Numerical Analysis
Theory

Christian Rossi

Academic Year 2023-2024

Abstract

Contents

1 Introduction **2**
1.1 Numerical analysis and errors 2
1.2 Floating point 3

2 Nonlinear equations **6**

Chapter 1

Introduction

1.1 Numerical analysis and errors

Numerical analysis is the field of mathematics dealing with methods to find the solutions of certain mathematical problems with an electronic calculator. It is the intersection between math and computer science.

On the other hand, scientific computing also deals with the model formalization and so it needs also engineering knowledge.

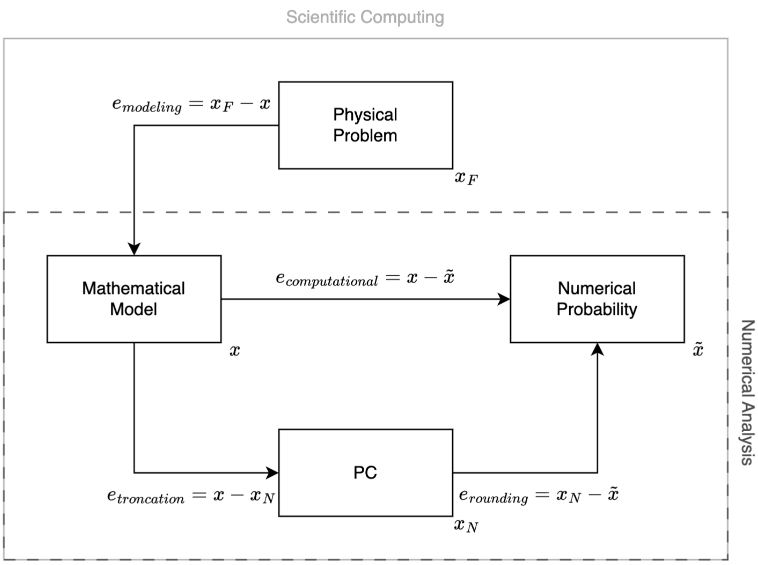


Figure 1.1: Difference between numerical analysis and scientific computing

As it is possible to see from the diagram every step of the computation have to deal with errors. The possible types of errors are:

- Absolute: $|x - \tilde{x}|$
- Relative: $\dfrac{|x - \tilde{x}|}{|x|}$, where $x \neq 0$

The relative error is more precise because it compares the error with the measure quantity.

Example

Let us consider $x = 100$ and $\tilde{x} = 100.1$. The errors in this case are:

$$e_{abs} = |x - \tilde{x}| = |100 - 100.1| = 0.1$$

$$e_{rel} = \frac{|x - \tilde{x}|}{|x|} = \frac{|100 - 100.1|}{|100|} = 0.001$$

Let us consider $x = 0.2$ and $\tilde{x} = 0.1$. The errors in this case are:

$$e_{abs} = |x - \tilde{x}| = |0.2 - 0.1| = 0.1$$

$$e_{rel} = \frac{|x - \tilde{x}|}{|x|} = \frac{|0.2 - 0.1|}{|0.2|} = 0.5$$

The result are that the measures have the same absolute error (10%), but the relative error is much grater in the second example (50% vs 0.1%). This result proves that the relative error is the most precise. ■

1.2 Floating point

A calculator can only handle a finite quantity of numbers and compute a finite number of operations. For those reason the set of the real numbers $\mathbb{R}$ is indeed represented by a finite set of machine numbers $\mathbb{F} = \{-\tilde{a}_{min}, \dots, \tilde{a}_{max}\}$ called floating points numbers. The function used to find the corresponding value in $\mathbb{F}$ to a number in $\mathbb{R}$ is $fl(x)$ that does an operation called truncation and rounding.

The set $\mathbb{F} = \mathbb{F}(\beta, t, L, U)$ is characterized by four parameters β, t, L and U such that every real number $fl(x) \in \mathbb{F}$ can be written as:

$$fl(x) = (-1)^s m\beta^{e-t} = (-1)^s (a_1 a_2 \dots a_t)_\beta \beta^{e-t}$$

where:

- $\beta \geq 2$ is the basics, an integer that determines the numeric system.

- $m = (a_1 a_2 \dots a_t)_\beta$ is the mantissa, $(0 < m < \beta^t - 1)$ where t is the number of digits such that $0 < a_1 \leq \beta - 1$ and $0 \leq a_i \leq \beta - 1$ for $i = 2, \dots, t$.
- $e = \mathbb{Z}$ is the exponent such that $L < e < U$, with $L < 0$ and $U > 0$.
- $s = \{0, 1\}$ is the sign.

In the definition of the numbers in the mantissa set we have to set the constraint $a_1 \neq 0$ to ensure the uniqueness of the representation.

The set of floating points has the following characteristic values:

- Machine epsilon, that is the distance between one and the smallest floating point number greater than one, and it is equal to:

$$\epsilon_M = \beta^{1-t}$$

- Round-off error, that is the relative error that is committed when substituting $x \in \mathbb{R} - \{0\}$ with his corresponding $fl(x) \in \mathbb{F}$. It is limited by:

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2}\epsilon_M$$

  where $x \neq 0$.

- Cardinality of the floating point set:

$$\#\mathbb{F} = 2\beta^{t-1}(\beta - 1)(U - L + 1) + 1$$

  where:

  - 2 is needed to consider both positive and negative numbers.
  - β^{t-1} is the cardinality of values that can be taken by all digits.
  - $(\beta - 1)$ is the cardinality of values that can be taken by a_1.
  - $(U - L + 1)$ considers all the possible variations for the exponent.
  - 1 is needed to consider also the zero.

- The biggest and the smallest numbers in the set are found with the formula:

$$x_{min} = \beta^{L-1}$$

$$x_{max} = \beta^U(1 - \beta^{-t})$$

Example

In MATHLAB the floating point set is defined with the following variables:

$$(\beta = 2, t = 53, L = -1021, U = 1024)$$

With the command *eps* we can find the machine epsilon, that in MATLAB case is:

$$\epsilon_M = 2.22 \cdot 10^{-16}$$

With the command *realmin* and *realmax* we can find the smallest and the largest numbers representable that are equal to:

$$x_{min} = 2.225073858507201 \cdot 10^{-308}$$

$$x_{max} = 1.797693134862316 \cdot 10^{308}$$

■

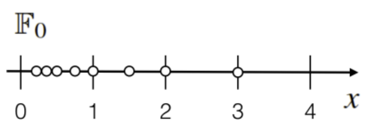
Since not all the numbers in the $\mathbb{R}$ set are also in the $\mathbb{F}$ set, in the second one there is no continuity. It is possible to demonstrate that while we are increasing the values of the numbers we are also increasing the distance between two consecutive numbers in $\mathbb{F}$.

Example

Let us consider the floating number set $\mathbb{F}(2, 2, -1, 2)$. The characteristic values of this set are:

- $\epsilon_M = \beta^{1-t} = 0.5$.
- $x_{min} = \beta^{L-1} = 0.25$.
- $x_{max} = \beta^U(1 - \beta^t) = 3$.
- $\#\mathbb{F} = 2\beta^{t-1}(\beta - 1)(U - L + 1) + 1 = 16$.

The exponent can have the values $-1, 0, 1$ and 2. The mantissa will be like $(a_1 a_2)_\beta$ because $t = 2$. The possible positive values are reported in the figure.



e	-1	0	1	2
$m = (10)_2 = 2$	$\frac{1}{4}$	$\frac{1}{2}$	1	2
$m = (11)_2 = 3$	$\frac{3}{8}$	$\frac{3}{4}$	$\frac{3}{2}$	3

■

The other important aspect is that the passage between the two sets causes the loss of two important properties such as associativity end the neutral number for the sum.

Chapter 2

Nonlinear equations