Numerical Methods

chapter 1:
FINITE-PRECISION ARITHMETIC

Mrs. N.BOUSAHBA

n.bousahba@univ-chlef.dz

Numerical Errors

1- Introduction

2- Sources of Error in Numerical Methods

3- How to round numbers

4- Floating-point arithmetic and rounding error

5- Significatifs Numbers

6- Measurement of Error

1- Introduction :

Numerical analysis is a branch of applied mathematics focused on the development of numerical tools and methods for calculating approximations of solutions to mathematical problems that would be difficult, or even impossible, to obtain through analytical means.

These methods are intended to be implemented and executed on machines.

1- Introduction :

The solutions to problems calculated by a numerical method are not exact; they are subject to errors, which necessitates an analysis to ensure the accuracy and relevance of the results they provide.

2- Sources of Error in Numerical Methods :

An approximate numerical result has no meaning unless accompanied by an estimate of the error made between the exact and approximate results; without that, we cannot say anything.

We can consider several ways to measure the error E between an approximate value x* and an exact value x.

2- Sources of Error in Numerical Methods :

1- Rounding Errors

2- Approximation Errors

3- Data (Measurement) Errors

2- Sources of Error in Numerical Methods :

1- Rounding Errors :

Difference between the represented or calculated value of a number and its exact mathematical value.

Rounding errors generally arise when exact numbers are represented in a system that is unable to express them exactly.

Rounding errors propagate during calculations with approximate values, which can increase the error in the final result.

- Every calculator operates with finite precision (with a limited number of digits):

Example : $4/3 = 1{,}33333.. = 1{,}\overline{3}$

With 6 significant digits : $4/3 = 1{,}33333$

$$\text{loss} = \frac{1}{3} * 10^{-5}$$

2- Sources of Error in Numerical Methods :

1- Rounding Errors :

Value	Representation	Approximate value	Error
1/7		0.142857	$1/7 * 10^{-6}$
Ln 2	0.693147180559945..	0.693147	0.000000180559945
	1.414213562373095..	1.41421	0.000003562373095
e	2.71828182845904590452353..	2. 718281828459045	0.00000000000000090452353..
	3.1415926535897932384626..	3.141592653589793	0.00000000000000002384626..

2- Sources of Error in Numerical Methods :

2- Approximation Errors :

These are truncation, approximation, or discretization errors:

- Stop the infinite series expansion of an analytical solution to allow for its evaluation.
- Stop an iterative process as soon as an iterate meets a given criterion within a specified tolerance.
- Approach the solution of a partial differential equation at a finite number of points.

2- Sources of Error in Numerical Methods :

3-Data (Measurement) Errors :

These are errors attributable to an imperfect knowledge of the data of the problem we are trying to solve..

- Physical measurements subject to experimental constraints..
- The data itself comes from an approximate calculation.
- Truncated or rounded data.

These are errors imposed from the outside, but they can have a significant impact on the results.

4.How to round numbers :

- Truncation (Cut) : Cut off the digits of a number after the desired decimal place..

Example : Truncate the number π = 3.141592653589.. after the fourth decimal place :

π = 3.141592653589.. =3.1415

- Rounding: round to the nearest value by adding 5 to the first decimal place after the decimal you want to keep.

By rounding up (6th decimal place) : π = 3.141592653589.. = 3.141593

By rounding down (2nd decimal place) : π = 3.141592653589.. = 3.14

4- Floating-point arithmetic and rounding error :

The IEEE 754 standard :

In the IEEE 754 standard, a floating-point number is always represented by a triplet (s, e, m):

Format	w	t	Total
Simple	8 bits e : -126 , +127	23 bits m : 1 ≤ m < 2	32 bits
Double	11 e : -1022, +1023	52 bits m : 1 ≤ m < 2	64 bits
Quadruple	15 e : -16382, +16383	112 bits m : 1 ≤ m < 2	128 bits

4- Floating-point arithmetic and rounding error :

The IEEE 754 standard :

$N = \frac{7}{250} = 0{,}028 = 1{,}792 * 2^{-6} \Rightarrow s = 0 \quad e = -6 \quad m = 1{,}792$

$e = E - 2^{w-1} + 1 \Rightarrow E = e + 2^{w-1} - 1 = -6 + 2^{w-1} - 1 \Rightarrow E = -7 + 2^{w-1}$

$m = M * 2^{-t} + 1 \Rightarrow M = (m - 1) * 2^{t} = 0{,}792 * 2^{t}$

Simple precision :

$E = -7 + 2^{7} = 121$

$M = 0{,}792 * 2^{23} = 6643777{,}536$

$Error = 0{,}536 \times 2^{-23} = 6{,}389 * 10^{-8}$

Double precision :

$E = -1 + 2^{10} = 1017$

$M = 0{,}792 * 2^{52} = 3566850904877432{,}832$

$Error = 0{,}832 \times 2{-}52 = 1{,}847 * 10^{-16}$

5- Significant Numbers :

The number of significant figures indicates the precision of a physical measurement.

Example: The measurement 0028.500 m contains 5 significant numbers.

5- Significant Numbers :

Rules for determining significant figures in a measurement :

All digits (1, ..., 9) that make up a number are considered significant.

All zeros located in the middle are significant, except for those that are farthest to the left.

All digits in a number written in scientific notation are significant except for the power of 10, which is not counted: 2.5×10^4 (2 significant figures).

Example :

Value	0,10	7300	0,0073	3,5889	0,0009	0023,2	3000,05	$0,0203x10^5$
Number of significant Numbers	2	4	2	5	1	3	6	3

5- Significant Numbers :

Rules for determining the significant numbers of the result of an operation :

- **Multiplication and division:** The result has a precision equivalent to that of the term with the lowest precision.

Example :

- 2,689 et 3,6 x 10^5 = 9,6804 x 10^5 ≃ 9,7 x 10^5.
- 500/100 = 5 = 5,00

- **Addition and subtraction:** The precision of the result is equivalent to that of the least precise number (focus is on the number of digits after the decimal point, not on the number of significant figures).

Example :

- 256,3 + 1,89 = 258,19 ≃ 258,2

5- Significant Numbers :

Loss of significant numbers :

Example1 :

Let $\pi = 3.141592653589793\ldots$

If we work with 8 significant numbers: $\pi = 3.1415927$

Let $x = \pi - 3.1415$

$x = 3.1415927 - 3.1415 = 0.0000927 = 9.27\ 10^{-5}$.

Loss of 5 significant numbers.

5- Significant Numbers :

Loss of significant numbers :

Example2 :

$$\text{Let } A = \frac{KN}{KD} = \frac{\pi - 3{,}1415}{10^4(\pi - 3{,}1415) - 0{,}927}$$

- Loss of 8 significant figures : Error DBZ
- With 9 numbers, we obtain: A= -0,1853
- With 10 numbers, we obtain : A= -0,197134

6- Measurement of error :

Let X be an approximate value of a variable, and $\bar{X}$ its exact value..

Absolute error:

$\Delta X = |X - \bar{X}|$ is called absolute error

Relative error :

$\rho(X) = |\frac{X - \bar{X}}{\bar{X}}|$ est appelé erreur relative

Example :

Let the value $X = \frac{1}{7}$.

Its approximate value represented with 6 numbers is X=0,142857

Its exact value is $\bar{X} = \frac{1}{7} = 0,\overline{142857}$

$$\Delta X = |X - \bar{X}| = |0.142857 - \frac{1}{7}| = \frac{1}{7} * 10^{-6}$$

$$\rho(X) = |\frac{X - \bar{X}}{\bar{X}}| = 10^{-6} = 0.0001\%$$