Numerical methods course

chapiter 1:

Basics of Numerical Analysis and Scientific Computing

By Dr, Asmaa Ourdighi

USTO MB

2023-2024

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- **Motivation**
- Floating point arithmetic and rounding errors
- Methods of error calculation

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- **Motivation**

« Numerical analysisis the study of Algorithms for the problem of continious mathematics » - Lloyd N. Trefethen

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- Floating point arithmetic and rounding errors
    - Machine representation of numbers
        - In this section we introduce the notions of mantissa, exponent and how numbers are represented on a calculator or computer.
        - Standard form is a way of writing numbers. It can be used to represent large numbers that include decimal values (this is also often called scientific notation).

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- Floating point arithmetic and rounding errors
- Machine representation of numbers
  - **Real decimal numbers**
    - When using standard form, you could write the number 34.567 as:

$$\overbrace{0.34567}^{\text{Mantissa}} \times 10^{\overset{\text{Exponent}}{2}}$$

    - 10^2 is the same as 10x10, which equals 100. Multiplying 0.34567 by 100 gives the original value of 34.567.
    - The decimal point has moved 2 places to the left to create a Mantissa.
    - Multiplying by 10^2 would shift the decimal point two places to the right and recreate the original value. In this example, the value 2 is referred to as the exponent.

Numerical methods course

chapitre 1:

General information on numerical analysis and scientific computing

- Floating point arithmetic and rounding errors
- Machine representation of numbers
  - **binary digits**
  - For example

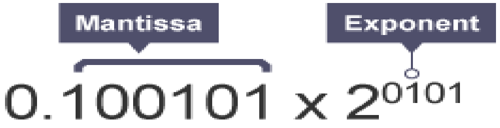
$$39 = 32 + 4 + 2 + 1 = 2^5 + 2^2 + 2^1 + 2^0 = (100111)_2,$$
$$3.625 = 2^1 + 2^0 + 2^{-1} + 2^{-3} = (11.101)_2 = (1.1101)_2\ 2^1$$

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- Floating point arithmetic and rounding errors
- Machine representation of numbers
  - Floating point representation
- The number 18.5 can be converted to binary to give us:
- **10010.1** => 16 + 2 + 0.5 = 18.5
- This is because when we convert 18.5 from denary to binary we get:
- To store this as a floating-point number, we need to move the decimal point so that it appears after the most significant bit, In this case we move the decimal point five places, so our number becomes:
- **0.100101 x 2** ^{5}
- But we also need to store the value 5 (seen as 2^5) as a binary number, so it would become:
- **0.100101 x 2** ^{0101}

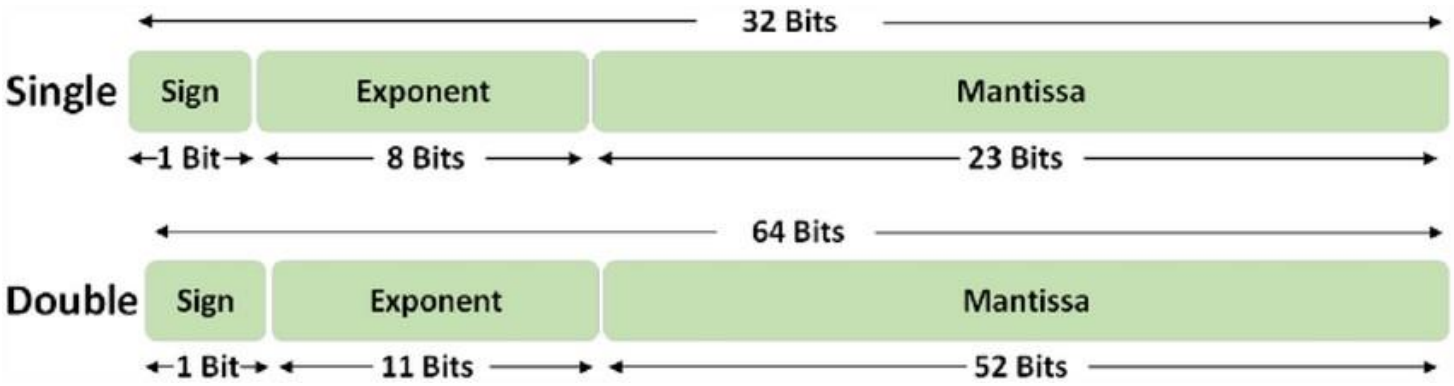


Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- Floating point arithmetic and rounding errors
- Machine representation of numbers
  - Floating point representation
  - **IEEE 754 norme**



Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- **Example:** Here's the representation of –6.75 in both single-precision and double-precision IEEE 754 formats.
- **1. Single Precision (32 bits)**
- **Step 1: Convert to Binary**
- *Absolute value*: 6.75 in binary is 110.11.
- **Step 2: Normalize**
- *Normalize to:* 1011×2^2
- **Step 3: Sign Bit**
- *Sign (S):* 1 (negative)
- **Step 4: Exponent**
- Actual exponent: 2
- Biased exponent: $2 + 127 = 129$
- In binary: 10000001
- **Step 5: Mantissa**
- Mantissa (without leading 1): 10110000000000000000000 (23 bits)
- **Final Representation**
-
- Putting it all together:
- **| 1 | 10000001 | 10110000000000000000000**

Numerical methods course

chapter 1:

General information on numerical analysis and scientific computing

- **2. Double Precision (64 bits)**
- **Step 1: Convert to Binary**
- *Absolute value*: 6.75 in binary is 110.11.
- **Step 2: Normalize**
- *Normalize to:* 1011×2^2
- **Step 3: Sign Bit**
- *Sign (S)*: 1 (negative)
- **Step 4: Exponent**
- Actual exponent: 2
- Biased exponent: $2 + 1023 = 1025$
- In binary: 10000001001 (11 bits)
- **Step 5: Mantissa**
- Mantissa (without leading 1): 1011000000000000000000000000000000000000000000000000 (52 bits)
- **Final Representation**
-
- Putting it all together:
- **| 1 | 10000001001 | 1011000000000000000000000000000000000000000000000000 |**

Numerical methods course

chapter 1:

General information on numerical analysis and scientific computing

- Introduction on Error calculation

Error calculation is the process used to find the significance of errors in a given dataset or set of results.

Numerical methods course

chapitre 1:

General information on numerical analysis and scientific computing

- **Methods of error calculation**

Errors in experimental measurements can be expressed in several different ways; the most common are absolute error , relative error and percentage error .

Numerical methods course

chapter 1:

General information on numerical analysis and scientific computing

- **Absolute error**

Absolute error is an expression of how far a measurement is from its actual or expected value. It is reported using the same units as the original measurement. As the true value may not be known, the average of multiple repeated measurements can be used in place of the true value.

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- **Absolute error**

Absolute error

- $Absolute\ Error = \ |\mathrm{Exact\ Value} - \mathrm{Approximate\ Value}|$

- For example, if the exact value of a number is 3.14159 and the computed value is 3.14, the absolute error is: |3.14159−3.14|=0.00159

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- **Relative error**

Relative error (sometimes called proportional error) expresses how large the absolute error is as a portion of the total value of the measurement.

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- **Relative error**
- **Relative error**

$$\text{Relative Error} = \frac{|\text{Exact Value}-\text{Approximate Value}|}{|\text{Approximate Value}|}$$

Using the previous example, the relative error for the values 3.14159 (exact) and 3.14 (approximate) would be:

- $\text{Relative Error} = \frac{|0.00159|}{|3.14159|} \approx 0.000506$

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- **Percentage error**

When the relative error is expressed as a percentage, it is called a **percentage error**.

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- **Stability and Error Analysis of Numerical Methods and Problem Conditioning**

To develop reliable and accurate numerical algorithms, it is essential to integrate error analysis, stability analysis, and considerations of problem conditioning:

- **Algorithm Selection.**
  - **Understanding the Problem**
  - **Error Analysis**
  - **Stability Analysis**
  - **Complexity and Efficiency**
  - **Specific Characteristics of Algorithms**
  - **Problem Conditioning**
  - **Availability of Libraries and Tools**
- **Refinement and Adaptation**
- **Validation and Testing**

Numerical methods course

chapiter 1:

General information on numerical analysis and scientific computing

- **Stability and Error Analysis of Numerical Methods and Problem Conditioning**

Error analysis and stability are vital for ensuring the reliability and accuracy of numerical algorithms. A comprehensive understanding of these concepts, combined with an awareness of problem conditioning, allows developers to create robust numerical methods suitable for a wide range of applications. By systematically addressing these aspects, we can enhance the performance and trustworthiness of numerical computations in science and engineering