1}{15}$$

So $0.\overline{1011}_2 = \frac{2}{3} + \frac{1}{15} = \frac{11}{15}$

Repeating Decimals: Special Cases

- What if we have an n-digit (bit) integer in base-2 that contains only 1's?
- From geometric series:

$$\underbrace{\{1111111\}}_{n}{}_2 = \boxed{2^n - 1}$$

Maximum value that can be represented by n bits!

- For any n-digit number in base-r that contains only $(r-1)$'s: $r^n - 1$
- Likewise, for n-digits of a binary fraction, we can apply geometric series

$$\{0.1111111\}_2 = \sum_{i=1}^{n} {2^{-1}}^{i} = \frac{2^{-1} - {2^{-1}}^{n+1}}{1 - 2^{-1}} = 1 - 2^{-n}$$

Binary Addition and Multiplication

- Binary addition and multiplication use the same rules that you learned in grade school for decimal addition and multiplication.
- Why do we care about these two operations?
    - They are typically the only ones that are directly supported by CPU hardware

What is $101101_2 + 1001_2$?

$$\begin{array}{r} 101101 \\ +\quad 1001 \\ \hline 110110 \end{array}$$

What is 11101×101?

$$\begin{array}{r} 11101 \\ \times \quad 101 \\ \hline 11101 \\ 000000 \\ +\ 1110100 \\ \hline 10010001 \end{array}$$

Subtraction Using 2's Complement

- Subtraction is just addition with a negative number: $x - y = x + (-y)$
- Objective: Design a method to represent a negative number such that we can use the addition hardware for subtraction.
- For this we use 2's complement
- To calculate 2's complement of a number:
  - Calculate 1's complement of the number by flipping all the bits
  - Add 1 to 1's complement to get 2's complement of
- We can now do $x - y$ by doing $x + 2'\text{s complement}(y)$
- In this system, the most significant bit (MSB) is the sign bit
  - 1 is negative (-) and 0 is positive (+)
  - If the result is negative, calculate 2's complement of the result to get its value. If it's positive, leave it alone
  - Carry out bit is discarded

2's Complement Subtraction Examples

Use 2's complement to solve $56_{10} - 7_{10} = 49_{10}$

$56_{10} = 00111000_2$

$7_{10} = 00000111_2$

$-7_{10} = 11111001_2$ (2's Complement)

$$\begin{array}{r} {\scriptstyle 1111\ 1\quad\quad} \\ 00111000 \\ +\ 11111001 \\ \hline 1\,|\,\boxed{00110001} \end{array} = 49_{10}$$

MSB = 0 $\boxed{(+)}$

Use 2's complement to solve $18_{10} - 30_{10} = -12_{10}$

$18_{10} = 00010010_2$

$30_{10} = 00011110_2$

$-30_{10} = 11100010_2$ (2's Complement)

$$\begin{array}{r} {\scriptstyle 1\quad\ } \\ 00010010 \\ +\ 11100010 \\ \hline 11110100 \end{array}$$

MSB = 1 $\boxed{(-)}$

Take 2's complement of result

$\boxed{00001100}$