

Introduction

CSE 541

Numerical methods

- Solving scientific/engineering problems using computers.
- Root finding, Chapter 3
- Polynomial Interpolation, Chapter 4
- Differentiation, Chapter 4
- Integration, Chapters 5 and 6
- Systems of linear equations, Chapters 7 and 8
- Pseudo-random numbers and Monte Carlo Integration, Chapter 13

Root finding

$$f(x) = 3x^{10} + 9x^9 - 25x^6 + 3x^2 + 4$$

Find all roots (zeros) of $f(x)$ in the range $(0,1)$.

That is, find all numbers x in interval $(0,1)$

such that $f(x) = 0$.

Differentiation

$$f(x) = 10x^{10} \sin x + 9x^9 - \sqrt{5x^6 + 3x^2 + 4}$$

What is the derivative of $f(x)$ at $x = 2.4$?

Integration

$$f(x) = 10x^{10} \sin x + 9x^9 - \sqrt{5x^6 + 3x^2 + 4}$$

$$\int_0^{10} f(x) dx = ?$$

System of linear equations

$$\begin{cases} 6x_1 - 2x_2 + 3x_3 + 8x_4 = 10 \\ 12x_1 + 4x_2 + 7x_3 - 4x_4 = 20 \\ 4x_1 + 7x_2 + 4x_3 + 4x_4 = 30 \\ -5x_1 - 2x_2 + 6x_3 - 8x_4 = 40 \end{cases}$$

Find x_1, x_2, x_3, x_4 that satisfy the system of linear equations.

Number Representation and Errors

Readings: Sec. 1.1; Chap. 2

Numerical Errors

α : an exact value

$\bar{\alpha}$: an approximate value of α

- The **error** of $\bar{\alpha}$ as an approximation of α is $\alpha - \bar{\alpha}$
- The **absolute error** of $\bar{\alpha}$ as an approximation of α is $|\alpha - \bar{\alpha}|$
- The **relative error** = $\left| \frac{\text{error}}{\text{exact value}} \right| = \frac{|\alpha - \bar{\alpha}|}{|\alpha|}$
- Relative error is more significant than absolute error.

Example:

- $\alpha_1 = 1000$ miles, $\bar{\alpha}_1 = 999$ miles, a.e. = 1 mile, r.e. = 0.1%.
- $\alpha_2 = 1$ meter, $\bar{\alpha}_2 = 0.9$ meters, a.e.= 0.1 meters, r.e = 10%.

Chopping and Rounding

$$\pi = 3.1415926535897\dots$$

- A number has more than n digits.
How to approximate it with only n digits?
- Chopping: 3.1415, 3.14159, 3.141592, 3.1415926
- Rounding: 3.1416, 3.14159, 3.141593, 3.1415927
- Chop to n decimal places:
- Absolute error $\leq 10^{-n}$.
- Round to n decimal places:
- Absolute error $\leq 0.5 \times 10^{-n}$.
- $|\text{Rounding error}| \leq |\text{Chopping error}|$

Rounding

$$x = 123.45$$

$$y = 9876.1235$$

- Round x to 1 decimal place: 123.4 or 123.5?
- Round y to 3 decimal places: 9876.123 or 9876.124?
- Rounding up or down, depending on which one will produce smaller absolute error.
- Convention: when rounding up and rounding down will produce the same absolute rounding error, **round so that the last digit is even.**
- Round x to 1 decimal place:
- Round y to 3 decimal places:

Rounding binary numbers

- When rounding up and rounding down will produce the same absolute rounding error, **round so that the last digit is even (0)**.
- Round $(1.10100)_2$ to 3 binary digits: $(1.10)_2$
- Round $(1.10101)_2$ to 3 binary digits: $(1.11)_2$
- Round $(1.10101)_2$ to 4 binary digits: $(1.101)_2$
- Round $(1.10101)_2$ to 5 binary digits: $(1.1010)_2$

Representation of numbers in different bases

$$(a_n a_{n-1} \dots a_1 a_0 . b_1 b_2 b_3 \dots)_\beta = \sum_{i=0}^n a_i \cdot \beta^i + \sum_{i=0}^{\infty} b_i \cdot \beta^{-i}$$

$$(523.48)_{10} = 5 \cdot 100 + 2 \cdot 10 + 3 \cdot 1 + 4 \cdot 10^{-1} + 8 \cdot 10^{-2}$$

$$(101.101)_2 = 1 \cdot 4 + 0 \cdot 2 + 1 \cdot 1 + \frac{1}{2} + \frac{0}{4} + \frac{1}{8}$$

Numbers are typically represented in computers in

- base 2 (binary)
- base 8 (octal)
- base 16 (hexadecimal)

Converting an integer from decimal to binary

1. Repeatedly divide the number by 2 until the quotient is 0.
2. Concatenate all the remainders.

206	
103	0
51	1
25	1
12	1
6	0
3	0
1	1
0	1

$$(206)_{10} = (11001110)_2$$

Why?

$$x = (a_n a_{n-1} \dots a_1 a_0)_2 = (a_n a_{n-1} \dots a_1)_2 \cdot 2 + a_0$$

$$a_0 = x \bmod 2 \quad x \leftarrow x \operatorname{div} 2 = (a_n a_{n-1} \dots a_1)_2$$

$$a_1 = x \bmod 2 \quad x \leftarrow x \operatorname{div} 2 = (a_n a_{n-1} \dots a_2)_2$$

$$a_2 = x \bmod 2 \quad x \leftarrow x \operatorname{div} 2 = (a_n a_{n-1} \dots a_3)_2$$

Converting an integer from binary to decimal

$$\begin{aligned}(11001110)_2 &= 0 \cdot 2^0 + 1 \cdot 2^1 + 1 \cdot 2^2 + 1 \cdot 2^3 \\ &\quad + 0 \cdot 2^4 + 0 \cdot 2^5 + 1 \cdot 2^6 + 1 \cdot 2^7 \\ &= 2 + 4 + 8 + 64 + 128 \\ &= 206\end{aligned}$$

Converting a fraction from decimal into binary

Repeatedly multiply the number by 2 and remove the integer parts, which form the fraction in binary.

0.372

0.744

1.488

0.976

1.952

1.904

1.808

⋮

$$(0.372)_{10} = (0.010111\dots)_2$$

Why?

$$x = (0.c_1c_2c_3\dots)_2$$

$$2x = (c_1.c_2c_3\dots)_2$$

c_1 = integer part of $2x$ expressed in base 2
= integer part of $2x$ expressed in base 10

Scientific notation for a none-zero number

In decimal:

$$\pm d_0.d_1d_2d_3\dots\times 10^n \quad \text{where } d_0 \neq 0$$

For example:

$$292.1234 = 2.921234 \times 10^2$$

In binary:

$$\pm b_0.b_1b_2b_3\dots\times 2^m \quad \text{where } b_0 \neq 0$$

For example:

$$1011.1011 = 1.0111011 \times 2^3$$

Sign, significand or mantissa, base, exponent

More examples

Ordinary decimal notation

1.234

2000

2000.

$\overbrace{0.00\dots00}^{30}91093826 \text{ kg}$

$59736\overbrace{00\dots00}^{20} \text{ kg}$

Scientific notation

1.234×10^0

2×10^3

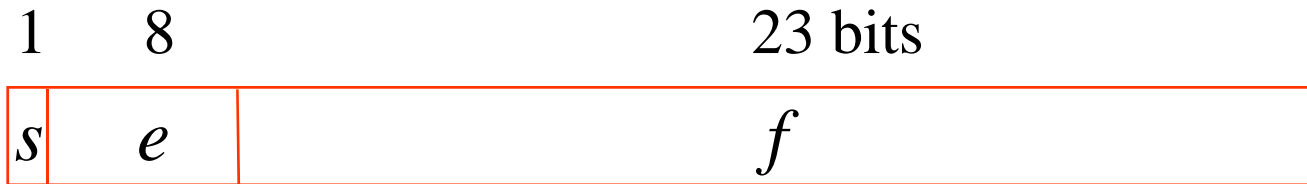
2.000×10^3

$9.1093826 \times 10^{-31} \text{ kg}$ (an electron)

$5.9736 \times 10^{24} \text{ kg}$ (Earth's mass)

Machine representation

IEEE single-precision (32 bits) floating point standard:



sign bit s : 1 bit 0 = +, 1 = -

biased exponent e : 8 bits actual exponent is $e - 127$

fraction f : 23 bits denoting $1.f$

The word represents: $(-1)^s \times 2^{e-127} \times (1.f)_2$

Special Cases

- When $e = 0$

$$0 \quad 00000000 \quad 00000000000000000000000000000000 = 0$$

$$1 \quad 00000000 \quad 00000000000000000000000000000000 = -0$$

- When $e = 255$

$$0 \quad 11111111 \quad 00000000000000000000000000000000 = \infty$$

$$1 \quad 11111111 \quad 00000000000000000000000000000000 = -\infty$$

- When $e = 0$ or 255 but $f \neq 0$: special meanings.

Example

Consider the number -52.234375 :

- Integer part: $52 = (110100)_2$
- Fraction part $.234375 = (.001111)_2$
- So, $52.234375 = (110100.001111)_2 = (1.10100001111)_2 \times 2^5$
- The exponent is 5. We want $e - 127 = 5$,
so $e = 132 = (10000100)_2$
- The machine representation is:
1 10000100 101000011110000000000000

Example

Consider the number $(1.0)_2$:

- $(1.0)_2 = (1.0)_2 \times 2^0$
- $f = 0$
- $e = 0 + 127 = 127 = (01111111)_2$
- The machine representation of 1.0 is:
0 01111111 00000000000000000000000000000000

Machine numbers

- Only a finite number of numbers can be exactly represented in a machine.
- These number are called **machine numbers**.
- In a single-precision machine, how many machine numbers are there (not counting special cases)?
- Other numbers: approximated by closest machine numbers.

Question:

$$x \leftarrow 0.1 \quad (\text{in decimal})$$

How is x represented in a single-precision machine?

$$\begin{aligned} (0.1)_{10} &= (0.0\ 0011\ 0011\ 0011\ 0011\ 0011\ \dots)_2 \\ &\approx (1.1\ 0011\ 0011\ 0011\ 0011\ 0011\ 01)_2 \times 2^{-4} \end{aligned}$$

0 01111011 10011001100110011001101

Addition:

$$x = (1.0001\ 1111\ 0000\ 1111\ 0000\ 111)_2 \times 2^8$$

$$y = (1.1111\ 0000\ 0000\ 0000\ 0001\ 111)_2 \times 2^4$$

$$x + y = ?$$

$$x = (1.0001\ 1111\ 0000\ 1111\ 0000\ 111)_2 \times 2^8$$

$$y = (0.0001\ 1111\ 0000\ 0000\ 0000\ 0001\ 111)_2 \times 2^8$$

$$x + y = (1.0011\ 1110\ 0000\ 1111\ 0000\ 1111\ 111)_2 \times 2^8$$

$$\approx (1.0011\ 1110\ 0000\ 1111\ 0001\ 000)_2 \times 2^8$$

$$1 + 2^{-24} = ?$$

- Mathematically, $1 + 2^{-24} = 1.\overbrace{000\dots000}^{23}1$.
- In a single-precision machine:

$$1 = (1.0000000000000000000000000000)_2 \times 2^0$$

$$2^{-24} = (1.0000000000000000000000000000)_2 \times 2^{-24}$$

$$= (0.0000000000000000000000000000\mathbf{1})_2 \times 2^0$$

$$1 + 2^{-24} = (1.0000000000000000000000000000\mathbf{1})_2 \times 2^0$$

$$= (1.0000000000000000000000000000)_2 \times 2^0$$

$$= 1$$

$$1 + (2^{-24} + 2^{-47}) = ?$$

Machine-dependent.

One possibility:

$$1 = (1.00000000000000000000000000000000)_2 \times 2^0$$

$$2^{-24} + 2^{-47} = (1.00000000000000000000000000000001)_2 \times 2^{-24}$$

$$= (0.00000000000000000000000000000001000\dots0001)_2 \times 2^0$$

$$1 + 2^{-24} + 2^{-47} = (1.00000000000000000000000000000001000\dots0001)_2 \times 2^0$$

$$= (1.00000000000000000000000000000001)_2 \times 2^0$$

$$> 1$$

Significance of Machine Epsilon

- Why do we want to know the machine epsilon?
- For instance, consider the approximation

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

- By choosing h smaller and smaller, we would expect to get better and better estimate of $f'(x)$.
- However, if h gets so small that $x+h = x$ in the machine, then $f(x+h) = f(x)$ and we will get $f'(x) \approx 0$.
- We require $x+h > x \Rightarrow 1+h/x > 1 \Rightarrow h/x \geq \varepsilon \Rightarrow h \geq x\varepsilon$.

Significant Digits

1204 4 significant digits

56.780 5 significant digits

0.0067 2 significant digits

0.006700 4 significant digits

1.006700 7 significant digits

2000 1 significant digit

2.0×10^3 2 significant digits

2.00×10^3 3 significant digits

2.000×10^3 4 significant digits

Significant Digits

- **Nonzeros** are always significant.
- **Zeros** between nonzero digits are significant.
- Leading **zeros** are not significant.
- Trailing **zeros** are significant if the decimal point is specified.
- For instance, 200.00 has 5 significant digits; 200 has 1.

Significant Digits

- Let \tilde{N} be an approximation to N .
- \tilde{N} is said to approximate N to k significant digits if \tilde{N} and N are equal when rounded to k digits (but not equal if rounded to more digits).
- Example: $x = 0.123456$
 - **0.123**321 approximates x to 3 significant digits.
 - **0.1235**12 approximates x to 4 significant digits.
- Include only significant digits when writing an approximation. In the above example, write $x \approx 0.123$ or $x \approx 0.1235$, **not** $x \approx 0.123321$ or $x \approx 0.123512$.

Significant Digits

- There are \tilde{N} books in the library.
- $\tilde{N} = 2.000 \times 10^3$, implying exactly 2000.
- $\tilde{N} = 2.00 \times 10^3$, implying 1995 ~ 2005.
- $\tilde{N} = 2.0 \times 10^3$, implying 1950 ~ 2050.
- $\tilde{N} = 2 \times 10^3$, implying 1500 ~ 2500.

Significant Digits

- Regard the least significant digit of approximate numbers as doubtful.
- Allow only one doubtful digit

$$\begin{array}{r} 4.7832 \\ 1.234 \\ + 2.02 \\ \hline 8.0372 \\ \downarrow \text{rounding} \\ 8.04 \end{array}$$

Significant Digits

- Subtraction may cause loss of significance.
- Both numbers have 5 significant digits.

The answer has only 3.

$$\begin{array}{r} 1.0236 \\ - 0.97268 \\ \hline 0.05092 \\ \downarrow \text{rounding} \\ 0.0509 \end{array}$$

Significant Digits

The answer must be rounded off to 2 significant figures, since 1.6 only has 2 significant figures.

$$\begin{array}{r} 2.8723 \\ \times 1.6 \\ \hline 4.59568 \\ \Downarrow \text{rounding} \\ 4.6 \end{array}$$

Significant Digits

The answer must be rounded off to 3 significant figures, since 45.2 only has 3 significant figures.

$$\begin{array}{r} 45.2 \\ \times 6.3578 \\ \hline 7.1093775 \\ \Downarrow \text{rounding} \\ 7.11 \end{array}$$

Loss of Significant Digits

- When x and y are nearly equal, the subtraction $x - y$ can cause loss of significant digits and lead to serious errors in many computations.
- This source of error is known as subtractive cancellation.
- To demonstrate this, suppose the machine has 5 decimal digits of accuracy.

- Exact values:

$$x = 0.3721448693$$

$$y = 0.3720214371$$

- 5 decimal digits of accuracy:

$$\tilde{x} = 0.37214$$

$$\tilde{y} = 0.37202$$

Loss of Significant Digits (contd)

- relative error of $\tilde{x} = |(x - \tilde{x}) / x| = 1.3 \times 10^{-5}$
relative error of $\tilde{y} = |(y - \tilde{y}) / y| = 3.9 \times 10^{-6}$
- Exact difference: $x - y = .0001234322$
- Approximation: $\tilde{x} - \tilde{y} = .00012$
- The relative error of $\tilde{x} - \tilde{y}$ is:

$$\frac{|(x - y) - (\tilde{x} - \tilde{y})|}{|x - y|} = \frac{.0000034322}{.0001234322} = 3 \times 10^{-2}$$

Avoiding Loss of Significance in Subtraction

- Locate the "sensitive" subtraction and eliminate it, if possible, by analytical manipulation.
- For instance, suppose the values of

$$f(x) = x - \sin x$$

are required near $x = 0$.

- Since $\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$, $\sin x \approx x$ when $x \approx 0$ and computing $f(x)$ directly will cause serious loss of accuracy.

- For $x = 0.6666666667 \times 10^{-1}$

$$\sin x = 0.6661729492 \times 10^{-1}.$$

$$f(x) = x - \sin x$$

$$= 0.0004937175 \times 10^{-1}$$

$$= 0.4937175000 \times 10^{-4}$$

Correct $f(x) = 0.4937174327 \times 10^{-4}$.

- 3 **spurious zeros** added which are a loss of significance.

- Approximate $\sin x$ with its Taylor series :

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

- $f(x) = x - \sin x = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \dots \approx \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!}$

- For $x = 0.6666666667 \times 10^{-1}$,

$$\frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} = 4.937174328 \times 10^{-5}.$$

- Correct $f(x) = 4.937174327 \times 10^{-5}$.

Another Example

- $f(x) = 1 - \sin x$
- Values of $f(x)$ for x near $\pi / 2$ are needed.
- $\sin x \approx 1$ for $x \approx \pi / 2$.
- Instead of computing $f(x) = 1 - \sin x$ directly, use

$$1 - \sin x = \frac{(1 - \sin x)(1 + \sin x)}{1 + \sin x} = \frac{1 - \sin^2 x}{1 + \sin x} = \frac{\cos^2 x}{1 + \sin x}$$