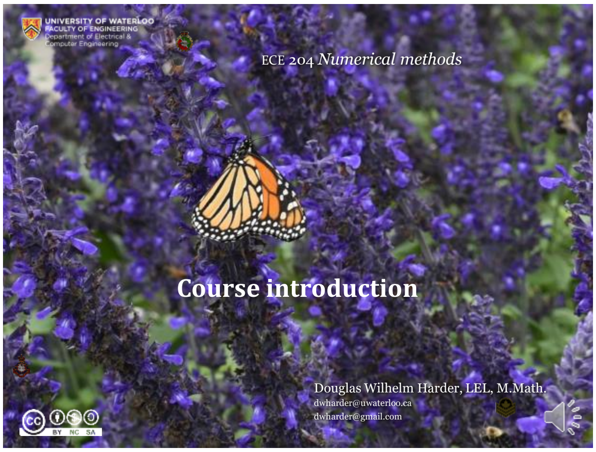


1

Introduction

- In this topic, we will
  - Give motivation for this course
  - Describe the difference between discrete mathematics and calculus (continuous mathematics)
  - Look at issues of trying to use computers to solve problems in calculus
  - Describe absolute and relative error
  - Describe the decimal and binary representation of real numbers
  - Define rounding and significant digits
  - Describe the difference between accuracy and precision
  - Given an overview of the balance of this course

2

Background

- Engineering design involves the application of mathematics and science to solve real-world problems
  - A failure of a design may cause harm to life, health, the environment or the finances of a client or the public
- The real world is modeled through equations and differential equations
  - Conservation and other physical laws
  - Relationship between forces
  - The superposition principle

3

Background

- Only the most trivial problems can be solved exactly
  - If a problem can be solved exactly, what is the need for an engineer?
- Problems can almost always be solved using a myriad of different approaches
  - Seldom, if ever, is there only one ideal approach
  - Each approach will have benefits and drawbacks that must be quantified and analyzed

4

Background

- In determining the most appropriate approach, it is necessary to test the various solutions
  - For example,
    - How robust is a solution
    - Is the solution sensitive to initial conditions?

5

5

Background

- One possible solution:
  - Build multiple instances of your design and test them
  - This is exceptionally costly
- Alternatively, if we have a mathematical description, can we not implement and simulate it, instead?

6

6

Background

- For example, a circuit is described by Maxwell's equations
  - This involves partial differential equations
  - Using wires, this effectively restricts these equations to one dimension
  - These partial differential equations can thus be simplified to differential equations
  - The use of linear circuit elements such as capacitors, resistors, inductors and memristors together with alternating current can further simplify the solutions to these differential equations to algebraic equations
  - Transistors, as well, may also be described linearly under the conditions of small-signal model
  - More complex models may still be simulated using differential equations

7

7

Background

- The response of a digital circuit can be described by a system of linear equations
  - This can involve millions of linear equations in an equal number of unknowns
  - Solving such a system cannot be done analytically in either a reasonable amount of time or memory
- Consequently, we will approximate such systems to find approximate solutions
  - Such solutions use numerical algorithms

8

8

Discrete mathematics and calculus

- In your courses on discrete mathematics and logic, you were introduced to
  - Boolean logic
  - Set theory
  - Combinatorics
  - Graph theory
- In your courses on continuous mathematics (calculus), you were introduced to
  - Differentiation
  - Integration
  - Measure
  - Infinite series

Discrete mathematics and calculus

- Concepts in discrete mathematics can be modeled using integers
  - Large discrete systems will sometimes be approximated using continuous systems
    - For example, statistics
- Concepts in calculus use real numbers
  - Real numbers can only be stored approximated
  - Computers use floating-point numbers
  - Every floating-point number represents a rational number
    - We approximate all possible real numbers using a finite number of rational approximations

Approximating the derivative

- For example, from calculus, we know that

$$\frac{d}{dx} f(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

- Let's try this out in C++:

```
double diff( double f( double ), double x, double h ) {
    return (f( x + h ) - f( x ))/h;
}
```

- This can be called with

```
std::cout << diff( std::sin, 1.0, 0.001 ) << std::endl;
```


Approximating the derivative

- Trying a few examples:

```
Enter a value of x: 1.0
The derivative of sine at x = 1 is 0.5403023058681398
Enter a value of h (0.0 to quit): 0.1
The derivative is approximately 0.4973637525353891
Enter a value of h (0.0 to quit): 1e-5
The derivative is approximately 0.5402980985058647
Enter a value of h (0.0 to quit): 1e-10
The derivative is approximately 0.5403022473871033
Enter a value of h (0.0 to quit): 1e-15
The derivative is approximately 0.5551115123125783
Enter a value of h (0.0 to quit): 1e-16
The derivative is approximately 0
Enter a value of h (0.0 to quit): 0.0
```

Approximating the derivative

- Obviously, we have an issue here…
  - The first step is to be able to describe how *good* an approximation is to the exact solution
  - In the next topic, we will focus on the design of floating-point numbers
  - The balance of the course will be focused on finding algorithms that avoid the issues with floating-point numbers

Error, absolute error and relative error

- If x_{approx} is an approximation of a value x, we write $x \approx x_{\text{approx}}$ and
$$x = x_{\text{approx}} + \varepsilon$$
- Consequently, the *error* ε is always:
$$\varepsilon = x - x_{\text{approx}}$$
- Usually, however, we may refer to the *absolute error*:
$$\varepsilon_{\text{abs}} = \left| x - x_{\text{approx}} \right|$$
- We may also refer to the *relative error* and *percent relative error*:
$$\varepsilon_{\text{rel}} = \frac{\left| x - x_{\text{approx}} \right|}{|x|} \qquad \varepsilon_{\text{rel}} \cdot 100\% = \frac{\left| x - x_{\text{approx}} \right|}{|x|} \cdot 100\%$$

Error, absolute error and relative error

- For example, from calculus,
$$f(x+h) = f(x) + f^{(1)}(x)h + \frac{1}{2} f^{(2)}(\xi)h^2$$
- An approximation of $f(x + h)$ is
$$f(x) + f^{(1)}(x)h$$
- The error of the approximation is
$$\frac{1}{2} f^{(2)}(\xi)h^2$$
- The absolute error is
$$\left| \frac{1}{2} f^{(2)}(\xi)h^2 \right| = \frac{1}{2} \left| f^{(2)}(\xi) \right| h^2$$

Error, absolute error and relative error

- For example, from calculus,
$$f(x+h) = f(x) + f^{(1)}(x)h + \frac{1}{2} f^{(2)}(\xi)h^2$$
- The relative error, assuming $f(x + h) \neq 0$, is
$$\frac{\left| \frac{1}{2} f^{(2)}(\xi)h^2 \right|}{\left| f(x+h) \right|} = \frac{1}{2} \frac{\left| f^{(2)}(\xi) \right|}{\left| f(x+h) \right|} h^2$$

Approximations of π

Approximation	Absolute error	Relative error	Percent relative error
3.14	0.001593	0.0005070	0.05 %
22/7	0.001264	0.0004025	0.04 %
355/113	0.0000002668	0.00000008491	0.000008491 %

Note, we will always describe the absolute and relative error to four digits.

17

17

Which do we prefer?

- It is easier to discuss absolute error over error
  - Neither, however, are unit-free:
    - If you were told that the absolute error was 10 m
      - This is great if you're measuring the distance to the Moon
      - Somewhat sub-optimal if you're measuring the width of a wire
  - The relative error is unit-free:
    - A relative error of 0.01 means that the approximation is good to one part in a hundred, regardless of the magnitude of the exact value
    - A relative error of 0.01 is a 1% relative error
  - The relative error is not defined for approximating zero, but if you're approximating zero and you know it…
- Note that the Taylor series gives an absolute error

18

18

Decimal numbers

- A non-zero real number is written as a decimal number when it is in the form

$$d_0.d_1d_2d_3d_4d_5\cdots\times 10^e$$

- Here we have
  - d_0 is a non-zero digit
  - Each other d_k is a decimal digit
  - The exponent e is an integer
- This is usually contrasted with fractional form for rational numbers

19

19

Decimal numbers

- If e is small in magnitude, we may simply move the decimal point:

$$d_0.d_1d_2d_3d_4d_5\cdots\times 10^3 = d_0d_1d_2d_3.d_4d_5\cdots$$

$$d_0.d_1d_2d_3d_4d_5\cdots\times 10^{-4} = 0.000d_0d_1d_2d_3d_4d_5\cdots$$

- For example,

$$1.532\times 10^2 = 153.2$$

$$1.8\times 10^{-3} = 0.0018$$

20

20

Decimal numbers

- If $d_k = 0$ for all $k > n$, we will say

$$d_0.d_1d_2d_3d_4d_5 \cdots d_n \times 10^e$$

has a *terminating decimal representation*

21

21

Decimal numbers

- If $d_k = 0$ for all $k > n$, we will say

$$d_0.d_1d_2d_3d_4d_5 \cdots d_n \times 10^e$$

represents an $n + 1$ digit decimal number

- For example,

  3.140 is a 4-digit decimal number

  $6.62607015 \times 10^{-34}$ is a 9-digit decimal number

22

22

Binary numbers

- A non-zero number is written as a binary number when it is in the form

$$1.b_1b_2b_3b_4b_5 \cdots \times 2^e$$

The base and exponent is usually still written in decimal

- Here
  - The one leading bit must be non-zero; that is, 1
  - Each b_k in the mantissa is a binary digit or "bit"
  - The exponent e is an integer

23

23

Binary numbers

- If e is small in magnitude,

  we may simply move the radix point:

$$1.b_1b_2b_3b_4b_5 \cdots \times 2^3 = 1b_1b_2b_3.b_4b_5 \cdots$$

$$1.b_1b_2b_3b_4b_5 \cdots \times 2^{-4} = 0.000b_0b_1b_2b_3b_4b_5 \cdots$$

- For example,

We will add a subscript "2" to show a number is in binary

$$1.001101 \times 2^3 = 1001.101_2$$

$$1.10011 \times 2^{-2} = 0.0110011_2$$

We can also use `0b1001101` and `0b0.110011`

24

24

Binary numbers

- If $b_k = 0$ for all $k > n$, we will say

$$1.b_1b_2b_3b_4b_5 \cdots b_{n-1}b_n \times 2^e$$

has a *terminating binary representation*

25

Binary numbers

- If $b_k = 0$ for all $k > n$, we will say

$$b_0.b_1b_2b_3b_4b_5 \cdots b_n \times 2^e$$

represents an $n + 1$ bit binary number

- For example,

  1.010 is a 4-bit binary number

  $1.00111101001000 \times 2^{-34}$ is a 15-bit binary number

26

Rounding

- Most real numbers have infinitely many digits
  - These require an infinite number of digits to store
  - Even most terminating decimal representations have far more digits than we care about
    3.1415926535897932384626433832795028841971693993751
  - However, we cannot store an infinite number of digits
  - In fact, the most efficient means of implementing such numbers is with a fixed number of digits
- We will need to represent all such numbers by a decimal or binary number with a fixed *n* digits in the mantissa

27

Rounding

- Which $n + 1$ digit number (n digits in the mantissa) do we use to approximate any real number?
  - The $n + 1$ digit number that has the smallest absolute error
  - We will say that we *round* a real number x to the closest $n + 1$ digit decimal or binary number

28

Rounding

- Rules for decimal rounding:

$$d_0.d_1d_2d_3d_4d_5\cdots d_n\boxed{d_{n+1}d_{n+2}\cdots}$$

  - To round a decimal representation to $n + 1$ digits:
    - If the digit d_{n+1} is 0, 1, 2, 3 or 4, just drop all subsequent digits
    - If the digit d_{n+1} is 5, 6, 7, 8 or 9, but not exactly 5000⋯, we will
      - Drop all subsequent digits
      - Add "1" to d_n, possibly resulting in a carry
- For example,

  1.534982 rounded to three digits is 1.53
  1.534982 rounded to four digits is 1.535
  1.534982 rounded to five digits is 1.5350
  1.534982 rounded to six digits is 1.53498

29

29

Rounding

- Rules for binary rounding:

$$b_0.b_1b_2b_3b_4b_5\cdots b_n\boxed{b_{n+1}b_{n+2}\cdots}$$

  - To round a binary representation to $n + 1$ bits:
    - If the next bit $b_{n+1} = 0$, just drop all subsequent bits
    - If the next bit $b_{n+1} = 1$ but not exactly 1000⋯, we will
      - Drop all subsequent bits
      - Add "1" to the last bit b_n, possibly resulting in a carry
- For example,

  1.1011011 rounded to three bits is 1.11
  1.1011011 rounded to four bits is 1.110
  1.1011011 rounded to five bits is 1.1011
  1.1011011 rounded to six bits is 1.10111

30

30

Rounding

- We skipped two cases:

$$d_0.d_1d_2d_3d_4d_5\cdots d_n5000\cdots$$

$$b_0.b_1b_2b_3b_4b_5\cdots b_n1000\cdots$$

- Both these numbers are half-way between two $n + 1$ digit numbers
  - Do we *round down* (truncate) or *round up* (truncate and add)?
  - If we choose one of these options, we will introduce a bias into calculations
  - Thus we choose one 50% of the time,
    and the other the other 50% of the time:
    - If d_n is odd, we increment it, otherwise we leave it
    - If $b_n = 1$, we increment it, otherwise we leave it

31

31

Rounding

- Note that we have only discussed rounding in our formal representation

$$d_0.d_1d_2d_3d_4d_5\cdots\times 10^e$$

$$1.b_1b_2b_3b_4b_5\cdots\times 2^e$$

- If the decimal/radix point is anywhere else,
  we count the digits starting at the most significant digit:

  0.0005838125 rounded to 4 digits is 0.0005838
  108513.829 rounded to 3 digits is 109000,
  but it's clearer if we present it as 1.09×10^5
  0.00011000101_2 rounded to 5 bits is 0.00011001_2
  111001010.001_2 rounded to 4 bits is 111000000_2,
  but it's clearer if we present it as 1.110×2^8

32

32

Course introduction

Significant digits

- Another colloquial means of describing the relative error is to use the concept of *significant digits*

$$\varepsilon_{\text{rel}} = \frac{|x - x_{\text{approx}}|}{|x|} \le 5\cdot 10^{-d}$$

  – We will never calculate this explicitly
    - Instead, we will use the number of significant digits to give a rough estimate of the relative error
      – 1 significant digit is a relative error no greater than 50%
      – 2 significant digits is a relative error no greater than 5%
  – A rough approximation is as follows:
    - x_{approx} approximates x to d significant digits if both x and x_{approx} rounded to d digits agree in all digits

33

33

Course introduction

Significant digits

- From the definition, we can find a formula to calculate the number of significant digits given the relative error:

$$\varepsilon_{\text{rel}} \le 5\cdot 10^{-d}$$

$$\frac{\varepsilon_{\text{rel}}}{5} \le 10^{-d}$$

$$\log_{10}\left(\frac{\varepsilon_{\text{rel}}}{5}\right) \le -d$$

$$\left\lceil \log_{10}\left(\frac{\varepsilon_{\text{rel}}}{5}\right)\right\rceil = -d$$

$$-\left\lceil \log_{10}\left(\frac{\varepsilon_{\text{rel}}}{5}\right)\right\rceil = d$$

34

34

Course introduction

Approximating the square root of two

- The ancient Babylonians were aware that:

$$\text{If } x < \sqrt{2} \text{ then } \frac{2}{x} > \sqrt{2}$$

$$\text{If } x = \sqrt{2} \text{ then } \frac{2}{x} = \sqrt{2}$$

$$\text{If } x > \sqrt{2} \text{ then } \frac{2}{x} < \sqrt{2}$$

  – Thus, if x approximates $\sqrt{2}$, it follows the average of x and $2/x$ must be a better approximation

$$\frac{1}{2}\left(x + \frac{2}{x}\right) = \frac{x}{2} + \frac{1}{x}$$

35

35

Course introduction

Approximating the square root of two

Approximation	Absolute error	Relative error	Percent relative error	Significant digits
1	0.4142	0.2929	29.29 %	1
1.5	0.08579	0.06066	6.066 %	1
1.416666666	0.002453	0.001735	0.1735 %	3
1.414215686	0.000002124	0.000001502	0.0001502 %	6

Note, we will always round the absolute and relative error to four digits.

36

36

Precision versus accuracy

- The concepts of precision and accuracy are usually introduced via references to firearms
  - The Tikka T3X TACT A1 is a more precise firearm than the AK-47

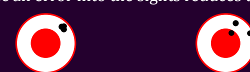
37

Precision versus accuracy

- Suppose we take these firearms to the range
  - The Tikka T3X TACT A1 is a more precise firearm than the AK-47

  - Give each rifle to a novice, the precision will be reduced:

  - Introduce an error into the sights reduces the accuracy



38

Precision versus accuracy

- When solving a problem numerically,
  we will use one or more different algorithms
- We will describe an algorithm through its accuracy and its precision
  - In general, all of our algorithms are parameterized by at least one value:
    - A value of h that may be made arbitrarily small
    - An integer n that may be made arbitrarily large
  - In approximating a solution x,
    - An algorithm is *accurate* if as our parameters are adjusted, the absolute error is correspondingly reduced
    - One algorithm is more *precise* than another if the absolute error for the first algorithm is generally less than the absolute error of another

39

Precision versus accuracy

- Suppose you are reading a sensor four times,
  and each value from the sensor will produce a random value in the range $[a, b]$
- What is a good approximation of a?
  - How about $\min\{x_1, x_2, x_3, x_4\}$?
  - This is actually a good approximation of
    $$\frac{4a+b}{5} = a + \frac{b-a}{5}$$
  - Similarly, $\max\{x_1, x_2, x_3, x_4\}$ is a good approximation of
    $$\frac{a+4b}{5} = b - \frac{b-a}{5}$$
  - This gives us two equations:
    $$4a + b = 5\min\{x_1, x_2, x_3, x_4\}$$
    $$a + 4b = 5\max\{x_1, x_2, x_3, x_4\}$$

40

Precision versus accuracy

- This gives us a system of two linear equations in two unknowns:

$$4a + b = 5\min\{x_1, x_2, x_3, x_4\}$$
$$a + 4b = 5\max\{x_1, x_2, x_3, x_4\}$$

- Using linear algebra, we may now solve for both a and b:

$$a = \frac{4\min\{x_1, x_2, x_3, x_4\} - \max\{x_1, x_2, x_3, x_4\}}{3}$$
$$b = \frac{4\max\{x_1, x_2, x_3, x_4\} - \min\{x_1, x_2, x_3, x_4\}}{3}$$

41

41

Precision versus accuracy

- Consequently,

$$\min\{x_1, x_2, x_3, x_4\}$$

  is a less accurate approximation of a than

$$\frac{4\min\{x_1, x_2, x_3, x_4\} - \max\{x_1, x_2, x_3, x_4\}}{3}$$

- The mathematics is beyond the scope of this course, but the first is more precise than the second
  - It has a smaller *standard deviation*

42

42

Precision versus accuracy

- Suppose we are uniformly sampling from the interval [5, 25]
- We could take 10 readings:

  8, 18, 16.8, 18.8, 13.4, 13 21, 15, 18, 9.6
  - We see the minimum of these, 8, is an okay approximation to 5
- We could take 20 readings:

  14.2, 16.4, 5.6, 10.8, 9.4, 24.8, 19.6, 13.8, 5.4, 14.6, 22.6, 24.6, 12.6, 8.4, 14.2, 8.8, 11.8, 23.6, 22.6, 14.2
  - We see the minimum of these, 5.4, is a more accurate approximation of 5

- As n becomes larger, it seems that the minimum is a better approximation of the lower bound 5

43

43

Precision versus accuracy

- If we modify the other formula for more samples, we:

$$a \approx \min\{x_1, \ldots, x_n\}$$
$$a \approx \frac{n\min\{x_1, \ldots, x_n\} - \max\{x_1, \ldots, x_n\}}{n-1}$$

- For the previous examples, we have

$$\min\{x_1, \ldots, x_{10}\} = 8 \quad \frac{10\min\{x_1, \ldots, x_{10}\} - \max\{x_1, \ldots, x_{10}\}}{9} = \frac{10 \cdot 8 - 18.8}{9} = 6.8$$
$$\min\{x_1, \ldots, x_{20}\} = 5.4 \quad \frac{20\min\{x_1, \ldots, x_{20}\} - \max\{x_1, \ldots, x_{20}\}}{19} = \frac{20 \cdot 5.4 - 24.8}{19} \approx 4.379$$

44

44

Course introduction

Looking ahead

- Thus, given n samples from $[a, b]$
  - The minimum of the n samples is not as accurate as our linear combination of the minimum and maximum
  - The minimum of the n samples is more precise than our linear combination of the minimum and maximum
  - As we increase n, both formulas become more precise
- Try this yourself, suppose we have different $[a, b]$:
  - What are your estimates of a for five samples:
    –5.98, –1.94, –6.28, –2.72, –5.86
  - What are your estimates of a for ten samples:
    –6.22, –4.62, –2.80, –5.66, –2.94, –4.32, –2.44, –2.34, –4.40, –4.11

– These were uniformly sampled from [–6.3, –1.3]

45

45

Course introduction

In this courses…

- The next topic in this course is to look at floating-point numbers
  - Floating-point numbers only approximate real numbers
  - There will be issues with them
    - Recall that we could not even precisely approximate the derivative using them
- The balance of this course will be looking at numerical algorithms in order to:
  - Approximating the value of an expression
  - Approximating solutions to algebraic equations and systems
  - Approximating solutions to analytic equations and systems
  - Optimization

46

46

Course introduction

Summary

- Following this topic, you now
  - Understand the purpose of this course
  - Are aware of the differences between discrete mathematics and continuous mathematics (calculus)
  - Have observed that floating-point numbers cause issues
  - Know the ideas behind:
    - Absolute and relative errors
    - Decimal and binary representations of numbers
    - Rounding and significant digits
  - Understand the concepts of accuracy and precision
  - Have an overview of what will be covered in the upcoming lectures

47

47

Course introduction

References

[1] https://en.wikipedia.org/wiki/Rounding
[2] https://en.wikipedia.org/wiki/Accuracy_and_precision
[3] https://en.wikipedia.org/wiki/Significant_figures
[4] https://en.wikipedia.org/wiki/Approximation_error

48

48

Acknowledgments

None so far.

Colophon

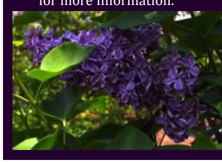
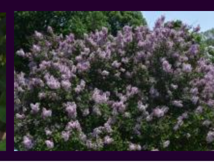
These slides were prepared using the Cambria typeface. Mathematical equations use Times New Roman, and source code is presented using `Consolas`. Mathematical equations are prepared in MathType by Design Science, Inc.

Examples may be formulated and checked using Maple by Maplesoft, Inc.

The photographs of flowers and a monarch butter appearing on the title slide and accenting the top of each other slide were taken at the Royal Botanical Gardens in October of 2017 by Douglas Wilhelm Harder. Please see

https://www.rbg.ca/

for more information.


Disclaimer

These slides are provided for the ECE 204 *Numerical methods* course taught at the University of Waterloo. The material in it reflects the author's best judgment in light of the information available to them at the time of preparation. Any reliance on these course slides by any party for any other purpose are the responsibility of such parties. The authors accept no responsibility for damages, if any, suffered by any party as a result of decisions made or actions based on these course slides for any other purpose than that for which it was intended.