PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
Ministry of Higher Education and Scientific Research
University of Laghouat
Faculty of Sciences
Department of Material Sciences

Numerical Analysis

- Theorical lessons
- Exercises with solutions
- Exercise series
- Practical work

This course is written by

Dr. Mohamed Lamine MOSTEFAI

LMD

Laghouat, 2023–2024

$\mathcal{UATL}$
2024

Contents

Introduction

Numerical analysis and scientific computing play a large role in applied mathematics and the design of industrial products; however, it is only one link in a long chain that mobilizes many and varied intellectual resources to arrive at the design, at best within the time limits, of the desired product. A study and design process can be represented very schematically by the following procedure:

- Mathematical modeling and numerical simulation
- Numerical algorithm, numerical methods for solving linear systems and nonlinear, optimization
- Polynomial interpolation, Numerical integration.
- Numerical resolution of ordinary differential equations (O.D.E): Euler methods, Runge-Kutta methods...
- Approximation of partial differential equations (P.D.E): Finite differences, finite elements, finite volumes...
- Computer calculation ...
- Experimentation.
- Exploitation of products.

This courses on the theory and application of numerical approximation techniques. This course is designed primarily for junior-level mathematics, science, and engineering majors who have completed at least the study in first-year courses in the speciallities science and technic offered at universities

Familiarity with the fundamentals of analysis, linear algebra and differential equations is useful, but there is sufficient introductory material on these topics so that courses in these subjects are not needed as prerequisites.

In some cases, the mathematical analysis underlying the development of approximation techniques was given more emphasis than the methods; in others, the emphasis was reversed.

The courses may be used as a reference for beginning graduate level courses in engineering, programmation and computer science. We have adapted the course to fit these diverse users without compromising our original purpose: To introduce modern approximation techniques; to explain how, why, and when they can be expected to work; and to provide a foundation for further study of numerical analysis and scientific computing.

In such course, students learn to identify the types of problems that require numerical techniques for their solution and see examples of the error propagation that can occur when numerical methods are applied. They accurately approximate the solution of problems that cannot be solved exactly and learn typical techniques for estimating error bounds for the approximations. The remainder of the text then serves as a reference for methods not considered in the course.

Modeling and numerical approximation see their applications in different areas, for example:

- Aircraft design (aerodynamics, composite materials ...)

- Car design (aerodynamics, flow in engines, spit tests, optimal control, structure (tires, bodywork,) ....
- Petroleum engineering: understanding the migration of hydrocarbons, improving the production of oil fields, ...
- Mathematical biology: epidemic propagation, mathematical model in cardiology, cancer, dental tissues, pulmonology, ...
- Inventory management, finance, road traffic
- Environment: air, water, soil pollution
- Weather: Model the World
- And many other applications...

This module, which relates to applied mathematics, allows the student to:

- Know how to approach a physical problem that can be analytically solved from a numerical point of view.
- Address analytically unsolvable problems numerically.

In this course, we are interested in numerical analysis; this discipline itself can be considered as divided into two major themes:

- Some notions of error
- Numerical algorithms: solving large sparse linear systems, polynomial approximation, integration numerical, numerical resolution of ODEs (with Cauchy and Dirichlet conditions), optimization
- Numerical calculation of eigenvalues and eigenvectors.
- Least squares approximations.
- Numerical simulation and programming.

Chapter 1

Notions of error

1.1 Introduction

In the title of this chapter, the word **error** is not pus in the sense of fault. It concerns only unavoidable errors. They can be classified into three categories:

1. Errors related to the inaccuracy of the **physical measurements** or to the result of an approximate calculation.
2. **Rounding errors:** These are errors due to the fact that the machine can only represent real numbers with a finite number of digits.

   Example 1.1.1
   $$\frac{1}{3} = 0.333333...; \ \pi = 3.14.....$$
3. **Approximation or discretization errors:** These are the errors that are made, for example, when calculating an integration of a finite sum, a derivative using finite differences or the sum of an infinite series using a finite number of its terms (this is sometimes referred to as a trancature error).

   Example 1.1.2
   $$\sum_{i\in\mathbb{N}} \frac{1}{i!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots \approx e^1,$$
   $$\frac{1}{n^2}\sum_{i=0}^{n} i \approx \int_0^1 x dx, \qquad when \qquad n \longrightarrow +\infty,$$
   $$\frac{b-a}{n}\sum_{i=0}^{n} f\left(a + i\frac{(b-a)}{n}\right) = \int_a^b f(x)dx, \qquad n \longrightarrow +\infty.$$

1.2 Error evaluation

Let's first recall some basic concepts if X is a quantity to calculate (Example: $\sqrt{2}$), and $X^\star$ the calculated value (Example: $X^\star = 1.4\cdots$).
We say that

a) $X - X^\star$ is the error and $|E| = |X - X^\star|$ is the absolute error.

   Example 1.2.1 *If $X = 2.224$ and $X^\star = 2.223$ then the absolute error is given by*
   $$X - X^\star = 2.224 - 2.223 = 0.001,$$
   and the absolute error is
   $$|E| = |X - X^\star| = 0.001.$$

b) $E_r = \left|\dfrac{X - X^\star}{X_r}\right|$ is the relative error, X_r is a reference value for X (provided $X_r \neq 0$ or X_r is not close to zero). In general, we take $X_r = X^\star$.

Example 1.2.2 *If $X = 2.224$ and $X^\star = 2.223$ then, if we take $X_r = X^\star$, the relative error is*

$$E_r = \left|\frac{X - X^\star}{X_r}\right| = \frac{|X - X^\star|}{|X_r|} = \frac{0.001}{2.223} = 4.496 \times 10^{-4}.$$

c) The percentage relative error is defined by $E_p = E_r \times 100$.

Remark 1.1 *In general, we write*

$$|X - X^\star| \leq \Delta X \qquad \text{and} \qquad |E| \approx \Delta X, \qquad E_r \approx \frac{\Delta X}{|X^\star|}$$

such as $|X| \approx |X^\star|$.

Remark 1.2 *In practice it is impossible to evaluate the absolute error because X is often unknown therefore, we introduce the notion of the bound upper part of this error noted ΔX and we have $E = |X - X^\star| \leq \Delta X$ which allows to write*

$$X = X^\star \pm \Delta X.$$

1.3 Operations on errors

Let X_1 the calculated value of a value, X_2 the computed value of another value, approaching by $X_1^\star$, $X_2^\star$ respectively. We have

Addition (+)

We set $X = X_1 + X_2$, and we suppose that X approximated by $X^\star$ such that $X^\star = X_1^\star + X_2^\star$, so

$$\begin{aligned} E = X - X^\star &= (X_1 + X_2) - (X_1^\star + X_2^\star) \\ &= (X_1 - X_1^\star) + (X_2 - X_2^\star) \\ &= E_1 + E_2, \end{aligned}$$

implies that $|E| \leq \Delta X_1 + \Delta X_2$ and $E_r \approx \dfrac{\Delta X_1 + \Delta X_2}{|X_1^\star + X_2^\star|}$.

Remark 1.3 *If $X = X_1 - X_2$ then*

$$E = E_1 + E_2 \leq \Delta X_1 + \Delta X_2$$

and

$$E_r \approx \frac{\Delta X_1 + \Delta X_2}{|X_1^\star - X_2^\star|}$$

with $X_1^\star \neq X_2^\star$.

Product ($\times$)

We set $X = X_1 X_2$ approximated by $X^\star = X_1^\star X_2^\star$, the absolute error

$$\begin{aligned} |E| &= |X - X^\star| \\ &= |X_1 X_2 - X_1^\star X_2^\star| \\ &= |X_1 X_2 - X_1 X_2^\star + X_1 X_2^\star - X_1^\star X_2^\star| \\ &= |X_1(X_2 - X_2^\star) + X_2^\star(X_1 - X_1^\star)| \\ &\leq |X_1||X_2 - X_2^\star| + |X_2^\star||X_1 - X_1^\star| \end{aligned}$$

since $X_1 \approx X_1^\star$, we deduce that

$$|E| \quad \leq \quad |X_1^\star|\Delta X_2 + |X_2^\star|\Delta X_1.$$

The relative error

$$E_r = \frac{|X_1^\star|\Delta X_2 + |X_2^\star|\Delta X_1}{|X_1^\star X_2^\star|} = \frac{\Delta X_2}{|X_2^\star|} + \frac{\Delta X_1}{|X_1^\star|},$$

since $X_1 \approx X_1^\star$ and $X_2 \approx X_2^\star$.

Quotient $(\div)$

We set $X = X_1/X_2$ approximated by $X^\star = X_1^\star/X_2^\star$.
Absolute error

$$\begin{aligned} |E| &= |X - X^\star| \\ &= \left|\frac{X_1}{X_2} - \frac{X_1^\star}{X_2^\star}\right| \\ &= \left|\frac{X_1X_2^\star - X_2X_1^\star}{X_2X_2^\star}\right| \\ &= \left|\frac{X_2^\star(X_1 - X_1^\star) + X_1^\star(X_2 - X_2^\star)}{X_2X_2^\star}\right| \\ &\leq \frac{|X_2^\star|\Delta X_1 + |X_1^\star|\Delta X_2}{|X_2^\star|} \end{aligned}$$

Relative error:

$$E_r = \frac{|E|}{|X_1^\star/X_2^\star|} = \frac{|X_2^\star|\Delta X_1 + |X_1^\star|\Delta X_2}{|X_2^\star|^2}\left|\frac{X_2^\star}{X_1^\star}\right| = \frac{\Delta X_1}{|X_1^\star|} + \frac{\Delta X_2}{|X_2^\star|}.$$

In summary:

The operator	The absolute error $E \approx \Delta X$	The relative error E_r
$X_1 \pm X_2$	$\Delta X_1 + \Delta X_2$	$\frac{\Delta X_1 + \Delta X_2}{\lvert X_1^\star \pm X_2^\star\rvert}$
X_1X_2	$\lvert X_1^\star\rvert\Delta X_2 + \lvert X_2^\star\rvert\Delta X_1$	$\frac{\Delta X_1}{\lvert X_1^\star\rvert} + \frac{\Delta X_2}{\lvert X_2^\star\rvert}$
$\frac{X_1}{X_2}$	$\frac{\lvert X_1^\star\rvert\Delta X_2 + \lvert X_2^\star\rvert\Delta X_1}{\lvert X_2^\star\rvert^2}$	$\frac{\Delta X_1}{\lvert X_1^\star\rvert} + \frac{\Delta X_2}{\lvert X_2^\star\rvert}$

Example 1.3.1 *Calculate the ebsolute error, the relative error of X such that $X = X_1 + X_2$ and $X_1 = 3.792 \pm 0.5 \times 10^{-3} = X_1^\star \pm \Delta X_1$ and $X_2 = 2.814 \pm 0.5 \times 10^{-3} = X_2^\star \pm \Delta X_2$.*

Correction 1.3.1 *Absolute error:*

$$E_{X_1+X_2} = |E_{X_1} + E_{X_2}| = 0.5 \times 10^{-3} + 0.5 \times 10^{-3} = 0.1 \times 10^{-2} \leq 0.5 \times 10^{-2},$$

implies that:

$$X_1 + X_2 = X_1^\star + X_2^\star \pm E_{X_1+X_2} = 6.606 \pm 0.1 \times 10^{-2}.$$

The relative error:

$$E_r = \frac{\Delta X}{|X^\star|} = \frac{0.1 \times 10^{-2}}{6.606} = 0.00015.$$

1.4 Representation of formulas

Equivalent formulas can give different results, the result can be improved by using an equivalent mathematical formula requiring different operations.

Example 1.4.1 *If we consider the numbers* $\sqrt{7001}$ *and* $\sqrt{7000}$.
In floating-point arithmetic with 8 *digits, we have*

$$\begin{aligned}\sqrt{7001} &= 0.83671979 \times 10^2,\\ \sqrt{7000} &= 0.83666003 \times 10^2.\end{aligned}$$

So,

$$\begin{aligned}\sqrt{7001} - \sqrt{7000} &= (0.83671979 - 0.83666003) \times 10^2\\ &= 0.59760000 \times 10^2.\end{aligned}$$

But,

$$\frac{1}{\sqrt{7001} + \sqrt{7000}} = \frac{1}{0.16733798 \times 10^3} = 0.59759297 \times 10^{-2}.$$

1.5 Numerical instability

If the errors introduced in the intermediate steps have a negligible effect on the final result, we will say that the calculation or the algorithm is numerically stable, if small changes in the data lead to small changes in the result, otherwise we will say that the The algorithm is numerically unstable.

Example 1.5.1 *We want to calculate the value of*

$$I_n = \int_0^1 \frac{x^n}{a+x} dx,$$

where a *is a constant greater than* 1*, for several values of* n*, to do this we will express* I_n *recursively, i.e. we will express* I_n *as function of* n *and* I_{n-1}*:*

$$\begin{aligned}I_n &= \int_0^1 \frac{x^{n-1}(x+a-a)}{a+x} dx\\ &= \int_0^1 x^{n-1} dx - a\int_0^1 \frac{x^{n-1}(x)}{a+x} dx\\ &= \frac{1}{n} - aI_{n-1}.\end{aligned}$$

By recurrence, we have

$$I_n = \sum_{i=0}^{n-1} \frac{(-a)^i}{n-i} + (-a)^n I_0.$$

As $I_0 = \ln\left(\frac{1+a}{a}\right)$*, we can compute* I_n *for all values of* n*. But the algorithm is numerically unstable because any error in the calculation of* $I_0 = \ln\left(\frac{1+a}{a}\right)$ *will propagate.*
Indeed if we denote by $I_0^\star$ *the approximate value of* I_0 *and if* $I_0^\star = I_0 + \epsilon$ *so*

$$I_0^\star = \sum_{i=0}^{n-1} \frac{(-a)^i}{n-i} + (-a)^n (I_0 + \epsilon),$$

so,

$$|I_n - I_0^\star| \geq a^n \epsilon.$$

1.6 Conclusion

There are actually different sources of errors. We can classify them into three categories:

a) Errors related to the printing of physical measurements or to the result of an approximate calculation.

b) Errors related to the algorithm used.

c) Calculation errors related to the machine.

1.7 Exercises with solutions

Exercise 1.1 *Let the values X_1, X_2 be such that*

$$\begin{aligned} X_1 &= 3.254 \pm 0.5 \times 10^{-3} \\ X_2 &= 3.568 \pm 0.5 \times 10^{-3} \end{aligned}$$

1. *Complete the following table:*

	the absolute error $\lvert E\rvert$	*the relative error* $\lvert E_r\rvert$
$X_1 \pm X_2$		
$X_1 \times X_2$		
X_1/X_2		

2. *Write values $X_1 \pm X_2$, $X_1 \pm X_2$, $X_1 \times X_2$, X_1/X_2 in the form $X = X^* \pm \Delta X$.*

Exercise 1.2 *Let the value X be such that*

$$X = X^* \pm \Delta X$$

1. *Determine $Y = \sqrt{X}$, $Z = X^n$, $n \in \mathbb{N}$, $V = \sqrt[3]{X^2}$, $U = X^\alpha$, $\alpha \in \mathbb{R}$*

2. *Calculate $W = X^{\frac{2}{3}}$ with $X = X^* \pm \Delta X = 8 \pm 0.1$*

Exercise 1.3 *We consider the equation*

$$x^2 - 1634x + 2 = 0. \tag{1.7.1}$$

1. *Solve the equation 1.7.1 by performing the calculations with $N = 10$ significant digits. (Use the discriminant)*

2. *Comment on the result obtained and suggest another method of calculation to circumvent the problem posed.*

1.7.1 Solutions

Solution 1.1 *We have the values X_1, X_2 such that*

$$\begin{aligned} X_1 &= 3.254 \pm 0.5 \times 10^{-3} = X_1^* \pm \Delta X_1 \\ X_2 &= 3.568 \pm 0.5 \times 10^{-3} = X_2^* \pm \Delta X_2 \end{aligned}$$

1. *Complete the following table:*

	the absolute error $\lvert E\rvert$	*the relative error* E_r
$X_1 \pm X_2$	$\Delta X_1 + \Delta X_2 = 10^{-3}$	$E_r(X_1 + X_2) = \frac{\lvert E\rvert}{X_1^* + X_2^*} = 0.17 \times 10^{-3}$, $E_r(X_1 - X_2) = \frac{\lvert E\rvert}{X_1^* - X_2^*} = 1.46 \times 10^{-3}$
$X_1 \times X_2$	$X_1^* \Delta X_2 + X_2^* \Delta X_1 = 2.911 \times 10^{-3}$	$E_r(X_1 \times X_2) = \frac{\lvert E\rvert}{X_1^* \times X_2^*} \approx 0.35 \times 10^{-3}$
$\frac{X_1}{X_2}$	$\frac{X_1^* \Delta X_2 + X_2^* \Delta X_1}{(X_2^*)^2}$	$E_r(X_1/X_2) = \frac{\lvert E\rvert}{X_1^*/X_2^*} \approx 0.35$

2. *We write the values $X_1 \pm X_2$, $X_1 \pm X_2$, $X_1 \times X_2$, X_1/X_2 in the form $X = X^* \pm \Delta X$:*

$$\begin{aligned} X_1 + X_2 &= X_1^* + X_2^* \pm \Delta(X_1 + X_2) \\ &= 5.822 \pm 10^{-3} \\ X_1 - X_2 &= X_1^* - X_2^* \pm \Delta(X_1 - X_2) \\ &= 0.686 \pm 10^{-3} \\ X_1 \times X_2 &= X_1^* \times X_2^* \pm \Delta(X_1 \times X_2) \\ &= 8.36 \pm 2.911 \times 10^{-3} \\ \frac{X_1}{X_2} &= \frac{X_1^*}{X_2^*} \pm \left(\frac{X_1}{X_2}\right) \\ &= 1.27 \pm 0.44. \end{aligned}$$

Solution 1.2 *We have*

$$X = X^* \pm \Delta X$$

1. *Determine $Y = \sqrt{X}$, $Z = X^n$, $n \in \mathbb{N}$, $V = \sqrt[3]{X^2}$, $U = X^\alpha$, $\alpha \in \mathbb{R}$*

   (a) *We have*

   $$Y = \sqrt{X} \Longleftrightarrow Y^2 = X,$$

   then

   $$\Delta Y^2 = \Delta X \Longleftrightarrow 2Y^* \Delta Y = \Delta X \Longleftrightarrow \Delta Y = \frac{\Delta X}{2Y^*}$$

   and we have $Y^* = \sqrt{X^*}$ then $Y = Y^* \pm \Delta Y = \sqrt{X^*} \pm \frac{\Delta X}{2Y^*}$.

   (b) *We have*

   $$Z = \sqrt[n]{X} \Longleftrightarrow Z^n = X,$$

   then

   $$\Delta Z^n = \Delta X \Longleftrightarrow n(Z^*)^{n-1}\Delta Z = \Delta X \Longleftrightarrow \Delta Z = \frac{\Delta X}{n(Z^*)^{n-1}}$$

   and we have $Z^* = \sqrt[n]{X^*}$ then $Z = Z^* \pm \Delta Z = \sqrt[n]{X^*} \pm \frac{\Delta X}{n(Z^*)^{n-1}}$.

   (c) *We have $V = \sqrt[3]{X^2} \Longleftrightarrow V^3 = X^2$ so $\Delta V^3 = \Delta X^2$ then*

   $$\Delta V^3 = \Delta(V \times V^2) = V^* \Delta V + (V^2)^* \Delta V = V^*(2V^* \Delta V) + (V^2)^* \Delta V = 3(V^*)^2 \Delta V$$

   we obtain

   $$3(V^*)^2 \Delta V = 2X^* \Delta X \Longrightarrow \Delta V = \frac{2X^* \Delta X}{3(V^*)^2}$$

   (d) *We have $U = X^\alpha$ then $\frac{\Delta U}{U^*} = \alpha \frac{\Delta X}{X^*} \Longleftrightarrow \Delta U = \alpha \frac{\Delta X}{X^*}$ so*

   $$U = U^* + \Delta U = U^* + \alpha \frac{\Delta X}{X^*}.$$

2. *Computing $W = X^{\frac{2}{3}}$ with $X = X^* \pm \Delta X = 8 \pm 0.1$.*
   We have $W = X^{\frac{2}{3}}$ then $\Delta W = \frac{2X^* \Delta X}{3(W^*)^2}$ and as $W^* = (X^*)^{2/3}$ implies that $(W^*)^2 = X^*(X^*)^{1/3}$ so

$$\Delta W = \frac{2X^* \Delta X}{3X^*(X^*)^{1/3}} = \frac{2\Delta X}{3(X^*)^{1/3}},$$

we obtain

$$\Delta W = \frac{2}{3} \times \frac{0.1}{(8)^{\frac{2}{3}}} \approx 0.333... \times 10^{-1} \leq 0.5 \times 10^{-1}.$$

Finally,

$$W = W^* \pm \Delta W = 4 \pm 0.5 \times 10^{-1}.$$

Solution 1.3 *We consider the equation*

$$x^2 - 1634x + 2 = 0,\ a = 1,\ b = -1634,\ c = 2. \tag{1.7.2}$$

1. *We are going to solve the equation 1.7.2 by performing the calculations with $N = 10$ significant digits. (Use discriminant).*

   We have,

   $$\Delta = b^2 - 4ac = (-1634)^2 - 4(1)(2) = 2669948 \Longrightarrow \sqrt{\Delta} = 1633.99755201775 > 0.$$

   and

   $$x_1 = \frac{-b + \sqrt{\Delta}}{2a} = \frac{-2c}{b + \sqrt{\Delta}} = x_1',$$
   $$x_2 = \frac{-b - \sqrt{\Delta}}{2a} = \frac{-2c}{b - \sqrt{\Delta}} = x_2',$$

   so that

   $$x_1 = 1633.99877600888,$$
   $$x_2 = 0.001223991125.$$

   and

   $$x_1' = 1633.99877593067,$$
   $$x_2' = 0.00122391124.$$

2. *We have $x_1 = x_1'$ and $x_2 = x_2'$ but we note $(x_1)^* \approx (x_1')^*$ and $(x_2)^* \approx (x_2')^*$. Equivalent formulas may give different results.*

1.8 Exercises without solutions

Exercise 1.1 1. *Calculate the absolute and relative error, then give the exact numbers in the sum $A + B$ where*

$$A = 3.792 \pm 0.5 \times 10^{-3}, \qquad B = 2.814 \pm 0.5 \times 10^{-3}.$$

2. *The same question for $A - B$.*

3. *Calculate the value $P = u \times v$ such that $u^* = 369.7$ and $v^* = 0.0042131$. If all exact numbers (give absolute error-relative error and exact numbers).*

4. *The same question for $q^* = \dfrac{3.219 + 2.73 \times 1.812}{8.7193}$.*

5. *Compute $u = \dfrac{x^2 y}{z}$ where $x = 8 \pm 0.08, \quad y = 5 \pm 0.1$ and $z = 10 \pm 0.1$*

6. *Compute $y = \sqrt[3]{x^2}$ where $x = 8000 \pm 3$.*

7. *Compute* $u = \dfrac{x+y}{z}$ *where*

$$x = 35.4 \pm 0.5 \times 10^{-1}, \quad y = 0.0123 \pm 0.5 \times 10^{-4}, \quad z = 227.143 \pm 0.5 \times 10^{-3}.$$

Exercise 1.2 *Using the Taylor polynomial of degree* 9 *of the function* $f(x) = e^x$ *and* 3 *digit floating-point arithmetic without rounding, calculate* e^{-5} *with the following formulas:*

a) $e^{-5} \approx \sum_{i=0}^{9} \dfrac{(-5)^i}{i!}$.

b) $e^{-5} = \dfrac{1}{e^5} \approx \dfrac{1}{\sum_{i=0}^{9} \dfrac{5^i}{i!}}$.

c) Taking $e^{-5} = 6.74 \times 10 - 3$*, which formula of a) or b) gives the best precision? justify your answer.*

Exercise 1.3 *We want to calculate the surface of a disk* $S = \pi R^2$ *where* $R = 2.3400$, $\pi = 3,1416$. *Assuming that all digits of* R *and* π *are correct.*

1. *Estimate the absolute and relative errors of* S.
2. *Calculate* S *by rounding to the number of exact significant digits.*

Exercise 1.4 *Consider the approximation* $\Delta f(x) \approx |df(x)|$.

1. *Show that* $\Delta \ln(x) = \frac{\Delta x}{x}$.
2. *Deduce the relative error of a power* $u = x^n$ *is such that* $\Delta u = n\Delta x$.
3. *Determine absolute error and relative error of nth root* $v = \sqrt[n]{x}$.
4. *Let* $x = 22.123002$ *and* $y = 1.252468$ *where all significant digits are correct. Calculate* $\ln(\frac{x}{y})$ *and deduce the value of* $\sqrt[3]{\frac{x}{y}}$ *round to the last nth significant digits.*

Exercise 1.5 *Let* T *be the period of the small oscillations of the pendulum which is given by* $T = 2\pi\sqrt{\frac{l}{g}}$ *where* l *is the length of the pendulum and* g *is the gravity. Suppose that the measurements made on* T *and* l *gave the following results:*

$$T = T^* \pm \Delta T = 1.936 \pm 0.002\, s \;\; \text{and} \;\; l = l^* \pm \Delta l = 92.95 \pm 0.10\, cm.$$

1. *Calculate the gravity* g *by rounding to the last significant digit.*
2. *Give the relative error on* g *in percentage.*

Exercise 1.6

1. *Verify that*

$$\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}, \quad x \geq 0.$$

– *For* $x = 5000$*, compute* $a = \sqrt{5001} - \sqrt{5000}$, $b = \frac{1}{\sqrt{5001}+\sqrt{5000}}$ *and* $a - b$ *then analyze the results.*

2. *This calculate the roots of the equation* $x^2 + 111.11x + 1.2121 = 0$*, in an arithmetic to* 5 *significant digits (base* 10*)by using the formulas*

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x'_{1,2} = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}$$

and analyze the results.

Chapter 2

Numerical solutions of nonlinear systems

2.1 Introduction

The determination of the exact roots of the equation $f(x) = 0$ with $x \in \mathbb{R}$, when the expression of the function f is very complex, is a difficult problem to solve by analytical methods. Practically, to solve this problem one must resort to approximate methods. These have proven to be very effective and widely used in practice. They give the roots with as much precision as you want and they are easily done on a computer.

The purpose of this chapter is to develop some methods. More precisely, we study the following methods Dichotomy (Bissection), successive approximations and Newton-Raphson. The convergence of these methods will also be studied.

2.2 Bisection method (dichotomy)

Before approaching this method let us recall the theorem of the intermediate value which is at the base of this one.

Theorem 2.1 *[2] Let f be a function defined and continuous on a bounded closed interval $[a, b]$ of $\mathbb{R}$. Then for any real θ belonging to $f([a, b])$, there exists a real $c \in [a, b]$ such that $\theta = f(c)$. Moreover if f is strictly monotone then the point c is unique.*

The special case of intermediate value theorem for $\theta = 0$ is given by

Theorem 2.2 *[2] Let f be a function defined and continuous on an interval $[a, b]$ and verifying $f(a) \times f(b) < 0$, then there exists a real $c \in [a, b]$ such that $f(c) = 0$. Moreover if f is strictly monotone then the point c is unique.*

To determine the solution of the equation $f(x) = 0$, where f is continuous on $[a, b]$ and $f(a) \times f(b) < 0$, we proceed as follows:

1. if $f(a) \times f(x_0) \leq 0$, then $\overline{x} \in [a, x_0]$.
2. if $f(x_0) \times f(b) \leq 0$, then $\overline{x} \in [x_0, b]$.
3. Note the new interval containing $\overline{x} \in [a_1, b_1]$ where

$$a_1 = \begin{cases} a \text{ if } \overline{x} \in [a, x_0] \\ x_0 \text{ if } \overline{x} \in [x_0, b] \end{cases} \quad \text{and} \quad b_1 = \begin{cases} x_0 \text{ if } \overline{x} \in [a, x_0] \\ b \text{ if } \overline{x} \in [x_0, b]. \end{cases}$$

4. By iterating this process, we obtain a sequence of values $x_0 = \frac{a+b}{2}$, $x_1 = \frac{a_1+b_1}{2} \cdots x_n = \frac{a_n+b_n}{2}$ which verify the inequality $|\overline{x} - x_n| \leq \frac{b-a}{2^{n+1}}$, (see Theorem 2.3).

Remark 2.1 *The last inequality shows us that when $n \to \infty$ we get $x_n \to \overline{x}$ the solution of $f(x) = 0$. Also allows us to estimate in advance the number of iterations necessary to approach $\overline{x}$ with a given precision ϵ.*

For example to have an error not exceeding ϵ, it suffices that

$$\begin{aligned}
\frac{(b-a)}{2^{n+1}} \leq \epsilon &\implies \ln\left(\frac{(b-a)}{2^{n+1}}\right) \leq \ln(\epsilon) \\
&\implies \ln\left(\frac{2^{n+1}}{(b-a)}\right) \geq \ln(\frac{1}{\epsilon}) \\
&\implies (n+1)\ln(2) \geq \ln\left(\frac{(b-a)}{\epsilon}\right) \\
&\implies n \geq \frac{\ln\left(\frac{(b-a)}{\epsilon}\right)}{\ln(2)} - 1 = \frac{\ln\left(\frac{(b-a)}{\epsilon}\right) - \ln(2)}{\ln(2)} \\
&\implies n \geq \frac{\ln\left(\frac{(b-a)}{2\epsilon}\right)}{\ln(2)}.
\end{aligned}$$

If $\epsilon = 10^{-5}$, $b = 2$, $a = 1$, then $n \geq 15.6$. This means that the number of iterations necessary is $n = 16$.
It is important to note that the number of iterations needed given by the above formula is, in many cases, an overestimate of the actual number of iterations needed.

Bisection method algorithm

Let f be a continuous function on $[a, b]$ such that $f(a) \times f(b) < 0$. In order to find an approximate solution of $f(x) = 0$ we proceed as follows:

- **Step** 1**:** Set $a_0 = a$ and $b_0 = b$.
- **Step** 2**:** Set $i = 0$.
- **Step** 3**:** Set $x_i = (a_i + b_i)/2$.
- **Step** 4**:** If x_i is a satisfactory approximation, go to step 10. Otherwise, go to step 5.
- **Step** 5**:** If $f(x_i) \times f(a_i) > 0$, go to step 6. If $f(x_i) \times f(a_i) < 0$, go to step 8.
- **Step** 6**:** Set $a_{i+1} = x_i$ and $b_{i+1} = b_i$.
- **Step** 7**:** Add 1 to i and go to step 3.
- **Step** 8**:** Set $a_{i+1} = a_i$ and $b_{i+1} = x_i$.
- **Step** 9**:** Add 1 to i and go to step 3.
- **Step** 10**:** End.

Theorem 2.3 *[2] Let f be a fonction continues on $[a, b]$ such that $f(a) \times f(b) < 0$. The dichotomy (bisection) method generates a sequence $\{x_n\}_{n=0}^{\infty}$ which converges to the solution $\overline{x}$ of $f(x) = 0$.*

Proof. We have

$$b_0 - a_0 = b - a,\ b_1 - a_1 = \frac{(b-a)}{2},\ b_2 - a_2 = \frac{\frac{(b-a)}{2}}{2}, \cdots,\ b_n - a_n = \frac{(b-a)}{2^n}.$$

Since the solution $\overline{x} \in [a_n, b_n]$ and $x_n = \frac{a_n + b_n}{2} \in [a_n, b_n]$, so we have $\overline{x} \in [a_n, b_n] = [a_n, x_n] \cup [x_n, b_n]$. This leads

$$|x_n - \overline{x}| \leq \frac{(b_n - a_n)}{2} = \frac{\left(\frac{(b-a)}{2^n}\right)}{2} = \frac{(b-a)}{2^{n+1}}.$$

By passing to the limit for $n \to \infty$, we get $x_n \to \overline{x}$. ■

Example 2.2.1 *Let $f(x) = x^3 + 4x^2 - 10$, $x \in [1, 2]$. Find an approximate value of the solution $\overline{x}$ of $f(x) = 0$ by the bisection method.*

Correction 2.2.1 *We have $f(x) = x^3 + 4x^2 - 10$ is continuous on $[1, 2]$ and $f(1) \times f(2) = (-5)(14) < 0$. So by theorem 2.2, there exists at least one $\overline{x} \in [1, 2]$ such that $f(\overline{x}) = 0$. Since f is an increasing function on $[1, 2]$, $\overline{x}$ is therefore the only solution of $f(x) = 0$. Let's look for an approximation of it.*

Applying the bisection algorithm, we get the following array of values:

n	a_n	b_n	x_n	$f(x_n)$
0	1.0	2.0	1.5	2.375
1	1.0	1.5	1.25	−1.79687
2	1.25	1.5	1.375	0.16211
3	1.25	1.375	1.3125	−0.84839
4	1.3125	1.375	1.34375	−0.35098
5	1.34375	1.375	1.359375	−0.09641
6	1.359375	1.375	1.36328125	−0.03215
7	1.359375	1.36328125	1.36328125	−0.03215
8	1.36328125	1.36328125	1.365234375	0.000072
9	1.36328125	1.365234375	1.364257813	−0.01605
10	1.364257813	1.365234375	1.364746094	−0.00799
11	1.364746094	1.365234375	1.364990235	−0.000396

After 12 iterations we can say that $x_{11} = 1.364990235$ approximates $\overline{x}$ with an error $|\overline{x} - x_{11}| \leq \frac{(b_{11} - a_{11})}{2} = 0.00024$.

Remark 2.2 *The disadvantage with this method is that the convergence is slow. One uses it to make start other more powerful methods (Newton, successive iterations).*

2.3 Method of successives approximations (of the fixed point)

This method of determining the approximate roots of $f(x) = 0$, $x \in [a, b]$ consists of

1. Replace the equation $f(x) = 0$ by an equivalent equation $x = g(x)$.
2. Create the sequence of numbers $\{x_n\}_{n=0}^{\infty}$ where $x_n = g(x_{n-1})$, $n = 1, 2, ...$ and x_0 is rough approximation of $\overline{x}$ solution of $f(x) = 0$ found, for example, by the dichotomy method.
3. Take x_n, when n is large enough, as an approximation to the root $\overline{x}$ of $f(x) = 0$.

Remark 2.3 *If the sequence $\{x_n\}_{n=0}^{\infty}$ converges where $x_n = g(x_{n-1})$ and g is continuous, then it must converge to a solution $\overline{x}$ of $x = g(x)$.*

Indeed

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} g(x_{n-1}) = g(\lim_{n\to\infty} x_{n-1}).$$

This means that

$$\xi = g(\xi), \quad \text{where} \quad \xi = \lim_{n\to\infty} x_n$$

that is, $\xi = \lim x_n$ is a root of $g(x) = x$.

Algorithm of successive approximations

Given the equation $x = g(x)$ and a rough approximation x_0, then to improve this one must proceed as follows:

- **Step 1:** Set $i = 1$.
- **Step 2:** Set $x_i = g(x_{i-1})$.

- **Step** 3**:** If x_i the approximation is satisfactory, go to step 5. Otherwise, go to step 4.
- **Step** 4**:** Add 1 to i and go to step 2.
- **Step** 5**:** End.

Remark 2.4 *The sequence generated by the fixed point method can diverge and therefore the n^{th} iterations cannot be a good approximation of the root $\overline{x}$, it can even be very different from that- this. So to make the application of this method possible, the function g must satisfy certain conditions that we are going to cite.*

Definition 2.1 *Let g be a functionn defined on $[a,b]$, then the point $\overline{x} \in [a,b]$ such that $g(\overline{x}) = \overline{x}$ is called fixed point of g.*

Theorem 2.4 *[2] Let $g : [a,b] \longrightarrow [a,b]$ be a differentiable function on $[a,b]$ such that $|g'(x)| \leq k < 1, \forall x \in [a,b]$. Then the sequence $\{x_n\}_{n=0}^{\infty}$ defined by $x_n = g(x_{n-1})$, $n = 1, 2, ...$ converges independently of the initial value x_0 converges to the unique fixed point $\overline{x}$ of the function g.*

Proof.

- 1^{st} **part of the proof** let us show that g has a fixed point $\overline{x} \in [a,b]$.

  Suppose that $g(a) \neq a$ and $g(b) \neq b$ (if $g(a) = a$ or $g(b) = b$ then a is a fixed point or b is a fixed point). So $g(a) > a$ and $g(b) < b$, because $g(a)$ and $g(b) \in [a,b]$. Let $h(x) = g(x) - x$ then the function h is continuous on $[a,b]$ and $h(a) = g(a) - a > 0$, and $h(b) = g(b) - b < 0$. So by the intermediate value theorem, there exists $\overline{x} \in [a,b]$ such that $h(\overline{x}) = 0$ i.e. $h(\overline{x}) = g(\overline{x}) - \overline{x} = 0$ therefore $\overline{x}$ is a fixed point of the function g.

  Let us now show that this point $\overline{x}$ is unique. To do this suppose there are two fixed points $\overline{x}_1$ and $\overline{x}_2$ of g with $\overline{x}_1 \neq \overline{x}_2$. Then by the mean theorem, there exist $\xi \in [a,b]$ such that

  $$\frac{g(\overline{x}_1) - g(\overline{x}_2)}{\overline{x}_1 - \overline{x}_2} = g'(\xi).$$

  This implies that

  $$|\overline{x}_1 - \overline{x}_2| = |g'(\xi)||\overline{x}_1 - \overline{x}_2|,$$

  so that

  $$|\overline{x}_1 - \overline{x}_2| \leq k|\overline{x}_1 - \overline{x}_2| < |\overline{x}_1 - \overline{x}_2|.$$

  This last inequality cannot be true, so $\overline{x}_1$ must be equal to $\overline{x}_2$.

- 2^{nd} **part of the proof** Let us show that the sequence $\{x_n\}_{n=0}^{\infty}$ with $x_n = g(x_{n-1})$, $n = 1, 2, ...$ converges to this fixed point $\overline{x}$ of the function g.

  Since $\overline{x} \in [a,b]$ and $x_n = g(x_{n-1}) \in [a,b]$, then by the mean value theorem we have

  $$|g(\overline{x}) - g(x_n)| = |\overline{x} - x_n||g'(\xi)|$$

  This implies that

  $$|\overline{x} - x_{n+1}| = |g'(\xi)||\overline{x} - x_n|$$

  or

  $$|\overline{x} - x_n| = |g'(\xi)||\overline{x} - x_{n-1}|$$

  where $\xi \in [a,b]$ and $n = 1, 2, ...$ this implies that

  $$|\overline{x} - x_n| \leq k|\overline{x} - x_{n-1}| \leq k^2|\overline{x} - x_{n-2}| \leq \cdots \leq k^n k|\overline{x} - x_0|.$$

  Since $|k| < 1$ we get $k^n \longrightarrow 0$ when $n \longrightarrow \infty$ and consequently $x_n \longrightarrow \overline{x}$ the fixed point of g.

■

Corollary 2.1 *If g satisfies the hypotheses of theorem 2.4, then*

$$|x_n - \overline{x}| \leq \frac{k^n}{1-k}|x_1 - x_0|, n \geq 1.$$

Proof. We have

$$g(x_n) - g(x_{n-1}) = x_{n+1} - x_n. \tag{2.3.1}$$

But according to the mean value theorem, there exist $\xi \in [a, b]$ such that

$$g(x_n) - g(x_{n-1}) = g'(\xi)(x_n - x_{n-1}). \tag{2.3.2}$$

From (2.3.1) and (2.3.2) we see that $x_{n+1} - x_n = g'(\xi)(x_n - x_{n-1})$, hence

$$|x_{n+1} - x_n| \leq k|x_n - x_{n-1}|,\ n = 1, 2, ...$$

so

$$|x_{n+1} - x_n| \leq k|x_n - x_{n-1}| \leq k^2|x_{n-1} - x_{n-2}| \leq ... \leq k^n|x_1 - x_0|.$$

Note that for $m > n \geq 1$, we have

$$|x_m - x_n| = |x_m - x_{m-1} + x_{m-1} - x_{m-2} + ... + x_{n+1} - x_n|.$$

By using triangular inequality, we get

$$|x_m - x_n| \leq |x_m - x_{m-1}| + |x_{m-1} - x_{m-2}| + ... + |x_{n+1} - x_n|,$$

we obtain

$$|x_m - x_n| \leq k^{m-1}|x_1 - x_0| + k^{m-2}|x_1 - x_0| + ... + k^n|x_1 - x_0|,$$

which implies that

$$|x_m - x_n| \leq k^n(1 + k + k^2 + ... + k^{m-1-n})|x_1 - x_0|.$$

When $m \longrightarrow \infty$ and $x_m \longrightarrow \overline{x}$ then

$$|\overline{x} - x_n| \leq \frac{k^n}{1-k}|x_1 - x_0|.$$

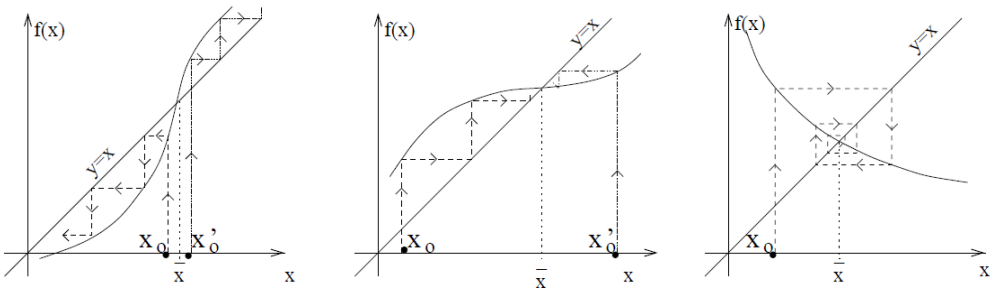
■

Remarks 2.1 *1. The last inequality allows us to estimate in advance the number of iterations to approximate $\overline{x}$ with a given precision ϵ.*

2. It also shows us that if $|k|$ less than 1 the method of successive approximations converges quickly.

3. Practically, we stop the iterative process as soon as the difference, in absolute value, of two approximations succeedssives x_n and x_{n-1} does not exceed a certain precision imposed in advance. That is $|x_n - x_{n-1}| \leq \epsilon$.

4. Geometrically, the method of successive approximations can be well understood by examining the following figures: This figure shows us how the sequence x_n generated by the method of



successive approximations tends the solution $\overline{x}$ of $g(x) = x$.

Example 2.3.1 *Let $f(x) = x^3 + 4x^2 - 10$, $x \in [1, 2]$. Find an approximate value of the solution $\overline{x}$ of $f(x) = 0$ by the fixed point method.*

Correction 2.3.1 *In the example 2.2.1, we have shown that $f(x) = x^3 + 4x^2 - 10 = 0$ has a single solution $\overline{x} \in [1, 2]$. The equation $f(x) = 0$ can be transformed into an equivalent equation $g(x) = x$, where $g(x) = \sqrt{\frac{10}{(x+4)}}$. In effect,*

$$\begin{aligned} x^3 + 4x^2 - 10 = 0 &\iff x^2(x+4) = 10 \\ &\iff x = \left(\frac{10}{(x+4)}\right)^{1/2} = g(x). \end{aligned}$$

Before applying the fixed point algorithm, let us ensure that the sequence $\{x_n\}_{n=0}^{\infty}$, where

$$x_n = g(x_{n-1}) = \left(\frac{10}{(x_{n-1}+4)}\right)^{1/2}, \; n = 1, 2, \ldots \text{ and } x_0 = 1.5$$

converges to the fixed point of $g(x)$ with $x \in [1, 2]$. We have $g'(x) = -\sqrt{10}/(4+x)^{3/2} < 0$ with $x \in [1, 2]$ this means g is decreasing over $[1, 2]$ and therefore $g(2) = 1.29 \leq g(x) \leq g(1) = 1.41$, $\forall x \in [1, 2]$. That is, $g(x) \in [1, 2]$, $\forall x \in [1, 2]$. Since g is continuous on $[1, 2]$ and $g'(x) = -\sqrt{10}/(4+x)^{3/2}$ implies that $|g'(x)| \leq 5/10(5)^{3/2} = 0.14$, $\forall x \in [1, 2]$, then g satisfies the conditions of theorem 2.4. Therefore $x_n = g(x_{n-1})$ converges to the fixed point $\overline{x}$ of $g(x)$ and $x_0 \in [1, 2]$ in particular $x_0 = 1.5$. For this choice we have

- 1^{st} ***iteration:*** $x_1 = g(1.5) = 1.3483$
- 2^{nd} ***iteration:*** $x_2 = g(x_1) = 1.3673$
- 3^{rd} ***iteration:*** $x_3 = g(x_2) = 1.3649$
- 4^{th} ***iteration:*** $x_4 = g(x_3) = 1.3652$
- 5^{th} ***iteration:*** $x_5 = g(x_4) = 1.3652$

Let's stop at this 5^{th} and accept it as a "good" approximation of the fixed point of g.

Remark 2.5 *We have $x^3 + 4x^2 - 10 = 0$ is also equivalent to $x = x - x^3 - 4x^2 + 10$. In this case $g(x) = x - x^3 - 4x^2 + 10$ implies that $g'(x) = 1 - 3x^2 - 8x$ comes $|g'(x)| > 1$, $\forall x \in [1, 2]$. In this case the theorem 2.4 does not converge the sequence $\{x_n = g(x_{n-1})\}_{n=1}^{\infty}$. With this choice of $g(x) = x - x^3 - 4x^2 + 10$, the fixed point algorithm generates a divergent sequence. Indeed*

$$\begin{aligned} x_1 = g(x_0) = g(1.5) &= -0.875, \\ x_2 = g(x_1) = g(-0.875) &= 6.73, \\ x_3 = g(x_2) = g(6.73) &= -469.26, \\ x_4 = g(x_3) = g(-469.26) &= 1.02 \times 10^8. \end{aligned}$$

2.4 Newton-Raphson method

This method is based on Taylor's theorem. Let x_n be the n^{th} iteration approximating the exact solution $\overline{x}$ of $f(x) = 0$, where $f \in C^2([a, b])$ and $f'(x_n) \neq 0$, then according to Taylor's formula we have

$$f(\overline{x}) = 0 = f(x_n) + (\overline{x} - x_n)f'(x_n) + (\overline{x} - x_n)^2 \frac{f''(\xi)}{2},$$

where ξ is a number between $\overline{x}$ and x_n. Assuming that x_n is very "near" to $\overline{x}$, this last equation gives us

$$\overline{x} \approx x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 0, 1, 2, \ldots$$

This sequence, if it converges, must converge to the solution $\overline{x}$ of $f(x) = 0$.

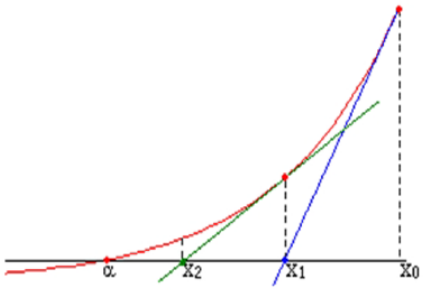
Newton-Raphson method algorithm

Let $f(x) = 0$ and x_0 be an initial approximation of the exact solution $\overline{x}$ of $f(x) = 0$

- **Step** 1**:** Set $i = 1$.
- **Step** 2**:** Set $x_i = x_{i-1} - \frac{f(x_{i-1})}{f'(x_{i-1})}$.
- **Step** 3**:** If x_i the approximation is satisfactory, go to step 5. Otherwise, go to step 4.
- **Step** 4**:** Add 1 to i and go to step 2.
- **Step** 5**:** End.

Remarks 2.2 *1. Practically, we stop the iterative process as soon as the difference, in absolute value, of two successive approximations x_n and x_{n-1} does not exceed a certain precession ϵ imposed in advance. That is $|x_n - x_{n-1}| \leq \epsilon$.*

2. Geometrically, the Newton-Raphson method is equivalent to replacing a small arc of the curve $y = f(x)$ by the tangent leading through a certain point $(x_0, f(x_0))$ of this curve.



The Newton-Raphson formula can be obtained by determining the point of intersection of the tangent at $(x_n, f(x_n))$, $n = 0, 1, 2, ...$ with the axis Ox. The equation for this is given by

$$y = f(x_n) + (x - x_n)f'(x_n). \tag{2.4.1}$$

Hence,

$$0 = f(x_n) + (x_{n+1} - x_n)f'(x_n) \Longrightarrow x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},\ n = 0, 1, ...$$

Theorem 2.5 *[2] If a function f defined on $[a, b]$ is such that*

1. *$f(a) \times f(b) < 0$.*
2. *$f'(x)$ and $f''(x)$ are nonzero and keep constant signs on $[a, b]$.*

Then the Newton-Raphson method generates a sequence which converges to the unique solution of $f(x) = 0$, starting from the approximation x_0 verifying $f(x_0) \times f''(x_0) > 0$.

Proof. Without loss of generality, suppose that $f(a), f(b) \in \mathbb{R}$, $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$, and $f''(x) > 0$, $x \in [a, b]$. With this assumption and the inequality $f'(x_0) \times f''(x_0) > 0$ we deduce that $f(x_0) > 0$. Choose $x_0 = b$ and show, by induction, that $x_n = \overline{x}$, $(n = 0, 1, 2, ...)$. We show this by induction.

If $x_0 = b > \overline{x}$. Assume that $x_n > \overline{x}$ and show that $x_{n+1} > \overline{x}$.

The exact solution $\overline{x}$ can be written as $\overline{x} = x_n + (\overline{x} - x_n)$. Applying Taylor's formula, we get

$$0 = f(\overline{x}) = f(x_n) + f'(x_n)(\overline{x} - x_n) + \frac{f''(\xi)}{2}(\overline{x} - x_n)^2, \text{ where } \overline{x} < \xi < x_n,$$

consequently,

$$f(\overline{x}) = f(x_n) + f'(x_n)(\overline{x} - x_n) = -\frac{f''(\xi)}{2}(\overline{x} - x_n)^2 < 0,$$

because $f''(x) > 0$, $\forall x \in [a, b]$ by hypothesis. So $f(x_n) + f'(x_n)(\overline{x} - x_n)$ is less than zero. This implies that $\overline{x} < x_n - \frac{f(x_n)}{f'(x_n)} = x_{n+1}$. Since by hypothesis $f'(x) > 0$, $\forall x \in [a, b]$ implies that f is increasing on $[a, b]$ and therefore $\overline{x} < x_n$ implies that $f(\overline{x}) < f(x_n)$. But $f(\overline{x}) = 0$ so $f(x_n) > 0$. Note that the sequence $\{x_n\}_{n=0}^{\infty}$ is decreasing. In effect,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} < x_n,$$

because $\frac{f(x_n)}{f'(x_n)} > 0$, so $\{x_n\}_{n=0}^{\infty}$ is a decreasing sequence bounded below, it must converge to a limit $\xi = \lim\limits_{n\to\infty} x_n$. Let us show that this limit ξ, is a solution of $f(x) = 0$. We have

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \Longrightarrow \lim_{n\to\infty} x_n = \lim_{n\to\infty} x_{n-1} - \frac{f(\lim\limits_{n\to\infty} x_{n-1})}{f'(\lim\limits_{n\to\infty} x_{n-1})},$$

we obtain

$$\xi = -\frac{f(\xi)}{f'(\xi)} + \xi \Longrightarrow \frac{f(\xi)}{f'(\xi)} = 0 \Longrightarrow f(\xi) = 0.$$

Finally ξ is a solution of $f(x) = 0$. ■

Example 2.4.1 *Let $f(x) = x^3 + 4x^2 - 10 = 0$, $x \in [1, 2]$. Using Newton's method, find an approximate value of the exact solution of $f(x) = 0$.*

Correction 2.4.1 *We have $f'(x) = 3x^2 + 8x$. So starting from $x_0 = 1.5$ and applying the Newton-Raphson formula, we get*

$$x_n = x_{n-1} - \frac{(x_{n-1}^3 + 4x_{n-1}^2 - 10)}{(3x_{n-1}^2 + 8x_{n-1})},\ n = 1, 2, ...$$

For $n = 1$ we have

$$x_1 = x_0 - \frac{(x_0^3 + 4x_0^2 - 10)}{(3x_0^2 + 8x_0)} = 1.5 - \frac{((1.5)^3 + 4(1.5)^2 - 10)}{(3(1.5)^2 + 8(1.5))} = 1.3733333.$$

The solution is given by $x_1 = 1.37333$.

2.5 Exercises with solutions

Exercise 2.1 *Let $f(x) = x^3 - x - 1$, $x \in \mathbb{R}$.*

1. *Check that f admits a separate root in $[1, 2]$.*
2. *Calculate the number of iterations so that the error made with the dichotomy method is less than 0.5×10^{-3}.*

Exercise 2.2 *Let f be the function defined by*

$$f(x) = \frac{1}{x} - 3,\ x \in \mathbb{R}.$$

(I) Show that

1. *If $f(\alpha) = 0$, then $\alpha = \frac{1}{3}$.*

2. *According to the Newton-Raphson formula we have the recurrence formula*

$$x_{n+1} = 2x_n - 3x_n^2. \tag{2.5.1}$$

(II) We assume that the Newton-Raphson algorithm converges

1. *Show that* $\lim_{n\longrightarrow\infty} x_n = \frac{1}{3}$.
2. *Show that the iterations given by 2.5.1 satisfy the equation*

$$\left(x_{n+1} - \frac{1}{3}\right) = -3\left(x_n - \frac{1}{3}\right)^2.$$

Exercise 2.3 *Let* $f(x) = x^3 + 3x - 1$ *and* $g(x) = x^3 + 3x - 3$ *with* $x \in \mathbb{R}$.

1. *Show that the equations* $f(x) = 0$ *and* $g(x) = 0$ *each admit a real root* α *(respectively* β*) in all* $\mathbb{R}$ *and* $\alpha, \beta \in]0, 1[$.
2. *Show that the sequences* $x_{n+1} = \frac{1}{x_n^2+3}$ *and* $y_{n+1} = \frac{3}{y_n^2+3}$ *converge respectively to* α, β *for all* $x_0, y_0 \in [0, 1]$.
3. *For* $x_0 = y_0 = 0.5$*, determine the sufficient number of iterations* n_0 *and* n_1 *to have* $|x_n - \alpha| \leq 10^{-6}$, $\forall n \geq n_0$ *and* $|y_n - \beta| \leq 10^{-6}$, $\forall n \geq n_1$.
4. *Explain the difference between* n_0 *and* n_1 *then compare with the number* n_2 *of iterations sufficient by dichotomy to have the same precision.*
5. *Calculate an approximation of* α *to within* 10^{-6} *by the Newton-Raphson algorithm* $z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}$ *starting at* $z_0 = 0.5$ *and compare with* x_6.

Exercise 2.4 *Let* $f(x) = xe^x - 2x$ *and* $x \in \mathbb{R}$.

1. *Show that the function* f *admits a unique minimum* α *in the interval* $[0, 1]$.
2. *Show that the Newton-Raphson algorithm makes it possible to determine this minimum* α*. Write the Newton-Raphson sequence* x_n.
3. *We pose* $x_0 = 1$*, calculate* x_5.

Exercise 2.5 *Let* f *be the function defined for* $x > 0$ *by* $f(x) = \ln(x) + x^2 + 2x - 5$.

1. *Show that the equation* $f(x) = 0$ *admits a root* α *and only one that we will locate between two consecutive integers* a *and* $a + 1$.
2. *Show that the iteration* $x_{n+1} = x_n - f(x_n)$; $x_0 \in [a, a + 1]$ *does not converge to* α.
3. *For which values of* λ*, the iteration* $x_{n+1} = x_n - \lambda f(x_n)$; $x_0 \in [a, a + Does1]$ *converge to* α*?. Determine the value* λ_0 *of* λ *for which* $\min_{x\in[a,a+1]} |1 - \lambda f'(x)|$ *is minimal.*
4. *The iteration* $x_{n+1} = \frac{5-\ln(x_n)}{x_n+2}$; $x_0 \in [a, a + 1]$ *does it converge to* α*?.*
5. *Write the Newton-Raphson algorithm applied to* $f(x) = 0$*. For which values of* x_0 *does it converge to* α*?.*

2.5.1 Solutions

Solution 2.1 *We have* $f(x) = x^3 - x - 1$, $x \in \mathbb{R}$.

1. *Check that* f *admits a separate root in* $[1, 2]$*:*
   (a) f *is continuous in* $[1, 2]$ *since* f *is a polynomial.*
   (b) f *is increasing (monotonic) in* $[1, 2]$ *because*

$$f'(x) = 3x^2 - 1 > 0 \Longrightarrow x^2 > \frac{1}{3} \Longrightarrow x \in]-\infty, -\frac{1}{\sqrt{3}}[\cup] + \frac{1}{\sqrt{3}}, +\infty[.$$

(c) $f(1)f(2) < 0$.

2. *We have* $|x_n - \alpha| \leq \frac{b-a}{2^{n+1}} \leq \epsilon$, *so that*

$$\begin{aligned}
\frac{1}{2^{n+1}} \leq \epsilon &\iff -(n+1)\ln(2) \leq \ln(\epsilon) \\
&\iff -(n+1) \leq \frac{\ln(\epsilon)}{\ln(2)} \\
&\iff -n \leq \frac{\ln(\epsilon)}{\ln(2)} + 1 \\
&\iff n \geq -\frac{\ln(\epsilon)}{\ln(2)} - 1.
\end{aligned}$$

We take

$$\begin{aligned}
n_0 &= E\left(-\frac{\ln(\epsilon)}{\ln(2)} - 1\right) + 1 \\
&= E\left(-\frac{\ln(0.5 \times 10^{-3})}{\ln(2)} - 1\right) + 1 \\
&= E\left(\frac{\ln(2) + 3\ln(10)}{\ln(2)} - 1\right) + 1 \\
&= E\left(\frac{3\ln(2) + 3\ln(5)}{\ln(2)}\right) + 1 \\
&= E\left(3 + 3\frac{\ln(5)}{\ln(2)}\right) + 1.
\end{aligned}$$

Solution 2.2 *We have*

$$f(x) = \frac{1}{x} - 3,\ x \in \mathbb{R}.$$

(I) Shows that

1. *If* $f(\alpha) = 0$, *then* $\alpha = \frac{1}{3}$.

$$f(\alpha) = 0 \Longrightarrow \frac{1}{\alpha} - 3 = 0 \Longrightarrow \frac{1}{\alpha} = 3 \Longrightarrow \alpha = \frac{1}{3}.$$

2. *We have, according to the Newton-Raphson formula*

$$\begin{aligned}
x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \\
&= x_n - \frac{\left(\frac{1}{x_n} - 3\right)}{\left(-\frac{1}{x_n^2}\right)} \\
&= x_n + x_n - 3x_n^2
\end{aligned}$$

(II) Assuming that the Newton-Raphson algorithm converges

1. *We will show that* $\lim_{n\longrightarrow\infty} x_n = \frac{1}{3}$. *We have* x_n *converges* $\iff \lim_{n\longrightarrow\infty} x_n = \lim_{n\longrightarrow\infty} x_{n+1} = l$, *so*

$$\begin{aligned}
l = 2l - 3l^2 &\iff 3l^2 - l = 0 \\
&\iff l(3l-1) = 0 \\
&\iff l = 0,\ \textit{(impossible)};\ l = \frac{1}{3}
\end{aligned}$$

and as $f(\alpha) = 0$ *for* $\alpha = \frac{1}{3}$ *then* $\lim_{n\longrightarrow\infty} x_n = \frac{1}{3}$.

2. *We will show that the iterations given by 2.5.1 satisfy the equation*

$$\left(x_{n+1}-\frac{1}{3}\right)=-3\left(x_n-\frac{1}{3}\right)^2.$$

According to 2.5.1, we have $x_{n+1}=2x_n-3x_n^2$ *so*

$$\begin{aligned} x_{n+1}-\frac{1}{3} &= 2x_n-3x_n^2-\frac{1}{3} \\ &= -3\left(x_n^2-\frac{2}{3}x_n+\frac{1}{9}\right) \\ &= -3\left(x_n-\frac{1}{3}\right)^2. \end{aligned}$$

Solution 2.3 *We have* $f(x)=x^3+3x-1$ *and* $g(x)=x^3+3x-3$ *with* $x\in\mathbb{R}$.

1. *We have* f *and* g *are continuous strictly increasing on* $\mathbb{R}$ *because* $f'(x)=g'(x)=3(x^2+1)>0,\ \forall x\in\mathbb{R}$. *Moreover* $f(0)f(1)=g(0)g(1)=-3<0$ *so there are unique* $\alpha,\beta\in\mathbb{R}$ *such that* $f(\alpha)=0$ *and* $g(\beta)=0$, *furthermore* $\alpha,\beta\in]0,1[$.

2. *The functions* $\phi(x)=\frac{1}{x^2+3}$ *and* $\psi(x)=\frac{3}{x^2+3}$ *verify*

$$\begin{aligned} f(x)=0 &\Longleftrightarrow x=\phi(x),\ \forall x\in[0,1], \\ g(x)=0 &\Longleftrightarrow x=\psi(x),\ \forall x\in[0,1], \end{aligned}$$

and we have $\psi(x)=3\phi(x)$ *on* $[0,1]$, ϕ *and* ψ *are continuous and increasing on* $[0,1]$, *so*

$$\phi([0,1])=([\phi(1),\phi(0)])=\left[\frac{1}{4},\frac{1}{3}\right]\subset[0,1],$$

and

$$\psi([0,1])=([\psi(1),\psi(0)])=\left[\frac{3}{4},1\right]\subset[0,1].$$

On the other hand

$$\phi'(x)=-\frac{2x}{(x^2+3)^2},\ |\phi''(x)|=\frac{6(1-x^2)}{(x^2+3)^3}\geq 0,\ \forall x\in[0,1],\ \text{ so } \sup_{x\in[0,1]}|\phi'(x)|=|\phi'(1)|=\frac{1}{8},$$

$$\psi'(x)=-\frac{6x}{(x^2+3)^2},\ |\psi''(x)|=\frac{18(1-x^2)}{(x^2+3)^3}\geq 0,\ \forall x\in[0,1],\ \text{ so } \sup_{x\in[0,1]}|\psi'(x)|=|\psi'(1)|=\frac{3}{8}.$$

According to the fixed point theorem, the sequences $x_{n+1}=\phi(x_n)$ *and* $y_{n+1}=\psi(y_n)$ *respectively converge to* α *and* β *for all* x_0 *and* $y_0\in[0,1]$. *For the calculation of* x_n *see table* 1.

3. *Since* $|x_n-\alpha|\leq\frac{L^n}{1-L}|x_1-x_0|\leq\epsilon$ *requires*

$$n\geq\frac{\ln\left(\dfrac{(1-L)\epsilon}{|x_1-x_0|}\right)}{\ln(L)}.$$

For $\epsilon=10^{-6}$, $L=\frac{1}{8}$, $x_0=0.5$, $x_1=\frac{1}{3.25}$, *we find* $n_0=6$.
For $\epsilon=10^{-6}$, $L=\frac{3}{8}$, $y_0=0.5$, $y_1=\frac{1}{3.25}$, *we find* $n_1=14$.

By dichotomy $|x_n-\alpha|\leq\frac{(b-a)}{2^n}\leq\epsilon$ *requires* $n\geq\frac{\ln\left(\frac{b-a}{\epsilon}\right)}{\ln(2)}$ *so that* $a=0$, $b=1$, $\epsilon=10^{-6}$ *we find* $n_2=20$.

4. *For* $x_0 = y_0 = 0.5$ *and* $\epsilon = 10^{-6}$*, we have* $L_1 = \frac{1}{8} \leq L_1 = \frac{3}{8} \leq \frac{1}{2}$ *therefore* $n_0 = 6 < n_1 = 14 < n_2 = 20$ *even if ,in the case ,* (x_n) *and* (y_n) *do not converge to the same root.*

5. *We have* $z_{n+1} = z_n - \dfrac{f(z_n)}{f'(z_n)} = \dfrac{2z_n^3 + 1}{3(z_n^2 + 1)}$; $z_0 = 0.5$ *gives the* z_n *of the table* 2 *to compare with those of the table* 1.

n	x_n
0	0.5
1	0.30769230
2	0.32313575
3	0.32212170
4	0.32218961
5	0.32218506
6	0.32218537

n	z_n
0	0.5
1	0.33333333
2	0.3222222
3	0.32218535
4	0.32218533
5	0.32218533
6	0.32218533

On the left is the table 1 and on the right is the table 2.

Solution 2.4 1. *We have* $f'(x) = (x+1)e^x - 2$*, for all* x *in* $\mathbb{R}$*,* $f'(0) = -1 < 0$*,* $f'(1) = 2e^1 - 2 > 0$ *and* $f''(x) = (x+2)e^x > 0$*,* $\forall x \in [0,1]$*, so, by the intermediate value theorem and strict monotonicity,* $f(x) = 0$ *has a unique root and only one* $\alpha \in]0,1[$.

Moreover, f *is increasing on* $[0,1]$*, so* $f(x) < f(\alpha) = 0$ *and* $f(x) > f(\alpha) = 0$*, for all* x *in* $]0,1]$*, so* $f(\alpha) = 0$ *then* f *is decreasing on* $[0,\alpha[$ *and increasing on* $]\alpha,1]$*. It follows that* $x = \alpha$ *is a minimum of* f *over* $[0,1]$ *and it is unique.*

2. *We have* $f'(x) = (x+2)e^x \neq 0$*,* $\forall x \in [0,1]$*, the algorithm* $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$; $x_0 \in [0,1]$ *is well defined, so it allows us to determine* α *for well-chosen* x_0 *(local convergence). Here it converges to* α *for all* $x_0 \in [0,1]$ *because* $f \in C^2([0,1])$*,* $f(0) \times f(1) < 0$*,* $f'(x) > 0$*, on* $[0,1]$*,* $f'''(x) = (x+3)e^x > 0$*,* $\forall x \in [0,1]$ *and* $c = 0$ *such that*

$$|f(c)| = \min\{|f'(0)|, |f'(1)|\},$$

$\frac{|f(c)|}{|f'(c)|} \leq b - a \Longleftrightarrow \frac{1}{2} \leq 1$.

3. *We have*

$$\begin{cases} x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = \frac{x_n^2 + x_n - 1 + 2e^{-x_n}}{x_n + 2}, \\ x_0 = 1. \end{cases} \tag{2.5.2}$$

gives the x_n *of the table*

n	x_n
0	1
1	0.57858629
2	0.40127775
3	0.37531258
4	0.37482269
5	0.37482252
6	0.37482252

Solution 2.5 1. *As* f *is continuous for* $x > 0$ *and* $f(1) = -1 < 0$*,* $f(2) = \ln(2) + 3$ *then there exist* $\alpha \in]1,2[$*:* $f(\alpha) = 0$*. On the other hand,* $f'(x) = \frac{1}{x} + 2x + 2 > 0$*,* $\forall x \in [1,2]$ *so* α *is unique.*

2. *Let* $\phi(x) = x - f(x) = 5 - x - x^2 - \ln(x)$ *hence* $\phi'(x) = 1 - 2x - \frac{1}{x}$ *and* $|\phi'|'(x) = 2 - \frac{1}{x^2} > 0$, $\forall x \in [1,2]$ *therefore* $\min_{x\in[1,2]} |\phi'(x)| = |\phi'(1)| = 4$ *because* $|\phi'|'$ *is increasing, hence*

$$|x_{n+1} - \alpha| = |\phi(x_n) - \phi(\alpha)| = |\phi'(\xi_n)| \cdot |x_n - \alpha| = 4|x_n - \alpha|.$$

then

$$|x_n - \alpha| \geq 4^n |x_0 - \alpha| \longrightarrow +\infty \text{ if } x_0 \neq \alpha, \text{ so } x_n \nrightarrow \alpha,\ n \longrightarrow \infty$$

3. *We have* $f'(x) = \frac{1}{x} + 2x + 2 > 0$, $f''(x) = -\frac{1}{x^2} + 2$ *on* $[1,2]$ *and*

$$M = \max_{x\in[1,2]} f'(x) = f'(2) = \frac{13}{2},\ m = \min_{x\in[1,2]} f'(x) = f'(1) = 5,$$

so that

$$\max_{x\in[1,2]} |1 - \lambda f'(x)| = \max\{|1 - \lambda m|, |1 - \lambda M|\},\ \text{if} \lambda > 0$$

for $\lambda < 0$, $1 - \lambda f' > 1$ *and* $x_n \nrightarrow \alpha$ *as in reponse of question 2. Effect*

$$m \leq f' \leq M \Longrightarrow \lambda m \leq \lambda f' \leq \lambda M,$$

and

$$\lambda > 0 \Longrightarrow 1 - \lambda M \leq 1 - \lambda f' \leq 1 - \lambda m,$$

moreover

$$|1 - \lambda M| < 1 \Longleftrightarrow \quad -1 < \tfrac{13}{2}\lambda - 1 < 1 \quad \Longleftrightarrow 0 < \lambda < \frac{4}{13}$$

and

$$|1 - \lambda m| < 1 \Longleftrightarrow \quad -1 < 1 - 5\lambda < 1 \quad \Longleftrightarrow 0 < \lambda < \frac{2}{5}$$

so $\max_{x\in[1,2]} |1 - \lambda f'| < 1$ *if* $0 < \lambda < \frac{4}{13}$ *and for these values of* λ, *where* $\max_{x\in[1,2]} |1 - \lambda f'|$ *is minimal at* $\lambda_0 \in]0, \frac{4}{13}[$, *when* $1 - \lambda_0 M = \lambda_0 m - 1$ *(examine the graph of* $\lambda \longrightarrow |1 - \lambda M|$ *and* $\lambda \longrightarrow |1 - \lambda m|$*).*

4. *We set* $\phi(x) = \frac{5 - \ln(x)}{x+2}$ *so* $\phi'(x) = \frac{\ln(x) - \frac{2}{x} - 6}{(x+2)^2} < 0$, $\forall x \in [1,2]$. *Because the numerator* $\psi(x) = \ln(x) - \frac{2}{x} - 6$ *is increasing and* $\psi(1) = -8 < 0$ *and* $\psi(2) = \ln(2) - 7 < 0$ *therefore* $\psi(x) < 0$, $\forall x \in [1,2]$ *and then*

$$\begin{aligned} |\phi'|(x) &= \frac{6 + \frac{2}{x} - \ln(x)}{(x+2)^2},\ \forall x \in [1,2], \\ |\phi'|'(x) &= -\frac{\frac{(x+2)^2}{x^2} + 2\big(6 + \frac{2}{x}\big) - \ln(x)}{(x+2)^3} < 0,\ \forall x \in [1,2], \end{aligned}$$

so $\sup_{x\in[1,2]} |\phi'(x)| = |\phi'(1)| = \frac{8}{9} <$*1. On the other hand,*

$$\phi([1,2]) = [\phi(2), \phi(1)] = \left[\frac{5 - \ln(2)}{4}, \frac{5}{3}\right] = [1.075..., 1.666...] \subset [1,2].$$

By the fixed point theorem, the sequence $x_{n+1} = \phi(x_n)$; $x_0 \in [1,2]$ *converges to the root* α *of* $f(x) = 0$ *because* $x = \phi(x) \Longleftrightarrow f(x) = 0$, $\forall x \in [1,2]$.

5. *As* $f'(x) = \frac{1}{x} + 2x + 2 > 0$, $\forall x \in [1,2]$ *we have*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = \frac{6 + x_n^2 - \ln(x_n)}{2 + 2x_n + \dfrac{1}{x_n}},\ n \geq 0,\ x_0 \in [1,2].$$

This algorithm converges for all $x_0 \in [1,2]$ *because*

- $f \in C^2([1,2])$.
- $f'(x) \neq 0$, $\forall x \in [1,2]$.
- $f''(x) = 2 - \frac{1}{x^2} \geq 0$, $\forall x \in [1,2]$.
- $f(1) \times f(2) < 0$.
- $f'(1) = 5$, $f'(2) = \frac{13}{2}$ *so* $c = 1$, *where* $|f'(c) = \min\{|f'(1)|, |f'(2)|\}$ *and* $\left|\frac{f(c)}{f'(c)}\right| \leq b - a \Longleftrightarrow \frac{2}{5} \leq 1$.

The global convergence theorem of the Newton-Raphson algorithm implies that $\lim_{n\to\infty}(x_n) = \alpha$, $\forall x_0 \in [1,2]$.

2.6 Exercises without solutions

Exercise 2.1 *Consider in* $\mathbb{R}$ *the equation* $f(x) = x^3 - 4x + 1 = 0$.

1. *Show that the equation* $f(x) = 0$ *has a unique solution in* $[1,3]$.
2. *Find the number of iterations necessary to obtain the approximate solution of the equation* $f(x) = 0$ *to within* 0.1 *with the dichotomy method on* $[1,3]$.
3. *Find the approximate solution of the equation* $f(x) = 0$ *up to* 0.1 *with the dichotomy method on* $[1,3]$.

Exercise 2.2 *We consider in the interval* $[1,2]$ *the equation* $\phi(x) = e^{x^2} - 3 = 0$.

1. *Show that* ϕ', ϕ'' *keep constant signs in* $[1,2]$.
2. *Show that* $x_0 = 2$, *the Newton-Raphson method applied to the equation* $\phi(x) = 0$ *converges.*
3. *Starting from the initial approximation* $x_0 = 2$, *find the approximate solution of the equation* $\phi(x) = 0$ *to within* 0.2 *using the stop test* $|x_{n+1} - x_n| \leq \epsilon$ *with the Newton-Raphson method.*

Exercise 2.3 *Let* f *be the function defined on* $\mathbb{R}_+$ *by*

$$\phi(x) = \frac{x}{1 - \ln(x)}, \forall x > 0, \phi(0) = 0, \text{ and } \alpha = e^{(1-\sqrt{5})/2}.$$

1. *Is the fixed point theorem applicable to the function* ϕ *on* $[0,\alpha]$*?*
2. *Show that the iteration* $x_{n+1} = \phi(x_n)$ *converges to* $l = 0$ *for all* $x_0 \in [0,\alpha]$

Exercise 2.4 *We are looking for the real roots of the equation* $x^2 = \ln(1+x)$.

1. *Show that the function* $F(x) = x^2 - \ln(1+x)$ *has two roots, one obvious that we will give, and the other that we note* $\overline{x}$ *(which we want to approximate in the following).*
2. *Locate* $\overline{x}$ *in a range of length* $1/4$.
3. *Write Newton's method relative* x_0. *Give a choice of the condition initial of the iterations of Newton* x_0 *which ensures the convergence of the process.*
4. *Consider the following successive approximation methods* (a) $x_{n+1} = \sqrt{\ln(1+x_n)}$ *and* (b) $x_{n+1} = e^{x_n^2} - 1$ *Specify whether they converge or diverge. In case of convergence indicate a choice of the initial condition* x_0.

Exercise 2.5 *Let the function* $F(x) = x^4 - 2x^3 + 1$, *we propose to find the real roots of* F *by the method of successive approximations (fixed point method that we will apply correctly), then by Newton's method.*

1. *(a) Show that* F *has two real roots, one of which is obvious. will give, the other* $\overline{x}$ *that we will locate in an interval* I *of length* $1/2$.

(b) *Study the convergence to* $\overline{x}$*, of the following three iterative methods* $x_0 \in I$ *given,*

$$x_{n+1} = x_n^4 - 2x_n^3 + x_n + 1;$$
$$x_{n+1} = \frac{-1}{(x_n - 2)^{1/3}};$$
$$x_{n+1} = 2 - \frac{1}{x_n^3}.$$

(c) *If one of these methods converges use it to determine* $\overline{x}$ *to within* 10^{-2}.

2. (a) *Write Newton's method for* F.

(b) *Check the global convergence theorem of this method on an interval that you will give.*

(c) *Give an explicit value of* x_0 *which ensures the convergence of the Newton's method to* $\overline{x}$.

Chapter 3

Solving linear systems

3.1 Introduction

Linear systems, $AX = b$ with $A \in M_n(\mathbb{R})$ and $X, b \in \mathbb{R}^n$, often arise in mathematical problems; solving boundary problems, physics and engineering.

Let us determine, for example, the potentials at each node of the electrical circuit above.

Solution: By applying Kirchoff's law, we get

- **Node 1:** $\frac{(v_1-v_a)}{3} + \frac{(v_1-v_2)}{5} + \frac{(v_1-v_b)}{6} + \frac{(v_1-v_b)}{2} = 0,$
- **Node 2:** $\frac{(v_2-v_1)}{5} + \frac{(v_2-v_3)}{10} + \frac{(v_2-v_4)}{1} = 0,$
- **Node 3:** $\frac{(v_3-v_2)}{10} + \frac{(v_3-v_4)}{7} = 0,$
- **Node 4:** $\frac{(v_4-v_2)}{1} + \frac{(v_4-v_3)}{7} + \frac{(v_4-v_b)}{4} = 0,$

hence

$$\begin{cases} 36v_1 - 6v_2 & = & 10v_a + 20v_b \\ -2v_1 + 13v_2 - v_3 - 10v_4 & = & 0 \\ -7v_2 + 17v_3 - 10v_4 & = & 0 \\ -28v_2 + 39v_4 - 4v_3 & = & 7v_b \end{cases}$$

this system can be put in the following matrix form

$$AX = b \Longleftrightarrow \begin{pmatrix} 36 & -6 & 0 & 0 \\ -2 & 13 & -1 & -10 \\ 0 & -7 & 17 & -10 \\ 0 & -28 & -4 & 39 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} 10v_a + 20v_b \\ 0 \\ 0 \\ 7v_b \end{pmatrix}.$$

In this course, we will see two main classes of methods for solving linear systems (these methods are classified into three main categories)

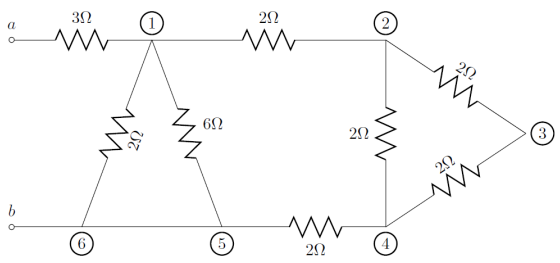


Figure 3.1: Part of electrical circuit

1. Direct methods (method of Cramer, Gauss, Cholesky, Crout, Doulittle, ...).
2. Iterative methods (Jacobi, Gauss-Seidel, relaxation methods).

3.2 Generals about matrices

The set of all matrices with m rows n columns with coefficients in field $\mathbb{K}$ ($\mathbb{R}$ or $\mathbb{C}$) is a vector space noted $M_{m,n}(\mathbb{K})$ of dimension $m \times n$. Let $A \in M_{m,n}(\mathbb{K})$

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \qquad a_{11}, a_{12}, \cdots, a_{mn} \in \mathbb{K}.$$

If $m = n$, we say that A is a square matrix

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

When $a_{ij} = 0$ for all $i, j = 1, ..., n$ we say that A is the null matrix, if

$$A = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \qquad a_{ij} = 0, \qquad \forall i, j.$$

Denote by I_n the neutral matrix defined as

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \qquad \text{equivalently,} \qquad a_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

- Let $A \in M_n(\mathbb{R})$ be such that $A = (a_{ij})_{1 \leq i \leq n; 1 \leq j \leq n}$, then A^t is a transpose matrix of A is defined $A^t =^t A = (a_{ji})_{1 \leq i \leq n; 1 \leq j \leq n}$ as

$$A^t = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix}.$$

Example 3.2.1 *Let*

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

The transpose matrix of A is

$$A^t = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}.$$

- A is a symmetric matrix if $A^t = A$ (i.e. if $a_{ij} = a_{ji}$).

Example 3.2.2 *Let*

$$A = \begin{pmatrix} \sqrt{2} & 1 & 7 \\ 1 & \sqrt{2} & 2 \\ 7 & 2 & \sqrt{3} \end{pmatrix},$$

then

$$A^t = \begin{pmatrix} \sqrt{2} & 1 & 7 \\ 1 & \sqrt{2} & 2 \\ 7 & 2 & \sqrt{3} \end{pmatrix}.$$

We see that $A = A^t$, this means that A is symmetric.

- A is an antisymmetric matrix if: $A^t = -A$.
- A is an invertible matrix if there exists a matrix, denoted by A^{-1}, such that $AA^{-1} = A^{-1}A = I_n$.
- A is said to be diagonal if $a_{ii} \neq 0$ and $a_{ij} = 0 \qquad (i \neq j)$

$$A = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}.$$

Example 3.2.3

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

- L is called lower triangular if $l_{ij} = 0$ for $j > i$.

$$L = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix}.$$

- U is said to be upper triangular if $u_{ij} = 0$ for $i > j$.

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & l_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix}.$$

- Trace of A, denoted by $\mathrm{tr}(A)$, is the sum of all diagonal elements that is $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$.
- Determinant of A denoted $\det(A)$

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

is defined as

$$\det(A) = ad - bc$$

- A is invertible if and only if $\mathrm{A} \neq 0$.

3.2.1 Cramer's method

Let $A \in M_{n,n}(\mathbb{R})$, $X, b \in \mathbb{R}^n$. We consider the linear system $AX = b$ such that $X^t = (x_1, x_2, \cdots, x_n)$ and $b^t = (b_1, b_2, \cdots, b_n)$.

$$AX = b \Longleftrightarrow \left\{ \begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & b_2 \\ \vdots & \vdots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & = & b_n \end{array} \right. \tag{3.2.1}$$

If $\det(A) \neq 0$, then this system admits a unique solution such that

$$x_i = \frac{\det(A_i)}{\det(A)} = \frac{\begin{vmatrix} a_{11} & \cdots & b_1 & \cdots & a_{1n} \\ a_{21} & \cdots & b_2 & \cdots & a_{2n} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \cdots & b_n & \cdots & a_{nn} \end{vmatrix}}{\begin{vmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2i} & \cdots & a_{2n} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \cdots & a_{ni} & \cdots & a_{nn} \end{vmatrix}}; \quad i = 1, ..., n.$$

Example 3.2.4 *We consider the following system*

$$\left\{ \begin{array}{l} 2x + 5y = 1 \\ -3x + y = 2 \end{array} \right. \Longleftrightarrow AX \begin{pmatrix} 2 & 5 \\ -3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

This system has a unique solution since $\det(A) \neq 0$. *By using the Cramer's method, we get*

$$\left\{ \begin{array}{l} x = \dfrac{\begin{vmatrix} 1 & 5 \\ 2 & 1 \end{vmatrix}}{\begin{vmatrix} 2 & 5 \\ -3 & 1 \end{vmatrix}} = \dfrac{1 - 10}{2 + 15} = \dfrac{-9}{17}. \\ \\ y = \dfrac{\begin{vmatrix} 2 & 1 \\ -3 & 2 \end{vmatrix}}{\begin{vmatrix} 2 & 5 \\ -3 & 1 \end{vmatrix}} = \dfrac{4 + 3}{2 + 15} = \dfrac{7}{17}. \end{array} \right.$$

Remark 3.1 *If* $A \in M_n(\mathbb{R})$ *and* $n > 5$ *solving system 3.2.1 by Cramer's method is very difficult.*

3.3 Direct Methods

3.3.1 Ascent method

We propose to solve the matrix equation

$$AX = b,$$

Assume that the matrix A is invertible. When A is an upper (or lower) triangular matrix, solving the system is immediate

$$AX = b \quad \Longleftrightarrow \quad \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

$$\Longleftrightarrow \quad \left\{ \begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots\cdots + a_{1n}x_n & = & b_1 \\ a_{22}x_2 + \cdots\cdots + a_{2n}x_n & = & b_2 \\ \ddots \quad \vdots & \vdots & \vdots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n & = & b_{n-1} \\ a_{nn}x_n & = & b_n. \end{array} \right.$$

We successively calculate x_n from the last equation, then x_{n-1} from the penultimate and so on. We obtain,

$$\left\{\begin{array}{lcl} x_n & = & b_n/a_{nn} \\ x_{n-1} & = & (b_n - a_{n-1,n}x_n)/a_{n-1,n-1} \\ & \vdots & \\ x_1 & = & (b_1 - a_{12}x_2 - \cdots - a_{1n}x_n)/a_{11}. \end{array}\right.$$

The lift method extends to triangular matrices by blocks

$$\begin{array}{rcl} x_n & = & b_n/a_{nn} \\ x_i & = & (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii}, \qquad i = n-1, ..., 1. \end{array}$$

Example 3.3.1 *Let the system*

$$\begin{array}{rcl} AX = b & \Longleftrightarrow & \left(\begin{array}{ccc} 2 & 1 & 3 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{array}\right)\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) = \left(\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right) \\ & \Longleftrightarrow & \left\{\begin{array}{rcl} 2x_1 + x_2 + 3x_3 & = & 1 \\ 2x_2 - x_3 & = & 1 \\ 3x_3 & = & 1 \end{array}\right. \end{array}$$

we get,

$$\left\{\begin{array}{l} x_3 = \dfrac{1}{3}, \\ x_2 = (1 + x_3)/2 = (1 + 1/3)/2 = \dfrac{4}{6} = \dfrac{2}{3}, \\ x_1 = (1 - x_2 - 3x_3)/3 = (1 - 2/3 - 3/3)/3 = -\dfrac{2}{9}. \end{array}\right.$$

Remark 3.2 *The problem is then to construct a triangular matrix by change of base.*

3.3.2 Elimination of Gauss

Gauss elimination ($n = 3$): For the clarity of the explanations, we apply the Gaussian method in the case when $n = 3$. Assume that $a_{11} \neq 0$.

$$\left\{\begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 & : & L_1^{(1)} \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 & : & L_2^{(1)} \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 & : & L_3^{(1)}. \end{array}\right.$$

In order to eliminate the term $a_{21}x_1$ in the line $L_2^{(1)}$ thanks to the combination $L_2^{(1)} - \dfrac{a_{21}}{a_{11}}L_1^{(1)}$ and to eliminate the term $a_{31}x_1$, combination $L_3^{(1)} - \dfrac{a_{31}}{a_{11}}L_1^{(1)}$ which gives the system $A^{(2)}X = b^2$, hance

$$\left\{\begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 & : & L_1^{(2)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 = b_2^{(2)} & : & L_2^{(2)} \\ a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 = b_3^{(2)} & : & L_3^{(2)}. \end{array}\right.$$

where $a_{ij}^{(2)} = a_{ij} - \dfrac{a_{i1}}{a_{11}}a_{1j}$, $j = 2.3$ and $b_j^{(2)} = b_j - \dfrac{a_{i1}}{a_{11}}b_i$.
Then, assuming that $a_{22} \neq 0$, we eliminate the term $a_{32}^{(2)}x_2$ in $L_3^{(2)}$ with the combination $L_3^{(2)} - \dfrac{a_{32}^{(2)}}{a_{22}^{(2)}}L_2^{(2)}$, which gives the system $A^{(3)}X = b^3$ equivalent to the initial system

$$\left\{\begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 & : & L_1^{(2)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 = b_2^{(2)} & : & L_2^{(2)} \\ a_{33}^{(3)}x_3 = b_3^{(3)} & : & L_3^{(3)}. \end{array}\right.$$

where $a_{33}^{(3)} = a_{33}^{(2)} - \dfrac{a_{32}^{(2)}}{a_{22}^{(2)}} a_{23}^{(2)}$ and $b_3^{(3)} = b_3^{(2)} - \dfrac{a_{32}^{(2)}}{a_{22}^{(2)}} b_2^{(2)}$. The resulting system is therefore triangular.

Gauss elimination ($n \in \mathbb{N}^\star$): The algorithm consists in replacing at each step the matrix A by a matrix $A^{(k)}$ whose $k-$th first column vectors correspond to the beginning of a triangular matrix. at $(n+1)-$th first columns of $A^{(k)}$

$$AX = b \sim A^{(1)}X = b^{(1)} \sim \cdots \sim A^{(n)}X = b^{(n)}.$$

such that $A^{(n)}$ is an upper triangular matrix

$$\begin{array}{l}
\text{For} \quad k = 1, \cdots, n-1 \\
\text{For} \quad i = k+1, \cdots, n \\
\left.\begin{array}{l} m_{ik} = \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} + m_{ik} a_{kj}^{(k)} \end{array}\right\} a_{ik}^{(k+1)} = 0, \qquad i = k+1, \cdots, n \\
\text{For} \quad j = k+1, \cdots, n \\
\qquad b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)} \\
\text{end} \\
\text{end} \\
\text{end}
\end{array}$$

Example 3.3.2 *Consider the system of equations*

$$\left\{\begin{array}{lcl} x_1 + 2x_2 + 3x_3 = 0 & : & L_1^{(1)} \\ 4x_1 + 5x_2 + 6x_3 = 3 & : & L_2^{(1)} \\ 7x_1 + 8x_2 + x_3 = 6 & : & L_3^{(1)}. \end{array}\right. \iff \left\{\begin{array}{lcl} x_1 + 2x_2 + 3x_3 = 0 & : & L_1^{(2)} \\ -3x_2 - 6x_3 = 3 & : & L_2^{(2)} \\ -6x_2 - 20x_3 = 6 & : & L_3^{(2)}. \end{array}\right.$$

$$\iff \left\{\begin{array}{lcl} x_1 + 2x_2 + 3x_3 = 0 & : & L_1^{(2)} \\ -3x_2 - 6x_3 = 3 & : & L_2^{(2)} \\ -8x_3 = 0 & : & L_3^{(3)}. \end{array}\right.$$

$$\iff \left\{\begin{array}{lcl} x_1 & = & 2 \\ x_2 & = & -1 \\ x_3 & = & 0. \end{array}\right.$$

Example 3.3.3 *Consider the linear system*

$$\begin{pmatrix} 2 & 8 & 4 \\ 2 & 10 & 6 \\ 1 & 8 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \iff \begin{pmatrix} 1 & 4 & 2 \\ 0 & 2 & 2 \\ 0 & 4 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix}$$

$$\iff \begin{pmatrix} 1 & 4 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix}$$

$$\iff \begin{pmatrix} 1 & 4 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ -1/8 \end{pmatrix}.$$

We find the solution of the system by going back to the equation $z = -\dfrac{1}{8}$, so $y = \dfrac{1}{8}$ and $x = \dfrac{1}{4}$.

Remark 3.3 *For a matrix of order n, Gauss's method requires $n(n-1)(2n+5)/6$ additions, $(n-5)(2n+5)n/6$ multiplications and $n(n+1)/2$ divisions, i.e. a total of $(4n^3 + 9n^2 - 7n)/6$ elementary operations.*
By using the Cramer mules, we would have $(n+1)(n!-1)$ additions, $(n+1)(n-1)n!$ multiplications and n divisions.

For $n = 10$, the Gauss-Jordan method, we don't replace A as in the Gauss method but replace A with the identity.
For $n = 10$, Gauss's method requires 805 operations against 39916800 operations for resolution by Gabriel Cramer's formulas, $x_i = \det(A_i)/\det(A)$ where A_i is the matrix made up of the elements a_{ij} under the column i where we place the elements of the vector b.

3.3.2.1 Gauss-Jordan method

In the Gauss-Jordan method, the aim is not to triangulate A as in the Gauss method, but to replace A by the identity.

Example 3.3.4 *We have*

$$\begin{aligned}
\begin{pmatrix} 2 & 8 & 4 \\ 2 & 10 & 6 \\ 1 & 8 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} &\iff \begin{pmatrix} 1 & 4 & 2 \\ 0 & 2 & 2 \\ 0 & 4 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix} \\
&\iff \begin{pmatrix} 1 & 4 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & -4 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix} \\
&\iff \begin{pmatrix} 1 & 4 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ -1/8 \end{pmatrix} \\
&\iff \begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ -1/8 \end{pmatrix} \\
&\iff \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/4 \\ 1/8 \\ -1/8 \end{pmatrix},
\end{aligned}$$

then $(x, y, z) = \left(\frac{1}{4}, \frac{1}{8}, -\frac{1}{8}\right)$.

Example 3.3.5 *Denote that the matrix of columns of A and identity matrix I respectively by $[A|I]$, after using Gauss elimination, we obtain $[I|A^{-1}]$, the matrix of columns ofidentity matrix I and A^{-1} respectively, we note $[A|I] \longrightarrow [I|A^{-1}]$. Taking the same example again, we write*

$$\begin{aligned}
[A|I] &= \left(\begin{array}{ccc:ccc} 2 & 8 & 4 & 1 & 0 & 0 \\ 2 & 10 & 6 & 0 & 1 & 0 \\ 1 & 8 & 2 & 0 & 0 & 1 \end{array}\right) \\
&\sim \left(\begin{array}{ccc:ccc} 1 & 4 & 2 & 1/2 & 0 & 0 \\ 0 & 2 & 2 & -1 & 1 & 0 \\ 0 & 4 & 0 & -1/2 & 0 & 1 \end{array}\right) \\
&\sim \left(\begin{array}{ccc:ccc} 1 & 0 & -2 & 5/2 & -2 & 0 \\ 0 & 1 & 1 & -1/2 & 1/2 & 0 \\ 0 & 0 & -4 & 3/2 & -2 & 1 \end{array}\right) \\
&\sim \left(\begin{array}{ccc:ccc} 1 & 0 & 0 & 7/4 & -3 & -1/2 \\ 0 & 1 & 0 & -1/8 & 1 & 1/4 \\ 0 & 0 & 1 & -3/8 & -1/2 & 1/4 \end{array}\right).
\end{aligned}$$

So that,

$$[I|A^{-1}] = \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 7/4 & -3 & -1/2 \\ 0 & 1 & 0 & -1/8 & 1 & 1/4 \\ 0 & 0 & 1 & -3/8 & -1/2 & 1/4 \end{array}\right) \Longrightarrow X = A^{-1}b$$

Remark 3.4 *If A is a real matrix, the Gauss-Jordan method requires $n(n^2-1)/2$ multiplications, $n(n^2-1)/2$ additions and $n(n+1)/2$ divisions.*

3.3.3 Factorization method

This is a way to write A as a product of two matrices

$$A = LU,$$

where

L lower triangular matrix, U upper triangular matrix,

therefore, the system $AX = b \Longrightarrow LUX = b$. Put $Y = UX$, then

$$\begin{cases} LY = b, & Y = ? \\ UX = Y, & X = ?. \end{cases}$$

Remark 3.5 *1. Every invertible matrix admits a LU factorization.*

2. This decomposition is unique.

Problem: How to find L and U? such as $A = LU$.

$$A = LU \sim \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & & a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots 0 & \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & & l_{nn} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & \cdots & \cdots & & u_{nn} \end{pmatrix}.$$

There are two methods

a) Crout's method $u_{ii} = 1, \qquad i = 1, \cdots, n$

b) Delittle's method $u_{ii} = 1, \qquad i = 1, \cdots, n$

3.3.3.1 Crout's method

Let $A = LU$, $u_{ii} = 1$, $i = 1, \cdots, n$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & & a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & & l_{nn} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & \cdots & \cdots & u_{1n} \\ 0 & 1 & \cdots & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & \cdots & \cdots & & 1 \end{pmatrix}.$$

So that,

$$a_{ij} = \sum_{k=1}^{n} l_{ik}u_{kj}, \qquad i = 1, \cdots, n,\ j = 1, \cdots, n.$$

- **Step $k = 1$:** Calculate the first column elements of L

$$\begin{cases} a_{11} = l_{11} \\ a_{21} = l_{21} \\ a_{31} = l_{31} \\ \quad\vdots \\ a_{i1} = l_{i1} \\ \quad\vdots \\ a_{n1} = l_{n1} \end{cases} \implies l_{i1} = a_{i1}, \qquad i = 1, \cdots, n.$$

- Compute the first line items of U

$$\begin{cases} a_{12} = l_{11}u_{12} \Longrightarrow u_{12} = \dfrac{a_{12}}{l_{11}} \\ \quad\vdots \\ a_{1i} = l_{11}u_{1i} \Longrightarrow u_{1i} = \dfrac{a_{1i}}{l_{11}} \\ \quad\vdots \\ a_{1n} = l_{11}u_{1n} \Longrightarrow u_{1n} = \dfrac{a_{1n}}{l_{11}} \end{cases} \Longrightarrow u_{1i} = \frac{a_{1i}}{l_{11}}, \quad i = 2, \cdots, n.$$

- By the same way, we arrive at step $k = i$

$$l_{ii}, l_{i+1,i}, \cdots, l_{ni},$$

so,

$$\begin{array}{ll} a_{ii} = \sum_{k=1}^{i} l_{ik}u_{ki} \Longrightarrow a_{ii} = \sum_{k=1}^{i-1} l_{ik}u_{ki} + l_{ii} \Longrightarrow l_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik}u_{ki} \\ \quad\vdots \qquad\qquad\qquad\qquad \vdots \\ a_{ri} = \sum_{k=1}^{i} l_{rk}u_{ki} \Longrightarrow a_{ri} = \sum_{k=1}^{i-1} l_{rk}u_{ki} + l_{ri} \Longrightarrow l_{ri} = a_{ri} - \sum_{k=1}^{i-1} l_{rk}u_{ki}, \quad r = i, \cdots, n. \end{array}$$

- Compute the line i of U

$$u_{ii} = 1, u_{i,i+1}, \cdots, u_{ir}, \cdots, u_{in}$$

so that

$$a_{i,i+1} = \sum_{k=1}^{i} l_{ik}u_{k,i+1} = \sum_{k=1}^{i-1} l_{ik}u_{k,i+1} + l_{ii}u_{i,i+1} \implies u_{i,i+1} = \frac{1}{l_{ii}}\left(a_{i,i+1} - \sum_{k=1}^{i-1} l_{ik}u_{k,i+1}\right),$$

hence

$$a_{i,r} = \sum_{k=1}^{i} l_{ik}u_{kr} = \sum_{k=1}^{i-1} l_{ik}u_{kr} + l_{ii}u_{ir} \implies u_{ir} = \frac{1}{l_{ii}}\left(a_{ir} - \sum_{k=1}^{i-1} l_{ik}u_{kr}\right), \qquad r = i+1, \cdots, n$$

we conclude that

$$l_{nn} = a_{nr} - \sum_{k=1}^{i-1} l_{ik}u_{kr}.$$

3.3.3.2 Delittle's method

Let $A = LU$, $l_{ii} = 1$, $i = 1, \cdots, n$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & & a_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots 0 & \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & \cdots & \cdots & & u_{nn} \end{pmatrix}.$$

we have

$$a_{ij} = \sum_{k=1}^{n} l_{ik}u_{kj}, \qquad i = 1, \cdots, n,\ j = 1, \cdots, n$$

- **Step $k = 1$:** Calculate the first line elements of U

$$\begin{cases} a_{11} = u_{11} \\ a_{12} = u_{12} \\ a_{13} = u_{13} \\ \quad \vdots \\ a_{1i} = u_{1i} \\ \quad \vdots \\ a_{1n} = u_{1n} \end{cases} \implies u_{1i} = a_{1i}, \qquad i = 1, \cdots, n.$$

- Compute the first column items of L

$$\begin{cases} a_{21} = u_{11}l_{21} \Longrightarrow l_{21} = \dfrac{a_{21}}{u_{11}} \\ \quad \vdots \\ a_{i1} = u_{11}l_{i1} \Longrightarrow l_{i1} = \dfrac{a_{i1}}{u_{11}} \\ \quad \vdots \\ a_{n1} = u_{11}l_{n1} \Longrightarrow l_{n1} = \dfrac{a_{n1}}{u_{11}} \end{cases} \Longrightarrow u_{1i} = \frac{a_{1i}}{l_{11}}, \quad i = 2, \cdots, n.$$

- At step $k = i$

$$u_{ii}, u_{i,i+1}, \cdots, u_{in},$$

so,

$$\begin{array}{ll} a_{ii} = \sum_{k=1}^{i} u_{ki}l_{ik} \Longrightarrow a_{ii} = \sum_{k=1}^{i-1} u_{ki}l_{ik} + u_{ii} \Longrightarrow u_{ii} = a_{ii} - \sum_{k=1}^{i-1} u_{ki}l_{ik} \\ \quad \vdots \qquad\qquad \vdots \\ a_{ir} = \sum_{k=1}^{i} u_{kr}l_{ik} \Longrightarrow a_{ir} = \sum_{k=1}^{i-1} u_{kr}l_{ik} + u_{ir} \Longrightarrow u_{ir} = a_{ir} - \sum_{k=1}^{i-1} u_{kr}l_{ik}, \quad r = i, \cdots, n \end{array}$$

- Compute the column i of L

$$l_{ii} = 1, l_{i+1,i}, \cdots, l_{ri}, \cdots, l_{ni}$$

hence,

$$a_{i+1,i} = \sum_{k=1}^{i} u_{ki}l_{i+1,k} = \sum_{k=1}^{i-1} u_{ki}l_{i+1,k} + u_{ii}l_{i+1,i} \implies u_{i+1,i} = \frac{1}{u_{ii}}\left(a_{i+1,i} - \sum_{k=1}^{i-1} u_{ki}l_{i+1,k}\right)$$

we obtain

$$a_{ri} = \sum_{k=1}^{i} u_{ki}l_{rk} = \sum_{k=1}^{i-1} u_{ki}l_{rk} + u_{ii}l_{ri} \implies l_{ri} = \frac{1}{u_{ii}}\left(a_{ri} - \sum_{k=1}^{i-1} u_{ki}l_{rk}\right), \qquad r = i+1, \cdots, n$$

and

$$u_{nn} = a_{rn} - \sum_{k=1}^{i-1} u_{ki}l_{rk}.$$

Example 3.3.6 *Consider the system $AX = b$ such that*

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 4 \\ 3 & 3 & 4 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

find the solution of system $AX = b$ by Crout's method ($A = LU$).

Correction 3.3.1 *We consider*

- *L is a lower triangular matrix.*
- *U is an upper triangular matrix, such that $u_{ii} = 1, \qquad i = 1, \cdots, n.$*

We have,

$$A = LU \sim \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 4 \\ 3 & 3 & 4 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{pmatrix}.$$

- 1^{st} ***step:*** *$k = 1$: find 1^{st} column of L*

$$\begin{cases} 1 = a_{11} = l_{11} \\ 0 = a_{21} = l_{21} \\ 3 = a_{31} = l_{31} \end{cases}$$

find 1^{st} row of U

$$\begin{aligned} 1 = a_{12} = l_{11}u_{12} &\implies u_{12} = 1 \\ 1 = a_{13} = l_{11}u_{13} &\implies u_{13} = 1 \end{aligned}$$

- 2^{nd} ***step:*** *$k = 2$: find 2^{nd} column of U*

$$\begin{aligned} 2 = a_{22} = l_{21}u_{12} + l_{22} \times 1 &\implies l_{22} = 2 \\ 3 = a_{32} = l_{31}u_{12} + l_{32} \times 1 &\implies l_{32} = 0 \end{aligned}$$

find 2^{nd} row of U

$$4 = a_{23} = l_{21}u_{13} + l_{22}u_{23} \implies u_{23} = 0$$

- 3^{rd} ***step:*** *$k = 3$: find 3^{rd} column of L*

$$4 = a_{33} = l_{31}u_{13} + l_{32}u_{23} + l_{33} \times 1 \implies l_{33} = 1.$$

We have

$$AX = b \Longleftrightarrow L\underbrace{UX}_{=Y} = b \Longleftrightarrow \begin{cases} LY = b \\ UX = Y \end{cases}$$

so that,

$$LY = b \Longrightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \Longrightarrow \begin{cases} y_1 = 1 \\ y_2 = \dfrac{1}{2} \\ y_3 = -2 \end{cases}$$

hence

$$UX = Y \Longrightarrow \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ \dfrac{1}{2} \\ -2 \end{pmatrix} \Longrightarrow \begin{cases} x_1 = -\dfrac{3}{2} \\ x_2 = \dfrac{9}{2} \\ x_3 = -2 \end{cases}$$

Exercise 3.1 *Taking the same system for Delittle's method.*

3.3.3.3 Cholesky's method

We will write A in the form

$$A = RR^t,$$

such that R is a lower triangular matrix and R^t is the transpose matrix of R, therefore

$$AX = b \Longleftrightarrow RR^tX = b \Longleftrightarrow \left\{ \begin{array}{ll} RY = b, & Y =? \\ R^tX = Y, & X =? \end{array} \right.$$

Definition 3.1 *Let $A \in M_n(\mathbb{R})$, we say that A is a positive definite symmetric matrix if*

1. $A = A^t$,
2. $\det A_{kk} > 0, \qquad \forall 1 \leq k < n.$

Theorem 3.1 *[2] If $A \in M_n(\mathbb{R})$ is a positive definite symmetric matrix then there exists a lower triangular matrix R such that $A = RR^t$*

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & & a_{nn} \end{pmatrix} = \begin{pmatrix} R_{11} & 0 & \cdots & \cdots & 0 \\ R_{21} & R_{22} & 0 & \cdots 0 & \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ R_{n1} & R_{n2} & \cdots & & R_{nn} \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} & \cdots & \cdots & R_{1n} \\ 0 & R_{22} & \cdots & \cdots & R_{2n} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & \cdots & \cdots & & R_{nn} \end{pmatrix}.$$

- 1^{st} **step** $(k = 1)$ the elements of 1^{st} column of R

$$\begin{array}{rcl} a_{11} = R_{11}^2 & \Longrightarrow & R_{11} = \sqrt{a_{11}}, \\ a_{12} = R_{11}R_{21} & \Longrightarrow & R_{21} = \dfrac{a_{12}}{R_{11}}, \\ & \vdots & \\ a_{1j} = R_{11}R_{j1} & \Longrightarrow & R_{j1} = \dfrac{a_{1j}}{R_{11}}, \\ & \vdots & \\ a_{1n} = R_{11}R_{n1} & \Longrightarrow & R_{n1} = \dfrac{a_{1n}}{R_{11}}. \end{array}$$

  In summary

$$\left\{ \begin{array}{l} R_{11} = \sqrt{a_{11}}, \\ R_{j1} = \dfrac{a_{1j}}{R_{11}} = \dfrac{a_{1j}}{\sqrt{a_{11}}}, \qquad j = 2, \cdots, n. \end{array} \right.$$

- 2^{nd} **step** $(k = 2)$ the elements of 2^{nd} column of R

$$R_{22}, R_{32}, \cdots, R_{n2}.$$

  We have

$$\begin{array}{rcl} a_{22} & = & R_{21}^2 + R_{22}^2 \Longrightarrow R_{22} = \sqrt{a_{22} - R_{21}^2}, \\ a_{23} & = & R_{21}R_{31} + R_{22}R_{32} \Longrightarrow R_{32} = \dfrac{1}{R_{22}}(a_{23} - R_{21}R_{32}), \\ & \vdots & \\ a_{2n} & = & R_{21}R_{n1} + R_{22}R_{n2} \Longrightarrow R_{n2} = \dfrac{1}{R_{22}}(a_{2n} - R_{21}R_{n2}). \end{array}$$

- i^{th} **step** $(k = i)$ the elements of i^{th} column of R

$$R_{ii}, R_{i+1,i}, \cdots, R_{ni}.$$

We have

$$a_{ii} = \sum_{k=1}^{i-1} R_{kk}^2 + R_{ii}^2 \implies R_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} R_{kk}^2}.$$

If $r > i$ then

$$a_{ir} = \sum_{k=1}^{i-1} R_{ik}R_{rk} + R_{ii}R_{ri} \implies R_{ri} = \frac{1}{R_{ii}}(a_{ri} - \sum_{k=1}^{r-1} R_{ik}R_{rk}),$$

such as $r = i+1, \cdots, n$
for $i = 2, \cdots, n-1$

$$R_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} R_{ik}^2}$$

for $r = i+1, \cdots, n$

$$\begin{aligned} R_{ri} &= \frac{1}{R_{ii}}(a_{ri} - \sum_{k=1}^{i-1} R_{ik}R_{rk}) \\ R_{nn} &= \sqrt{a_{nn} - \sum_{k=1}^{i-1} R_{nk}^2}. \end{aligned}$$

Example 3.3.7 *Let the system $AX = b$ be such that*

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 5 & 3 \\ 1 & 3 & 11 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Correction 3.3.2 *The matrix A is symmetric ($A = A^t$) and positive definite* $\det(A_{kk}) > 0$

$$\det(A_{11}) = 1 > 0, \qquad \det(A_{22}) = 4 > 0,$$

and

$$\det(A_{33}) > 0, \qquad (\det(A_{33}) = \det(A) = \det(RR^t) = \prod_{i=1}^{n} R_{ii}^2).$$

We have

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 5 & 3 \\ 1 & 3 & 11 \end{pmatrix} = \begin{pmatrix} R_{11} & 0 & 0 \\ R_{21} & R_{22} & 0 \\ R_{31} & R_{32} & R_{33} \end{pmatrix} \begin{pmatrix} R_{11} & R_{21} & R_{31} \\ 0 & R_{22} & R_{32} \\ 0 & 0 & R_{33} \end{pmatrix}$$

After multiplying the two matrices on the right, we equal the product of the first line of the lower trigonometric matrix by the columns of the upper trigonometric matrix

$$\begin{cases} 1 = a_{11} = R_{11}^2 & \implies R_{11} = 1, \\ 1 = a_{12} = R_{11}R_{21} & \implies R_{21} = 1, \\ 1 = a_{13} = R_{11}R_{31} & \implies R_{31} = 1, \end{cases}$$

Then we equal the product of the other two lines of the lower triangular matrix by the columns of the upper triangular matrix

$$\begin{cases} 5 = a_{22} = R_{21}^2 + R_{22}^2 & \implies R_{22} = 1, \\ 3 = a_{23} = R_{21}R_{31} + R_{22}R_{32} & \implies R_{32} = 1, \\ 11 = a_{13} = R_{31}^2 + R_{32}^2 + R_{33}^2 & \implies R_{31} = 1, \end{cases}$$

Then we write the linear problem as two problems with two triangular matrices

$$AX = b \Longleftrightarrow RR^tX = b \Longleftrightarrow \left\{ \begin{array}{ll} RY = b, & Y =? \\ R^tX = Y, & X =? \end{array} \right.$$

We solve the first problem

$$RY = b \Longleftrightarrow \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 3 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \Longrightarrow \left\{ \begin{array}{l} y_1 = 1 \\ y_2 = 0 \\ y_3 = 0 \end{array} \right.$$

By solving the second problem, we have found the solutions to the problem

$$R^tX = Y \Longleftrightarrow \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \left\{ \begin{array}{l} y_1 = 1 \\ y_2 = 0 \\ y_3 = 0 \end{array} \right.$$

3.4 Iterative Methods

In general, iterative methods are used if $A \in M_n(\mathbb{R})$ with $n \geq 100$, and if there are rounding errors.

- Let the system $AX = b \Longleftrightarrow X = BX + C$ such that $b, C \in \mathbb{R}^n$, $A, B \in M_n(\mathbb{R})$
- We construct the following

$$\left\{ \begin{array}{l} X^{(0)} \in \mathbb{R}^n \\ X^{(k+1)} = BX^{(k)} + C, \qquad k = 0, 1, \cdots . \end{array} \right.$$

  If the sequence vector converges to the exact solution of linear system $AX = b$.

3.4.1 General case

In iterative methods, the system $AX = b$ admits a unique solution X such that $A \in M_n(\mathbb{R})$ and $b \in \mathbb{R}^n$. We write A in the form

$$A = M - N,$$

such that $\det(M) \neq 0$, hence

$$\begin{array}{rcl} AX = b & \Longleftrightarrow & (M - N)X = b \\ & \Longleftrightarrow & MX = NX + b \\ & \Longleftrightarrow & X = M^{-1}NX + M^{-1}b \\ & \Longleftrightarrow & X = BX + C \quad \text{where} \quad B = M^{-1}N, \quad \text{and} \quad C = M^{-1}b \end{array}$$

so,

$$\left\{ \begin{array}{l} X^0 \text{ given}, \\ X^{(k+1)} = BX^{(k)} + C, \qquad k = 0, 1, \cdots . \end{array} \right.$$

when the sequence $X^{(k)}$ converges, i.e. $\lim_{k \longrightarrow +\infty} X^{(k)} = X$, we say that the method is convergent.

3.4.2 The convergence study

The basic principle of such methods is to generate a sequence of vectors $X^{(k)}$ (the iterates) converging to the solution X of the linear system $AX = b$.
Most iterative methods are of the following form.
Starting from an arbitrary vector $X^{(0)}$, we generate a sequence $(X^{(k)})_k$ defined by

$$X^{(k+1)} = BX^{(k)} + C, \tag{3.4.1}$$

with B is matrix of $M_{n,n}(K)$, $C \in \mathbb{R}^n$ or $\mathbb{C}^n$.

Definition 3.2 *An iterative method of the form (3.4.1) is said to be convergent if for all $X^{(0)}$, we have*

$$X^{(k)} \longrightarrow X, \text{ when } k \longrightarrow \infty$$

and the limit checks $AX = b$. ($AX = b$ is then equivalent to $X = BX + C$).

Let $\rho(B)$ designate the spectral radius of the matrix B. For the converge of iterative method we have the following theorem.

Theorem 3.2 *For an iterative method of the form (3.4.1) to be convergent, it is necessary and sufficient that $\rho(B) < 1$ i.e.*

$$\lim_{k\longrightarrow\infty} B^k \Longleftrightarrow \rho(B) < 1.$$

Proof. See, for example, [2]. ■

Definition 3.3 *Let $B \in M_n(\mathbb{R})$, we define the norm $\|\rho((B)\|$ of matrix B*

$$\begin{aligned} \| B \|_1 &= \sum_{i=1}^{n}\sum_{j=1}^{n}(|b_{i,j}|), \\ \| B \|_2 &= \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(b_{i,j})^2}, \\ \| B \|_\infty &= \max_{1\leq i,j\leq n}(|b_{i,j}|). \end{aligned}$$

Corollary 3.1 *If $\| B \| < 1$, then $\{X^k\}_{k=0}^{\infty}$, such that $X^{(k+1)} = BX^{(k)} + C$, $k > 0$, converge, for all $X^{(0)} \in \mathbb{R}^n$, to vector $X \in \mathbb{R}^n$. and we have inequalities*

$$\| X^{(k)} - X \| \leq \| B \|^k \| X^{(0)} - X \| \tag{3.4.2}$$

and

$$\| X^{(k)} - X \| \leq \frac{\| B \|^k}{1 - \| B \|} \| X^{(1)} - X^{(0)} \| . \tag{3.4.3}$$

Proof. Firstly, for all $k \in \mathbb{R}$, we have

$$\begin{aligned} X^{(k)} - X &= (BX + C) - (BX^{(k-1)} + C) \\ &= B(X - X^{(k-1)}) \\ &= B((BX + C) - (BX^{(k-1)} + C)) \\ &= B^2(X - X^{(k-2)}), \end{aligned}$$

we obtain

$$X^{(k)} - X = B^k(X - X^{(0)}),$$

implies that

$$\| X^{(k)} - X \| \leq \| B \|^k \| X^{(0)} - X \|;$$

because $\| B \| < 1$, with $k \longrightarrow \infty$ we obtain $\| B \|^k \longrightarrow 0$ then $\lim_{k\longrightarrow\infty} \| X^{(k)} - X \| = 0$, after that $\lim_{k\longrightarrow\infty} X^{(k)} = X$ solution of $X = BX + C$.

We prove $\| X^{(k)} - X \| \leq \frac{\| B \|^k}{1 - \| B \|} \| X^{(1)} - X^{(0)} \| .$

Let $p \geq 1$, we have

$$\begin{aligned} \| X^{(k+p)} - X^{(k)} \| &= \| X^{(k+p)} - X^{(k+p-1)} + X^{(k+p-1)} - X^{(k+p-2)} + X^{(k+p-2)} \\ &\qquad - \cdots\cdots - X^{(k+1)} + X^{(k+1)} - X^{(k)} \| \\ &\leq \| X^{(k+p)} - X^{(k+p-1)} \| + \| X^{(k+p-1)} - X^{(k+p-2)} \| + \cdots\cdots + \| X^{(k+1)} - X^{(k)} \| . \end{aligned}$$

Then

$$\begin{aligned} \| X^{(k+p)} - X^{(k+p-1)} \| &\leq \| (BX^{(k+p-1)} + C) - (BX^{(k+p-2)} + C) \| \\ &\leq \| B(X^{(k+p-1)} - X^{(k+p-2)}) \| \\ &\leq \| B \| \| X^{(k+p-1)} - X^{(k+p-2)} \|, \end{aligned}$$

by recurrence

$$\| X^{(k+p)} - X^{(k+p-1)} \| \leq \| B \|^{p-1} \| X^{(k+1)} - X^{(k)} \|, p \geq 1.$$

After that

$$\| X^{(k+p)} - X^{(k)} \| \leq \| X^{(k+1)} - X^{(k)} \| + \| B \| \| X^{(k+1)} - X^{(k)} \| + \cdots + \| B \|^{p-1} \| X^{(k+1)} - X^{(k)} \|$$

then

$$\begin{aligned} \| X^{(k+p)} - X^{(k)} \| &\leq (1 + \| B \| + \| B \|^2 + \cdots + \| B \|^{p-1}) \| X^{(k+1)} - X^{(k)} \| \\ &= \frac{1 - \| B \|^p}{1 - \| B \|} \| X^{(k+1)} - X^{(k)} \| . \end{aligned}$$

If $p \longrightarrow +\infty$ on obtain

$$\| X^{(k+p)} - X^{(k)} \| \leq \frac{1}{1 - \| B \|} \| X^{(k+1)} - X^{(k)} \|,$$

then

$$\begin{aligned} \| X^{(k+p)} - X^{(k)} \| &\leq \frac{1}{1 - \| B \|} \| B \| \| X^{(k)} - X^{(k-1)} \| \leq \frac{1}{1 - \| B \|} \| B \|^2 \| X^{(k-1)} - X^{(k-2)} \| \\ &\leq \cdots \leq \frac{\| B \|^k}{1 - \| B \|} \| X^{(1)} - X^{(0)} \|, \end{aligned}$$

after that

$$\| X^{(k+p)} - X^{(k)} \| \leq \frac{\| B \|^k}{1 - \| B \|} \| X^{(1)} - X^{(0)} \| .$$

■

Remark 3.6 *If* $\| X - X^{(k)} \| \leq \epsilon$ *we obtain*

$$\begin{aligned} \frac{\| B \|^k}{1 - \| B \|} \| X^{(1)} - X^{(0)} \| \leq \epsilon &\implies \| B \|^k \leq \frac{(1 - \| B \|)\epsilon}{\| X^{(1)} - X^{(0)} \|} \\ &\implies k \ln(\| B \|) \leq \ln \left(\frac{(1 - \| B \|)\epsilon}{\| X^{(1)} - X^{(0)} \|} \right) \\ &\implies k \geq \frac{\ln \left(\frac{(1 - \| B \|)\epsilon}{\| X^{(1)} - X^{(0)} \|} \right)}{\ln(\| B \|)}. \end{aligned}$$

We choose

$$k = \left[\frac{\ln \left(\frac{(1 - \| B \|)\epsilon}{\| X^{(1)} - X^{(0)} \|} \right)}{\ln(\| B \|)} \right] + 1.$$

Corollary 3.2 *If* $\|B\| < 1$ *then* $\rho(B) < 1$ *and the method (3.4.1) converges.*

Proof. See, for example, [2]. ■

Definition 3.4 *We say A is with strictly dominant diagonal, that is to say that the diagonal element is strictly superior in absolute value to the sum of non-diagonal elements of the same line :*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

Proposition 3.1 *When the matrix A is with strictly dominant diagonal, the Jacobi method converges.*

Proof. See, for example, [2]. ■

3.4.3 Jacobi method

Let $AX = b$ be the linear system such that

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & a_{nn} \end{pmatrix}$$

We put

$$A = D - E - F$$

such as

$$D_{ij} = \begin{cases} a_{ij}, & i = j \\ 0, & i \neq j \end{cases}, \qquad E_{ij} = \begin{cases} -a_{ij}, & i > j \\ 0, & i \leq j \end{cases}, \qquad F_{ij} = \begin{cases} -a_{ij}, & i < j \\ 0, & i \geq j \end{cases}.$$

In other words,

$$D = \begin{pmatrix} a_{11} & 0 & \cdots & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & a_{nn} \end{pmatrix}, \qquad -E = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & 0 \end{pmatrix},$$

and

$$-F = \begin{pmatrix} 0 & a_{12} & \cdots & \cdots & a_{1n} \\ 0 & 0 & \ddots & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & a_{n-1,n} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

Formula using coordinates We have

$$AX = b \iff \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_2 \\ \qquad \vdots \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases}$$

$$\iff \begin{cases} x_1 = \dfrac{1}{a_{11}}\left(b_1 - \sum\limits_{j=2}^{n} a_{ij}x_j\right) \\ x_2 = \dfrac{1}{a_{22}}\left(b_2 - \sum\limits_{j=1, j\neq 2}^{n} a_{2j}x_j\right) \\ \qquad \vdots \\ x_n = \dfrac{1}{a_{nn}}\left(b_n - \sum\limits_{j=1}^{n-1} a_{ij}x_j\right) \end{cases}$$

we write algorithm of Jacobi

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{a_{11}}\left(b_1 - \sum\limits_{j=2}^{n} a_{ij}x_j^{(k)}\right) \\ x_2^{(k+1)} = \dfrac{1}{a_{22}}\left(b_2 - \sum\limits_{j=1,j\neq 2}^{n} a_{2j}x_j^{(k)}\right) \\ \quad\vdots \\ x_n^{(k+1)} = \dfrac{1}{a_{nn}}\left(b_n - \sum\limits_{j=1}^{n-1} a_{ij}x_j^{(k)}\right) \end{cases} \tag{3.4.4}$$

so that,

$$\begin{aligned} X^{(0)} &= (x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)},) \qquad \text{given,} \\ x_i^{(k+1)} &= \frac{1}{a_{ii}}\left(b_i - \sum_{j=1,j\neq i}^{n} a_{ij}x_j^{(k)}\right), \quad k = 0, 1, 2, \ldots. \end{aligned}$$

Remark 3.7 *We have*

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \underbrace{\sum_{j<i} a_{ij}\, x_j^{(k)}}_{E} - \underbrace{\sum_{j>i} a_{ij}\, x_j^{(k)}}_{F}\right) \qquad i = 1, \cdots, n \quad \textit{and} \quad D_{ij}^{-1} = \frac{1}{a_{ii}}$$

Example 3.4.1 *Consider the system*

$$\begin{pmatrix} 4 & 2 & 1 \\ -1 & 2 & 0 \\ 2 & 1 & 4 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 9 \end{pmatrix}$$

Correction 3.4.1 *We put*

$$\begin{cases} x = 1 - \dfrac{y}{2} - \dfrac{z}{4} \\ y = 1 + \dfrac{x}{2} \\ z = \dfrac{9}{4} - \dfrac{x}{2} - \dfrac{y}{4} \end{cases}$$

implies that

$$\begin{cases} x^{(k+1)} = 1 - \dfrac{y^{(k)}}{2} - \dfrac{z^{(k)}}{4} \\ y^{(k+1)} = 1 + \dfrac{x^{(k)}}{2} \\ z^{(k+1)} = \dfrac{9}{4} - \dfrac{x^{(k)}}{2} - \dfrac{y^{(k)}}{4} \end{cases} \qquad k = 0, 1, \cdots \tag{3.4.5}$$

Let $X^{(0)} = (0, 0, 0)$ be the initial vector, by taking $k = 0, 1, 2, \ldots$ in (3.4.5) and after the calculation

$$\begin{aligned} x^{(1)} &= (1, 1, 9/4) \\ x^{(2)} &= (-1/16, 3/2, 3/2) \\ x^{(3)} &= (-1/8.3/2.61/32) \\ x^{(4)} &= (5/128, 15/16, 265/128) \\ x^{(5)} &= (7/512, 261/256, 511/256), \end{aligned}$$

the sequence $(X^{(k)})$ converges to the solution of the system $(0, 1, 2)$.

3.4.4 Gauss-Seidel method

To accelerate the speed of convergence, one can improve the method of Jacobi by a method called method of Gauss-Seidel

Matrix form We have

$$A = D - E - F = (D - E) - F$$

hence

$$AX = b \Longleftrightarrow (D - E)X - FX = b$$

since $D - E$ is inversible lower triangular matrix ($\det(D - E) = \prod_{i=1}^{n} a_{ii} \neq 0$), then

$$\begin{aligned} X &= (D - E)^{-1}FX + (D - E)^{-1}b \\ &= BX + C. \end{aligned}$$

Finally,

$$\begin{aligned} &X^{(0)} \quad \text{given,} \\ &X^{(k+1)} = BX^{(k)} + C \quad \text{with} \quad B = (D - E)^{-1}F \quad \text{and} \quad C = (D - E)^{-1}b. \end{aligned}$$

Coordinates form We have by using the Jacobi method

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right), \quad i = 1, \cdots, n$$

The Gauss-Seidel method

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right), \quad i = 1, \cdots, n$$

Example 3.4.2 *Let's apply the Gauss-Seidel method to the system form example (3.4.1)*

$$\begin{cases} x^{(k+1)} = 1 - \dfrac{y^{(k)}}{2} - \dfrac{z^{(k)}}{4} \\ y^{(k+1)} = 1 + \dfrac{x^{(k)}}{2} \\ z^{(k+1)} = \dfrac{9}{4} - \dfrac{x^{(k)}}{2} - \dfrac{y^{(k)}}{4} \end{cases} \qquad k = 0, 1, \cdots$$

Correction 3.4.2 *We write the system for each $k = 0, 1, 2, ...$ as*

$$\begin{cases} x^{(k+1)} = 1 - \dfrac{y^{(k)}}{2} - \dfrac{z^{(k)}}{4} \\ y^{(k+1)} = 1 + \dfrac{x^{(k+1)}}{2} \\ z^{(k+1)} = \dfrac{9}{4} - \dfrac{x^{(k+1)}}{2} - \dfrac{y^{(k+1)}}{4} \end{cases} \qquad k = 0, 1, \cdots$$

Starting from the point $X^{(0)} = (0, 0, 0)$, we successively calculate

$$\begin{aligned} X^{(1)} &= (1, 3/2, 11/8) \\ X^{(2)} &= (-3/32.61/64.527/256) \\ X^{(3)} &= (9/1024, 2047/2048, 16349/8192). \end{aligned}$$

This set of points converges to the exact solution $(0, 1, 2)$. The Gauss-Seidel method improves the speed of converges.

Remark 3.8 • *Let $X = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$, for example, we have*

$$\begin{aligned} \| X \|_1 &= \sum_{i=1}^{n} |x_i|, \\ \| X \|_2 &= \sqrt{\sum_{i=1}^{n} x_i^2}, \\ \| X \|_\infty &= \max_i \Big(\sum_{j=1,\ldots,n} |x_{ij}| \Big). \end{aligned}$$

- *Denote that $J = D^{-1}(E+F)$ the matrix of Jacobi algorithm and $\mathcal{L} = (D-E)^{-1}F$ the matrix of Gauss-Seidel algorithm.*

Example 3.4.3 *Let $AX = b$ be the linear system*

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \qquad X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

1. *Write the Jacobi and Gauss-Seidel algorithms of the linear system $AX = b$.*
2. *Study the convergences of these algorithms.*
3. *Calculate J, $\|J\|_\infty, \rho(J)$ and $\mathcal{L}$, $\|\mathcal{L}\|_\infty, \rho(\mathcal{L})$.*
4. *For $X^{(0)} = (0,0,0)^T$ and $\epsilon = 10^{-6}$, calculate the numbers of iterations k_1 and k_2 where we effected two algorithms respectively $\|X^{(k)} - X\|_\infty \leq \epsilon$, $\forall k \geq k_1$ by Jacobi and $\|X^{(k)} - X\|_\infty \leq \epsilon$, $\forall k \geq k_2$ by Gauss-Seidel.*

Correction 3.4.3 *1. The algorithms of Jacobi and Gauss-Seidel applied on $AX = b$*

$$(Jacobi) \begin{cases} x_1^{(k+1)} &= (x_1^{(k)} + 1)/2, \\ x_2^{(k+1)} &= (x_1^{(k)} + x_3^{(k)} + 2)/3, \qquad X^{(0)} \in \mathbb{R}^3, \text{ given} \\ x_3^{(k+1)} &= (x_2^{(k)} + 1)/2. \end{cases}$$

and

$$(Gauss - Seidel) \begin{cases} x_1^{(k+1)} &= (x_1^{(k)} + 1)/2, \\ x_2^{(k+1)} &= (x_1^{(k+1)} + x_3^{(k)} + 2)/3, \qquad X^{(0)} \in \mathbb{R}^3, \text{ given} \\ x_3^{(k+1)} &= (x_2^{(k+1)} + 1)/2. \end{cases}$$

2. The matrice A is dominant strictement diagonal ($2 > 1+0$, $3 > 1+1$, $2 > 0+1$).

3. We calculate J, $\|J\|_\infty, \rho(J)$ and $\mathcal{L}$, $\|\mathcal{L}\|_\infty, \rho(\mathcal{L})$

$$J = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 \end{pmatrix}, \qquad \mathcal{L} = \begin{pmatrix} 0 & 1/2 & 0 \\ 0 & 1/6 & 1/3 \\ 0 & 1/12 & 1/6 \end{pmatrix}.$$

so that,

$$\begin{aligned} \|J\|_\infty &= 2/3, & \rho(J) &= 1/\sqrt{3}, \\ \|\mathcal{L}\|_\infty &= 1/2, & \rho(\mathcal{L}) &= 1/3. \end{aligned}$$

Indeed, $\det(J - \lambda I) = \lambda^3 - \lambda/3$ *and* $\det(\mathcal{L} - \lambda I) = \lambda^3 - \lambda^2/3$ *or* $\rho(\mathcal{L}) = \rho^2(J)$

*4. We have **Jacobi** algorithm $X^{(0)} = (0,0,0)^T$, $X^{(1)} = (\frac{1}{2}, \frac{2}{3}, \frac{1}{2})^T$, $\|X^{(1)} - X^{(0)}\|_\infty = \frac{2}{3}$ and $\epsilon = 10^{-6}$, hence*

$$k_1 \geq \ln\left(\frac{(1 - \|J\|_\infty)\epsilon}{\|X^{(1)} - X^{(0)}\|_\infty}\right) / \ln(\|J\|_\infty) = 35.782, \text{ so that } k_1 = 36.$$

***Gauss-Seidel** algorithm $X^{(0)} = (0,0,0)^T$, $X^{(1)} = (\frac{1}{2}, \frac{5}{6}, \frac{11}{12})^T$, $\|X^{(1)} - X^{(0)}\|_\infty = \frac{11}{12}$ and $\epsilon = 10^{-6}$, then*

$$k_2 \geq \ln\left(\frac{(1 - \|\mathcal{L}\|_\infty)\epsilon}{\|X^{(1)} - X^{(0)}\|_\infty}\right) / \ln(\|\mathcal{L}\|_\infty) = 20.806, \text{ so that } k_1 = 21.$$

3.4.5 Relaxation method

It is based on this decomposition, let $\omega \neq 0$, the matrix A is written $A = \left(\frac{D}{\omega} - E\right) + \left(D - \frac{D}{\omega} - F\right)$.
The system $Ax = b$ is written

$$\left(\frac{D}{\omega} - E\right)X = \left(\frac{1-\omega}{\omega}D + F\right)X + b.$$

The relaxation method is given by

$$\begin{cases} X^{(0)} \text{ given} \\ \left(\frac{D}{\omega} - E\right)X^{(k+1)} = \left(\frac{1-\omega}{\omega}D + F\right)X^{(k)} + b, \end{cases}$$

we get

$$\Big(D - \omega E\Big)X^{(k+1)} = \Big((1-\omega)D + \omega F\Big)X^{(k)} + \omega b.$$

The iteration matrix is then written

$$\mathcal{L}_\omega = (D - \omega E)^{-1}((1-\omega)D + \omega F).$$

The components of the vector $X^{(k+1)} = (x_1, x_2, \cdots, x_k, \cdots, x_n)$ are solutions of

$$\begin{cases} X^{(0)} = (x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)},), \text{ donnée} \\ x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \dfrac{\omega}{a_{ii}}\Big(b_i - \sum\limits_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum\limits_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big), \quad i = 1, \cdots, n. \end{cases}$$

For $\omega = 1$, we get the Gauss-Seidel method.

Remark 3.9 *We have $\omega \in]0, 2[$ because relaxation algorithme converge if $\det(\mathcal{L}_\omega) \leq 1$ such that*

$$\begin{aligned} \det(\mathcal{L}_\omega) &= \det((D - \omega E)^{-1}((1-\omega)D + \omega F) \\ &= \det((D - \omega E)^{-1})\det(((1-\omega)D + \omega F)) \\ &= \frac{1}{\prod_{i=1}^n a_{ii}}(1-\omega)^n \prod_{i=1}^n a_{ii} \\ &= (1-\omega)^n = \prod_{i=1}^n \lambda_i \leq \rho^n(\mathcal{L}_\omega) < 1, \end{aligned}$$

we obtain

$$|1 - \omega| < 1 \Longrightarrow -1 < 1 - \omega < 1 \Longrightarrow 0 < \omega < 2,$$

then

$$\omega \in]0, 2[.$$

Remark 3.10 *We choose*

$$\omega = \frac{2}{\sqrt{1 - (\rho(J))^2}} \in]0, 2[,$$

with $\rho(J)$ designate the spectral radius of the Jacobi matrix.

Example 3.4.4 *By the relaxation method, we obtain*

$$\begin{pmatrix} 7 & -2 & -1 \\ 1 & -6 & 2 \\ 2 & -1 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 10 \\ 0 \\ 13 \end{pmatrix}$$

we write

$$\begin{cases} x_1 = \dfrac{1}{7}\Big(10 + 2x_2 + x_3\Big) \\ x_2 = \dfrac{1}{6}\Big(x_1 + 2x_3\Big) \\ x_3 = \dfrac{1}{5}\Big(13 - 2x_1 + x_2\Big) \end{cases}$$

implies that

$$\begin{cases} x_1^{(k+1)} = (1-\omega)x_1^{(k)} + \dfrac{\omega}{7}\Big(10 + 2x_2^{(k)} + x_3^{(k)}\Big) \\ x_2^{(k+1)} = (1-\omega)x_2^{(k)} + \dfrac{\omega}{6}\Big(x_1^{(k+1)} + 2x_3^{(k)}\Big) \\ x_3^{(k+1)} = (1-\omega)x_3^{(k)} + \dfrac{\omega}{5}\Big(13 - 2x_1^{(k+1)} + x_2^{(k+1)}\Big) \end{cases}$$

Let $X^{(0)} = (0,0,0)$ *the initial vector and* $\omega = 0.25$ *we calculate* $X^{(1)}$

$$\begin{cases} x_1^{(1)} = \dfrac{10}{7}(0.25) = 0.3571 \\ x_2^{(1)} = \dfrac{0.25}{7}(0.3571) = 0.0148 \\ x_3^{(1)} = 0.615 \end{cases}$$

3.5 Exercises with solutions

Exercise 3.1 *Consider the system* $AX = b$ *such that*

$$A = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 4 & -2 \\ -1 & -3 & 5 \end{pmatrix}, X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } X^0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

1. *Solve this system using*
   a) *The elimination of Gauss.*
   b) *Crust Decomposition* ($A = LU$, *with* $u_{ii} = 1$, $i = 1, 2, 3.$).
2. *Have the matrix A a factorization by Cholesky's method? Why?.*
3. *Verify that the matrix A is strictly diagonal dominant, what do you conclude?.*
4. *Write the iteration algorithm using*
   a) *Jacobi method, then calculate* $X^{(1)}$, $X^{(2)}$.
   b) *Gauss-Seidel method, then calculate* $X^{(1)}$, $X^{(2)}$.

Exercise 3.2 *Consider the linear system* $AX = b$, *such that*

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}, X = \begin{pmatrix} x \\ y \\ z \\ t \end{pmatrix} \text{ and } b = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

1. *Find the solution of this system by Gaussian method.*
2. *Show that A has a Cholesky factorization.*
3. *Determine the lower triangular matrix R such that* $A = RR^T$.
4. *Solve the linear system* $AX = b$ *by Cholesky's method.*

Exercise 3.3 *Let $AX = b$ be the linear system*

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 5 & 2 \\ 0 & 2 & 10 \end{pmatrix}, \qquad b = \begin{pmatrix} 2 \\ 8 \\ 12 \end{pmatrix}, \qquad X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

1. *Solve the system $AX = b$, using Gaussian elimination.*
2. *Determine the Doolittle factorization $A = LU$, $l_{ii} = 1$.*
3. *Deduce the factorization $A = LDL^t$, $l_{ii} = 1$.*
4. *Check that A admits a Cholesky factorization RR^t; determine R.*
5. *Write the Jacobi and Gauss-Seidel algorithms applied to $AX = b$.*

3.5.1 Solutions

Solution 3.1 *1. Solve this system using*

a) The elimination of Gauss

$$\left\{\begin{array}{ll} 3x_1 - x_2 - x_3 = 1 & L_1^{(1)} \\ -x_1 + 4x_2 - 2x_3 = 1 & L_2^{(1)} \\ -x_1 - 3x_2 + 5x_3 = 1 & L_3^{(1)} \end{array}\right. \longrightarrow \left\{\begin{array}{ll} 3x_1 - x_2 - x_3 = 1 & L_1^{(1)} = L_1^{(2)} \\ \frac{11}{3}x_2 - \frac{7}{3}x_3 = \frac{4}{3} & L_2^{(1)} + \frac{1}{3}L_1 = L_2^{(2)} \\ -\frac{10}{3}x_2 + \frac{14}{3}x_3 = \frac{4}{3} & L_3^{(1)} + \frac{1}{3}L_1^{(1)} = L_{(3)}^2 \end{array}\right.$$

$$\longrightarrow \left\{\begin{array}{ll} 3x_1 - x_2 - x_3 = 1 & L_1^{(2)} = L_1^{(3)} \\ \frac{11}{3}x_2 - \frac{7}{3}x_3 = \frac{4}{3} & L_2^{(2)} = L_2^{(3)} \\ \frac{84}{33}x_3 = \frac{84}{33} & L_3^{(2)} + \frac{10}{11}L_2^{(2)} = L_3^{(3)}. \end{array}\right.$$

According to the lift method, we have $x_3 = 1, \quad x_2 = 2 \quad$ and $\; x_1 = 1$.

b) Crout Decomposition ($A = LU$, with $u_{ii} = 1$, $i = 1, 2, 3.$). We have

$$A = LU \iff \begin{pmatrix} 3 & -1 & -1 \\ -1 & 4 & -2 \\ -1 & -3 & 5 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

$$\iff \left\{\begin{array}{l} l_{11} = 3, \qquad l_{12} = -1, \qquad l_{13} = -1, \\ l_{11}u_{12} = -1, \qquad l_{11}u_{13} = -1, \\ l_{21}u_{12} + l_{22} = 4, \qquad l_{21}u_{13} + l_{22}u_{23} = -1, \\ l_{31}u_{12} + l_{32} = -3, \qquad l_{31}u_{13} + l_{32}u_{23} + l_{33} = 5 \end{array}\right.$$

$$\iff \left\{\begin{array}{l} l_{11} = 3, \qquad l_{12} = -1, \qquad l_{13} = -1, \\ u_{12} = -1/3, \qquad u_{13} = -1/3, \\ l_{22} = 11/3, \qquad u_{23} = -7/11, \\ l_{32} = -10/3, \qquad l_{33} = 84/33. \end{array}\right.$$

In summary,

$$LU = \begin{pmatrix} 3 & 0 & 0 \\ -1 & 11/3 & 0 \\ -1 & -10/3 & 84/33 \end{pmatrix} \begin{pmatrix} 1 & -1/3 & -1/3 \\ 0 & 1 & -7/11 \\ 0 & 0 & 1 \end{pmatrix}.$$

We solve this system by Crout's method as follows

$$AX = b \Longleftrightarrow LUX = b \Longleftrightarrow \left\{\begin{array}{l} LY = b \to Y = ? \\ and \\ UX = Y \to X = ? \end{array}\right.$$

so,

$$LY = b \Longleftrightarrow \begin{pmatrix} 3 & 0 & 0 \\ -1 & 11/3 & 0 \\ -1 & -10/3 & 84/33 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

According to the ascent method

$$(y_1, y_2, y_3) = \left(\frac{1}{3}, \frac{4}{11}, 1\right).$$

so,

$$LY = b \Longleftrightarrow \begin{pmatrix} 3 & -1/3 & 1/3 \\ 0 & 1 & -7/11 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 4/11 \\ 1 \end{pmatrix}.$$

According to the ascent method

$$(x_1, x_2, x_3) = \Big(1, 1, 1\Big).$$

2. *The matrix A does not have a factorization by the Cholesky method since*

$$A^t = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 4 & -3 \\ -1 & -2 & 5 \end{pmatrix} \neq \begin{pmatrix} 3 & -1 & -1 \\ -1 & 4 & -2 \\ -1 & -3 & 5 \end{pmatrix} = A$$

then the matrix A is not symmetric. Hence A does not have a Cholesky factorization.

3. *The matrix A is strictly diagonal dominant*

$$\begin{aligned} |a_{11}| = 3 > 2 = |-1| + |-1| = |a_{12}| + |a_{13}| \\ |a_{22}| = 4 > 3 = |-1| + |-2| = |a_{21}| + |a_{23}| \\ |a_{33}| = 5 > 4 = |-1| + |-3| = |a_{31}| + |a_{32}| \end{aligned}$$

then the Jacobi algorithm and the Gauss-Seidel algorithm are converge to the exact solution of this system.

4. *Write the iteration algorithm using*

   a) *Jacobi method, then calculate $X^{(1)}$, $X^{(2)}$.*

$$\begin{cases} x_1^{(n+1)} = \dfrac{1}{3}(1 + x_2^{(n)} + x_3^{(n)}) \\ x_2^{(n+1)} = \dfrac{1}{4}(1 + x_1^{(n)} + 2x_3^{(n)}) \\ x_3^{(n+1)} = \dfrac{1}{5}(1 + x_1^{(n)} + 3x_2^{(n)}) \end{cases}$$

   and as $X^{(0)} = (0, 0, 0)$ then $\left(\frac{1}{3}, \frac{1}{4}, \frac{1}{5}\right)$ and $X^{(2)} = \left(\frac{29}{60}, \frac{26}{60}, \frac{5}{12}\right)$

   b) *Gauss-Seidel method, then calculate $X^{(1)}$, $X^{(2)}$.*

$$\begin{cases} x_1^{(n+1)} = \dfrac{1}{3}(1 + x_2^{(n)} + x_3^{(n)}) \\ x_2^{(n+1)} = \dfrac{1}{4}(1 + x_1^{(n+1)} + 2x_3^{(n)}) \\ x_3^{(n+1)} = \dfrac{1}{5}(1 + x_1^{(n+1)} + 3x_2^{(n+1)}) \end{cases}$$

   and since $X^{(0)} = (0, 0, 0)$ then $X^{(1)} = \left(\frac{1}{3}, \frac{1}{3}, \frac{7}{15}\right)$ and $X^{(2)} = \left(\frac{3}{5}, \frac{19}{30}, \frac{7}{10}\right)$.

Solution 3.2 *1. Solve this system by using Gaussian elimination*

$$(A \mid b) = \left(\begin{array}{cccc:c} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \end{array}\right) \sim \left(\begin{array}{cccc:c} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 & 1 \\ 0 & 1 & 2 & 2 & 2 \\ 0 & 1 & 2 & 3 & 3 \end{array}\right)$$

$$\sim \left(\begin{array}{cccc:c} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 & 2 \end{array}\right)$$

$$\sim \left(\begin{array}{cccc:c} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right).$$

According to the lift method, we have: $t = 1$, $z = 0$, $y = 0$ *and* $x = 0$.

2. The matrix A have a Cholesky decomposition because first, the matrix A is symmetric

$$A = A^T \Leftrightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

In addition A is positive if

$$\det(A_{11}) = 1 > 0, \quad \det(A_{22}) = 1 > 0 \quad \textit{and} \quad \det(A_{33}) = 1 > 0,$$

then the matrix A admits a decomposition by Cholesky's method.

3. We write A in the form RR^T *such that R is a lower triangular matrix We have*

$$A = RR^T \Leftrightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix} = \begin{pmatrix} r_{11} & 0 & 0 & 0 \\ r_{21} & r_{22} & 0 & 0 \\ r_{31} & r_{32} & r_{33} & 0 \\ r_{41} & r_{42} & r_{43} & r_{44} \end{pmatrix} \begin{pmatrix} r_{11} & r_{21} & r_{31} & r_{41} \\ 0 & r_{22} & r_{32} & r_{42} \\ 0 & 0 & r_{33} & r_{43} \\ 0 & 0 & 0 & r_{44} \end{pmatrix}$$

$$= \begin{pmatrix} r_{11}^2 & r_{11}r_{21} & r_{11}r_{31} & r_{11}r_{41} \\ r_{11}r_{21} & r_{21}^2 + r_{22}^2 & r_{21}r_{31} + r_{22}r_{32} & r_{21}r_{41} + r_{22}r_{42} \\ r_{11}r_{31} & r_{21}r_{31} + r_{22}r_{32} & r_{31}^2 + r_{32}^2 + r_{33}^2 & r_{31}r_{41} + r_{32}r_{42} + r_{33}r_{43} \\ r_{11}r_{41} & r_{21}r_{41} + r_{22}r_{42} & r_{41}r_{31} + r_{42}r_{32} + r_{43}r_{33} & r_{41}^2 + r_{42}^2 + r_{43}^2 + r_{44}^2 \end{pmatrix}$$

so,

$$\begin{cases} r_{11}^2 = 1, \\ r_{11}r_{21} = 1, \\ r_{11}r_{31} = 1, \\ r_{11}r_{41} = 1, \\ r_{21}^2 + r_{22}^2 = 2, \\ r_{21}r_{31} + r_{22}r_{32} = 2, \\ r_{21}r_{41} + r_{22}r_{42} = 2, \\ r_{31}^2 + r_{32}^2 + r_{33}^2 = 3, \\ r_{31}r_{41} + r_{32}r_{42} + r_{33}r_{43} = 3, \\ r_{41}^2 + r_{42}^2 + r_{43}^2 + r_{44}^2 = 4. \end{cases}$$

so that,

$$r_{11}=1,\quad r_{21}=1,\quad r_{31}=1,\quad r_{41}=1,\quad r_{22}=1,\quad r_{32}=1.$$
$$r_{41}=1,\quad r_{33}=1,\quad r_{43}=1,\quad ,r_{44}=1.$$

we obtain,

$$R=\begin{pmatrix}1&0&0&0\\1&1&0&0\\1&1&1&0\\1&1&1&1\end{pmatrix}$$

4. We will solve the system by the Cholesky method
We have

$$AX=b\Rightarrow RR^TX=b\Rightarrow\begin{cases}RY=b & \Rightarrow & Y=? & (1)\\ R^TX=Y & \Rightarrow & X=? & (2)\end{cases}$$

firstly, we determine the solution system (1)

$$(1)\Leftrightarrow\begin{pmatrix}1&0&0&0\\1&1&0&0\\1&1&1&0\\1&1&1&1\end{pmatrix}\begin{pmatrix}x'\\y'\\z'\\t'\end{pmatrix}=\begin{pmatrix}1\\2\\3\\4\end{pmatrix}\Rightarrow\begin{cases}x'=1\\y'=1\\z'=1\\t'=1\end{cases}$$

secondly, we determine the solution system (2)

$$(2)\Leftrightarrow\begin{pmatrix}1&1&1&1\\0&1&1&1\\0&0&1&1\\0&0&0&1\end{pmatrix}\begin{pmatrix}x\\y\\z\\t\end{pmatrix}=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}\Rightarrow\begin{cases}x=0\\y=0\\z=0\\t=1\end{cases}$$

finally, the solution of linear system is obtained $X=(0,0,0,1)$.

Solution 3.3 *Let* $AX=b$ *be the linear system*

$$A=\begin{pmatrix}1&1&0\\1&5&2\\0&2&10\end{pmatrix},\qquad b=\begin{pmatrix}2\\8\\12\end{pmatrix},\qquad X=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}.$$

1. We solve the system $AX=b$ *by using Gaussian elimination.*

$$(A\vdots b)=\left(\begin{array}{ccc:c}1&1&0&2\\1&5&2&8\\0&2&10&12\end{array}\right)\sim\left(\begin{array}{ccc:c}1&1&0&2\\0&4&2&6\\0&2&10&12\end{array}\right)$$
$$\sim\left(\begin{array}{ccc:c}1&1&0&2\\0&4&2&6\\0&0&9&9\end{array}\right)$$
$$\Rightarrow\begin{cases}x_3=1,\\x_2=1,\\x_1=1.\end{cases}$$

2. The factorization $A=LU$, $l_{ii}=1$

$$U=\begin{pmatrix}1&1&0\\0&4&2\\0&0&9\end{pmatrix},\ L=\begin{pmatrix}1&0&0\\1&1&0\\0&1/2&1\end{pmatrix}.$$

3. *We have $A = A^t$, so A is symmetric, we have $A = LU = LDL^t$ such that*

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1/2 & 1 \end{pmatrix}, D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{pmatrix}.$$

4. *As $A = A^t$, $\det(A_{11}) = 2 > 0$ and $\det(A_{22}) = 5 > 0$, So A is symmetric positive definite then A has a Cholesky decomposition. Hence $A = LDL^t = (LD^{\frac{1}{2}})(D^{\frac{1}{2}}L^t) = RR^t$, then*

$$\begin{aligned} R = LD^{\frac{1}{2}} &= \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{1} & 0 & 0 \\ 0 & \sqrt{4} & 0 \\ 0 & 0 & \sqrt{9} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 3 \end{pmatrix}. \end{aligned}$$

5. *The Jacobi and Gauss-Seidel algorithms applied to $AX = b$*

$$(Jacobi) \begin{cases} x_1^{(k+1)} &= (2 - x_2^{(k)}), \\ x_2^{(k+1)} &= (8 - x_1^{(k)} - 2x_3^{(k)})/5, \\ x_3^{(k+1)} &= (12 - x_2^{(k)})/10. \end{cases} \quad X^{(0)} \in \mathbb{R}^3, \text{ given}$$

And

$$(Gauss - Seidel) \begin{cases} x_1^{(k+1)} &= (2 - x_2^{(k)}), \\ x_2^{(k+1)} &= (8 - x_1^{(k+1)} - 2x_3^{(k)})/5, \\ x_3^{(k+1)} &= (12 - x_2^{(k+1)})/10. \end{cases} \quad X^{(0)} \in \mathbb{R}^3, \text{ given}$$

3.6 Exercises without solutions

Exercise 3.1 *Consider the following linear system*

$$\begin{cases} 3x_1 - 2x_2 + x_3 &= 2 \\ 2x_1 + x_2 + x_3 &= 7 \\ 4x_1 - 3x_2 + 2x_3 &= 4 \end{cases}$$

1. *Solve this system by Gaussian method.*
2. *Factor the matrix A of the system into the product LU where L is a matrix trangular lower (with 1s on the main diagonal) and triangular U superior, then solve this system.*

Exercise 3.2 *Consider the linear system $AX = B$ where*

$$A = \begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 2 \\ 4 & -3 & 0 \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad \text{and } B = \begin{pmatrix} 2 \\ 5 \\ -1 \end{pmatrix}$$

Factoring the matrix A produces LU then solving the system (with $u_{ii} = 1$ such that $i = 1, 2, 3$.).

Exercise 3.3 *Let $\alpha, \beta \in \mathbb{R}$ be the system $AX = b$ such that:*

$$A = \begin{pmatrix} \alpha & \beta & 1 \\ 1 & \alpha & 1 \\ 1 & 1 & \alpha \end{pmatrix}, X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, b = \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix}.$$

1. *Which conditions on α and β for A possed a Cholesky decomposition*
2. *Posing $\alpha = 2$:*

(a) Write A in the form RR^T such that R is a lower triangular matrix.

(b) Solve this system by Cholesky's method.

(c) Deduce a factorization of A in the form LU with $l_{ii} = 1$; $i = 1, 2, 3$.

Exercise 3.4 *We consider the system (S) defined by*

$$(S)\begin{cases} 4x_1 + x_2 + x_3 = 9, \\ 2x_1 - 9x_2 = -16, \\ -6x_2 - 8x_3 = -34. \end{cases}$$

1. *Write the system (S) in matrix form $Ax = b$ and show that it has a unique solution.*
2. *Show that the Gauss-Seidel algorithm converges for this system.*
3. *Write the iteration algorithm using*
   a) Jacobi method, then calculate $X^{(1)}, X^{(2)}$ with $X^{(0)} = (0, 0, 0)$.
   b) Gauss-Seidel method, then calculate $X^{(1)}, X^{(2)}$ with $X^{(0)} = (0, 0, 0)$.
4. *Estimate the number of iterations needed to approximate the solution of this system to within* 0.001 *using the Jacobi (and Gauss-Seidel) method.*

Chapter 4

Interpolation and approximation polynomial

4.1 Introduction

Let f be an application from $\mathbb{R}$ in $\mathbb{R}$, of which we know $n+1$ points $(x_i, f(x_i))$, for $i = 0, 1, \cdots, n$. The purpose of the interpolation problem is to determine a simple p function to calculate, such as

$$p(x_i) = f(x_i), \qquad i = 0, 1, \cdots, n$$

The points $(x_i, f(x_i))$ are called interpolation or support points. The most commonly used p functions are polynomials of fractions rationals, sums of exponentials etc.
In this chapter, we have a function f known for example only by its values at certain points, and we try to replace or approximate f by a simpler function, most often by a polynomial.
Interpolation: consists in finding a polynomial which passes exactly through the given points.

4.2 Lagrange method

Let $f : [a, b] \longrightarrow \mathbb{R}$ known in $(n+1)$ distinct points x_0, x_1, $\cdots$, x_n of the interval $[a, b]$. It is a matter of constructing a polynomial P of degree less than or equal to n; this polynomial is given by

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x), \qquad \text{with} \quad L_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}.$$

Remark 4.1 *1. The polynomial P_n is called the Lagrange interpolation polynomial of the function f at the points x_0, x_1, $\cdots$, x_n.*

2. The polynomials $L_i(x)$ are called Lagrange basis polynomials associated with these points.

Example 4.2.1 *Let f be defined by*

x_i	1	2	3
$y_i = f(x_i)$	0	2	3

Correction 4.2.1 *From Lagrange's interpolation:*

$$\begin{aligned} L_0(x) &= \frac{(x-1)}{(0-1)} \times \frac{(x-2)}{(0-2)} = \frac{1}{2}(x-1)(x-2), \\ L_1(x) &= \frac{(x-0)}{(1-0)} \times \frac{(x-2)}{(1-2)} = -x^2 + 2x, \\ L_2(x) &= \frac{(x-0)}{(2-0)} \times \frac{(x-1)}{(2-1)} = \frac{1}{2}(x^2 - x), \end{aligned}$$

so,

$$\begin{aligned} P_2 = \sum_{i=0}^{2} f(x_i)L_i(x) &= 0\left[\frac{1}{2}(x-1)(x-2)\right] + 2(-x^2+2x) + \frac{3}{2}(x^2-x) \\ &= -\frac{1}{2}x^2 + \frac{5}{2}x. \end{aligned}$$

4.3 Divided differences method

The interpolation polynomial by the method of divided differences or **Newton's method** of the function f at the distinct points $x_0,\ x_1,\ \cdots,\ x_n$ is given by

$$P_n = \sum_{i=0}^{n} f[x_0, x_1, \cdots, x_i] \prod_{k=0}^{i-1}(x - x_k),$$

where $f[\cdot]$ denotes the divided differences of f defined by

$$\begin{aligned} f[x_i] &= f(x_i), \qquad i = 0,\ 1,\ 2, \cdots, n \\ \text{and } f[x_0, x_1, \cdots, x_k] &= \frac{f[x_1, x_1, \cdots, x_k] - f[x_0, x_1, \cdots, x_{k-1}]}{x_k - x_0}. \end{aligned}$$

Remark 4.2 *By convention* $\prod_{k=0}^{i-1}(x - x_k) = 1$ *if* $i \leq 1$.

Divided difference table: allows to compute inductively divided differences of a function according to the following scheme

x_i	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	$\cdots$	$\cdots$
x_0	$f[x_0]$					
		$f[x_0, x_1]$				
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$			
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$	$\cdots$	$\cdots$
x_2	$f[x_2]$		$f[x_1, x_2, x_3]$	$\vdots$		
		$f[x_2, x_3]$	$\vdots$			
x_3	$f[x_3]$	$\vdots$	$f[x_{n-2}, x_{n-1}, x_n]$			
$\vdots$	$\vdots$	$f[x_{n-1}, x_n]$				
x_n	$f[x_n]$					

where

$$\begin{aligned} f[x_0, x_1] &= \frac{f[x_1] - f[x_0]}{x_1 - x_0}, \\ f[x_0, x_1, x_2] &= \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}, \\ f[x_0, x_1, x_2, x_3] &= \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}. \end{aligned}$$

Example 4.3.1 *Let*

x_i	0	1	3
$y_i = f(x_i)$	0	2	8

Correction 4.3.1 *The table of divided differences is given as*

x_i	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$
0	*0*		
		2	
1	2		*1/3*
		3	
3	8		

Therefore

$$\begin{aligned} P_2(x) &= f[x_0] + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\ &= 0 + 2(x - 0) + \frac{1}{3}(x - 0)(x - 1) \\ &= 2x + \frac{1}{3}x^2 - \frac{1}{3}x \\ &= \frac{1}{3}x^2 + \frac{5}{3}x. \end{aligned}$$

4.3.1 Newton's formula using decreasing divided differences

Let $n + 1$ points in the interval $[a, b]$ such that $x_i = x_0 + ih$ where $i = 0, 1, \cdots, n$ with $h > 0$ and f function defines at x_0, x_1, $\cdots$, x_n by

$$f(x_i) = y_i, \qquad i = 0, \ldots, n.$$

so,

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\cdots$	$\cdots$
x_0	y_0					
		Δy_0				
x_1	y_1		$\Delta^2 y_0$			
		Δy_1		$\Delta^3 y_0$	$\cdots$	$\cdots$
x_2	y_2		$\Delta^2 y_1$	$\vdots$		
		Δy_2	$\vdots$	$\Delta^3 y_{n-3}$	$\cdots$	$\cdots$
x_3	y_3	$\vdots$	$\Delta^2 y_{n-2}$			
$\vdots$	$\vdots$	Δy_{n-1}				
x_n	y_n					

such as

$$\begin{aligned} \Delta y_0 &= y_1 - y_0, \\ \Delta y_i &= y_{i+1} - y_i, \end{aligned}$$

and

$$\begin{aligned} \Delta^2 y_0 &= \Delta y_1 - \Delta y_0, \\ \Delta^3 y_0 &= \Delta^2 y_1 - \Delta^2 y_0. \end{aligned}$$

The interpolating polynomial of f defined by

$$\begin{aligned} P_n(x) = y_0 + (x - x_0)\frac{\Delta y_0}{1!h} &+ (x - x_0)(x - x_1)\frac{\Delta^2 y_0}{2!h^2} \\ &+ \cdots + (x - x_0)(x - x_1)\cdots(x - x_{n-2})(x - x_{n-1})\frac{\Delta^n y_0}{n!h^n}. \end{aligned}$$

Example 4.3.2 *We define f by the following array*

x_i	4	6	8
y_i	1	3	8

Find P_2 and calculate $f(5)$.

Correction 4.3.2 *We have*

x_i	$y_i = f(x_i)$	Δy_i	Δy_i
4	$\boxed{1}$		
		$\boxed{2}$	
6	3		$\boxed{3}$
		5	
8	8		

As $h = 2$, we obtain

$$\begin{aligned} P_2(x) &= y_0 + (x - x_0)\frac{\Delta y_0}{1!h} + (x - x_0)(x - x_1)\frac{\Delta^2 y_0}{2!h^2} \\ &= 1 + (x - 4)\frac{2}{2} + (x - 4)(x - 6)\frac{3}{2!2^2}, \end{aligned}$$

so,

$$P_2(x) = \frac{3}{8}x^2 - \frac{11}{4}x + 6.$$

Hence

$$f(5) \approx P_2(5) = \frac{3}{8}(5)^2 - \frac{11}{4}(5) + 6.$$

4.3.2 Newton's formula using increasing divided differences

Let $n + 1$ points in the interval $[a, b]$ such that $x_i = x_0 + ih$ where $i = 0, 1, \cdots, n$ and f function defined at $x_0, x_1, \cdots, x_n$ by

$$f(x_i) = y_i, \qquad i = 0, \cdots, n.$$

Here is the following table

x_i	y_i	∇y_i	$\nabla^2 y_i$	$\nabla^3 y_i$	$\cdots$	$\cdots$
x_0	y_0					
		∇y_1				
x_1	y_1		$\nabla^2 y_2$			
		∇y_2		$\nabla^3 y_3$	$\cdots$	$\cdots$
x_2	y_2		$\nabla^2 y_3$	$\vdots$		
		∇y_3	$\vdots$	$\nabla^3 y_n$	$\cdots$	$\cdots$
x_3	y_3	$\vdots$	$\nabla^2 y_n$			
$\vdots$	$\vdots$	∇y_n				
x_n	y_n					

where

$$\begin{aligned} \nabla y_1 &= y_1 - y_0, \\ \nabla y_{i+1} &= y_{i+1} - y_i, \end{aligned}$$

and

$$\begin{aligned} \nabla^2 y_1 &= \nabla y_1 - \nabla y_0, \\ \nabla^3 y_1 &= \nabla^2 y_1 - \nabla^2 y_0. \end{aligned}$$

The interpolating polynomial of f defined by

$$\begin{aligned} P_n(x) = y_n + (x-x_n)\frac{\nabla y_n}{1!h} &+ (x-x_n)(x-x_{n-1})\frac{\nabla^2 y_n}{2!h^2} \\ &+ \cdots + (x-x_n)(x-x_{n-1})\cdots(x-x_1)(x-x_0)\frac{\nabla^n y_0}{n!h^n} \end{aligned}$$

Example 4.3.3 *We define f by the following table*

x_i	4	6	8
y_i	1	3	8

Find P_2 and calculate $f(5)$.

Correction 4.3.3 *We have*

x_i	$y_i = f(x_i)$	∇y_i	∇y_i
4	1		
		2	
6	3		$\boxed{3}$
		$\boxed{5}$	
8	$\boxed{8}$		

As $h = 2$ we obtain

$$\begin{aligned} P_2(x) &= y_2 + (x-x_2)\frac{\nabla y_2}{1!h} + (x-x_2)(x-x_1)\frac{\nabla^2 y_2}{2!h^2} \\ &= 8 + (x-8)\frac{5}{2} + (x-8)(x-6)\frac{3}{2!2^2}, \end{aligned}$$

so,

$$P_2(x) = \frac{3}{8}x^2 - \frac{11}{4}x + 6.$$

Hence

$$f(5) \approx P_2(5) = \frac{3}{8}(5)^2 - \frac{11}{4}(5) + 6.$$

4.4 The Error Study

The purpose of the interpolation being to replace the evaluation of $f(x)$ by that of $P_n(x)$, it is important to know the error

$$E_n(x) = f(x) - P_n(x), \qquad x \in [a,b].$$

Theorem 4.1 *Let $f : [a,b] \longrightarrow \mathbb{R}$, $n+1$ times continuously differentiable and P_n the Lagrange interpolation polynomial at the point $x_0, x_1, \cdots, x_n$ of $[a,b]$, then*

$$\exists \alpha \in [a,b] : \qquad |f(x) - P_n(x)| = \frac{f^{(n+1)}(\alpha)}{(n+1)!}\prod_{i=0}^{n}(x-x_i).$$

Remark 4.3 *If f unknown, we cannot specify the error exactly, but we can find an approximate error if M_{n+1} known such that*

$$M_{n+1} = \max_{a \le x \le b} |f^{(n+1)}(x)|,$$

and we have

$$|f(x) - P_n(x)| \le \frac{M_{(n+1)}}{(n+1)!}\prod_{i=0}^{n}(x-x_i), \tag{4.4.1}$$

Example 4.4.1 *Let $f(x) = e^x$, for $x \in [a, b]$, we have $M_{n+1} = e^b$. On the other hand, we have*

$$\prod_{i=0}^{n}(x - x_i) \leq (b - a)^{n+1},$$

for any choice of $n + 1$ points x_i. From the estimate (4.4.1)

$$\forall x \in [a, b]; \qquad |f(x) - P_n(x)| \leq \frac{(b - a)^{n+1}}{(n + 1)!}e^b.$$

In particular

$$\lim_{n \longrightarrow +\infty} |f(x) - P_n(x)| = 0,$$

therefore, in the case of the exponential function, the more points taken, the better the interpolation.

Remark 4.4 *The interpolation error $E(x) = f(x) - P_n(x)$, $x \in [a, b]$ $(a \leq x_0 \leq \ldots \leq x_n \leq b)$, using divided differences, whatever the formula for P_n, since it is unique is given by*

$$E(x) = f[x_0, x_1, \ldots, x_n, x](x - x_0)(x - x_1) \cdots (x - x_n).$$

This general formula can be modified when the function f is $(n+1)$ times continuously differentiable on the interval $[a, b]$, which is the smallest interval containing the x_i, $i = 0, ..., n$, $f \in C^{n+1}([a, b])$.

4.5 Hermite's method

Hermite's interpolation is a generalization of Lagrange's interpolation by making not only f and P_N coincide at nodes x_i. let x_0, x_1, $\cdots$, x_n, $(n+1)$ distinct points of the interval $[a, b]$ and $y = f(x)$ a function defined on the same interval adding the derivatives (y_0, y_0'), (y_1, y_1'), $\cdots$, (y_n, y_n'). In this case, there is one and only one polynomial such that $P_N(x_i) = y_i$ and $P_N'(x_i) = y_i'$ with $N = 2n+1$ and $i = 0$, 1, $2, \cdots$, n. The polynomial is given by

$$P_N(x) = \sum_{i=0}^{n} H_i(x)f(x_i) + \sum_{i=0}^{n} K_i(x)f'(x_i),$$

where

$$\begin{cases} H_i(x) &= [1 - 2(x - x_i)L_i'(x_i)]L_i^2(x), \\ K_i(x) &= (x - x_i)L_i^2(x), \\ L_i(x) &= \prod\limits_{j=0, j\neq i}^{n} \left(\dfrac{x - x_j}{x_i - x_j}\right). \end{cases}$$

Hermite interpolation error Let $f \in C^{(2n+2)}([a, b])$, and $P_N(x)$ be the Hermite interpolation polynomial of f on the points $(x_i, f(x_i))$, for $i = 0, \cdots , n$. For all $x \in [a, b]$, there exists $\xi \in [a, b]$ such that the error $f(x) - P_N(x)$ is

$$E(x) = \frac{(\gamma_{n+1}(x))^2}{(2n + 2)!} f^{(2n+2)}(\xi),$$

where $\gamma_{n+1}(x) = \prod\limits_{j=0}^{n}(x - x_j)$.

Example 4.5.1 *Calculate the Hermite polynomial Q such that*

$$Q(0) = f(0),\ Q'(0) = f'(0),\ Q(5) = f(5),\ Q'(5) = f'(5),$$

and

$$f(x) = \frac{1}{1 + x^2}.$$

Deduce the value of $Q(4)$, compare $f(4)$ to $Q(4)$.

Correction 4.5.1 *We have*

x_i	0	5
$f(x_i)$	1	$\frac{1}{26}$
$f(x_i)$	0	$-\frac{5}{338}$

For $i = 0, 1$ we obtain

$$\begin{aligned} L_0(x) &= \frac{x-x_1}{x_0-x_1} = \frac{x-5}{(0-5)} = -\frac{1}{5}(x-5) = -\frac{x}{5}+1 \Longrightarrow L_0'(x) = -\frac{1}{5}, \\ L_1(x) &= \frac{x-x_0}{x_1-x_0} = \frac{x-0}{(5-0)} = \frac{1}{5}(x) = \frac{x}{5} \Longrightarrow L_1'(x) = \frac{1}{5}, \end{aligned}$$

then

$$\begin{aligned} H_0(x) &= [1-2(x-x_0)L_0'(x)]L_0^2(x) \\ &= \left[1-2(x-0)\Big(-\frac{1}{5}\Big)\right]\left(1+\frac{x^2}{25}-\frac{2x}{5}\right) \\ &= \left(1+\frac{2x}{5}\right)\left(\frac{x^2}{25}-\frac{2x}{5}+1\right) \\ &= \frac{x^2}{25}-\frac{2x}{5}+1+\frac{2x^3}{125}-\frac{4x^2}{25}+\frac{2x}{5} \\ &= \frac{2}{125}x^3-\frac{3}{25}x^2+1, \end{aligned}$$

and

$$\begin{aligned} H_1(x) &= [1-2(x-x_1)L_1'(x)]L_1^2(x) \\ &= \left[1-2(x-5)\Big(\frac{1}{5}\Big)\right]\left(\frac{x^2}{25}\right) \\ &= \left(1-\frac{2}{5}x+2\right)\frac{x^2}{25} \\ &= \frac{x^2}{25}-\frac{2x^3}{125}+\frac{2x^2}{25} \\ &= -\frac{2}{125}x^3+\frac{3}{25}x^2, \end{aligned}$$

after that

$$\begin{aligned} K_0(x) &= (x-x_0)L_0^2(x) = x\left(1-\frac{x}{5}\right)^2 = x-\frac{2x^2}{5}+\frac{x^3}{25}, \\ K_1(x) &= (x-x_1)L_1^2(x) = (x-5)\left(\frac{x}{5}\right)^2 = -\frac{x^2}{5}+\frac{x^3}{25}, \end{aligned}$$

finally

$$\begin{aligned} Q(x) &= \sum_{i=0}^{1} H_i(x)f(x_i) + \sum_{i=0}^{1} K_i(x)f'(x_i) \\ &= H_0(x)f(x_0) + H_1(x)f(x_1) + K_0(x)f'(x_0) + K_1(x)f'(x_1) \\ &= 1\left(\frac{2x^3}{125}-\frac{3x^2}{25}+1\right) + \frac{1}{26}\left(-\frac{2x^3}{125}+\frac{3x^2}{25}\right) + \frac{5}{338}\left(\frac{x^3}{25}+\frac{x^2}{5}\right) \\ &= \left(\frac{2}{125}-\frac{2}{26\times 125}-\frac{5}{25\times 338}\right)x^3 + \left(-\frac{3}{25}+\frac{3}{26\times 25}+\frac{5}{338\times 5}\right)x^2 \\ &= \frac{5}{338}x^3-\frac{19}{169}x^2+1. \end{aligned}$$

We deduce the value $Q(4) = 0.1479$ and we have $f(4) = \frac{1}{17} \approx 0.0588$ so that $|E| = |Q(4) - f(4)| = 0.0891$.

4.6 Exercises with solutions

Exercise 4.1 *Let $f(x) = 3^x$, for all $x \in \mathbb{R}$ and let $p(x)$ be the interpolation polynomial associated with the function f at the points $x_0 = 0$, $x_1 = 1$, and $x_2 = 2$.*

1. *Construct $p(x)$ by Lagrange's method.*
2. *Construct $p(x)$ by Newton's method.*
3. *Find a bound of the error $|f(x) - p(x)|$ uniform for all $x \in [0, 2]$.*

Exercise 4.2 *Let f be a function whose value is known at certain points x_0, x_1, x_2 and x_3 as indicated to the table of divided differences below*

x_i	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$
0	0			
		$\frac{1}{4}$		
1	$\frac{1}{4}$		α	
		$\frac{1}{2}$		γ
2	$\frac{3}{4}$		β	
		$\frac{1}{4}$		
3	1			

1. *Calculate α, β and γ.*
2. *Using this table, calculate an approximation of $f(1.5)$ using the interpolation polynomial at points x_0, x_1, x_2 and x_3.*
3. *Knowing that $\sup_{x\in[0,3]} f^{(4)}(x) \leq \dfrac{\pi^4}{162}$, give an increase in value absolute of the interpolation error in $x = 1.5$.*

Exercise 4.3 *Let the points $(-1, \alpha)$, $(0, \beta)$ and $(1, \alpha)$ where α and β are reals.*

1. a) *Determine the Lagrange polynomial P which interpolates the three points $(-1, \alpha)$, $(0, \beta)$ and $(1, \alpha)$.*
   b) *If $\alpha = \beta$, give the degree of the polynomial P.*
   c) *Show that P is even. Can we have P of degree 1?.*
2. *Let $\alpha = -1$ and $\beta = 1$, determine the interpolation polynomial P by Newton's method.*

Exercise 4.4 *Let $f(x) = \frac{1}{1+x^2}$, $\quad x \in [0, 1]$.*

1. *Determine the Lagrange interpolation polynomial of f at points $x_0 = 0$ and $x_1 = 1$*
2. *Determine the Hermite interpolation polynomial of f at points $x_0 = 0$ and $x_1 = 1$*
3. *Deduce two approximate values of $f(\frac{1}{2})$. Compare the results obtained and conclude.*

4.6.1 Solutions

Solution 4.1 1. *Lagrange method: We have $f(x) = 3^x$, $x \in [0, 2]$ so:*

x_i	0	1	2
$y_i = f(x_i)$	1	3	9

Let us calculate L_i, $i = 0, 1, 2$.

$$\begin{aligned} L_0(x) &= \frac{(x-1)(x-2)}{(0-1)(0-2)} = \frac{1}{2}(x^2 - 3x + 2), \\ L_1(x) &= \frac{(x-0)(x-2)}{(1-0)(1-2)} = 2x - x^2, \\ L_2(x) &= \frac{(x-0)(x-1)}{(2-0)(2-1)} = \frac{1}{2}x^2 - \frac{1}{2}x, \end{aligned}$$

so,

$$\begin{aligned} p(x) &= y_0L_0(x) + y_1L_1(x) + y_2L_2(x) \\ &= 1 \times \frac{1}{2}(x^2 - 3x + 2) + 3(2x - x^2) + 9\left(\frac{1}{2}x^2 - \frac{1}{2}x\right) \\ &= 2x^2 + 1. \end{aligned}$$

2. *Newton's method*

   The divided difference table is

x_i	$f[x_i]$	$f[x_i, x_j]$	$f[x_0, x_1, x_2]$
0	$f[x_0] = 3^0 = 1$		
		$f[x_0, x_1] = \dfrac{f[x_1] - f[x_0]}{x_1 - x_0} = 2$	
1	$f[x_1] = 3^1 = 3$		$f[x_0, x_1, x_2] = \dfrac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = 2$
		$f[x_1, x_2] = \dfrac{f[x_2] - f[x_1]}{x_2 - x_1} = 6$	
2	$f[x_2] = 3^2 = 9$		

   It follows that $p(x) = 1 + 2x + 2x(x-2) = 2x^2 + 1$.

3. *We have* $E(x) = |f(x) - p(x)| = \dfrac{f^(3)(\xi)}{3!}x(x-1)(x-2)$ *such that* $\xi \in [0,1]$.
   We have $f'(x) = \ln(3)3^x$, $f''(x) = (\ln(3))^2 3^x$ *and* $f^{(3)}(x) = (\ln(3))^3 3^x$ *as* $f^{(4)}(x) = (\ln(3))^4 3^x$ *then* $|f^{(3)}(x)| \leq 9(\ln(3))^3$, $\forall x \in [0,2]$.
   Moreover consider the function $g(x) = x(x-1)(x-2)$ *on* $[0,2]$ *then* $g'(x) = 3x^2 - 6x + 2$.
   Hence $g'(x) = 0 \Leftrightarrow 3x^2 - 6x + 2 = 0 \Leftrightarrow x_1 = \dfrac{1}{3}\sqrt{3} + 1 \in [0,2]$ *and* $x_2 = 1 - \dfrac{1}{3}\sqrt{3} \in [0,2]$.

   We then deduce $\max_{x\in[0,2]} |g(x)| = \max\{|g(x_1)|, |g(x_2)|\} = \dfrac{2}{9}\sqrt{3}$.
   In conclusion $\forall x \in [0,2]$ $|f(x) - p(x)| \leq \dfrac{1}{3}\sqrt{3}(\ln(3))^3$.

Solution 4.2 1. *We calculate* α, β *and* γ.

$$\begin{aligned} \alpha &= f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{1/2 - 1/4}{2 - 0} = \frac{1}{8} \\ \beta &= f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1} = \frac{1/4 - 1/2}{3 - 1} = -\frac{1}{8} \\ \gamma &= f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0} = \frac{-1/8 - 1/8}{3 - 0} = -\frac{1}{12} \end{aligned}$$

2. *Using this table, we calculate an approximation of* $f(1.5)$ *using the interpolation polynomial at points* x_0, x_1, x_2 *and* x_4.

$$\begin{aligned} P_3(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &\quad + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) \\ &= 0 + \frac{1}{4}(x - 0) + \frac{1}{8}(x - 0)(x - 1) - \frac{1}{12}(x - 0)(x - 1)(x - 2) \\ &= -\frac{1}{12}x^3 + \frac{3}{8}x^2 - \frac{1}{24}x, \end{aligned}$$

where

$$f(1.5) = f(\frac{3}{2}) \simeq P_3(\frac{3}{2}) = -\frac{1}{12}(\frac{3}{2})^3 + \frac{3}{8}(\frac{3}{2})^2 - \frac{1}{24}(\frac{3}{2}) = \frac{7}{32}.$$

3. *We give an increase of the absolute value of the interpolation error in* $x = 1.5$, *as* $\sup_{x\in[0,3]} f^{(4)}(x) \leq \frac{\pi^4}{162}$ *and we have*

$$E(x) = |f(x) - P_3(x)| \leq \frac{\sup_{x\in[0,3]} f^{(4)}(x)}{4!} \prod_{i=0}^{3}(x - x_i),$$

then

$$\begin{aligned} |E(1.5)| = \big|E(\frac{3}{2})\big| &= \frac{\pi^4}{(162)(4\times 3\times 2\times 1)}(\frac{3}{2}-0)(\frac{3}{2}-1)(\frac{3}{2}-2)(\frac{3}{2}-3) \\ &= \frac{\pi^4}{(162)(24)}(\frac{3}{2})(\frac{1}{2})(-\frac{1}{2})(-\frac{3}{2}) \\ &\leq \frac{\pi^4}{6912}. \end{aligned}$$

Solution 4.3 *Let the points* $(-1,\alpha)$, $(0,\beta)$ *and* $(1,\alpha)$ *where* α *and* β *are reals.*
We put

$$x_0 = -1, \quad x_1 = 0, \quad x_2 = 1, \quad y_0 = \alpha, \quad y_1 = \beta, \quad y_2 = \alpha.$$

1. *a) In Largange form the polynomial* P *is written*

$$P(x) = y_0L_0(x) + y_1L_1(x) + y_2L_2(x),$$

such that

$$\begin{aligned} L_0(x) &= \frac{(x-x1)(x-x2)}{(x0-x1)(x0-x2)} = \frac{x(x-1)}{(-1)(-2)} = \frac{1}{2}x(x-1), \\ L_1(x) &= \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x+1)(x-1)}{(1)(-1)} = -(x+1)(x-1), \\ L_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{x(x+1)}{(2)(1)} = \frac{1}{2}x(x+1), \end{aligned}$$

hence

$$P(x) = \frac{1}{2}\alpha x(x-1) - \beta(x+1)(x-1) + \frac{1}{2}\alpha x(x+1)) = (\alpha-\beta)x^2 + \beta.$$

b) If $\alpha = \beta$, $P(x) = \beta$, *which is a polynomial of degree* 0.

c) For all $x \in \mathbb{R}$ *we have* $P(-x) = P(x)$ *so* P *is even. The polynomial* P *cannot be of degree* 1 *because a polynomial of degree* 1 *is of the form* $a_0 + a_1x$ *which cannot be even.*

2. *Newton's Table*

x_i	y_i	Δy_i	$\Delta^2 y_i$
-1	-1		
		2	
0	1		-4
		-2	
1	-1		

The interpolation polynomial at points x_0, x_1 *and* x_2 *with* $h = 1$.

$$\begin{aligned} P_2(x) &= y_0 + \frac{\Delta y_0}{1!h}(x-x_0) + \frac{\Delta^2 y_0}{2!h^2}(x-x_0)(x-x_1) \\ &= -1 + 2(x+1) + \frac{-4}{2}(x+1)(x-0) \\ &= -1 + 2x + 2 - 2x^2 - 2x \\ &= -2x^2 + 1. \end{aligned}$$

Solution 4.4 *We have* $f(x) = \frac{1}{1+x^2}$, $x \in [0,1]$. *For* $i = 0, 1$ *we obtain*

x_i	0	1
$f(x_i)$	1	$\frac{1}{2}$
$f(x_i)$	0	$-\frac{1}{2}$

1. *We determine the Lagrange interpolation polynomial of* f *at points* $x_0 = 0$ *and* $x_1 = 1$, *we have*

$$\begin{aligned} L_0(x) &= \frac{x - x_1}{x_0 - x_1} = \frac{x-1}{0-1} = -x + 1, \\ L_1(x) &= \frac{x - x_0}{x_1 - x_0} = \frac{x-0}{1-0} = x, \end{aligned}$$

we get

$$\begin{aligned} P_1(x) &= L_0(x)f(x_0) + L_1(x)f(x_1) \\ &= 1(1-x) + \frac{x}{2} \\ &= -\frac{1}{2}x + 1. \end{aligned}$$

2. *We determine the Hermite interpolation polynomial of* f *at points* $x_0 = 0$ *and* $x_1 = 1$, *we have*

$$\begin{aligned} L_0(x) = -x + 1 &\implies L_0'(x) = -1, \\ L_1(x) = x &\implies L_1'(x) = 1, \end{aligned}$$

then

$$\begin{aligned} H_0(x) &= (1 - 2(x - x_0)L_0'(x))L_0^2(x) \\ &= (1 - 2(x-0)(-1))(x^2 + 1 - 2x) \\ &= 2x^3 - 3x^2 + 1 \end{aligned}$$

and

$$\begin{aligned} H_1(x) &= (1 - 2(x - x_1)L_1'(x))L_1^2(x) \\ &= (1 - 2(x-1)(1))(x^2) \\ &= -2x^3 + 3x^2, \end{aligned}$$

after that

$$\begin{aligned} K_0(x) &= (x - x_0)L_0^2(x) = x(1 + x^2 - 2x) = x^3 - 2x^2 + x, \\ K_1(x) &= (x - x_1)L_1^2(x) = (x-1)x^2 = x^3 - x^2, \end{aligned}$$

we obtain

$$\begin{aligned} Q_3(x) &= H_0(x)f(x_0) + H_1(x)f(x_1) + K_0(x)f'(x_0) + K_1(x)f'(x_1) \\ &= 2x^3 - 3x^2 + 1 + \frac{1}{2}(3x^2 - 2x^3) - \frac{1}{2}(x^3 - x^2). \end{aligned}$$

3. *We deduce two approximate values of* $f(\frac{1}{2})$. *Compare the results obtained and conclude*

$$\begin{aligned} f(\frac{1}{2}) &= \frac{4}{5} = 0.8, \\ P_1(\frac{1}{2}) &= 0.75, \\ Q_3(\frac{1}{2}) &= 0.8125, \end{aligned}$$

Hermite's method is more precise.

4.7 Exercises without solutions

Exercise 4.1 *Let f be defined by the table below*

x	0	2	3	5
$f(x)$	-1	2	9	87

1. *Determine the interpolation polynomial of f by the Lagrange method.*
2. *Construct the divided difference table of f and deduce its interpolation polynomial.*

Exercise 4.2 *Let $f(x) = \sin(\pi x)$, $x \in [0,1]$ and P be the interpolation polynomial of f in points $\frac{1}{6}, \frac{1}{2}, \frac{5}{6}$ over $[0,1]$.*

1. *Determine P using*
   (a) *The Lagrange method.*
   (b) *Newton's method (or divided differences).*
2. *Give the estimate of the interpolation error $E(x) = |f(x) - P(x)|$, $x \in [0,1]$.*

Exercise 4.3 *Let f be defined by*

$$f(x) = \frac{1}{1+x^2}, \qquad x \in [-2,+2].$$

1. *Determine the Lagrange interpolation polynomial for the support points abscissa: $-2,-1,0,1,2$.*
2. *Determine the interpolation polynomial by Newton's method (Divided difference method with constant step).*
3. *Discuss the interpolation error.*

Exercise 4.4 • *How accurately can we calculate $\sqrt{115}$ using Lagrange interpolation, if we take the points $x_0 = 100$, $x_1 = 121$, $x_2 = 144$.*

Exercise 4.5 *Let $f(x) = \frac{1}{1+x^2}$, $x \in [-5,5]$.*

1. a) *Compute the Hermite polynomial Q such that $Q(0) = f(0), Q'(0) = f'(0), Q(4) = f(4), Q'(4) = f'(4)$.*
   b) *Deduce the value of $Q(4)$, compare $f(3)$ to $Q(3)$.*
2. a) *Determine the Lagrange interpolation polynomial of f at points $x_0 = 0$ and $x_1 = 2$.*
   b) *Determine the Hermite interpolation polynomial of f at points $x_0 = 0$ and $x_1 = 2$.*
   c) *Deduce two approximate values of $f(\frac{3}{2})$. Compare the results obtained and conclude.*

Chapter 5

Numerical integration

5.1 Introduction

We want to evaluate the integral of a function f over an interval $[a, b]$ in $\mathbb{R}$. If we know its primitive F, then

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

But in many cases F cannot be known.

Example 5.1.1 *Let* $\int_0^1 e^{-x^3} dx$ *and* $\int_a^b \frac{\cos(x)}{x} dx$.

We propose to evaluate the integral $\int_a^b f(x)dx$ by subdividing the integration interval

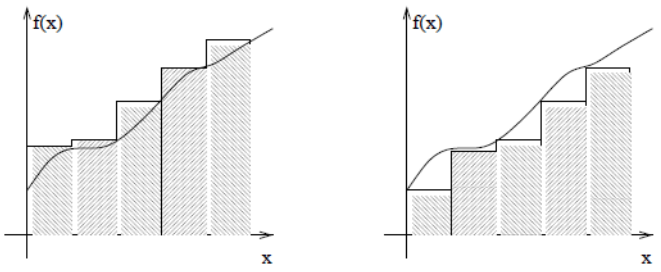
$$a = x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n = b,$$

and approximating f on each interval by a finite sum of the form

$$\int_a^b f(x)dx \quad \simeq \quad \sum_{i=0}^{n-1} a_i f(x_i).$$

5.2 Rectangle method

In the method of rectangles, we replace the function to be integrated f by a piecewise constant function $h(x)$ on each elementary interval $[x_i, x_{i+1}]$



- **Rectangles on the left,** we have $h(x) = f(x_i)$ for $x \in [x_i, x_{i+1}]$

$$\begin{aligned}
\int_a^b f(x)dx &= \int_{a=x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{n-1}}^{x_n=b} f(x)dx \\
&= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx \\
&\approx \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} h(x)dx \\
&= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x_i)dx \\
&= \sum_{i=0}^{n-1} f(x_i) \int_{x_i}^{x_{i+1}} dx \\
&= \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i),
\end{aligned}$$

  hence,

$$\int_a^b f(x)dx = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i).$$

- **Rectangles on the right,** we have $h(x) = f(x_{i+1})$ for $x \in [x_i, x_{i+1}]$

$$\begin{aligned}
\int_a^b f(x)dx &= \int_{a=x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{n-1}}^{x_n=b} f(x)dx \\
&= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx \\
&\approx \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} h(x)dx \\
&= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x_{i+1})dx \\
&= \sum_{i=0}^{n-1} f(x_{i+1}) \int_{x_i}^{x_{i+1}} dx \\
&= \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_{i+1}),
\end{aligned}$$

  hence,

$$\int_a^b f(x)dx = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_{i+1}).$$

- We note h the step of this subdivision i.e.

$$\forall i = 0, \cdots, n-1, \qquad h = x_{i+1} - x_i \qquad \text{and} \qquad x_i = a + ih, \qquad \text{with} \qquad h = \frac{b-a}{n},$$

  we have

$$\int_a^b f(x)dx = h \sum_{i=0}^{n-1} f(x_i),$$

or

$$\int_a^b f(x)dx = h\sum_{i=0}^{n-1} f(x_{i+1}).$$

- The error in the method of rectangles is given by the expression

$$\left|\int_a^b f(x)dx - h\sum_{i=0}^{n-1} f(a+ih)\right| \leq \frac{1}{2}\frac{(b-a)^2}{n}\sup_{x\in[a,b]} f'(x).$$

Remark 5.1 *Note that T_l for numerical integrate by rectangles on the left and T_r for numerical integrate by rectangles on the right.*

Example 5.2.1 *Let*

$$I = \int_1^2 e^x dx.$$

1. *Determine I by the method of rectangles on the left then on the right for $h = 0.1$ (step) "$n = 10$" then find the errors.*
2. *Find n for error less than $\epsilon = 10^{-4}$.*

Correction 5.2.1 1. *We have $h = \frac{2-1}{10} = \frac{1}{10} = 0.1$, so by the left rectangles method*

$$\begin{aligned} I &= \int_1^2 e^x dx \\ &\approx \frac{1}{10}\sum_{i=0}^{9} e^{1+0.1i} \\ &\approx \frac{e^1}{10}\sum_{i=0}^{9}(e^{0.1})^i \\ &\approx \frac{e^1}{10}\frac{e^1-1}{e^{0.1}-1} \\ &\approx 4.441127 = T_l. \end{aligned}$$

And the method of rectangles on the right

$$\begin{aligned} I &= \int_1^2 e^x dx \\ &\approx \frac{1}{10}\sum_{i=0}^{9} e^{1+0.1(i+1)} \\ &\approx \frac{e^{1.1}}{10}\sum_{i=0}^{9}(e^{0.1})^i \\ &\approx \frac{e^{1.1}}{10}\frac{e^1-1}{e^{0.1}-1} \\ &\approx 4.908204 = T_r. \end{aligned}$$

The errors

$$\begin{aligned} E_g &= \left|\int_1^2 e^x dx - T_l\right| \approx \left|e^2 - e^1 - \frac{e^1}{10}\frac{e^1-1}{e^{0.1}-1}\right| = 0.229647. \\ E_d &= \left|\int_1^2 e^x dx - T_r\right| \approx \left|e^2 - e^1 - \frac{e^{1.1}}{10}\frac{e^1-1}{e^{0.1}-1}\right| = 0.23743. \end{aligned}$$

2. *We find n for the error $\epsilon \leq 10^{-4}$, so*

$$\begin{aligned}\left|\int_1^2 e^x dx - \frac{1}{n}\sum_{i=0}^{n-1} f(x_i)\right| &\leq \frac{1}{2}\frac{(2-1)^2}{n}\sup_{x\in[1,2]} f'(x)\\ &\Leftrightarrow \frac{1}{2}\frac{(2-1)^2}{n}\sup_{x\in[1,2]} f'(x) \leq 10^{-4}\\ &\Leftrightarrow \frac{1}{n} < \frac{2\times 10^{-4}}{e^2}\\ &\Leftrightarrow n > \frac{e^2}{2}10^4\\ &\Leftrightarrow n = \left[\frac{e^2}{2}10^4\right] + 1,\end{aligned}$$

we can take $n = 36946$.

5.3 Trapeze method

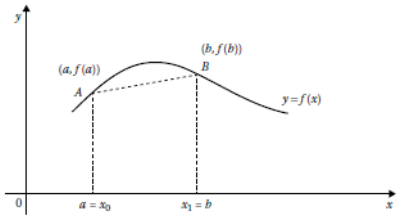
Let f be a continuous function on $[a,b]$, differentiable on $]a,b[$ and $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ a regular subdivision of the interval $[a,b]$ i.e, $x_{i+1} = x_i + h$ with $i = 1, \cdots, n$ and we have $h = \frac{b-a}{n}$ the step of this subdivision.

To numerically evaluate $I = \int_a^b f(x)\,dx$, using the trapezium method, we use the simple trapezium formula

$$S = \frac{Height}{2} \times (Small\_base + Large\_base).$$

In the simple case $n = 1, x_0 = a,\ x_1 = b$, we get

$$\int_a^b f(x)\,dx \simeq \frac{b-a}{2}\left(f(a) + f(b)\right).$$



We have the small base and large base correspond to $f(a)$ and $f(b)$ and height h such that $h = b-a$.

Example 5.3.1 *Let the integrals*

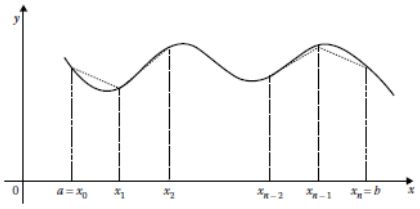
$$I = \int_0^4 x dx,\ J = \int_a^b x dx,\ K = \int_{\frac{\pi}{2}}^{\pi} \frac{\sin x}{x} dx.$$

Correction 5.3.1 *We have*

$$\begin{aligned} I &= \int_0^4 x dx \simeq \frac{4-0}{2}\left(f(0) + f(4)\right) = \frac{4-0}{2}(0+4) = 8.\\ J &= \int_a^b x dx \simeq \frac{b-a}{2}\left(f(a) + f(b)\right) = \frac{b-a}{2}(a+b).\\ K &= \int_{\frac{\pi}{2}}^{\pi} \frac{\sin x}{x} dx \simeq \frac{\pi - \frac{\pi}{2}}{2}\left(\frac{\sin\frac{\pi}{2}}{\frac{\pi}{2}} + \frac{\sin\pi}{\pi}\right) = \frac{\pi}{4}\left(\frac{2}{\pi}\right) = \frac{1}{2}.\end{aligned}$$

5.3.1 Generalized trapezium method

The area I included between $[a,b]$ and the graph of f can be be approximated by the sum of the areas of the n trapezoids induced by the points $x_0, x_1, ..., x_n$.



In the trapezium method, the function f is replaced on each interval $[x_i, x_{i+1}]$ with $0 \leq i \leq n-1$ by the line joining the points $(x_i, f(x_i))$ and $(x_{i+1}, f(x_{i+1}))$, then on each interval $[x_i, x_{i+1}]$

$$\int_{x_i}^{x_{i+1}} f(x)\,dx \simeq \frac{(x_{i+1}-x_i)}{2}\left(f(x_i)+f(x_{i+1})\right),$$

and generally, we have

$$\begin{aligned} I &= \int_a^b f(x)\,dx = \int_{x_0=a}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + ... + \int_{x_{n-1}}^{x_n=b} f(x)\,dx \\ &= \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)\,dx \\ &\approx \frac{h}{2}\sum_{i=0}^{n-1}\left(f(x_i)+f(x_{i+1})\right) \\ &= \frac{h}{2}\left(f(x_0)+f(x_1)\right)+\left(f(x_1)+f(x_2)\right)+\left(f(x_2)+f(x_3)\right)+...+\left(f(x_{n-1})+f(x_n)\right) \\ &= \frac{h}{2}\left(f(x_0)+2f(x_1)+2f(x_2)+...+2f(x_{n-1})+f(x_n)\right), \end{aligned}$$

hence

$$I = \int_a^b f(x)\,dx \simeq \frac{h}{2}\left(f(a)+2\sum_{i=1}^{n-1} f(x_i)+f(b)\right) = I_t,$$

this is the composite trapezium formula on the interval $[a,b]$.

Example 5.3.2 *Let the integral $I = \int_0^1 x^2\,dx$ evaluate I by the trapezium method with $n=3$ and estimate the errors.*

Correction 5.3.2 *Let $f(x)=x^2$, $a=0$, $b=1$, we take $n=3$ subdivisions. We have*

$$h = \frac{b-a}{n} = \frac{1}{3}, x_0 = a = 0, \quad x_1 = \frac{1}{3}, \quad x_2 = \frac{2}{3}, \quad x_3 = b = 1,$$

then

$$y_0 = f(x_0) = 0, \quad y_1 = f(x_1) = \frac{1}{9}, \quad y_2 = f(x_2) = \frac{4}{9} \text{ and } y_3 = f(x_3) = 1,$$

hence

$$\begin{aligned} I &= \int_0^1 x^2 dx \simeq \frac{h}{2}\left(y_0 + 2\left(y_1+y_2\right)+y_3\right) \\ &= \frac{\frac{1}{3}}{2}\left(0+2\left(\frac{1}{9}+\frac{4}{9}\right)+1\right) = \frac{19}{54} \simeq 0.351 \end{aligned}$$

- *Approximation error $I(f) = \int_0^1 x^2 dx = \frac{1}{3} \simeq 0.333$.*
- *Absolute error* $\Delta I(f) \simeq |0.351 - 0.333| = 0.054$
- *The relative error* $\simeq \frac{|0.351-0.333|}{0.333} = 5.4\%$.
- *On the other hand, if $n=6$, the relative error* $\simeq 1.5\%$.

5.3.2 The Error Study (Simple Trapezium Method)

Let $f : [a,b] \to \mathbb{R}$ be continuous. We are looking for the approximation error if $f \in C^2([a,b])$ simple trapezium method. We set

$$E_t = I(f) - I_t(f) = \int_a^b f(x)\,dx - \frac{b-a}{2}\left(f(a) + f(b)\right),$$

with f'' is continuous on $[a,b]$, so by the theorem of the intermediate values, there exist $\theta_1 \in [a,b]$ where

$$E_t(h) = -\frac{h^3}{12} f''(\theta_1),$$

and $E_t(h)$ implies that

$$|E_t| \leq \frac{(b-a)^3}{12} M_1,$$

where $M_1 = max_{\theta_1 \in [a,b]} \| f''(\theta_1) \|$. The simple trapezium formula on the interval $[a,b]$

$$I(f) = \int_a^b f(x)\,dx = \frac{b-a}{2}\left(f(a) + f(b)\right) - \frac{h^3}{12} f''(\theta),\ \theta \in]a,b[\,.$$

Example 5.3.3 *Let* $I(f) = \int_0^1 e^{-2x} dx$.

Correction 5.3.3 *Let* $f(x) = e^{-2x}$, $a = 0$, $b = 1$, $f''(\theta) = 4e^{-2\theta}$, $f'''(\theta) = -8e^{-2\theta} \leq 0$. *where* $M_1 = \max_{\theta \in [0,1]} |f''(\theta)| = 4$.

$$|E_t| \leq \frac{(1-0)^3}{12} 4 = \frac{1}{3} \simeq 0.333.$$

Remark 5.2 *The trapezium formula is exact if* f *is a polynomial of degree less than or equal to* 1.

Moreover, this formula is of degree of precision equal to 1.
Indeed, for $f(x) = x^2$, $n = 1$, $x_0 = a$, $x_1 = b$, we have

$$I = \int_a^b x^2 dx = \left[\frac{x^3}{3}\right]_a^b = \frac{b-a}{3}\left(b^3 - a^3\right) = \frac{b-a}{3}\left(a^2 + ab + b^2\right) \neq I_1(f) = \frac{b-a}{2}\left(a^2 + b^2\right).$$

5.3.3 The study of error (Generalized trapezium method)

Let $f : [a,b] \to \mathbb{R}$ be continue function such that the maximum M and the minimum m. so,

$$\forall y \in [m,M],\ \exists x \in [a,b], \quad f(x) = y.$$

We are looking for the approximation error if $f \in C^2([a,b])$ composite trapezium method. We set

$$E_t = I(f) - I_t(f) = \int_a^b f(x)\,dx - \frac{h}{2}\left(f(a) + 2\sum_{i=1}^{n-1} f(x_i) + f(b)\right),$$

with $x_i = x_0 + ih$, $(h = \frac{b-a}{n}\ )$ $i = 0, 1, ..., n$. f'' is continuous on $[a,b]$, so from the intermediate value theorem, there exist $\theta \in [a,b]$

$$E_t(h) = -\frac{nh^3}{12} f''(\theta) \to |R(h)| \leq \frac{nh^3}{12} M_2,$$

where $M_2 = max_{\theta \in [a,b]} \| f''(\theta) \|$. As $h = (\frac{b-a}{n})$, we get

$$|E_t(h)| \leq \frac{(b-a)^3}{12n^2} M_2.$$

The composite trapezium formula on the interval $[a,b]$

$$I(f) = \int_a^b f(x)\,dx = \frac{h}{2}\left(f(a) + 2\sum_{i=1}^{n-1} f(x_i) + f(b)\right) - \frac{nh^3}{12} f''(\theta),\ \theta \in]a,b[\,.$$

Example 5.3.4 *We consider the integral*

$$I = \int_1^2 \frac{1}{x} dx.$$

What number of sub-intervals n must be chosen to have a lower error to 10^{-4}.

Correction 5.3.4 *Remember that the associated error is written, if* $f \in C^2([a,b])$,

$$E_t = -\frac{(b-a)^3}{12n^2} f''(\theta), \ \theta \in [a,b].$$

The error is increased by

$$|E_t| \leq \frac{(b-a)^3}{12n^2} M_2.$$

where $M_2 = \max_{\theta\in[1,2]} |f''(\theta)| = \max_{\theta\in[1,2]} \frac{2}{\theta^3} = 2$. *Here we have*

$$|E_t| \leq \frac{1}{6n^2}.$$

For $|E| < 10^{-4}$ *it is necessary that*

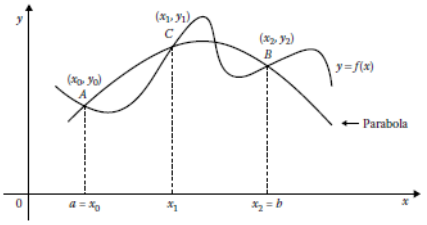
$$\frac{1}{6n^2} < 10^{-4},$$

so $n > \frac{10^2}{\sqrt{6}} \simeq 40.8$. *We take* $n = 41$ *subintervals, the quadrature error is less than* 10^{-4}.

5.4 Simpson's method

5.4.1 Simple Simpson's method

In this method, the function f is replaced by aquadratic polynomial defining an arc of parabola passing through three points.
As three dots induce two subdivisions, the number $n = 2$ of subdivisions must be taken even ($n = 2m$, $m = 1$).



In the simple case, we interpolate every 3 points $(a, f(a))$, $(\frac{a+b}{2}, f(\frac{a+b}{2}))$, $(b, f(b))$.
This is the first simple Simpson's formula on the interval $[a,b]$, such that

$$I = \int_a^b f(x)\, dx \simeq \frac{h}{3}\left[f(a) + 4f(\frac{a+b}{2}) + f(b)\right] = I_S,\ h = \frac{b-a}{2}.$$

Example 5.4.1 *Calculate, using Simpson's method, the integral*

$$I = \int_0^\pi \sin x^2 dx$$

Correction 5.4.1 *We have*

$$I = \int_0^\pi \sin x^2 dx \simeq \frac{\frac{\pi-0}{2}}{3}\left[f(0) + 4f(\frac{0+\pi}{2}) + f(\pi)\right] = \frac{\pi}{6}\left[0 + 4\times 1 + \sin \pi^2\right].$$

we obtain,

$$I_S = \frac{\pi}{6}\left(4 + \sin \pi^2\right).$$

5.4.2 Generalized Simpson's method

We have, for all interval $[x_{2i}, x_{2i+2}]$, $0 \leq i \leq n-1$, the integral

$$\int_{x_{2i}}^{x_{2i+2}} f(x)\,dx \simeq \frac{h}{3}\left[y_{2i} + 4y_{2i+1} + y_{2i+2}\right], \quad y_i = f(x_i),$$

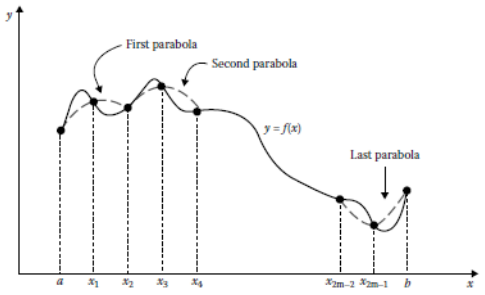
so that,

$$\begin{aligned}
\int_a^b f(x)\,dx &\simeq \int_{x_0=a}^{x_2} f(x)\,dx + \int_{x_2}^{x_4} f(x)\,dx + ... + \int_{x_{2m-2}}^{x_{2m}=b} f(x)\,dx \\
&= \frac{h}{3}(y_0 + 4y_1 + y_2) + \frac{h}{3}(y_2 + 4y_3 + y_4) + ... + \frac{h}{3}(y_{2m-2} + 4y_{2m-1} + y_{2m}) \\
&= \frac{h}{3}\left[y_0 + y_{n=2m} + 4\left(y_1 + y_3 + ... + y_{2m-1}\right) + 2\left(y_2 + y_4 + ... + y_{2m}\right)\right] \\
&= \frac{h}{3}\left[y_0 + y_n + 4\sum_{i=2k+2} y_i + 2\sum_{i=2k} y_i\right],
\end{aligned}$$

where

$$I = \int_a^b f(x)\,dx \simeq \frac{h}{3}\left[y_0 + y_n + 4\sum_{i\text{impair}} y_i + 2\sum_{i\text{pair}} y_i\right], \quad \text{with} \quad h = \frac{b-a}{n}.$$

This is Simpson's first composite on the interval $[a, b]$. We have
$I = \int_a^b f(x)\,dx \simeq \frac{h}{3}\left[f(a) + f(b) + 4\sum_i f(x_{2i+1}) + 2\sum_i f(x_{2i+2})\right] = I_S$. Where, $m = \frac{n}{2}$, $h = \frac{b-a}{n}$.



Example 5.4.2 *We apply the previous example, with $n = 4$.*

Correction 5.4.2 *We have $f(x) = x^2, a = 0, b = 1, h = 1/4$, according to Simpson's formula*

$$\int_0^1 f(x)\,dx \simeq \frac{h}{3}\left[y_0 + y_4 + 4\left(y_1 + y_3\right) + 2y_2\right] = \frac{1}{12}\left[0 + 1 + 4\left(\frac{1}{16} + \frac{9}{16}\right) + 2\left(\frac{4}{16}\right)\right],$$

so,

$$I_S(f) = \frac{1}{3}.$$

5.4.3 The Study of Error

The error associated with the Simpsons method is written, if $f \in C^4([a, b])$,

$$E_S(h) = I(f) - I_S(f) = -\frac{h^5}{90} f^{(4)}(\zeta_2), \quad \zeta_2 \in]a, b[.$$

We recall that the error associated with the generealized Simpson's method is written, if $f \in C^4([a, b])$,

$$\exists \zeta \in]a, b[, \quad E_S(h) = I(f) - I_{S,n}(f) = -\frac{nh^5}{180} f^{(4)}(\zeta).$$

Therefore,

$$|E_S(h)| = |I(f) - I_{S,n}(f)| \leq \frac{(b-a)^5}{180n^4} M_4.$$

Where $M_4 = \max_{\zeta \in [a,b]} \| f^{(4)}(\zeta) \|$.

Remark 5.3 *This implies that Simpson's formula is exact if f is a polynomial of degree less than or equal to 3. Moreover, this formula is of degree of precision equal to 3.*

Example 5.4.3 *Let $f(x) = x^4$, $n = 2$, $x_0 = a$, $x_1 = \frac{a+b}{2}$ and $x_2 = b$.*

Correction 5.4.3 *We have*

$$I(f) = \int_a^b x^4 dx = \left[\frac{x^5}{5}\right]_a^b = \frac{1}{5}\left(b^5 - a^5\right) = \frac{b-a}{5}\left(b^4 + ab^3 + a^2b^2 + a^3b + a^4\right).$$

While

$$I_2(f) = \frac{b-a}{6}\left(a^4 + 4\left(\frac{a+b}{2}\right)^4 + b^4\right) = \frac{b-a}{24}\left(5b^4 + 4ab^3 + 6a^2b^2 + 4a^3b + 5a^4\right)$$

We clearly see that $I(f) = I_{S,2}(f)$.

Example 5.4.4 *Let*

$$I = \int_0^1 \frac{e^x}{e^x + 1} dx, \qquad \text{and} \qquad x_0 = 0,\ x_1 = \frac{1}{2},\ x_2 = 1.$$

1. *Calculate the integral of I by:*
   a) *The trapezium method.*
   b) *The Simpson's method.*
2. *Estimate errors.*

Correction 5.4.4 1. a) *The trapezium method* $h = \frac{1}{2} = 0.5$

$$\begin{aligned} I_t &\simeq \frac{h}{2}\left(\frac{1}{2} + 2\frac{\sqrt{e}}{\sqrt{e}+1} + \frac{e}{e+1}\right) \\ &= \frac{1}{4}\left(\frac{1}{2} + \frac{2\sqrt{e}}{\sqrt{e}+1} + \frac{e}{e+1}\right). \end{aligned}$$

b) *The Simpson's method* $h = \frac{1}{2} = 0.5$

$$\begin{aligned} I_t &\simeq \frac{h}{6}\Big(f(0) + 2f(0.5) + f(1)\Big) \\ &= \frac{1}{12}\left(\frac{1}{2} + \frac{4\sqrt{e}}{\sqrt{e}+1} + \frac{e}{e+1}\right). \end{aligned}$$

2. *Errors, we have*

$$\begin{aligned} I &= \int_0^1 \frac{e^x}{e^x+1} dx \\ &= \ln(e^x + 1)\Big]_0^1 \\ &= \ln(e^1 + 1) - \ln(2) = 0.620114 \end{aligned}$$

so that

$$\begin{aligned} E_t &= |I - I_t| = |0.620114 - 0.618994| = 1.12 \times 10^{-3}, \\ E_S &= |I - I_S| = |0.620114 - 0.620149| = 3.5 \times 10^{-5}. \end{aligned}$$

5.5 Newton-Côtes method

The Newton-Côtes method is generalized the trapezoid method and the Simpson method (the function f is approximated by a polynomial of degree). The integral is evaluated according to the expression

$$\int_a^b f(x)dx \simeq a_0 f(x_0) + a_1 f(x_1) + \cdots + a_n f(x_n).$$

To determine the coefficients a_j, it suffices to write that the preceding relation is exact when f is a polynomial of degree less than or equal to n. By successively taking $g(x) = x^k$ for $k = 0, 1, 2, \cdots, n$, we obtain the following linear system

$$\begin{cases} a_0 + a_1 + \cdots + a_n & = & b - a = \int_a^b x^0 dx \\ a_0 x_0 + a_1 x_1 + \cdots + a_n x_n & = & \dfrac{b^2 - a^2}{2} = \int_a^b x^1 dx, \\ a_0 x_0 + a_1 x_1 + \cdots + a_n x_n & = & \dfrac{b^3 - a^3}{3} = \int_a^b x^2 dx, \\ \vdots & \vdots & \vdots \\ a_0 x_0 + a_1 x_1 + \cdots + a_n x_n & = & \dfrac{b^3 - a^3}{3} = \int_a^b x^2 dx. \end{cases}$$

The determinant of this system is a Vander-world determinant, which equals $(x_0 - x_1)(x_1 - x_2)\cdots(x_n - 0)$. When the points are regularly spaced, we obtain the Newton-Côtes formulas. For $n = 1$ (trapezium method)

$$\int_{x_0}^{x_1} f(x)dx \simeq \frac{h}{2}(f(x_0) + f(x_1)).$$

For $n = 2$ (Simpson's method)

$$\int_{x_0}^{x_1} f(x)dx \simeq \frac{h}{3}(f(x_0) + 4f(x_{1/2}) + f(x_1)).$$

Example 5.5.1 *Calculate the integral I by the Newton-Côtes method such as*

$$\int_{-1}^{1} f(x)dx = a_0 f(-1) + a_1 f(0) + a_2 f(1).$$

Correction 5.5.1 *We have*

$$\begin{aligned} k = 0, \quad g(x) = 1 &\implies \int_{-1}^{1} 1dx = 2 = a_0 + a_1 + a_2, \\ k = 1, \quad g(x) = x &\implies \int_{-1}^{1} xdx = -1a_0 + 1a_2 = -a_0 + a_2, \\ k = 2, \quad g(x) = x^2 &\implies \int_{-1}^{1} x^2 dx = \frac{2}{3} = 1a_0 + 1a_2. \end{aligned}$$

so,

$$\begin{cases} a_0 + a_1 + a_2 = 2, & (1) \\ -a_0 + a_2 = 0 & (2), \\ a_0 + a_2 = \dfrac{2}{3} & (3), \end{cases}$$

hence

$$\begin{aligned} (2) &\iff a_0 = a_2, \\ (3) &\iff 2a_0 = \frac{2}{3} \implies a_0 = \frac{1}{3} = a_2, \\ (1) &\iff a_1 + \frac{2}{3} = 2 \implies a_1 = \frac{4}{3}. \end{aligned}$$

Finally,

$$\begin{aligned}\int_{-1}^{1} f(x)dx &\simeq \frac{1}{3}f(-1)+\frac{4}{3}f(0)+\frac{1}{3}f(1))\\ &= \frac{1}{3}(f(-1)+4f(0)+f(1)).\end{aligned}$$

Remark 5.4 *We can write, if*

$$\begin{aligned} n=2 &\quad , \quad \textit{Simpson's method,}\\ n=1 &\quad , \quad \textit{Trapeze method,}\\ n=4 &\quad , \quad \textit{Villarceau's method,}\\ n=6 &\quad , \quad \textit{Hardy's method.}\end{aligned}$$

5.6 Exercises with solutions

Exercise 5.1 *Consider the integral I given by*

$$I=\int_0^1 e^{-x^2}\,dx.$$

1. *Approximate the integral I, for $n=8$, using*
   - *i) The trapezium method.*
   - *ii) Simpson's method.*
2. *Determine the number of necessary subdivisions of the integration intervals to evaluate to 10^{-5} by the trapezium method.*

Exercise 5.2 *We consider the numerical integration formula (M) given by*

$$(M)\qquad \int_0^1 f(x)dx \approx I_M=\frac{1}{8}\Big(f(0)+3f(\frac{1}{3})+3f(\frac{2}{3})+f(1)\Big).$$

1. *Deduce from the formula (M) a numerical integration formula on the interval $[a,b]$.*
2. *We divide the interval $[a,b]$ into $n=3m$, $m\in\mathbb{N}^*$ equal parts of length $h=\frac{b-a}{n}$. Generalize the formula obtained in the previous question.*
3. ***Applications:*** *Given the two integrals*

$$I_1=\int_{-1}^{1} x^2dx, \qquad I_2=\int_0^1 e^x dx$$

   a) *Compute the integrals I_1 and I_2 for $n=6$ by*
      - *i) by the trapezium formula,*
      - *ii) by Simpson's formula,*
      - *iii) by the formula obtained in question (2).*

   b) *Compare the results obtained in each case.*

5.6.1 Solutions

Solution 5.1 *Consider the integral I given by*

$$I=\int_0^1 e^{-x^2}\,dx.$$

As the function has no primitive, it is absolutely necessary to use a numerical method, we have

$$n=8,\ h=\frac{b-a}{n}=\frac{1-0}{8}=\frac{1}{8}.$$

1. *We approximate the integral I, for $n = 8$, using*

   i) The trapezium method

$$\begin{aligned} I &= \int_0^1 e^{-x^2}\,dx \\ &\approx \frac{h}{2}\Big(e^0 + 2\Big(e^{-(0.125)^2} + e^{-(0.25)^2} + e^{-(0.375)^2} + e^{-(0.5)^2} \\ &\qquad + e^{-(0.625)^2} + e^{-(0.75)^2} + e^{-(0.875)^2}\Big) + e^{-1.0}\Big) \\ &\approx 0.746809163. \end{aligned}$$

   ii) Simpson's method

$$\begin{aligned} I &= \int_0^1 e^{-x^2}\,dx \\ &\approx \frac{h}{3}\Big(e^0 + 4e^{-(0.125)^2} + 2e^{-(0.25)^2} + 4e^{-(0.375)^2} + 2e^{-(0.5)^2} + 4e^{-(0.625)^2} \\ &\qquad + 2e^{-(0.75)^2} + 4e^{-(0.875)^2} + e^{-1.0}\Big) \\ &\approx 0.7468261205. \end{aligned}$$

2. *We determine the number of necessary subdivisions of the integration intervals to evaluate to $\epsilon = 10^{-5}$ by the trapezium method. We have*

$$|E| < \frac{(b-a)^3}{12n^3} \sup_{x\in[0,1]} f''(x) < \epsilon = 10^{-5} \Longrightarrow n^3 > \frac{\sup_{x\in[0,1]} f''(x)}{12} 10^5,$$

   so

$$n = \left[\sqrt[3]{\frac{\sup_{x\in[0,1]} f''(x)}{12} 10^5}\right] + 1.$$

   As that

$$f'(x) = -2xe^{-x^2},\ f''(x) = (-2 + 4x^2)e^{-x^2},\ f'''(x) = (12x - 8x^3)e^{-x^2}.$$

   hence

$$f''(0) = \sup_{x\in[0,1]} f''(x) = 2e^{-0},$$

   we obtain,

$$n = \left[\sqrt[3]{\frac{2e^0}{12} 10^5}\right] + 1 = [129.09] + 1 = 130.$$

Solution 5.2 *We consider the numerical integration formula (M) given by*

$$(M) \qquad \int_0^1 f(x)dx \approx V(f) = \frac{1}{8}\Big(f(0) + 3f(\frac{1}{3}) + 3f(\frac{2}{3}) + f(1)\Big)$$

1. *We deduce the approximate integration formula over an interval $[a, b]$*
   Let $x = (b-a)y + a$, then

$$\begin{aligned} \int_a^b f(x)dx &= \int_0^1 f((b-a)y + a)(b-a)dy \\ &= (b-a)\int_0^1 g(y)dy \qquad (we\ set\ g(y) = f((b-a)y + a)) \\ &\approx \frac{b-a}{8}(g(0) + 3g(\frac{1}{3}) + 3g(\frac{2}{3}) + 3g(1)) \\ &= \frac{b-a}{8}\Big(f(a) + 3f\Big(\frac{b+2a}{3}\Big) + 3f\Big(\frac{a+2b}{3}\Big) + 3f(b)\Big). \end{aligned}$$

2. *We divide the interval* $[a,b]$ *into* $n=3m$, $m\in\mathbb{N}$ *equal parts of length* $h=\dfrac{b-a}{n}$
Let us generalize the formula obtained in the previous question. We set

$$x_j=a+jh \quad \textit{and} \quad y_j=f(x_j), \quad j=0,\cdots,n.$$

We have

$$\begin{aligned}
\int_a^b f(x)dx &= \int_{x_0}^{x_{3m}} f(x)dx \\
&= \sum_{i=0}^{m-1}\int_{x_{3i}}^{x_{3i+3}} f(x)dx \\
&\approx \frac{x_{3i+3}-x_{3i}}{8}\left(\sum_{i=0}^{m-1}(f(x_{3i})+3f\left(\frac{x_{3i+3}+2x_{3i}}{3}\right)+3f\left(\frac{2x_{3i+3}+x_{3i}}{3}\right)+f(x_{3i+3}))\right) \\
&= \frac{3h}{8}(y_0+3y_1+3y_2+y_3)+\frac{3h}{8}(y_3+3y_4+3y_5+y_6) \\
&\quad +\cdots+\frac{3h}{8}(y_{3m-3}+3y_{3m-2}+3y_{3m-1}+y_{3m}) \\
&= \frac{3h}{8}[y_0+y_{3m=n}+3(y_1+y_2+y_4+y_5+\cdots+y_{3m-2}+y_{3m-1}) \\
&\qquad 2(y_3+y_6+\cdots+y_{3m-3})] \\
&= \frac{3h}{8}\left(f(x_0)+f(x_n)+3\sum_{i=0}^{m-1}(y_{3i+1}+y_{3i+2})+2\sum_{i=0}^{m-1}y_{3i}\right).
\end{aligned}$$

3. ***Applications,*** *let the integral* $I_1=\int_0^1 x^2dx$. *Here* $f(x)=x^2$, $[a,b]=[-1,1]$

a) $n=6\Longrightarrow h=\dfrac{2}{6}=\dfrac{1}{3}$. *We set*

$x_0=-1$, $x_1=-\dfrac{2}{3}$, $x_2=\dfrac{1}{3}$, $x_3=0$, $x_4=\dfrac{1}{3}$, $x_5=\dfrac{2}{3}$, $x_6=1$, *and* $y_i=x_i^2$, $i=0,\cdots,6$.

i) By the trapezium formula

$$\begin{aligned}
I_1\approx I_T &= \frac{h}{2}(y_0+2(y_1+y_2+y_3+y_4+y_5)+y_6) \\
&= \frac{1}{6}(0+2(\frac{4}{9}+\frac{1}{9})+0+\frac{1}{9}+\frac{4}{9}+1) \\
&= \frac{38}{54}.
\end{aligned}$$

ii) By Simpson's formula $I_S=I_1=\frac{2}{3}$, *because* $f\in\mathbb{P}_2$ *and Simpson's formula has degree of precision* $p=3$.

iii) By using the formula obtained in question (1) $I_M=I_1=\frac{2}{3}$, *because* $f\in\mathbb{P}_2$ *and the formula* (M) *has degree of precision* $p=3$.

b) The (M) *and Simpson's formula is more precise than that of trapezoids because they have degree of precision* $p=3$ *and that of trapezoids has degree* $p=1$.

Consider the integral $I_2=\int_0^6 e^x dx$. *Here* $f(x)=e^x$, $[a,b]=[0,6]$

a) $n=6\Longrightarrow h=\frac{6}{6}=1$. *We set*

$x_0=0$, $x_1=1$, $x_2=2$, $x_3=3$, $x_4=4$, $x_5=5$, $x_6=6$, *and* $y_i=e^{x_i}$, $i=0,\cdots,6$.

i) By the trapezium formula

$$\begin{aligned}
I_2\approx I_T &= \frac{h}{2}(y_0+2(y_1+y_2+y_3+y_4+y_5)+y_6) \\
&= \frac{1}{2}(1+2(e+e^2+e^3+e^4+e^5)+e^6) \\
&\approx 435.42.
\end{aligned}$$

ii) By Simpson's formula

$$\begin{aligned} I_2 \approx I_S &= \frac{h}{3}(y_0 + 4(y_1 + y_3 + y_5) + 2(y_2 + y_4) + y_6) \\ &= \frac{1}{3}(1 + 4(e + e^3 + e^5) + 2(e^2 + e^4) + e^6) \\ &\approx 404.42\,. \end{aligned}$$

iii) By the formula obtained in question (2)

$$\begin{aligned} I_2 \approx I_M &= \frac{3h}{8}(y_0 + 3(y_1 + y_2 + y_4 + y_5) + 2(y_3 + y_4) + y_6) \\ &= \frac{3}{8}(1 + 3(e + e^2 + e^4 + e^5) + 2e^3 + e^6) \\ &\approx 406.48\,. \end{aligned}$$

b) We have

$$I_2 = \int_0^6 e^x dx = e^6 - 1.$$

So that

$$\begin{aligned} E_T &= |I_2 - I_T| = 32.99 \\ E_S &= |I_2 - I_S| = 1.99 \\ E_M &= |I_2 - I_M| = 4.05. \end{aligned}$$

Note that $E_S < E_M < E_T$. So the most precise method is that of Simpson and the least precise is that of the trapezoids.

5.7 Exercises without solutions

Exercise 5.3 *1. Give the formulas of the rectangles approximating $I = \int_a^b f(x)dx$.*

2. Find the error made by applying the method of rectangles, as well as the minimum number n of subdivisions of $[0,1]$ to have $\int_0^1 e^{x^2}dx$ to the nearest $1/100$.

3. Give the method of rectangles with midpoint.

4. Find the error made by applying the method of rectangles with midpoint. Same application as question 2.

Exercise 5.4 *Determine by the trapezium method then by Simpson method $\int_0^{\frac{\pi}{2}} f(x)dx$ on the basis of the following table*

x	0	$\pi/8$	$\pi/4$	$3\pi/8$	$\pi/2$
$f(x)$	0	0.382683	0.707107	0.923880	1

the points of support are those giving $\sin(x)$*, after that compare the results obtained with the exact value.*

Exercise 5.5 *We launch a rocket vertically from the ground and we measure during the first* 80 *seconds the acceleration γ*

$t(on\,s)$	0	10	20	30	40	50	60	70	80
$\gamma(on\,m/s^2)$	30	31.63	33.44	35.47	37.75	40.33	43.29	46.70	50.67

Calculate the speed V of the rocket at time $t = 80s$, by the Trapezoids then by Simpson.

Exercise 5.6 *Compute* $\arctan(3)$ *by trapeze and Simpson integration methods for $n = 6$.*
Indication: $\arctan(x) = \int_0^x \frac{1}{1+t^2}dt$.

Exercise 5.7 *1. Determine the number of necessary subdivisions of the integration interval* $[-\pi, \pi]$ *to evaluate to* 0.5×10^{-3}*, by Simpson's method, the integral*

$$\int_{-\pi}^{-\pi} \cos(x)dx.$$

2. Determine the number of necessary subdivisions of the integration interval $[-\pi, \pi]$ *to evaluate to within* 0.5×10^{-3}*, using the trapezium method the integrals*

$$\int_0^1 \frac{1}{1+e^x}dx, \int_0^1 \frac{1}{1+\sin(x)}dx.$$

Exercise 5.8 *Evaluate using the trapezium method, the integral* $\int_0^\pi \frac{\sin(x)}{x}dx$*, with error less than* 10^{-4}.

Exercise 5.9 *Let* $F(x) = \int_0^x te^{-t}dt$*. How many subdivisions of* $[0, 1]$ *are needed to evaluate* $F(1)$ *to* 10^{-8} *near using*

1. *the trapezium method.*
2. *the Simpsons method.*

Chapter 6

Numerical solution of ordinary differential equations

6.1 Differential equations

Definition 6.1 *A differential equation is a relation between a real variable t, a function y which depends on this variable, a certain number of its successive derivatives and a determined function f.*
A differential equation of order n can therefore be written in the form

$$f\big(t, y(t), y'(t), y''(t), ..., y^{(n)}(t)\big) = 0.$$

Definition 6.2 *A linear differential equation with constant coefficients of order $n \in \mathbb{N}^*$ is any equation of the form*

$$a_0 y + a_1 y' + a_2 y'' + ... + a_{n-1} y^{(n)} = b,$$

where $a_0, a_1, ..., a_{n-1}$ are reals. If $b = 0$ then, this equation is called homogeneous linear differential equation.

Linear equations of order 1 defined on an interval of the form $I = [t_0, +\infty[$ where $t_0 \in \mathbb{R}$ are of the form

$$a_0 y'(t) + a_1 y(t) = b_0, \qquad t \in I.$$

If $a_0 = 1$, then this equation is called a normalized linear differential equation.

Otherwise we can reduce to a normalized equation we divide by a_0.

Hence the equation becomes as follows

$$y'(t) + ay(t) = b, \qquad t \in I, \tag{6.1.1}$$

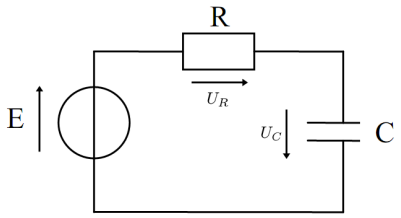
where $a = \dfrac{a_1}{a_0}$ and $b = \dfrac{b_0}{a_0}$. For the resolution of the equation (6.1.1), we need a unitial condition y_0. If we set $y_0 = y(t_0)$, we find the following Cauchy problem

$$\begin{cases} y'(t) = -ay(t) + b, & t \geq t_0, \\ y(t_0) = y_0, \end{cases} \tag{6.1.2}$$

6.1.1 Examples and motivations

Today, differential equations are extremely present in many scientific fields such as physics, finance, biology...
In this section, we give the following two examples

Example 6.1.1 ***(Electrical circuit)*** *We consider the electrical circuit RC described by the following figure*



Where U_C and U_R respectively represent the voltage of the electric current across capacitor C and resistor R under a constant voltage E.

According to the law of additivity of tensions (or mesh law), $E = U_R + U_C$ and according to Ohm's law $U_R = Ri$, where the intensity $i = \frac{dq}{dt} = \frac{d(C \cdot U_C)}{dt} = C\frac{dU_C}{dt}$, we find the differential equation of the circuit RC determined by the following form

$$E = RC\frac{dU_C}{dt} + U_C \Leftrightarrow \frac{dU_C(t)}{dt} + \frac{1}{RC}U_C(t) = \frac{E}{RC}.$$

Example 6.1.2 *(**Bank account balance**) Every year, a bank account is credited with interest at a rate of 5%. If we suppose that the customer of this account pays 50 for his rent then the balance y of this bank account in function of the year t would verify*

$$y' = \frac{5}{100}y - 50.$$

6.2 Analytical solution

In this subsection, we consider the Cauchy problem (6.1.2). Analytically, the solution of this system is given by the method of the variation of the constant (see [4]), in the following form

$$y(t) = y_0 e^{-a(t-t_0)} + \frac{b}{a}\Big(1 - e^{-a(t-t_0)}\Big), \qquad t \geq t_0. \tag{6.2.1}$$

Example 6.2.1 *We consider the following Cauchy problem*

$$\begin{cases} y'(t) &= -2y(t) + 6, \qquad t \geq 0, \\ y(0) &= -1. \end{cases}$$

By the constant variation method (see (6.2.1)), the solution is given by:

$$\begin{aligned} y(t) &= -e^{2t} - 3(1 - e^{2t}), \qquad t \geq 0 \\ &= 2e^{2t} - 3 \end{aligned}$$

6.3 The Cauchy problem

6.3.1 First-order equations

Let f be a definite function of $I \times \mathbb{R}$ with $I \subset \mathbb{R}$. The Cauchy problem consists in finding a function $y : I \longrightarrow \mathbb{R}$ solution of

$$\{\ y'(t) = f(t, y(t)), y(t_0) = y_0, \qquad t, t_0 \in I. \tag{6.3.1}$$

The condition $y(t_0) = y_0$ is an initial condition or the Cauchy condition.

If we suppose that the function f is continuous with respect to the two variables t, y and that f is uniformly Lipschitz with respect to y, that is to say that

$$\exists L > 0, \quad \forall t \in I, \quad \forall y_1, y_2 \in \mathbb{R}, \quad |f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|$$

then the Cauchy problem admits a unique solution $y \in C^1(I)$ (This is Cauchy's theorem), see [4].

The Cauchy problem is an evolution problem, i.e. from the initial condition, we can calculate the solution at time t as

$$y(t) = y(t_0) + \int_{t_0}^{t} f(s, y(s))ds, \tag{6.3.2}$$

and the solution at time t depends only on the solution at time $t_0 \leq s \leq t$.

The solution ((6.3.2)) does not give y explicitly except in simple cases.

6.3.2 Higher-order equations.

If we consider a higher order differential equation

$$\begin{cases} y^{(p)} = f(t, y, y', \cdots, y^{(p-1)}), \\ y(t_0) = y_0,\ y'(t_0) = y'_0,\ \cdots,\ y^{(p-1)}(t_0) = y_0^{p-1}, \end{cases}$$

we can reduce the problem to a system of first-order differential equations, passing $Y = (z_1, z_2, \cdots, z_p)$ and $y = z_1$ we get system

$$\begin{cases} z_1' &= z_2 \\ z_2' &= z_3 \\ &\vdots \\ z_{p-1}' &= z_p \\ z_p' &= f(t, z_1, z_2, \cdots, z_{p-1}, z_p) \end{cases}$$

and

$$z_1(t_0) = z_1^0,\ z_2(t_0) = z_2^0,\ \cdots,\ z_{(p-1)}(t_0) = z_{p-1}^0,$$

They are written in the following form

$$\begin{cases} Y'(t) = F(t, Y(t)) \\ Y(t_0) = Y_0 \end{cases}$$

Example 6.3.1 *Consider the second order differential equation*

$$y''(t) = 3y'(t) - ty(t).$$

Let's set $y = z_1$ and $z_1' = z_2$ then this equation can be reduced to the system

$$\begin{cases} z_1' = z_2 \\ z_2' = 3z_1' - tz_1 = 3z_2 - tz_1 \end{cases} \iff Y' = F(t, Y), \quad Y = (z_1, z_2).$$

6.3.3 Numerical approximation of the Cauchy problem

The goal is to study consistent and stable numerical methods allowing the calculation of good approximations of the exact solution of the ODE (6.3.1). The most famous method is that of Euler.

6.3.3.1 Euler's method

A numerical diagram is a diagram which gives us an approximation of the solution at a given point.

We give ourselves a subdivision of $I = [a, b]$ into N intervals of steps h

$$t_0 = a < t_1 < \cdots < t_N < t_{N+1} = b,$$

with $h = t_{N+1} - t_N$ the step of discretization is approached, in integrity

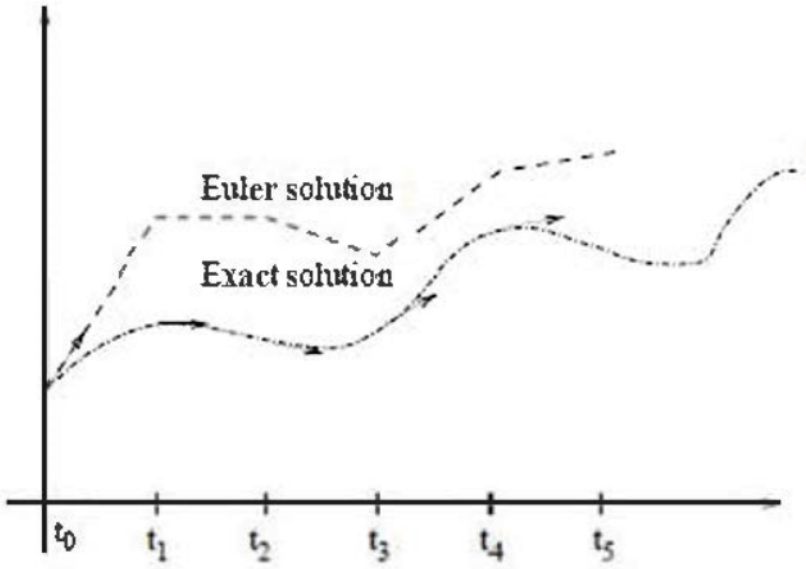
$$\int_{t_n}^{t_{n+1}} y'(t)dt = \int_{t_n}^{t_{n+1}} f(t, y(t))dt,$$

so,

$$y(t)]_{t_n}^{t_{n+1}} = \int_{t_n}^{t_{n+1}} f(t, y(t))dt,$$

Finally,

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt. \tag{6.3.3}$$



Explicit Euler scheme We replace f in (6.3.3) by $f(t_n, y(t_n))$, we get

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt \\ &\simeq y(t_n) + f(t_n, y(t_n)) \int_{t_n}^{t_{n+1}} dt \\ &= y(t_n) + (t_{n+1} - t_n) f(t_n, y(t_n)) \\ &= y(t_n) + hf(t_n, y(t_n)). \end{aligned}$$

We note $y(t_n) = y_n$ and $y(t_{n+1}) = y_{n+1}$ so

$$\begin{cases} y_{n+1} = y_n + hf(t_n, y_n) \\ y_0 \text{ given} \end{cases}$$

Implicit Euler scheme We replace f in (6.3.3) by $f(t_{n+1}, y(t_{n+1}))$, we get

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt \\ &\simeq y(t_n) + f(t_{n+1}, y(t_{n+1})) \int_{t_n}^{t_{n+1}} dt \\ &= y(t_n) + (t_{n+1} - t_n) f(t_{n+1}, y(t_{n+1})) \\ &= y(t_n) + hf(t_n, y(t_n)). \end{aligned}$$

We note $y(t_n) = y_n$ and $y(t_{n+1}) = y_{n+1}$ so that

$$\begin{cases} y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}) \\ y_0 \text{ given} \end{cases}$$

The $\theta-$Euler scheme We combine the two schemes (explicit and implicit)

$$\begin{cases} y_{n+1} = y_n + \theta f(t_n, y_n) + (1-\theta) f(t_{n+1}, y_{n+1}), & 0 \leq \theta \leq 1 \\ y_0 \quad \text{given} \end{cases}$$

It remains to choose an optimal θ.

Example 6.3.2 *Consider the following system*

$$\begin{cases} y' = -y, & t \in]0, T[\\ y(0) = 1, \end{cases}$$

Explicit Euler *We have* $h = \dfrac{T-0}{N} = \dfrac{T}{N}$ *and we can write*

$$y_{n+1} = y_n + hf(t_n, y_n), \qquad y_0 = 1$$

implies that

$$y_{n+1} = y_n - hy_n = (1-h)y_n,$$

by recurrence

$$y_{n+1} = (1-h)^N y_0 = (1-h)^N$$

so,

$$y_{n+1} = (1-h)^{\frac{T}{h}} \longrightarrow e^{-T}, \quad \textit{when} \quad h \longrightarrow 0.$$

Implicit Euler *We have* $h = \dfrac{T-0}{N} = \dfrac{T}{N}$ *and we write*

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}), \qquad y_0 = 1$$

implies that

$$y_{n+1} = y_n - hy_{n+1} \Longrightarrow y_{n+1} = \left(\frac{1}{1+h}\right) y_n,$$

by recurrence

$$y_{n+1} = \left(\frac{1}{1+h}\right)^N y_0 = \left(\frac{1}{1+h}\right)^N$$

so,

$$y_{n+1} = \left(\frac{1}{1+h}\right)^{\frac{T}{h}} \longrightarrow e^{-T}, \quad \textit{when} \quad h \longrightarrow 0.$$

6.3.4 Runge-Kutta method

6.3.4.1 Runge-Kutta 2 (RK2) method

According to the system (6.3.1) we have

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt,$$

we approach $\int_{t_n}^{t_{n+1}} f(t, y(t))dt$ by the trapezium method

$$\begin{aligned} \int_{t_n}^{t_{n+1}} f(t, y(t))dt &\simeq \frac{h}{2}\Big(f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))\Big) \\ &= \frac{h}{2}\Big(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\Big) \\ &= \frac{h}{2}\Big(f(t_n, y_n) + f(t_{n+1}, y_n + hf(t_n, y_n))\Big). \end{aligned}$$

We note $K_1 = f(t_n, y_n)$ and $K_2 = f(t_{n+1}, y_n + hf(t_n, y_n))$, so we get the algorithm of Runge-Kutta method (RK2)

$$\begin{cases} y_{n+1} = y_n + \dfrac{h}{2}(K_1 + K_2), & n = 0, 1, 2, ... \\ K_1 = f(t_n, y_n), \qquad K_2 = f(t_{n+1}, y_n + hK_1). \end{cases}$$

6.3.4.2 Runge-Kutta method 4 (RK4)

According to the system (6.3.1) we have

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt,$$

we approximate $\int_{t_n}^{t_{n+1}} f(t, y(t))dt$ by Simpson's method. For Simpson's family, we use the points

$$t_n,\ t_{n+1/2} = t_n + \frac{h}{2},\ t_{n+1} = t_n + h$$

The Runge-Kutta scheme (RK4)

$$\begin{aligned} k_1 &= hf(t_n, y_n), \\ k_2 &= hf(t_n + \frac{h}{2}, y_n + k_1), \\ k_3 &= hf(t_n + \frac{h}{2}, y_n + k_2), \\ k_4 &= hf(t_n + h, y_n + k_3), \end{aligned}$$

thus we obtain the algorithm of (RK4)

$$\begin{cases} y_{n+1} = y_n + \dfrac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), & n = 0.1, 2, ... \\ y_0 \quad \text{given} \end{cases}$$

6.3.4.3 The Error Study

Theorem 6.1 *[2] Suppose that the map $f(t, y)$ is continuous with respect to the two variables and Lipschtzian with respect to y uniformly with respect to t, and that $y \in C^2([a, b])$ We set $M_2 = \sup_{t\in[a,b]} |y''(t)|$, then the markup*

$$|y(t_i) - y_i| \leq (e^{L(b-a)} - 1)\frac{M_2}{2L}h.$$

6.4 Dirichlet's problem

6.4.1 The mathematical problem

Consider the following problem: is there a function $y(x)$ defined on $[a, b]$ with value in $\mathbb{R}$ which is twice continuously differentiable on $[a, b]$ and verifies

$$\begin{cases} y''(x) = f(x, y(x), y'(x)), & x \in [a, b] \quad (1) \\ y(a) = \alpha,\ y(b) = \beta & \quad (2) \end{cases} \tag{6.4.1}$$

This problem is called boundary problem or Dirichlet problem with α and β are constant.

This problem may not have a solution. Here are two examples

$$\begin{cases} y''(x) = -\pi^2 y(x), & x \in [0, 1] \\ y(0) = 0,\ y(1) = d. \end{cases} \tag{6.4.2}$$

The general solution of the differential equation of this problem is $y(x) = A\cos(\pi x) + B\sin(\pi x)$. The condition $y(0) = 0$ results $A = 0$. So $y(x) = B\sin(\pi x)$. The second condition $y(1) = d$ leads to $0 = B\sin(\pi) = d$. For $d \neq 0$, this problem has no solution. If $d = 0$, it has an infinite number of solutions.

Now, we consider the 2^{nd} example

$$\begin{cases} y'(x) = y(x), & x \in [a, b] \\ y(a) + Ay(b) = d, \end{cases} \tag{6.4.3}$$

where A and d are given constants.

The general solution of the differential equation is $y(x) = Ce^x$. The condition $y(a) + Ay(b) = d$ leads to $Ce^a + ACe^b = d$. If $d \neq 0$ and $A = -e^{a-b}$, there is no C such that $y(x) = Ce^x$ is a solution to the problem (6.4.3).
In what follows, we will assume that the problem (6.4.1) has a unique solution $y(x)$ and develop numerical methods to approach it.

Definition 6.3 *The differential equation* (1) *of the problem (6.4.1) is said to be linear if* $f(x, y, y')$ *has the form* $p(x)y'(x) + q(x)y(x) + r(x)$.
In this case, the boundary value problem (6.4.1) is said to be linear.

Theorem 6.2 *[2] The following boundary problem*

$$\begin{cases} y''(x) = p(x)y'(x) + q(x)y(x) + r(x), & x \in [a, b] \quad (1) \\ y(a) = \alpha,\ y(b) = \beta & (2) \end{cases} \tag{6.4.4}$$

has only one solution if

1. $p(x)$, $q(x)$ *and* $r(x)$ *are continuous on* $[a, b]$.
2. $q(x) > 0$ *on* $[a, b]$.

6.4.2 Finite difference method

Philosophy: The finite difference method for solving boundary value problems replaces each derivative in the differential equation with an appropriate difference-ratio approximation.

The difference quotients and the step h are chosen to have a truncation error of a given order. However, h cannot be chosen too small because in stabilities in the derivative approximation.

Model problem We choose as a model the homogeneous two-point problem, the following boundary problem

$$\begin{cases} u''(x) = f(x), & x \in]a, b[\\ u(0) = u(1) = 0, \end{cases} \tag{6.4.5}$$

where f is a continuous function on $[0, 1]$.
The unique solution of this problem is given by

$$u(x) = x \int_0^1 f(s)(1 - s)ds - \int_0^x f(s)(x - s)ds, \qquad x \in [0.1]. \tag{6.4.6}$$

Approximation We begin by choosing an integer $N \geq 1$ and we divide the interval $[0, 1]$ into $(N + 1)$ sub-intervals whose extremities are the points of the matrix

$$x = ih, \qquad i = 0, 1, \cdots, N + 1 \qquad \text{with} \qquad h = \frac{1}{N + 1}.$$

The step h is chosen constant to facilitate the use of the algorithms serving to solve the linear systems which result from the approximation and which involve a matrix $N \times N$.

Approximation of the problem (6.4.5) by finite differences The finite difference method applied to the problem (6.4.5) requires that we replace the second derivative $u''(x_i)$ by a relation to the differences at each of the interior points x_i for $i = 1, 2, \cdots, N$.

Using a Taylor expansion at point x_i, we can see that if u is of class C^4 on $[0, 1]$, we have the approximation

$$u''(x_i) = \frac{1}{h^2}[u(x_{i+1}) - 2u(x_i) + u(x_{i-1})] - \frac{h^2}{12}u^{(4)}(\xi_i), \qquad \xi_i \in]x_{i-1}, x_{i+1}[.$$

which is called the "centered difference formula" for $u''(x_i)$. Neglect the term containing ξ_i in the previous formula and denote u_i an approximation of u at the point x_i.

Taking into account the boundary conditions, we obtain the following finite difference method

$$\begin{cases} u_0 = 0, \qquad u_{N+1} = 0, \\ -u_{i-1} + 2u_i - u_{i+1} = h^2 f(x_i), \qquad i = 1, 2, \cdots, N. \end{cases} \tag{6.4.7}$$

Given the neglected term, the truncation error here is of order $O(h^2)$.

The preceding numerical scheme can be written in the following matrix form

$$AU_h = b_h, \tag{6.4.8}$$

where A is the symmetric tridiagonal matrix $N \times N$ given by

$$A = \text{tridiag}(-1, 2, -1) = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}$$

and U_h and b_h the vectors of $\mathbb{R}^N$

$$U_h = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix}, \qquad b_h = h^2 \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) \end{pmatrix}.$$

The problem (6.4.5) for $f(x) = \pi^2 \sin(\pi x)$ admits the unique exact solution $u(x) = \sin(\pi x)$ and its discretization by the finite difference method (6.4.8), for example, $h = \dfrac{1}{8}$ uses the points

$$x_0 = 0, x_1 = 0.125, x_2 = 0.250, x_3 = 0.375, x_4 = 0.500, x_5 = 0.625, x_6 = 0.750, x_7 = 0.875, x_8 = 1,$$

and leads to the system of linear equations

$$\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \end{pmatrix} = \frac{1}{64} \begin{pmatrix} f(0.125) \\ f(0.250) \\ f(0.375) \\ f(0.500) \\ f(0.625) \\ f(0.750) \\ f(0.875) \end{pmatrix}.$$

The calculations give for the second member vector b_h

$$b_h^t = (0.0590146, 0.109045, 0.142474, 0.154213, 0.142474, 0.109045, 0.0590146).$$

We obtain for solution of the tridiagonal system $A_h U_h = b_h$, the vector

$$U_h^t = (0.3876401, 0.7162656, 0.9358461, 1.0129526, 0.9358461, 0.7162656, 0.3876401).$$

So the approximate solution

$$(0, 0), (x_1, u_1), (x_2, u_2), (x_3, u_3), (x_4, u_4), (x_5, u_5), (x_6, u_6), (x_7, u_7), (1, 0),$$

and the exact solution $u(x) = \sin(\pi x)$, then the effective error

$$E_e = \sup_{i=1,..,7} \{|u(x_i) - u_i|.$$

To solve the system $A_h U_h = b_h$ using Thomas' algorithm.

Thomas' algorithm The Thomas algorithm used for the tridiagonal system calculation is written as follows

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, \qquad \forall i \in \{1, 2, \cdots, n\}$$

such that x_0 and x_{n+1} are given by the boundary conditions, the matrix form is written

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ a_1 & b_1 & c_1 & 0 & \cdots & 0 & 0 \\ 0 & a_2 & b_2 & c_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & a_n & b_n & c_n \\ 0 & 0 & \cdots & \cdots & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ \vdots \\ d_n \\ d_{n+1} \end{pmatrix}.$$

So let's simplify the system

$$\begin{pmatrix} b_1 & c_1 & 0 & \cdots & \cdots & 0 & 0 \\ a_2 & b_2 & c_2 & 0 & \cdots & 0 & 0 \\ 0 & a_3 & b_3 & c_3 & 0 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & \cdots & \cdots & 0 & a_n & b_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} d_1 - a_1 d_0 \\ d_2 \\ d_3 \\ \vdots \\ \vdots \\ d_{n-1} \\ d_n - c_c d_{n+1} \end{pmatrix} = \begin{pmatrix} d'_1 \\ d'_2 \\ d'_3 \\ \vdots \\ \vdots \\ d'_{n-1} \\ d'_n \end{pmatrix},$$

note that $d_0 = x_0$ and $d_{n+1} = x_{n+1}$.
According to the elimination of Gauss, we set

$$Q_i = \begin{cases} \dfrac{c_1}{a_1} & i = 1 \\ \dfrac{c_i}{b_i - Q_{i-1} a_i} & i = 2, 3, \cdots, n \end{cases}$$

so,

$$\begin{pmatrix} 1 & Q_1 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 1 & Q_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & Q_3 & 0 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & Q_{n-1} \\ 0 & 0 & \cdots & \cdots & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ \vdots \\ R_{n-1} \\ R_n \end{pmatrix}.$$

Finaly, we use the escalation method to solve this problem

$$\begin{cases} R_n & i = n \\ R_i - Q_i x_{i+1} & i = n-1, n-2, \cdots, 1. \end{cases}$$

6.4.3 Finite difference method for linear problem

Consider the second-order linear boundary system

$$\begin{cases} u''(x) = p(x)u'(x) + q(x)u(x) + r(x), & 0 \le x \le 1 \\ u(0) = \alpha, \qquad u(1) = \beta \end{cases} \tag{6.4.9}$$

where p, q, and r are given functions, continuous on $[0, 1]$, and α, β are two known real numbers.

To approach the solution of this problem by the method of finite differences, we start as above by dividing the interval $[0, 1]$ into $(N + 1)$ equal sub-intervals (to simplify).

This gives the points

$$x_i = ih, \qquad i = 0, 1, 2, \cdots, N+1 \qquad \text{where} \qquad h = \frac{1}{N+1}.$$

At the interior points, x_i, for $i = 1, 2, \cdots, N$, the differential equation to approximate is

$$u''(x_i) = p(x_i)u'(x_i) + q(x_i)u(x_i) + r(x_i).$$

Suppose that $u \in C^4(]x_{i-1}, x_{i+1}[)$.

By expanding u according to a Taylor polynomial of order three around the point x_i and evaluated at the points x_{i-1} and x_{i+1}, we obtain

$$\begin{aligned} u(x_{i+1}) &= u(x_i + h) \\ &= u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^3}{6}u^{(4)}(\xi_i^+), \end{aligned}$$

for some ξ_i^+ in $]x_i, x_{i+1}[$, and

$$\begin{aligned} u(x_{i-1}) &= u(x_i - h) \\ &= u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^3}{6}u^{(4)}(\xi_i^-), \end{aligned}$$

for some ξ_i^- in $]x_{i-1}, x_i[$.

Summing these relations, we get

$$u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + h^2u''(x_i) + \frac{h^2}{12}[u^{(4)}(\xi_i^+) + u^{(4)}(\xi_i^-)],$$

which gives for $u''(x_i)$.

$$u''(x_i) = \frac{1}{h^2}[u(x_{i+1}) - 2u(x_i) + u(x_{i-1})] - \frac{h^2}{12}[u^{(4)}(\xi_i^+) + u^{(4)}(\xi_i^-)].$$

Approximation of u'': The intermediate value theorem allows us to write the simplified expression

$$u''(x_i) = \frac{1}{h^2}[u(x_{i+1}) - 2u(x_i) + u(x_{i-1})] - \frac{h^2}{12}u^{(4)}(\xi_i), \qquad \xi \in]x_{i-1}, x_{i+1}[$$

said centered difference formula for $u''(x_i)$.

Approximation of u': A centered difference formula for $u'(x_i)$ is obtained by the same way

$$u'(x_i) = \frac{1}{2h}[u(x_{i+1}) - u(x_{i-1})] - \frac{h^2}{6}u^{(3)}(\eta_i), \qquad \eta \in]x_{i-1}, x_{i+1}[.$$

The use of the centered difference formulas in equation (6.4.9) leads to the relation

$$\begin{aligned} \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} &= p(x_i)\left[\frac{u(x_{i+1}) - u(x_{i-1})}{2h}\right] \\ &+ q(x_i)u(x_i) + r(x_i) - \frac{h^2}{12}[2p(x_i)u^{(3)}(\eta_i) - u^{(4)}(\xi_i)]. \end{aligned}$$

The previous relation of the terms containing the derivatives of u at the unknown points η_i and ξ_i and taking into account the boundary conditions $u(a) = \alpha$ and $u(b) = \beta$, we obtain the following finite difference method with a truncation error of order $O(h^2)$

$$\begin{aligned} &u_0 = \alpha,\ u_{N+1} = \beta \\ &\frac{-u_{i+1} + 2u_i - u_{i-1}}{h^2} + p(x_i)\frac{u_{i+1} - u_{i-1}}{2h} + q(x_i)u_i = -r(x_i),\ i = 1, 2, \cdots, N. \end{aligned}$$

This equation will be used in the form

$$-\left(1+\frac{h}{2}p(x_i)\right)u_{i-1}+(2+h^2q(x_i))u_i-\left(1-\frac{h}{2}p(x_i)\right)u_{i+1}=-h^2r(x_i),$$

which can be written as a linear system with a tridiagonal $N\times N$ matrix

$$A_hU_h=b_h, \tag{6.4.10}$$

where

$$A_h=\begin{pmatrix} 2+h^2q_1 & -1+\frac{h}{2}p_1 & 0 & \cdots & \cdots & 0 & 0\\ -1-\frac{h}{2}p_2 & 2+h^2q_2 & -1+\frac{h}{2}p_2 & 0 & \cdots & 0 & 0\\ 0 & -1-\frac{h}{2}p_2 & 2+h^2q_2 & -1+\frac{h}{2}p_2 & 0 & 0 & 0\\ 0 & 0 & \ddots & \ddots & \ddots & 0 & 0\\ 0 & 0 & 0 & -1-\frac{h}{2}p_{N-1} & 2+h^2q_{N-1} & -1+\frac{h}{2}p_{N-1} & 0\\ 0 & 0 & 0 & 0 & -1-\frac{h}{2}p_{N-1} & 2+h^2q_{N-1} & -1+\frac{h}{2}p_{N-1}\\ 0 & 0 & \cdots & \cdots & 0 & 2+h^2q_N & -1+\frac{h}{2}p_N \end{pmatrix},$$

$$U_h\begin{pmatrix} u_1\\ u_2\\ u_3\\ \vdots\\ \vdots\\ u_{N-1}\\ u_N \end{pmatrix},\qquad b_h=\begin{pmatrix} -h^2r_1+(1+\frac{h}{2}p_1)\alpha\\ -h^2r_2\\ -h^2r_3\\ \vdots\\ \vdots\\ -h^2r_{N-1}\\ -h^2r_N+(1+\frac{h}{2}p_N)\beta \end{pmatrix}$$

with $p_i=p(x_i)$, $q_i=q(x_i)$ and $r_i=r(x_i)$, $\quad i=1,\cdots,N$.
The diagonal $(a_{ii})_{i=1}^N$ is formed by the elements

$$a_{ii}=2+h^2q_i,\qquad i=1,2,\cdots,N.$$

Below the diagonal, we have

$$a_{i,i-1}=2+h^2p_i,\qquad i=2,\cdots,N.$$

Above this diagonal, we have the elements

$$a_{i,i+1}=2+h^2p_i,\qquad i=1,\cdots,N-1.$$

All other elements of A are zero.

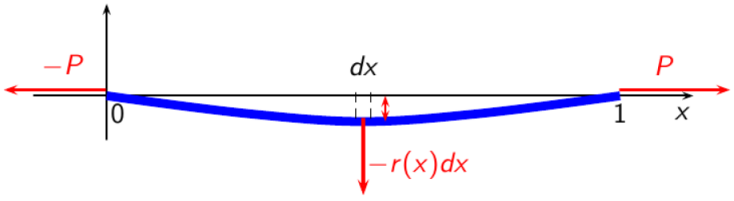
Theorem 6.3 *[2] Suppose p,q and r continuous on $[0,1]$. If $q(x)>0$ on $[0,1]$ then the linear tridiagonal system (6.4.10) has a unique solution for $h=\dfrac{2}{L}$ with*

$$L=\max_{0\le x\le 1}|p(x)|.$$

Application Given two functions $q,r\in C^0([0,1])$ and two constants α and β, find a function $u\in C^2([0,1])$ which verifies

$$\begin{cases} u''(x)=q(x)u(x)+r(x), & 0<x<1\\ u(\alpha)=0,\ u(1)=\beta. \end{cases}$$

This problem models, for example, the bending of a beam of length 1, stretched along its axis by a force P, subjected to a transverse load $-r(x)dx$ per unit length dx and simply pressed at its extremities 0 and 1 (see figure).



Then the bending moment $u(x)$ at the easting point x is the solution of a boundary value problem of the above type, with

$$q(x) = \frac{P}{E \cdot l(x)}$$

where

- E is the Young's modulus of the constituent material
- $l(x)$ the principal moment of inertia of the section of the beam at the point of abscissa x and with $\alpha = \beta = 0$.

If we suppose the function $q \geq 0$ on the interval $[0, 1]$, we can show that this problem has a solution. To illustrate the approximation of this type of problem by finite differences, we take in this problem $q \equiv 0$ and $r(x) = x$ with $\alpha = \beta = 0$. This gives the problem

$$\begin{cases} u''(x) = u(x) + x, & 0 < x < 1, \\ u(0) = 0,\ u(1) = 0, \end{cases}$$

whose exact solution is

$$u(x) = \frac{2e}{e^2 - 1}\sinh(x) - x.$$

Taking $h = \dfrac{1}{8}$ the system (6.4.10) is written with the matrix A and the second member b_h as follows

$$A_h = \begin{pmatrix} 2+h^2 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2+h^2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2+h^2 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2+h^2 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2+h^2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2+h^2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2+h^2 \end{pmatrix}$$

and U_h and b_h the vectors of $\mathbb{R}^N$

$$U_h = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \end{pmatrix}, \qquad b_h = -h^2 \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{pmatrix}.$$

so, A_h is the tridiagonal matrix $A_h = \text{tridiag}(-1, 2.015625, -1)$ and b_h is the vector

$$b_h^t = -(0.001953125, 0.00390625, 0.005859375, 0.0078125, 0.009765625, 0.0117875, 0.013671875).$$

This gives the condition

$$U_h^t = -(0.0183367, 0.0350068, 0.0483176, 0.0565240, 0.0578011, 0.0502157, 0.0316961).$$

Example 6.4.1 *Consider the following boundary value problem*

$$\begin{cases} y''(x) = \left(1 - \dfrac{x}{5}\right) y(x) + x, \\ y(1) = 2,\ y(3) = -1. \end{cases}$$

Find approximate values at points $x_1 = \dfrac{3}{2}$, $x_2 = 2$, $x_3 = \dfrac{5}{2}$ *so that* $h = 0.5$.

Correction 6.4.1 *The approximations* w_1, w_2, w_3 *of* $y(x)$ *at points* $x_1 = 1.5$, $x_1 = 2$, $x_1 =$*2.5 are the components of the following system solution*

$$\begin{pmatrix} a_1 & -c_1 & 0 \\ -b_2 & a_2 & -c_2 \\ 0 & -b_3 & a_3 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$$

where $a_i = 2 + h^2 q(x_i)$ *with* $i = 1, 2, 3$ *and* $b_i = 1 + \dfrac{h}{2} p(x_i)$ *with* $i = 2, 3$ *and* $c_i = 1 - \dfrac{h}{2} p(x_i)$ *with* $i = 1, 2$ *and we have*

$$\begin{aligned} r_1 &= -h^2 r(x_1) + \left(1 + \frac{h}{2} p(x_1)\right) w_0 \\ r_2 &= -h^2 r(x_2) \\ r_3 &= -h^2 r(x_3) + \left(1 + \frac{h}{2} p(x_3)\right) w_4 \end{aligned}$$

and where

$$q(x) = 1 - \frac{x}{5}, \qquad p(x) = 0, \qquad w_0 = 1, \qquad w_4 = 3.$$

We obtain

$$a_1 = 2 + (0.5)^2(1 - \frac{1.5}{5}) = 2.175,\ a_2 = 2 + (0.5)^2(1 - \frac{2}{5}) = 2.150,\ a_3 = 2 + (0.5)^2(1 - \frac{2.5}{5}) = 2.125,$$

and

$$b_2 = b_3 = 1,\ c_1 = c_2 = 1,$$

also that

$$r_1 = -(0.5)^2(1.5) + 2 = 1.625,\ r_2 = -0.5,\ r_3 = -(0.5)^2(2.5) - 1 = -1.625,$$

hence

$$\begin{pmatrix} 2.175 & -1 & 0 \\ -1 & 2.150 & -1 \\ 0 & -1 & 2.125 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$$

Resolving this system by Gauss's method (Thomas' algorithm), for example we have $w_1 = 0.552$, $w_2 = -0.424$ *and* $w_3 = -0.964$ *, we end the example.*

6.5 Exercises with solutions

Exercise 6.1 *Consider the Cauchy system*

$$\begin{cases} y'(t) = y(t) + e^{2t}, \\ y(0) = 2, \end{cases}$$

1. *Check that the analytical solution is* $y(t) = e^t + e^{2t}$.
2. *Make three iterations with* $h = 0.1$ *of explicit Euler method for this problem and calculate the error committed on* y_3 *by comparing the results with the analytical solution* $y(0, 3)$.

3. *Do three iterations with $h = 0.1$ of Runge-Kutta method 2 (RK2) and calculate the error committed on y_3 by comparing the results with the analytical solution $y(0,3)$.*

Exercise 6.2 *Use Euler's method to find the first four values of the solution y of the differential equation*

$$y' = \frac{y-t}{y+t},$$

which satisfies the initial condition $y(0) = 1$, taking the step $h = 10^{-1}$. Perform calculations with three exact decimal places.

Exercise 6.3 *We consider the following boundary value problem*

$$(P)\begin{cases} -u''(x) + u'(x) = 1, & 0 < x < 1 \\ u(0) = u(1) = 1, \end{cases}$$

1. *Write the discretization scheme of this problem by approximating the convection term $u'(x)$ by a centered derivation formula.*
2. *Give the matrix form of this schema.*

Exercise 6.4 *Consider the boundary value problem*

$$\begin{cases} u''(x) = u(x) + x,\ x \in [0,1] \\ u(0) = 0,\ u(1) = 0, \end{cases}$$

1. *Find the approximate values of u at the points $x_1 = 0.25$, $x_2 = 0.50$, and $x_3 = 0.75$ by the finite difference method ($h = 0.25$).*
2. *Calculate error committed on u_1, u_2, u_3, by comparing the results with the analytical solution $u(x_1)$, $u(x_2)$, $u(x_3)$, respectively.*

   (We give the exact solution $u(x) = \dfrac{2e}{e^2-1}\sinh(x) - x$ and $e \approx 2.71$).

6.5.1 Solutions

Solution 6.1 1. *We have $y(t) = e^t + e^{2t}$ then $y'(t) = e^t + 2e^{2t}$ as $y'(t) = y(t) + e^{2t}$ so $(e^t + 2e^{2t}) = (e^t + e^{2t}) + e^{2t}$ plus $y(0) = e^0 + e^{2(0)} = 2$. Then $y(t) = e^t + e^{2t}$ is the analytical solution of this problem.*

2. *We have $y' = y + e^{2t}$, $y(0) = 2$ and $h = 0.1$. We have that $t_0 = 0$, $y_0 = 2$ and $f(t_n, y_n) = y_n + e^{2t_n}$.*

$$\text{Explicit Euler}\begin{cases} y_{n+1} = y_n + hf(t_n, y_n) \\ t_{n+1} = t_n + h \end{cases}$$

$$\begin{aligned} y_0 &= 2 \\ y_1 &= 2 + (0.1) \times (y_0 + e^{2(0.1)}) = 2.3221... \\ y_2 &= 2.3221 + (0.1) \times (y_1 + e^{2(0.2)}) = 2.671... \\ y_3 &= 2.671 + (0.1) \times (y_2 + e^{2(0.3)}) = 3.053... \end{aligned}$$

The error is then given by

$$|y(0,3) - y_3| = |3.053 - 3.1700001557| = 0.1116788...$$

3. *As $y(t) = e^t + e^{2t}$ then $y(0,3) = 3.171977608$. Also, we have that $f(t_n, y_n) = y_n + e^{2t_n}$, that $y(0) = 2$ and therefore that $y_0 = 2$ and that $t_0 = 0$. We do 3 iterations with $h = 0.1$ by using the Runge-Kutta method 2 (RK2)*

$$\begin{cases} y_{n+1} = y_n + \dfrac{h}{2}(K_1 + K_2) & n = 1, ..., N-1 \\ K_1 = f(t_n, y_n),\ K_2 = f(t_n + h, y_n + K_1), \end{cases}$$

so,

$$\begin{aligned} \hat{y} = 2.3 &\implies y_1 = 2.2215688 \\ \hat{y} = 2.68081743 &\implies y2 = 2.712075889 \\ \hat{y} = 3.132465948 &\implies y3 = 3.1700001557 \end{aligned}$$

Then the error is given by

$$|y(0,3) - y_3| = |3.171977608 - 3.1700001557| = 0.00197745.$$

Solution 6.2 *For $h = 0.1$ the successive values of the independent variable will be $t_0 = 0$, $t_1 = 0.1$, $t_2 = 0.2$, $t_3 = 0.3$; $t_4 = 0.4$ with the differential equation*

$$y' = \frac{y-t}{y+t} = f(t,y),$$

which satisfies the initial condition $y(0) = 1$.

Let us calculate the respective values of the sought solution according to Euler's formula

$$\begin{aligned} y_{i+1} &= y_i + hf(t_i, y_i) \\ &= y_i + h\left(\frac{y_i - t_i}{t_i + y_i}\right), \end{aligned}$$

with $y_0 = 1$ and $i = 0, 1, 2, 3$.
For $i = 0$

$$\begin{aligned} y_1 &= y_0 + hf(t_0, y_0) \\ &= y_0 + h\left(\frac{y_0 - t_0}{t_0 + y_0}\right) \\ &= 1 + 0.1\left(\frac{1-0}{0+1}\right) \\ &= 1.1. \end{aligned}$$

For $i = 1$

$$\begin{aligned} y_2 &= y_1 + hf(t_1, y_1) \\ &= y_1 + h\left(\frac{y_1 - t_1}{t_1 + y_1}\right) \\ &= 1.1 + 0.1\left(\frac{1.1 - 0.1}{0.1 + 1.1}\right) \\ &= 1.183. \end{aligned}$$

For $i = 2$

$$\begin{aligned} y_3 &= y_2 + hf(t_2, y_2) \\ &= y_2 + h\left(\frac{y_2 - t_2}{t_2 + y_2}\right) \\ &= 1.183 + 0.1\left(\frac{1.183 - 0.2}{0.2 + 1.183}\right) \\ &= 1.254. \end{aligned}$$

For $i = 3$

$$\begin{aligned} y_4 &= y_3 + hf(t_3, y_3) \\ &= y_2 + h\left(\frac{y_3 - t_3}{t_3 + y_3}\right) \\ &= 1.254 + 0.1\left(\frac{1.254 - 0.3}{0.3 + 1.254}\right) \\ &= 1.315. \end{aligned}$$

Solution 6.3 *Consider the following boundary value problem*

$$(P)\begin{cases} -u''(x)+u'(x)=1, & 0<x<1 \\ u(0)=u(1)=1, \end{cases}$$

1. *Write the problem discretization scheme (P) by approximating the convection term $u'(x)$ by a centered derivation formula*

$$-\left(\frac{u_{i-1}-2u_i+u_{i+1}}{h^2}\right)+\left(\frac{u_{i+1}-u_{i-1}}{2h}\right)=1.$$

Hence the schema is written

$$-\left(\frac{1}{h^2}+\frac{1}{2h}\right)u_{i-1}+\frac{2}{h^2}u_i+\left(\frac{1}{2h}-\frac{1}{h^2}\right)u_{i+1}=1.$$
$$\textit{With} \qquad 1\le i\le n-1$$

2. *Let's give the matrix form of this schema*
We have the following system

$$\begin{cases} i=1, & -\left(\frac{1}{h^2}+\frac{1}{2h}\right)u_0+\frac{2}{h^2}u_1+\left(\frac{1}{2h}-\frac{1}{h^2}\right)u_2=1 \\ i=2, & -\left(\frac{1}{h^2}+\frac{1}{2h}\right)u_1+\frac{2}{h^2}u_2+\left(\frac{1}{2h}-\frac{1}{h^2}\right)u_3=1 \\ & \vdots \\ & \vdots \\ i=n-1, & -\left(\frac{1}{h^2}+\frac{1}{2h}\right)u_{n-2}+\frac{2}{h^2}u_{n-1}+\left(\frac{1}{2h}-\frac{1}{h^2}\right)u_n=1 \end{cases}$$

and since $u_0=u_n=0$, then
for $i=n-1$,

$$-\left(\frac{1}{h^2}+\frac{1}{2h}\right)u_0+\frac{2}{h^2}u_1+\left(\frac{1}{2h}-\frac{1}{h^2}\right)u_2=1\Longleftrightarrow\frac{2}{h^2}u_1+\left(\frac{1}{2h}-\frac{1}{h^2}\right)u_2=1,$$

for $i=1$,

$$-\left(\frac{1}{h^2}+\frac{1}{2h}\right)u_{n-2}+\frac{2}{h^2}u_{n-1}+\left(\frac{1}{2h}-\frac{1}{h^2}\right)u_n=1\Longleftrightarrow-\left(\frac{1}{h^2}+\frac{1}{2h}\right)u_{n-2}+\frac{2}{h^2}u_{n-1}=1.$$

So the matrix form of the scheme is

$$\begin{pmatrix} \frac{2}{h^2} & \left(\frac{1}{2h}-\frac{1}{h^2}\right) & 0 & \cdots & \cdots & 0 \\ -\left(\frac{1}{2h}+\frac{1}{h^2}\right) & \frac{2}{h^2} & \left(\frac{1}{2h}-\frac{1}{h^2}\right) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & -\left(\frac{1}{2h}+\frac{1}{h^2}\right) & \frac{2}{h^2} & \left(\frac{1}{2h}-\frac{1}{h^2}\right) & 0 \\ \vdots & \ddots & \ddots & \ddots & \frac{2}{h^2} & \left(\frac{1}{2h}-\frac{1}{h^2}\right) \\ 0 & \cdots & \cdots & 0 & -\left(\frac{1}{2h}+\frac{1}{h^2}\right) & \frac{2}{h^2} \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ u_N \end{pmatrix}=\begin{pmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1 \end{pmatrix}$$

Solution 6.4 *We consider the boundary value problem*

$$\begin{cases} u''(x)=u(x)+x,\ x\in[0,1] \\ u(0)=0,\ u(1)=0. \end{cases}$$

1. *We will find the approximate values of u at the points $x_1 = 0.25$, $x_2 = 0.50$, $x_3 = 0.75$ by the finite difference method.*

   We set x_0 and $x_4 = 1$ then $h = x_{i+1} - x_i = 0.25$ with $i = 0, 1, 2, 3, 4$. According to the finite difference method, we have

$$A_h = \begin{pmatrix} 2+h^2 & -1 & 0 \\ -1 & 2+h^2 & -1 \\ 0 & -1 & 2+h^2 \end{pmatrix}$$

   and

$$b_h = -h^2 \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

   So the tridiagonal matrix $A_h = \text{tridiag}(-1, 2.0625, -1)$ and b_h is the vector $b_h^t = -(0.0625, 0.125, 0.1875)$.

   In summary

$$A_h U_h = b_h \iff \begin{pmatrix} 2.0625 & -1 & 0 \\ -1 & 2.0625 & -1 \\ 0 & -1 & 2.0625 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = - \begin{pmatrix} 0.0625 \\ 0.125 \\ 0.1875 \end{pmatrix}$$

   This gives the solution $U_h^t = (u_1, u_2, u_3) = (-0.0350068, -0.0565240, -0.0578011)$.

2. *Errors are given by*

$$\begin{aligned} |u(x_1) - u_1| &= |-0.0341839 - (-0.0350068)| = 0.0008229... \\ |u(x_2) - u_2| &= |-0.0548089 - (-0.0565240)| = 0.0017151... \\ |u(x_3) - u_3| &= |-0.0474643 - (-0.0578011)| = 0.0103368... \end{aligned}$$

6.6 Exercises without solutions

6.6.1 Cauchy problems

Exercise 6.1 *Consider the differential equation with initial condition:*

$$\begin{cases} y'(t) = y(t) + t, & t \in [0, 1] \\ y(0) = 1. \end{cases}$$

Approach the solution of this equation in $t = 1$ using Euler's method by dividing the interval into 10 equal parts. Compare to the exact solution.

Exercise 6.2 *Consider the following Cauchy problem*

$$\begin{cases} y'(t) = 2t - y(t), & t \in [0, 1] \\ y(0) = 1. \end{cases}$$

1. *Find the exact solution to this problem.*
2. *Apply Euler's method to this problem, with $h = 0.1$ then evaluate the solution at $t = 0.3$. Compare to the exact solution.*

Exercise 6.3 *Approach the solution of the differential equation below in $t = 0.2$ using RK2, with a step $h = 0.2$*

$$\begin{cases} y'(t) = y(t) - \dfrac{2t}{y(t)}, \\ y(0) = 1. \end{cases}$$

Compare to the exact solution.

Exercise 6.4 *Consider the following Cauchy problem*

$$\begin{cases} y'(t) = t + y(t), & t \in [0, 2] \\ y(0) = 1.24, \end{cases}$$

which has the analytical solution

$$y(t) = 2.24\, e^t + (t - 1).$$

a) Solve numerically the same problem using the method

1. *of Euler, with a step* $h = 0, 2$.
2. *Runge-Kutta of order* 4, *with step* $h = 1$.

- *Compare at the point* $t = 1$, *the numerical values with the analytical value and give (in %) the relative error committed by each of the two methods.*

Exercise 6.5 *By giving the solutions of the differential equation below with the initial condition* $y(0) = 1$ *then* $y(0) = 1 + \epsilon$, ϵ *nonzero real, check that it leads to unstable patterns.*

$$y'(t) = 36y(t) - 37e^{-t}.$$

Exercise 6.6 *Write the differential equation modeling the movement of the simple pendulum. Apply the Euler method then the Taylor method of order* 2.
Hint: *We have* $ay''(t) + g\sin(y(t)) = 0$ *such that* a *is the acceleration and* g *is the gravity.*

Exercise 6.7 *Consider the second-order differential equation with initial conditions*

$$\begin{cases} y''(t) + 2y'(t) = 2y(t), & t \in [a, b] \\ y(a) = 1 \quad \text{and} \quad y'(a) = 2. \end{cases}$$

1. *Write this differential equation as a differential system of two differential equations of order one.*
2. *Apply Runge-Kutta method (RK2) to this system.*

6.6.2 Dirichlet problems

Exercise 6.8 *Consider the problem with boundary conditions*

$$\begin{cases} u''(x) = \pi^2 \sin(\pi x), & x \in]0, 1[\\ u(0) = 0, \quad u(1) = 0. \end{cases}$$

1. *Write the finite difference scheme of this problem for* $h = \frac{1}{4}$ *and* $h = \frac{1}{8}$.
2. *Compare approximate results with exact result.*

Exercise 6.9 *We consider the problem*

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), & 0 < x < 1 \\ u(0) = a, \quad u(1) = b. \end{cases} \tag{6.6.1}$$

where $c \in C([0, 1], \mathbb{R}_+)$, $f \in C([0, 1], \mathbb{R})$ *and* $(a, b) \in \mathbb{R}^2$.

1. *Give a finite difference discretization of this problem. We call* u_h *the approximate solution ie* $u_h = (u_1, \ldots, u_N)$ *where* u_i *is the discrete unknown in* x_i.
2. *We assume here that* $c = 0$. *Show that* $u_i > \min(a, b)$, *for all* $i = 1, \ldots, N$.

Exercise 6.10 *Consider the problem with boundary conditions*

$$\begin{cases} u''(x) = u(x) + x, & x \in]0, 1[\\ u(0) = u(1) = 0. \end{cases}$$

1. *Check that the exact solution is* $u(x) = \frac{2e}{e^2-1}\sinh(x) - x$.
2. *Find the finite difference approximation of this problem for* $h = \frac{1}{4}$ *and* $h = \frac{1}{8}$ *respectively.*
3. *Compare approximate results with exact result.*

Exercise 6.11 *Let us consider the problem with the following limits*

$$\begin{cases} y''(x) = \left(1 - \dfrac{x}{5}\right)y(x) + x, & x \in]1, 3[\\ y(1) = 2, \quad y(3) = -1. \end{cases}$$

- *Find the approximate values of* y *at the points* $x_1 = 1.5; x_2 = 2; x_3 = 2.5$.

Exercise 6.12 *We are interested in the following one-dimensional elliptic problem*

$$\begin{cases} -u''(x) + 2u(x) = x, & 0 < x < 1 \\ u(0) = 1, \quad u'(1) + u(1) = 0. \end{cases}$$

1. *Write a discretization of this finite difference problem for a uniform of sub-intervals.*
2. *Write the resulting linear system.*

Chapter 7

Numerical calculation of eigenvalues and eigenvectors

7.1 Preliminaries

Definition 7.1 *Let $A \in M_n(\mathbb{R})$, we say that $\lambda \in \mathbb{R}$ is an eigenvalue of A, if there exists a vector $X \in \mathbb{R}^n$ nonzero such that*

$$AX = \lambda X,$$

X is the eigenvector associated with the eigenvalue λ.

Example 7.1.1 *Let*

$$A = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{pmatrix}, \qquad \text{and} \qquad X = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}. \tag{7.1.1}$$

We have

$$AX = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 0 \end{pmatrix} = 2 \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = 2X = \lambda X, \tag{7.1.2}$$

so,$X^t = (2, 1, 0)$ is an eigenvector of A with eigenvalue $\lambda = 2$.

Definition 7.2 *The matrix characteristic polynomial A is written as follows:*

$$P_A(\lambda) = \det(A - \lambda I).$$

Remark 7.1 *1. If $n = 2$ then $A \in M_2(\mathbb{R})$ so that*

$$P_A(\lambda) = \det(A - \lambda I) = \lambda^2 - \operatorname{tr}(A)\lambda + \det(A),$$

the classic method

$$AX = \lambda X \Longleftrightarrow (A - \lambda I)X = 0_{\mathbb{R}^n} \Longrightarrow \det(A - \lambda I) = 0.$$

2. Let $A \in M_n(\mathbb{R})$, $n \in \mathbb{N}$ and the eigenvalue $\lambda_i \in \mathbb{R}^*$ of A associated by the eigenvector $V_i \in \mathbb{R}^n$ with $i = 1, ..., n$, we note that

$$\lambda = \lambda_i \longrightarrow V = V_i.$$

Example 7.1.2 *Let*

$$A = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{pmatrix}, \qquad (A - \lambda I) = \begin{pmatrix} 1 - \lambda & 2 & 2 \\ 0 & 2 - \lambda & 1 \\ -1 & 2 & 2 - \lambda \end{pmatrix},$$

then

$$\begin{aligned}\det(A-\lambda I) &= \begin{vmatrix} 1-\lambda & 2 & 2 \\ 0 & 2-\lambda & 1 \\ -1 & 2 & 2-\lambda \end{vmatrix} \\ &= (1-\lambda)\begin{vmatrix} 2-\lambda & 1 \\ 2 & 2-\lambda \end{vmatrix} - \begin{vmatrix} 2 & 2 \\ 2-\lambda & 1 \end{vmatrix} \\ &= (1-\lambda)((2-\lambda^2)-2)+2(2-\lambda)-2 \\ &= 4-8\lambda+5\lambda^2-\lambda^3,\end{aligned}$$

so that,

$$P_A(\lambda) = 4-8\lambda+5\lambda^2-\lambda^3,$$

the three roots of this polynomial are

- $\lambda = 1$,
- $\lambda = 2$ *(double root),*

the eigenvectors.

For $\lambda = 1$

$$\begin{aligned}AX-\lambda IX &= \begin{pmatrix} 1-\lambda & 2 & 2 \\ 0 & 2-\lambda & 1 \\ -1 & 2 & 2-\lambda \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} \\ &= \begin{pmatrix} 0 & 2 & 2 \\ 0 & 1 & 1 \\ -1 & 2 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix},\end{aligned}$$

Check for $\lambda = 1$ *If* $(x, y, z) = (-1, -1, 1)$, *we have*

$$AX = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{pmatrix}\begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} = 1\begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} = \lambda X,$$

for $\lambda = 2$

$$\begin{aligned}AX-\lambda IX &= \begin{pmatrix} 1-\lambda & 2 & 2 \\ 0 & 2-\lambda & 1 \\ -1 & 2 & 2-\lambda \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} \\ &= \begin{pmatrix} -1 & 2 & 2 \\ 0 & 0 & 1 \\ -1 & 2 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix},\end{aligned}$$

for $\lambda = 2$, *if* $(x, y, z) = (2, 1, 0)$, *we have*

$$AX = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = 2\begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \lambda X.$$

In summary,

$$A = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{pmatrix}$$

and

$$\lambda = 1 \longrightarrow X = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, \qquad \lambda = 2 \longrightarrow X = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \quad \textit{double root}$$

7.1.1 Diagonalization

The problem comes down to finding a diagonal matrix D

$$D = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n),$$

and a regular matrix P such that

$$A = PDP^{-1},$$

the λ_i are the eigenvalues of A and the columns of P the associated eigenvectors, where

$$P = (X_1, X_2, \cdots, X_n).$$

Remark 7.2 *In the previous example P is not invertible because 2 eigenvectors are identical.*

Example 7.1.3 *Let*

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 2 & -4 & 2 \end{pmatrix}.$$

Calculate λ_i, v_i, i=1,2,3 then write $A = PDP^{-1}$, calculate A^n.

Correction 7.1.1 *Eigenvectors and eigenvalues*

$$\det(A - \lambda I) = P_A(\lambda) = (\lambda - 3)(\lambda - 2)(\lambda - 1),$$

so that

$$\lambda = 3 \quad \longrightarrow \quad V_1 = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix},$$

$$\lambda = 2 \quad \longrightarrow \quad V_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\lambda = 1 \quad \longrightarrow \quad V_3 = \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}.$$

We want $A = PDP^{-1}$ and

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} -1 & 0 & -1 \\ -1 & 0 & 0 \\ 2 & 1 & 2 \end{pmatrix}, \quad P^{-1} = \begin{pmatrix} 0 & -1 & 0 \\ 2 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix},$$

then

$$A = PDP^{-1} \Longleftrightarrow \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 2 & -4 & 2 \end{pmatrix} = \begin{pmatrix} -1 & 0 & -1 \\ -1 & 0 & 0 \\ 2 & 1 & 2 \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ 2 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix}$$

finally

$$\begin{aligned} A^n &= (PDP^{-1})^n = \underbrace{(PDP^{-1}) \cdot (PDP^{-1}) \cdots (PDP^{-1})}_{n \text{ times}} \\ &= PD^nP^{-1}, \end{aligned}$$

such as

$$D^n = \begin{pmatrix} 3^n & 0 & 0 \\ 0 & 2^n & 0 \\ 0 & 0 & 1^n \end{pmatrix}.$$

Remark 7.3 *The applications of calculation of eigenvalues and eigenvectors*

- *in quantum mechanics (Hamiltonian matrices).*
- *in solid mechanics (resonance frequencies of a hamornic oscillator).*
- *in geology (study of earthquakes).*
- *in electronics.*
- *in statistics.*
- *in economics.....*

7.2 Numerical method for calculating eigenvalues and eigenvectors

In general, the Leverrier-Souriau method and Krylov method are used to numerically determine the characteristic polynomial and the eigenvectors and Bernoulli method to calculate the eigenvalues.

7.2.1 Leverrier-Souriau method

Let $A \in M_n(\mathbb{R})$ be the range n, or the characteristic polynomial

$$p_n(\lambda) = \lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + \cdots + a_{n-1}\lambda + a_n.$$

We determine $a_1, a_2, \cdots, a_n$ as follows

$$\begin{cases} B_0 = I, \quad a_1 = -\mathrm{tr}(AB_0) \\ B_1 = AB_0 + a_1 I, \quad a_2 = -\dfrac{1}{2}\mathrm{tr}(AB_1) \\ B_2 = AB_1 + a_2 I, \quad a_3 = -\dfrac{1}{3}\mathrm{tr}(AB_2) \\ \quad \vdots \\ B_{n-1} = AB_{n-2} + a_{n-1} I, \quad a_n = -\dfrac{1}{n}\mathrm{tr}(AB_{n-1}) \end{cases}$$

then

$$B_n = AB_{n-1} + a_n I \equiv 0,$$

hence

$$A^{-1} = -\frac{1}{a_n} B_{n-1}.$$

If λ_i is an eigenvalue of matrix A, then

$$V_i = \lambda_i^{n-1} B_0 + \lambda_i^{n-2} B_1 + \cdots + \lambda_i B_{n-2} + B_{n-1},$$

such that V_i the eigenvector (column i of P).

Example 7.2.1 *We assume the following matrix*

$$A = \begin{pmatrix} -1 & 3 & 3 \\ 3 & -1 & -3 \\ -3 & 3 & 5 \end{pmatrix}$$

find $P_A(\lambda)$ by the Leverrier-Souriau method.

Correction 7.2.1 *We have*

$$\begin{array}{rcl} AB_0 = AI = A \Longrightarrow a_1 = -\mathrm{tr}(AB_0) = -\mathrm{tr}(A) = -3, \\ \left.\begin{array}{rcl} B_1 & = & AB_0 + a_1 I \\ & = & A - 3I \end{array}\right\} \Longrightarrow a_2 = -\frac{1}{2}\mathrm{tr}(AB_1) = 0, \\ \left.\begin{array}{rcl} B_2 & = & AB_1 + a_2 I \\ & = & AB_1 \end{array}\right\} \Longrightarrow a_3 = -\frac{1}{3}\mathrm{tr}(AB_2) = 4, \end{array}$$

so the characteristic polynomial of A is

$$P_A(\lambda) = \lambda^3 - 3\lambda^2 + 4.$$

Remark that $\lambda_1 = -1$ is an eigenvalue of matrix A, because $P_A(-1) = 0$, then

$$\begin{array}{rcl} V_1 & = & (-1)^{3-1}B_0 + (-1)^{3-2}B_1 + B_2 \\ & = & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} -4 & 3 & 3 \\ 3 & -4 & -3 \\ -3 & 3 & 2 \end{pmatrix} + \begin{pmatrix} 4 & -6 & -6 \\ -6 & 4 & 6 \\ 6 & -6 & -8 \end{pmatrix} \\ & = & \begin{pmatrix} 9 & -9 & -9 \\ -9 & 9 & 9 \\ 9 & -9 & -9 \end{pmatrix} \end{array}$$

we write $V_1 = (9, -9, -9)^t$ the eigenvector of A (column 1 of P), and we have

$$A^{-1} = -\frac{1}{a_3}B_2 = -\frac{1}{4}\begin{pmatrix} 4 & -6 & -6 \\ -6 & 4 & 6 \\ 6 & -6 & -8 \end{pmatrix}$$

7.2.2 Krylov's method

Let $A \in M_n(\mathbb{R})$ be the range n, we search for the eigenvalues and vectors by the Krylov method, using the Hamilton-Cayley theorem

Theorem 7.1 *[2] Let $A \in M_n(\mathbb{R})$ and P_A be the characteristic polynomial*

$$P_A(A) = A^n + a_1 A^{n-1} + a_2 A^{n-2} + \cdots + a_{n-1}A + a_n I = 0_{M_n(\mathbb{R})},$$

hence

$$A^n = -\sum_{k=1}^{n} a_k A^{n-k}.$$

therefore $y^{(0)}$ the non-zero characteristic vector, we assume the sequence

$$\begin{array}{rcl} y^{(1)} & = & Ay^{(0)} \\ y^{(2)} & = & Ay^{(1)} = A^2 y^{(0)} \\ & \vdots & \\ y^{(n-1)} & = & Ay^{(n-2)} = A^{n-1}y^{(0)}. \end{array}$$

We obtain

$$\begin{array}{rcl} y^{(n)} & = & A^n y^{(0)} \\ & = & -\sum_{k=1}^{n} a_k A^{n-k} y^{(0)} \\ & = & -\sum_{k=1}^{n} a_k y^{(n-k)}. \end{array}$$

We take note that

$$y^{(0)} = \begin{pmatrix} y_1^{(0)} \\ \vdots \\ y_n^{(0)} \end{pmatrix}, \quad y^{(1)} = \begin{pmatrix} y_1^{(1)} \\ \vdots \\ y_n^{(1)} \end{pmatrix}, \quad \cdots, y^{(n)} = \begin{pmatrix} y_1^{(n)} \\ \vdots \\ y_n^{(n)} \end{pmatrix}.$$

The polynomial parameters P_A is the solution of the following linear system

$$\begin{pmatrix} y_1^{(n-1)} & y_1^{(n-2)} & \cdots & y_1^{(0)} \\ y_2^{(n-1)} & y_2^{(n-2)} & \cdots & y_2^{(0)} \\ \vdots & \vdots & \cdots & y_1^{(0)} \\ y_n^{(n-1)} & y_n^{(n-2)} & \cdots & y_n^{(0)} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = - \begin{pmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_n^{(n)} \end{pmatrix}.$$

We find the eigenvalues λ_i, assuming that $\lambda_i \neq \lambda_j$ with $i \neq j$ then the eigenvectors given by

$$V_i = \sum_{k=0}^{n-1} b_{n-k} y^{(k)}, \qquad i = 1, 2, 3, \cdots$$

such that b_i is the polynomial parameter

$$\frac{P_A(\lambda)}{\lambda - \lambda_i} = \sum_{k=0}^{n-1} b_{n-k} \lambda^k.$$

Example 7.2.2 *By Krylov's method, calculate the following matrix eigenvalues and vectors*

$$A = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

taking

$$y^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Correction 7.2.2 *We have*

$$\begin{aligned} y^{(1)} &= Ay^{(0)} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \\ y^{(2)} &= Ay^{(1)} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix} \\ y^{(3)} &= Ay^{(2)} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 6 \end{pmatrix}. \end{aligned}$$

The characteristic polynomial parameters is the following linear system solution

$$(y^{(2)}, y^{(1)}, y^{(0)}) \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = -y^{(3)},$$

either

$$\begin{pmatrix} 3 & 2 & 1 \\ 1 & 1 & 1 \\ 3 & 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = - \begin{pmatrix} 4 \\ 2 \\ 6 \end{pmatrix},$$

so that

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = - \begin{pmatrix} -4 \\ 6 \\ -4 \end{pmatrix},$$

then, the characteristic polynomial

$$P_A(\lambda) = \lambda^3 - 4\lambda^2 + 6\lambda - 4,$$

so,

$$\begin{aligned} \lambda_1 = 2 \quad &\Longrightarrow \quad \frac{P_A(\lambda)}{\lambda - 2} = \lambda^2 - 2\lambda + 2 \\ &\Longrightarrow \quad V_1 = y^{(2)} - 2y^{(1)} + 2y^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \end{aligned}$$

7.2.3 Bernoulli method

Let the algebraic equation

$$P_n(x) = \sum_{k=0}^{n} a_k x^{n-k} = 0, \tag{7.2.1}$$

whose roots are assumed to be distinct. It is associated with the difference equation

$$a_0 y_{n+k} + a_1 y_{n+k-1} + \cdots + a_n y_n = 0, \qquad k = 0, 1, 2, \cdots \tag{7.2.2}$$

which is a linear recurrence relation between $(n+1)$ elements of a sequence of numbers $y_0, y_1, \cdots, y_n$.

A sequence (z_k) is said to be a solution of (7.2.2) if $(n+1)$ consecutive elements of this sequence satisfy this equation. In the theory of finite difference equations, a recurrent equation admits several solutions. The equation (7.2.1) is called the characteristic equation of the equation (7.2.2).

If (7.2.1) admits n distinct roots $x_1, x_2, \cdots, x_n$, any sequence (z_k) solution has as elements

$$z_k = C_1 x_1^k + C_2 x_2^k + \cdots + C_n x_n^k, \; k = 1, 2, \cdots, C_1, C_2, \cdots = \text{constants}.$$

If we suppose that the equation (7.2.1) admits a root greater in modulus than all the others, for example $|x_1| > |x_2 \; vert > |x_3| > \cdots > |x_n|$. Two consecutive solutions y_k and y_{k+1} of (7.2.2) are written

$$\begin{aligned} y_k &= x_1^k \left[C_1 + C_2 \left(\frac{x_2}{x_1} \right)^k + \cdots + C_n \left(\frac{x_n}{x_1} \right)^k \right], \\ y_{k+1} &= x_1^{k+1} \left[C_1 + C_2 \left(\frac{x_2}{x_1} \right)^{k+1} + \cdots + C_n \left(\frac{x_n}{x_1} \right)^{k+1} \right], \end{aligned}$$

hence the report

$$\frac{y_{k+1}}{y_k} = x_1 \frac{C_1 + C_2 \left(\frac{x_2}{x_1} \right)^{k+1} + \cdots + C_n \left(\frac{x_n}{x_1} \right)^{k+1}}{C_1 + C_2 \left(\frac{x_2}{x_1} \right)^{k} + \cdots + C_n \left(\frac{x_n}{x_1} \right)^{k}}.$$

If we assume that $C_1 \neq 0$, we will have, $\lim_{k \longrightarrow \infty} \dfrac{y_{k+1}}{y_k} = x_1$ because each of the ratios $\dfrac{x_2}{x_1}, \dfrac{x_3}{x_1}, \cdots, \dfrac{x_n}{x_1}$ being less than 1 in modulus, $\left(\dfrac{x_2}{x_1} \right)^k, \left(\dfrac{x_3}{x_1} \right)^k, \cdots, \left(\dfrac{x_n}{x_1} \right)^k$ tend to zero when $k \longrightarrow \infty$.

Bernoulli's method therefore makes it possible to calculate the greatest root in modulus of the equation (7.2.1).

In practice we give ourselves n arbitrary numbers $y_0, y_1, \cdots, y_{n-1}$ from which we construct the sequence (y_n) solution of (7.2.2), of elements:

$$y_{n+k} = -\frac{1}{a_0}\Big(a_1 y_{n+k-1} + a_2 y_{n+k-2} + \cdots + a_n y_k\Big), \qquad k = 1, 2, 3, \cdots$$

then we evolve the reports $\frac{y_{n+1}}{y_n}, \cdots, \frac{y_{n+k+1}}{y_{n+k}}, \cdots$
Note that the solution sequence corresponding to $C_1 = C_2 = \cdots = C_n = 1$ is none other than the sequence (S_k) of elements

$$S_k = x_1^k + x_2^k + \cdots + x_n^k.$$

One can thus in this case use the relations of Newton for the construction of this sequence. We will then have

$$x_1 = \lim_{k \longrightarrow \infty} \frac{S_{k+1}}{S_k}.$$

Example 7.2.3 *Using Bernoulli's method, solve the equation*

$$2x^3 + 9x^2 - 33x + 14 = 0.$$

We have

$$x_1 = \lim_{p \longrightarrow \infty} \frac{y_p}{y_{p-1}}$$

where the y_p are determined from the recurrence relations

$$y_{k+3} = \frac{1}{2}(-9y_{k+2} + 33y_{k+1} - 14y_k),\ k = 0, 1, 2, \cdots$$

Taking as a starting point $y_0 = y_1 = 0$ and $y_2 = 1$, we obtain the following table

p	y_p	$\frac{y_p}{y_{p-1}}$
0	0	
1	0	
2	1	
3	−4.5	−4.5
4	36.75	−8.16666
5	−246.625	−6.71088
6	1747	−7.08641
7	−12191.15625	−6.97559
8	85423.4219	−7.00699
9	−597802.0548	−6.99810
10	4184933.802	−7.00053

it is clear that this sequence converges to the solution $x_1 = 7$. To determine the following roots, we calculate the quotient

$$\frac{P_3(x)}{x+7}$$

using Horner's scheme

$x = -7$	
2	
9	2
−33	−5
14	2
0	

then $\frac{P_3(x)}{x+7} = 2x^2 - 5x + 2$. *If we again apply Bernoulli's method to the polynomial $P_2(x)$, we have*

0	0	
1	1	
2	2.5	2.5
3	5.25	2.1
4	10.625	2.02380
5	21.3125	2.00588
6	42.65625	2.00146

it is clear that this sequence converges to the solution $x_2 = 2$, the last solution is obtained by dividing $P_2(x)$ by $(x = 2)$ as follows:

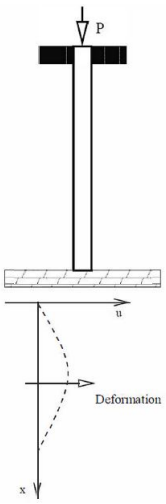
$x = 2$	
2	
-5	2
2	-1

Let $P_1(x) = 2x - 1$, the last solution is $x_3 = \dfrac{1}{2}$.

7.3 Application to the continuous case

The search for eigenvalues and eigenvectors is a problem that intervenes naturally in the study of structural dynamics. For example

Example 7.3.1 *Given a bar of length l, fixed at one end H. We applies a force P towards down in the direction of the axis. When P is low, no deformation of the rod. When P increases, a critical value $\overline{P}$ is reached from which the bar deforms. We denote $u(x)$ the displacement of the point located at the perpendicular abscissa at the axis of the rod.*



For small moves, u checks

$$\begin{cases} \frac{d}{dx}\left(a(x)\frac{du}{dx}\right) + Pu = 0, \\ u(0) = u(H) = 0, \end{cases} \tag{7.3.1}$$

with $a(x)$ depends on the characteristics of the bar (section, modulus of elasticity). If a =constant, then

$$\begin{cases} -u''(x) = Pu(x), \\ u(0) = u(H) = 0, \end{cases} \tag{7.3.2}$$

such that P is an eigenvalue. The analytics solutions are $u_k(x) = \sin\left(\dfrac{k\pi x}{H}\right)$, and $P_k = \dfrac{\pi^2 k^2}{H^2}$, $k \in \mathbb{N}^*$. Practically here, we are interested in the smallest eigenvalue $P_1 = \frac{\pi^2}{H^2}$.

If $a = a(x)$, there is no analytical solution, in general. We can then seek an approximate solution by discretizing the problem for example by the method finite differences:

Let $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = H$ be a regular subdivision of the interval $[0, H]$ i.e. $h = x_{i+1} - x_i$ with $i = 1, \cdots, N$ and we note $h = \frac{H-0}{N}$ the step of this subdivision and

$x_{i+\frac{1}{2}} = \frac{x_{i+1}+x_i}{2} = x_i + \frac{1}{2}h$ *so*

$$\begin{cases} \frac{1}{h^2}\Big(a_{i+\frac{1}{2}}(u_{i+1}-u_i) - a_{i-\frac{1}{2}}(u_i-u_{i-1})\Big) + Pu_i = 0,\ i = 1,..,N \\ u_0 = u_{N+1} = 0, \end{cases} \tag{7.3.3}$$

such as $a_{i+\frac{1}{2}} = a(x_{i+\frac{1}{2}})$, $a_{i-\frac{1}{2}} = a(x_{i-\frac{1}{2}})$ *and* $u_i = u(x_i)$ *with* $i = 1, \cdots, N$ *and* $u_0 = u(x_0)$, $u_{N+1} = u(x_{N+1})$.

This system is equivalent to $AU = \lambda U$ *with* $\lambda = P$ *the smallest eigenvalue and it designates the critical load.*

7.4 Exercises with solutions

Exercise 7.1 *Let* $A \in M_3(\mathbb{R})$ *such that*

$$A = \begin{pmatrix} 0 & 2 & -1 \\ 3 & -2 & 0 \\ -2 & 2 & 1 \end{pmatrix}.$$

1. *Calculate the characteristic polynomial using:*
   a) *The Leverrier-Souriau method.*
   b) *Krylov's method, we take* $y^{(0)} = (1,1,0)$ *the characteristic vector.*
2. *Determine the eigenvalues and eigenvectors of* A.
3. *Write* A *in the form* PDP^{-1}, *then calculate* A^n *with* $n \in \mathbb{N}$.

Exercise 7.2 *Let*

$$M = \begin{pmatrix} 4 & 6 & 0 \\ -3 & -5 & 0 \\ -3 & -6 & -5 \end{pmatrix}.$$

1. *Compute the characteristic polynomial* $P_M(\lambda)$ *by the de Leverier-Souriau method.*
2. *Deduce* M^{-1}.
3. *Calculate* λ_1, λ_2, λ_3 *by Bernoulli method.*
4. *Deduce* V_1, V_2, V_3 *such that* $MV_k = \lambda_k V_k$, $k = 1,2,3$.
5. *Determine a matrix* P *with* $M = PDP^{-1}$, $D = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$ *and* P^{-1}.

Exercise 7.3 *Let*

$$M = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 4 \end{pmatrix}$$

1. *Determine the caracteristic polynomial* $P_M(\lambda)$ *by the Leverier-Souriau method.*
2. *Deduce* λ_1, λ_2, λ_3 *and* V_1, V_2, V_3 *such that* $MV_k = \lambda_k V_k$, $k = 1,2,3$.
3. *Determine a matrix* P *with* $M = PDP^{-1}$ *and* $D = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$ *and* $P^{-1} = P^t$.

7.4.1 Solutions

Solution 7.1 *Let $A \in M_3(\mathbb{R})$ such that*

$$A = \begin{pmatrix} 0 & 2 & -1 \\ 3 & -2 & 0 \\ -2 & 2 & 1 \end{pmatrix}.$$

1. *Let be the characteristic polynomial of the matrix A*

$$p_3(\lambda) = \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3$$

We use the methods

a) The Leverrier-Souriau method. First, we have

$$B_0 = I_3 \Longrightarrow a_1 = -\mathrm{tr}(AB_0) - \mathrm{tr}(A) = -(-1) = 1$$

and like that

$$B_1 = AB_0 + a_1 I_3 = AI_3 + I_3 = \begin{pmatrix} 1 & 2 & -1 \\ 3 & -1 & 0 \\ -2 & 2 & 2 \end{pmatrix}$$

so,

$$AB_1 = \begin{pmatrix} 0 & 2 & -1 \\ 3 & -2 & 0 \\ -2 & 2 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 & -1 \\ 3 & -1 & 0 \\ -2 & 2 & 2 \end{pmatrix} = \begin{pmatrix} 8 & -4 & -2 \\ -3 & 8 & -3 \\ 2 & -4 & 4 \end{pmatrix}$$

so that,

$$a_2 = -\frac{1}{2}\mathrm{tr(AB_1)} = -\frac{1}{2}(8+8+4) = -10$$

and in addition

$$B_2 = AB_1 - 10I_3 = \begin{pmatrix} 8 & -4 & -2 \\ -3 & 8 & -3 \\ 2 & -4 & 4 \end{pmatrix} - \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix} = \begin{pmatrix} -2 & -4 & -2 \\ -3 & -2 & -3 \\ 2 & -4 & -6 \end{pmatrix}$$

hence,

$$AB_2 = \begin{pmatrix} 0 & 2 & -1 \\ 3 & -2 & 0 \\ -2 & 2 & 1 \end{pmatrix}\begin{pmatrix} -2 & -4 & -2 \\ -3 & -2 & -3 \\ 2 & -4 & -6 \end{pmatrix} = \begin{pmatrix} -8 & 0 & 0 \\ 0 & -8 & 0 \\ 0 & 0 & -8 \end{pmatrix}$$

we obtain

$$a_3 = -\frac{1}{3}\mathrm{tr(AB_2)} = -\frac{1}{3}(-8-8-8) = 8.$$

Finally,

$$\begin{aligned} p_3(\lambda) &= \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 \\ &= \lambda^3 + \lambda^2 - 10\lambda + 8. \end{aligned}$$

b) The Krylov's method (we take $y^{(0)} = (1,1,0)$ the characteristic vector). We have

$$y^{(1)} = Ay^{(0)} = \begin{pmatrix} 0 & 2 & -1 \\ 3 & -2 & 0 \\ -2 & 2 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$

and

$$y^{(2)} = Ay^{(1)} = \begin{pmatrix} 0 & 2 & -1 \\ 3 & -2 & 0 \\ -2 & 2 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ -2 \end{pmatrix}$$

and

$$y^{(3)} = Ay^{(2)} = \begin{pmatrix} 0 & 2 & -1 \\ 3 & -2 & 0 \\ -2 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \\ -2 \end{pmatrix} = \begin{pmatrix} 10 \\ -2 \\ 2 \end{pmatrix}.$$

We solve the following system

$$y^{(3)} = Ay^{(2)} = \begin{pmatrix} 2 & 2 & 1 \\ 4 & 1 & 1 \\ -2 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = - \begin{pmatrix} -10 \\ 2 \\ -2 \end{pmatrix}.$$

That's to say

$$\begin{cases} 2a_1 + 2a_2 + a_3 = -10 & (1) \\ 4a_1 + a_2 + a_3 = 2 & (2) \\ -2a_1 = -2 & (3) \end{cases}$$

According to the equation (3) *we have* $a_1 = 1$, *which implies*

$$\begin{cases} 2a_2 + a_3 = -12 & (4) \\ a_2 + a_3 = -2 & (5) \end{cases}$$

take (4) − (5), *we get*

$$a_2 = -10,$$

and from the equation (5) *we have*

$$a_1 = 8,$$

Finally,

$$\begin{aligned} p_3(\lambda) &= \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 \\ &= \lambda^3 + \lambda^2 - 10\lambda + 8. \end{aligned}$$

2. *We determine the eigenvalues and vectors of A, we note that:*

$$p_3(1) = 0$$

then $\lambda = 1$ *is an eigenvalue and we have*

$$\begin{aligned} p_3(\lambda) &= \lambda^3 + \lambda^2 - 10\lambda + 8 \\ &= (\lambda - 1)(\lambda^2 + 2\lambda - 8) \\ &= (\lambda - 1)(\lambda - 2)(\lambda + 4). \end{aligned}$$

then the eigenvalues of A are given as follows

$$\lambda_1 = -4,\ \lambda_2 = 2,\ \lambda_3 = 1,$$

We calculate the eigenvectors v_1, v_2 *and* v_3

$$\begin{aligned} \det(M - \lambda_1 I)v_1 = 0_{\mathbb{R}^3} &\implies \begin{pmatrix} 4 & 2 & 1 \\ 3 & 2 & 0 \\ -2 & 2 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ &\implies \begin{cases} 4x + 2y + z = 0 \\ 3x + 2y = 0 \\ -2x + 2y + 5z \end{cases} \\ &\implies (x, y, z) = \Big(-\frac{2}{3}y, y, \frac{2}{3}y \Big), \end{aligned}$$

takes $y = 1$ *so* $v_1 = \left(-\frac{2}{3}, 1, \frac{2}{3}\right)$.
And we calculate v_2 *as follows*

$$\begin{aligned}
\det(A - \lambda_2 I)v_2 = 0_{\mathbb{R}^3} &\implies \begin{pmatrix} -2 & 2 & 1 \\ 3 & -4 & 0 \\ -2 & 2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\
&\implies \begin{cases} -2x + 2y + z = 0 \\ 3x - 4y = 0 \\ -2x + 2y - 1z = 0 \end{cases} \\
&\implies (x, y, z) = \left(\frac{4}{3}y, y, 0\right)
\end{aligned}$$

takes $y = 1$ *so that* $v_2 = \left(\frac{4}{3}, 1, 0\right)$.
We calculate v_3 *as follows:*

$$\begin{aligned}
\det(A - \lambda_2 I)v_2 = 0_{\mathbb{R}^3} &\implies \begin{pmatrix} -1 & 2 & 1 \\ 3 & -3 & 0 \\ -2 & 2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\
&\implies \begin{cases} -x + 2y - z = 0 \\ 3x - 3y = 0 \\ -2x + 2y = 0 \end{cases} \\
&\implies (x, y, z) = \Big(y, y, y\Big),
\end{aligned}$$

takes $y = 1$ *so* $v_3 = \Big(1, 1, 1\Big)$.

3. *Write* A *in the form* PDP^{-1}, *then calculate* A^n *with* $n \in \mathbb{N}$. *We have*

$$P = (v_1, v_2, v_3) = \begin{pmatrix} -\frac{2}{3} & \frac{4}{3} & 1 \\ 1 & 1 & 1 \\ \frac{2}{3} & 0 & 1 \end{pmatrix}$$

we obtain

$$P^{-1} = -\frac{3}{8} \begin{pmatrix} 1 & -\frac{1}{3} & -\frac{2}{3} \\ -\frac{4}{3} & -\frac{4}{3} & -\frac{8}{9} \\ \frac{1}{3} & \frac{5}{3} & -2 \end{pmatrix}$$

hence

$$\begin{aligned}
A^n &= \underbrace{A \times A \times \cdots \times A}_{n \text{ times}} \\
&= \underbrace{(PDP^{-1}) \times (PDP^{-1}) \times \cdots \times (PDP^{-1})}_{n \text{ times}} \\
&= P \underbrace{D \times D \times \cdots \times D}_{n \text{ times}} P^{-1} \\
&= PD^nP^{-1}
\end{aligned}$$

such as

$$D^n = \begin{pmatrix} (-4)^n & 0 & 0 \\ 0 & 2^n & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Solution 7.2 *Let*

$$M = \begin{pmatrix} 4 & 6 & 0 \\ -3 & -5 & 0 \\ -3 & -6 & -5 \end{pmatrix}$$

1. *We calculate the characteristic polynomial $P_M(\lambda)$ by the Leverier-Souriau method. Consider the characteristic polynomial of the matrix A*

$$p_M(\lambda) = \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3$$

We use the Leverrier-Souriau method, first, we have

$$B_0 = I_3 \Longrightarrow a_1 = -\mathrm{tr}(AB_0) = -\mathrm{tr}(A) = -(4 - 5 - 5) = 6,$$

hence

$$B_1 = MB_0 + a_1 I_3 = MI_3 - 6I_3 = \begin{pmatrix} 10 & 6 & 0 \\ -3 & 1 & 0 \\ -3 & -6 & 1 \end{pmatrix}$$

so,

$$MB_1 = \begin{pmatrix} 4 & 6 & 0 \\ -3 & -5 & 0 \\ -3 & -6 & -5 \end{pmatrix}\begin{pmatrix} 10 & 6 & 0 \\ -3 & 1 & 0 \\ -3 & -6 & 1 \end{pmatrix} = \begin{pmatrix} 22 & 30 & 0 \\ -15 & -23 & 0 \\ 3 & 6 & -5 \end{pmatrix}$$

so that,

$$a_2 = -\frac{1}{2}\mathrm{tr(MB_1)} = -\frac{1}{2}(22 - 23 - 5) = 3,$$

and in addition

$$B_2 = MB_1 + 3I_3 = \begin{pmatrix} 22 & 30 & 0 \\ -15 & -23 & 0 \\ 3 & 6 & -5 \end{pmatrix} + \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 25 & 30 & 0 \\ -15 & -20 & 0 \\ 3 & 6 & -2 \end{pmatrix}$$

then

$$MB_2 = \begin{pmatrix} 4 & 6 & 0 \\ -3 & -5 & 0 \\ -3 & -6 & -5 \end{pmatrix}\begin{pmatrix} 25 & 30 & 0 \\ -15 & -20 & 0 \\ 3 & 6 & -2 \end{pmatrix} = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}$$

we obtain

$$a_3 = -\frac{1}{3}\mathrm{tr(MB_2)} = -\frac{1}{3}(10 + 10 + 10) = -10.$$

Finally,

$$\begin{aligned} p_M(\lambda) &= \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 \\ &= \lambda^3 + 6\lambda^2 + 3\lambda - 10. \end{aligned}$$

2. *Deduce M^{-1} As $B_3 = MB_2 + a_3 I = 0_{M_3(\mathbb{R})}$ such that $MB_2 = -a_3 I$ therefore $M^{-1} = -\frac{1}{a_3}B_2$ then*

$$M^{-1} = -\frac{1}{a_3}B_2 = +\frac{1}{10}\begin{pmatrix} 25 & 30 & 0 \\ -15 & -20 & 0 \\ 3 & 6 & -2 \end{pmatrix}$$

3. *Using Bernoulli's method, calculate λ_1, λ_2, λ_3, we have*

$$\lambda^3 + 6\lambda^2 + 3\lambda - 10 = 0.$$

We have

$$\lambda_1 = \lim_{p \longrightarrow \infty} \frac{y_p}{y_{p-1}}$$

where the value y_p are determined from the recurrence relations

$$y_{k+3} = -6y_{k+2} - 3y_{k+1} + y_k,\ k = 0, 1, 2, \cdots$$

Taking as a starting point $y_0 = y_1 = 0$ and $y_2 = 1$, we obtain the following table

p	y_p	$\frac{y_p}{y_{p-1}}$
0	0	
1	0	
2	1	
3	-6	-6
4	33	-5.5
5	-179	-5.424
6	969	-5.413
7	-5244	-5.411
8	29197	-5.56
9	-158481	-5.427
10	770460	-4.861

It is clear that this sequence converges to the solution $\lambda_1 = -5$. *We determine the following roots, we calculate the quotient*

$$\frac{P_3(\lambda)}{\lambda + 5} = \lambda^2 + \lambda - 2,$$

we solve the equation

$$\lambda^2 + \lambda - 2 = 0,$$

we get $\lambda_2 = -2$ *and* $\lambda_3 = 1$. *Then*

$$\begin{aligned} p_M(\lambda) &= \lambda^3 + 6\lambda^2 + 3\lambda - 10 \\ &= (\lambda - 1)(\lambda - 2)(\lambda - 5). \end{aligned}$$

4. *Determine the eigenvectors of* M, *we have the eigenvalues of* M *are given as follows*

$$\lambda_1 = 1,\ \lambda_2 = -2,\ \lambda_3 = -5,$$

We calculate the eigenvectors V_1, V_2 *and* V_3

$$\begin{aligned} V_1 &= \lambda_1^2 B_0 + \lambda_1^1 B_1 + \lambda_1^0 B_2 \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 10 & 6 & 0 \\ -3 & 1 & 0 \\ -3 & -6 & 1 \end{pmatrix} + \begin{pmatrix} 25 & 30 & 0 \\ -15 & -20 & 0 \\ 3 & 6 & -2 \end{pmatrix} \\ &= \begin{pmatrix} 36 & 36 & 0 \\ -18 & -18 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

then $V_1 = \frac{1}{\sqrt{5}}(2, -1, 0)^t$

$$\begin{aligned} V_2 &= \lambda_2^2 B_0 + \lambda_2^1 B_1 + \lambda_2^0 B_2 \\ &= 4\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - 2\begin{pmatrix} 10 & 6 & 0 \\ -3 & 1 & 0 \\ -3 & -6 & 1 \end{pmatrix} + \begin{pmatrix} 25 & 30 & 0 \\ -15 & -20 & 0 \\ 3 & 6 & -2 \end{pmatrix} \\ &= \begin{pmatrix} 9 & 18 & 0 \\ -9 & -18 & 0 \\ 9 & 18 & 0 \end{pmatrix} \end{aligned}$$

then $V_2 = \frac{1}{\sqrt{3}}(1, -1, 1)^t$

$$\begin{aligned} V_3 &= \lambda_3^2 B_0 + \lambda_3^1 B_1 + \lambda_3^0 B_2 \\ &= 25\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - 5\begin{pmatrix} 10 & 6 & 0 \\ -3 & 1 & 0 \\ -3 & -6 & 1 \end{pmatrix} + \begin{pmatrix} 25 & 30 & 0 \\ -15 & -20 & 0 \\ 3 & 6 & -2 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 40 \\ 18 & 36 & 18 \end{pmatrix} \end{aligned}$$

then $V_3 = (0,0,1)^t$.

5. *We determine* P *with* $M = PDP^{-1}$ *and* $D = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$ *and* $P^{-1} = P^t$. *As* M *is symmetric* ($M = M^t$) *and* $\|V_k\| = 1$, *with* $k = 1, 2, 3$. *Therefore*

$$P = (V_1, V_2, V_3) = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{3}} & 0 \\ -\frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{3}} & 0 \\ 0 & \frac{1}{\sqrt{3}} & 1 \end{pmatrix}$$

and

$$P^{-1} = P^t = \begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$D = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -5 \end{pmatrix}.$$

Solution 7.3 *Let*

$$M = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 4 \end{pmatrix}$$

1. *We calculate the polynomial caracteristic* $P_M(\lambda)$ *by Leverier-Souriau method. Let* p_M *be the polynomial carateristic of the matrix* A*:*

$$p_M(\lambda) = \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3$$

We use Leverrier-Souriau method. Firstly, we have

$$B_0 = I_3 \Longrightarrow a_1 = -\mathrm{tr}(AB_0) = -\mathrm{tr}(A) = -(2+2+4) = -8,$$

so as

$$B_1 = MB_0 + a_1 I_3 = MI_3 - 8I_3 = \begin{pmatrix} -6 & 1 & 1 \\ 1 & -6 & 1 \\ 1 & 1 & -4 \end{pmatrix}$$

then

$$MB_1 = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{pmatrix} -6 & 1 & 1 \\ 1 & -6 & 1 \\ 1 & 1 & -4 \end{pmatrix} = \begin{pmatrix} -10 & -3 & -1 \\ -3 & -10 & -1 \\ -1 & -1 & -14 \end{pmatrix},$$

hence

$$a_2 = -\frac{1}{2}\mathrm{tr(MB_1)} = -\frac{1}{2}(-10-10-14) = 17$$

and we have

$$B_2 = MB_1 + 17I_3 = \begin{pmatrix} -10 & -3 & -1 \\ -3 & -10 & -1 \\ -1 & -1 & -14 \end{pmatrix} + \begin{pmatrix} 17 & 0 & 0 \\ 0 & 17 & 0 \\ 0 & 0 & 17 \end{pmatrix} = \begin{pmatrix} 7 & -3 & -1 \\ -3 & 7 & -1 \\ -1 & -1 & 3 \end{pmatrix},$$

then

$$MB_2 = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{pmatrix} 7 & -3 & -1 \\ -3 & 7 & -1 \\ -1 & -1 & 3 \end{pmatrix} = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}$$

we obtain

$$a_3 = -\frac{1}{3}\mathrm{tr(MB_2)} = -\frac{1}{3}(10+10+10) = -10.$$

Finally

$$\begin{aligned} p_M(\lambda) &= \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 \\ &= \lambda^3 - 8\lambda^2 + 17\lambda - 10. \end{aligned}$$

2. *We determine the eigenvalues and eigenvectors of M. Remark that*

$$p_M(1) = 0$$

so $\lambda = 1$ is the eigenvalue of M and we have

$$\begin{aligned} p_M(\lambda) &= \lambda^3 - 8\lambda^2 + 17\lambda - 10 \\ &= (\lambda - 1)(\lambda - 2)(\lambda - 5), \end{aligned}$$

then the eigenvalues of matrix M

$$\lambda_1 = 1,\ \lambda_2 = 2,\ \lambda_3 = 5,$$

We calculate the eigenvectors V_1, V_2 and V_3

$$\begin{aligned} V_1 &= \lambda_1^2 B_0 + \lambda_1^1 B_1 + \lambda_1^0 B_2 \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} -6 & 1 & 1 \\ 1 & -6 & 1 \\ 1 & 1 & -4 \end{pmatrix} + \begin{pmatrix} 7 & -3 & -1 \\ -3 & 7 & -1 \\ -1 & -1 & 3 \end{pmatrix} \\ &= \begin{pmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \end{aligned}$$

then $V_1 = \dfrac{1}{\sqrt{2}}(1, -1, 0)^t$ and we have

$$\begin{aligned} V_2 &= \lambda_2^2 B_0 + \lambda_2^1 B_1 + \lambda_2^0 B_2 \\ &= 4\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + 2\begin{pmatrix} -6 & 1 & 1 \\ 1 & -6 & 1 \\ 1 & 1 & -4 \end{pmatrix} + \begin{pmatrix} 7 & -3 & -1 \\ -3 & 7 & -1 \\ -1 & -1 & 3 \end{pmatrix} \\ &= \begin{pmatrix} -1 & 1 & 1 \\ -1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \end{aligned}$$

then $V_2 = \dfrac{1}{\sqrt{3}}(1, 1, -1)^t$, and we have

$$\begin{aligned} V_2 &= \lambda_3^2 B_0 + \lambda_3^1 B_1 + \lambda_3^0 B_2 \\ &= 25\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + 5\begin{pmatrix} -6 & 1 & 1 \\ 1 & -6 & 1 \\ 1 & 1 & -4 \end{pmatrix} + \begin{pmatrix} 7 & -3 & -1 \\ -3 & 7 & -1 \\ -1 & -1 & 3 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 2 & 4 \\ 2 & 2 & 4 \\ 4 & 4 & 8 \end{pmatrix}, \end{aligned}$$

then $V_3 = \dfrac{1}{\sqrt{6}}(1, 1, 2)^t$.

3. *We determine P with $M = PDP^{-1}$ and $D = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$ et $P^{-1} = P^t$. As M is symetric ($M = M^t$) and $\|V_k\| = 1$, with $k = 1, 2, 3$ hence*

$$P = (V_1, V_2, V_3) = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & -\frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} \end{pmatrix}$$

and

$$P^{-1} = P^t = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \end{pmatrix}$$

and

$$D = \text{diag}(\lambda_1, \lambda_2, \lambda_3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

7.5 Exercises without solutions

Exercise 7.1 *Let the matrix*

$$A = \begin{pmatrix} 7 & 3 & -9 \\ -2 & -1 & 2 \\ 2 & -1 & -4 \end{pmatrix}.$$

1. *Show that the eigenvalues of A are $\lambda_1 = -2, \lambda_2 = 1$ and $\lambda_3 = 3$.*
2. *Specify the passing matrix P, what is the relationship between the matrices A, P, P^{-1} and D?.*
3. *After giving D^n, calculate A^n for all $n \in \mathbb{N}$.*

Exercise 7.2 *Let*

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{pmatrix},$$

find the characteristic polynomial of A by the "Leverrier-Sonriau" method and factorize this polynomial. Is the matrix A diagonalizable in $\mathbb{R}$?.

Exercise 7.3 *Let M be the following real matrix 3×3*

$$M = \begin{pmatrix} 0 & 2 & -1 \\ 3 & -2 & 0 \\ -2 & 2 & 1 \end{pmatrix}$$

1. *Determine the eigenvalues of M, using Krylov's method with $y^{(0)} = (1, 1, 0)^t$ the characteristic vector.*
2. *Show that M is diagonalizable.*
3. *Determine an eigenvalue basis and P passing matrix.*
4. *We have $M = PDP^{-1}$, for $k \in \mathbb{N}$ express M^k as a function of D^k, then calculate M^k.*

Exercise 7.4 *We consider the polynomial*

$$P_3(x) = -x^3 + 8x^2 - 17x + 10.$$

1. *Apply Descartes' sign rule to this polynomial.*
2. *Calculate the roots of this polynomial by Bernoulli's method, with $y_0 = y_1 = 0$ and $y_2 = 1$.*

Chapter 8

Least Squares Approximations

8.1 Introduction

Let E be a real vector space such that $\mathbb{P}_n \subset E$, $f \in E$, where $\mathbb{P}_n$ is the space of polynomials of lower degree or equal to n. The polynomial approximation in the sense of least squares consists in determining the polynomial $P^* \in \mathbb{P}_n$ which verifies

$$\| f - P^* \| = \min_{P \in \mathbb{P}_n} \| f - P \|,$$

P^* is called the best least squares approximant of f in $\mathbb{P}_n$.

8.2 Preliminaries

Let E be a pre-Hilbertian vector space i.e. E a vector space endowed by the scalar product $\langle \cdot, \cdot \rangle$ which is an application

$$\begin{array}{rcl} \langle \cdot, \cdot \rangle : E \times E & \longrightarrow & \mathbb{R} \\ (f, g) & \longrightarrow & \langle f, g \rangle \end{array}$$

satisfying

1. Positif

$$\langle f, f \rangle \geq 0, \forall f \in E$$

2. Well-defined

$$\langle f, f \rangle = 0 \Leftrightarrow f = 0_E,$$

3. Symmetric

$$\langle f, g \rangle = \langle g, f \rangle, \forall f, g \in E$$

4. Linear

$$\langle f, \lambda g + \mu h \rangle = \lambda \langle f, g \rangle + \mu \langle f, h \rangle \qquad \forall f, g, h \in E, \forall \lambda, \mu \in \mathbb{R},$$

and the application

$$\begin{array}{rcl} \| \cdot \| : E & \longrightarrow & \mathbb{R} \\ f & \longrightarrow & \| f \| = \sqrt{\langle f, f \rangle} \end{array}$$

is called the norm on E associated with the scalar product $\langle \cdot, \cdot \rangle$.

Definition 8.1 *A complete pre-Hilbert space for the norm associated with this scalar product is called Hilbert space.*

Definition 8.2 *We say that E is a complete space if every Cauchy sequence of E is convergent in E.*

Definition 8.3 *Let $\{\varphi_0, \varphi_1, \cdots, \varphi_n\}$ be a basis of E*

- *The family $\{\varphi_0, \varphi_1, \cdots, \varphi_n\}$ is called orthogonal basis of E if*

$$\langle \varphi_i, \varphi_j \rangle, \forall i, j = 0, ..., n \ \text{ and } \ i \neq j$$

- *The family $\{\varphi_0, \varphi_1, \cdots, \varphi_n\}$ is called orthonormal basis of E if*

$$\langle \varphi_i, \varphi_j \rangle = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases}$$

Example 8.2.1 *1. Let $E = \mathbb{R}^n$, $\mathbf{x} = (x_1, \cdots, x_n)$, $\mathbf{y} = (y_1, \cdots, y_n) \in E$.*
The application

$$\begin{aligned} \langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n &\longrightarrow \mathbb{R} \\ (\mathbf{x}, \mathbf{y}) &\longrightarrow \langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i, \end{aligned}$$

defines a scalar product on $\mathbb{R}^n$ and the norm associated with this product is defined by

$$\| \mathbf{x} \| = \sqrt{\langle \mathbf{x}, \mathbf{y} \rangle} = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}.$$

2. Let $E = C([a, b])$ be the space of continuous functions, $f, g \in E$

$$\begin{aligned} \langle \cdot, \cdot \rangle : C([a,b]) \times C([a,b]) &\longrightarrow \mathbb{R} \\ (f, g) &\longrightarrow \langle f, g \rangle = \int_a^b f(x) g(x)\, dx, \end{aligned}$$

defines a scalar product on $C([a, b])$ and the norm associated with this product is defined by

$$\| f(x) \| = \sqrt{\langle f(x), f(x) \rangle} = \left(\int_a^b (f(x))^2 dx \right)^{\frac{1}{2}}.$$

3. The family $\{1, x, \frac{1}{2}(3x^2 - 1)\}$ is an orthonormal basis of $\mathbb{P}_2$ for the product scalar defined by

$$\langle f, g \rangle = \int_{-1}^{1} f(x) g(x)\, dx.$$

Definition 8.4 *(Best approximation) Let F be a normed vector space of E ($F \subseteq E$), we say that $\varphi^* \in F$ is the best approximation of $f \in E$ if:*

$$\| f - \varphi^* \| = \min_{\varphi \in F} \| f - \varphi \|.$$

Theorem 8.1 *[2] Let $\varphi^* \in F$ the best approximation of $f \in E$ be equivalent*

$$\langle f - \varphi^*, \varphi \rangle = 0, \quad \forall \varphi \in F.$$

Proposition 8.1 *the vector φ^* is unique.*

Proof. Let φ_1^* be the best approximation of $f \in E$, $\langle f - \varphi_1^*, \varphi\rangle = 0$ and let φ_2^* the best approximation of $f \in E$, $\langle f - \varphi_2^*, \varphi\rangle = 0$, we set $\varphi_1 = \varphi_1^* - \varphi_2^*$, so that

$$\begin{aligned} \| \varphi_1 \|^2 &= \langle \varphi_1, \varphi_1\rangle \\ &= \langle \varphi_1^* - \varphi_2^*, \varphi_1\rangle \\ &= \langle \varphi_1^* - \varphi_2^* + f - f, \varphi_1\rangle \\ &= \langle -(f - \varphi_1^*) + (f - \varphi_2^*), \varphi_1\rangle \\ &= -\langle f - \varphi_1^*, \varphi_1\rangle + \langle f - \varphi_2^*, \varphi_1\rangle \\ &= -0 + 0 = 0, \end{aligned}$$

hence

$$\| \varphi_1 \|^2 = 0 \Longrightarrow \| \varphi_1 \| = 0 \Longrightarrow \varphi_1 = 0 \Longrightarrow \varphi_1^* = \varphi_2^*.$$

So the best approximation is unique.

Creation of φ^* Let $(\varphi_1, \varphi_2, \cdots, \varphi_n)$ be the basic elements of F, we will look for $\varphi^* \in F$ the best approximation of $f \in E$, we have

$$\varphi^* = \sum_{k=1}^{n} a_k^* \varphi_k,$$

we are looking for $a_k^* \in \mathbb{R}$ with $k = 1, \cdots, n$. We have

$$\langle f - \varphi^*, \varphi\rangle = 0, \qquad \forall \varphi \in F,$$

such that

$$\langle f - \varphi^*, \varphi_j\rangle = 0, \qquad j = 1, \cdots, n$$

this implies that

$$\langle f - \sum_{k=1}^{n} a_k^* \varphi_k, \varphi_j\rangle = 0, \qquad j = 1, \cdots, n$$

so,

$$\langle f, \varphi_j\rangle = \sum_{k=1}^{n} a_k^* \langle \varphi_k, \varphi_j\rangle, \qquad j = 1, \cdots, n$$

this system of n equations and n unknowns $(a_k^*)_{1\leq k\leq n}$, according to the previous one the solution is unique since the best approximation is unique

$$\begin{pmatrix} \langle \varphi_1, \varphi_1\rangle & \langle \varphi_2, \varphi_1\rangle & \cdots & \langle \varphi_n, \varphi_1\rangle \\ \langle \varphi_1, \varphi_2\rangle & \langle \varphi_2, \varphi_2\rangle & \cdots & \langle \varphi_n, \varphi_2\rangle \\ \vdots & \vdots & \vdots & \vdots \\ \langle \varphi_1, \varphi_n\rangle & \langle \varphi_2, \varphi_n\rangle & \cdots & \langle \varphi_n, \varphi_n\rangle \end{pmatrix} \begin{pmatrix} a_1^* \\ a_2^* \\ \vdots \\ a_n^* \end{pmatrix} = \begin{pmatrix} \langle f, \varphi_1\rangle \\ \langle f, \varphi_2\rangle \\ \vdots \\ \langle f, \varphi_n\rangle \end{pmatrix},$$

the solution of a linear system gives $(a_k^*)_{1\leq k\leq n}$ and we denote by G the matrix of this system.

The matrix G is called the Gram matrix. As $(\varphi_1, \varphi_2, \cdots, \varphi_n)$ is independent linear therefore $\det(G) \neq 0$ then according to Cramer's theorem, the solution is unique. ■

Remark 8.1 *1. If the base elements $(\varphi_1, \varphi_2, \cdots, \varphi_n)$ are orthogonal i.e.*

$$\langle \varphi_i, \varphi_j\rangle = 0, \text{ with } i \neq j.$$

then

$$\langle f - \varphi^*, \varphi\rangle = 0,$$

so that,

$$\langle f, \varphi_j \rangle = \sum_{k=1}^{n} a_k^* \langle \varphi_k, \varphi_j \rangle,$$

where

$$\langle f, \varphi_j \rangle = a_j^* \langle \varphi_j, \varphi_j \rangle,$$

this implies that

$$\langle f, \varphi_j \rangle = a_j^* \parallel \varphi_j \parallel^2,$$

Finally,

$$a_j^* = \frac{\langle f, \varphi_j \rangle}{\parallel \varphi_j \parallel^2}.$$

2. *If the basis elements are orthogonal and orthonormal, then*

$$\langle \varphi_i, \varphi_j \rangle = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

so that

$$a_j^* = \langle f, \varphi_j \rangle,$$

then, the best approximation is

$$\varphi^* = \sum_{j=1}^{n} \langle f, \varphi_j \rangle \varphi_j.$$

8.2.1 Approximation error

We have

$$\parallel f - P^* \parallel = \sqrt{\parallel f \parallel^2 - \sum_{k=0}^{n} a_k^* \langle f, \varphi_k \rangle}.$$

Indeed,

$$\begin{aligned} \parallel f - P^* \parallel^2 &= \langle f - P^*, f - P^* \rangle \\ &= \langle f - P^*, f \rangle - \langle f - P^*, P^* \rangle \\ &= \langle f, f \rangle - \langle f, P^* \rangle \\ &= \parallel f \parallel^2 - \sum_{k=0}^{n} a_k^* \langle f, \varphi_k \rangle. \end{aligned}$$

Special case if the basis $\{\varphi_0, \varphi_1, \cdots, \varphi_n\}$ is orthonormal, we obtain

$$\parallel f - P^* \parallel = \sqrt{\parallel f \parallel^2 - \sum_{k=0}^{n} (a_k^*)^2}.$$

8.2.2 Gram-Schmidt algorithm

Let F be a finite dimensional space. Starting from any basis of F, denoted $\{\varphi_0, \varphi_1, \cdots, \varphi_n\}$, we can determine both an orthonormal basis, denoted $\{h_0, h_1, \cdots, h_n\}$, using the following Gram-Schmidt algorithm

$$\begin{aligned}
h_0 &= \frac{u_0}{\| u_0 \|}, \text{ where } u_0 = \varphi_0, \\
h_1 &= \frac{u_1}{\| u_1 \|}, \text{ where } u_1 = \varphi_1 - \langle \varphi_1, h_0 \rangle h_0, \\
h_2 &= \frac{u_2}{\| u_2 \|}, \text{ where } u_2 = \varphi_2 - \langle \varphi_2, h_0 \rangle h_0 - \langle \varphi_2, h_1 \rangle h_1, \\
&\vdots \\
h_k &= \frac{u_k}{\| u_k \|}, \text{ where } u_k = \varphi_k - \sum_{m=0}^{k-1} \langle \varphi_k, h_m \rangle h_m, \; 1 \leq k \leq n \\
&\vdots \\
h_n &= \frac{u_n}{\| u_n \|}, \text{ where } u_n = \varphi_n - \sum_{m=0}^{n-1} \langle \varphi_n, h_m \rangle h_m.
\end{aligned}$$

8.3 Application to the continuous case

Let I be the integral

$$I = \int_a^b w(x) h(x) dx, \qquad \forall h \in C([a, b]),$$

such that w is continuous and is called weight function, we define the scalar product in $C([a, b])$ as follows:

$$\langle f, g \rangle = \int_a^b w(x) f(x) g(x) dx,$$

we set the standard norm

$$\| f \| = \left(\int_a^b w(x) (f(x))^2 dx \right)^{\frac{1}{2}},$$

with w weight function, i.e. E_n under vector space of $C([a, b])$ of dimension n and generated by $\varphi_1, \varphi_2, \cdots, \varphi_n$.

Definition 8.5 *(Best continuous approximation) φ^* of E_n is said to be the best continuous least-squares approximation for f in $C([a, b])$ if*

$$\int_a^b w(x) [f(x) - \varphi^*(x)]^2 dx = \min_{\varphi \in E_n} \int_a^b w(x) [f(x) - \varphi(x)]^2 dx.$$

By theorem (8.1), we have

$$\langle f - \varphi^*, \varphi \rangle = 0, \qquad \forall \varphi \in E_n,$$

then for the base elements φ_j

$$\langle f, \varphi_j \rangle = \langle \varphi^*, \varphi_j \rangle,$$

hence

$$\varphi^* = \sum_{k=1}^{n} a_k^* \varphi_k,$$

we will look for a_k^* with $k = 1, 2, \cdots, n$, we have

$$\langle f, \varphi_j \rangle = \sum_{k=1}^{n} a_k^* \langle \varphi_k, \varphi_j \rangle,$$

this implies that

$$\int_a^b w(x) f(x) \varphi_j(x) dx = \sum_{k=1}^{n} a_k^* \int_a^b w(x) \varphi_k(x) \varphi_j(x) dx, \quad j = 1, \cdots, n$$

has a one solution.

Example 8.3.1 *Find the polynomial $p_1 \in \mathbb{P}_1$ which satisfies the best least-squares approximation for f such that $f(x) = x^3$ on the interval $[-1, 1]$ with $w(x) = 1$.*

Correction 8.3.1 *We have $\mathbb{P}_1 = \{\varphi_0(x), \varphi_1(x)\} = \{1, x\}$, so*

$$p_1^*(x) = \sum_{k=0}^{1} a_k^* \varphi_k = a_0^* \cdot 1 + a_1^* \cdot x,$$

then

$$\int_{-1}^{1} 1 \cdot [x^3 - (a_0^* + a_1^* x)]^2 dx = \min \int_{-1}^{1} [x^3 - (a_0 + a_1 x)]^2 dx,$$

we are looking for p_1^* such that

$$\langle f - \varphi^*, \varphi_j \rangle = 0, \quad j = 0, 1$$

implies that

$$\langle f - p_1^*, \varphi_j \rangle = 0, \quad j = 0, 1$$

we solve the system

$$\begin{cases} \int_{-1}^{1} 1 \cdot (x^3 - (a_0^* + a_1^* x)) \cdot 1 dx = 0, \\ \int_{-1}^{1} 1 \cdot (x^3 - (a_0^* + a_1^* x)) \cdot x dx = 0, \end{cases}$$

find a_0^*, a_1^*, we write the matrix form

$$\langle f, \varphi_j \rangle = \sum_{k=0}^{1} a_k \langle \varphi_k, \varphi_j \rangle, \quad j = 0, 1$$

hence,

$$\begin{pmatrix} \langle f, \varphi_0 \rangle \\ \langle f, \varphi_1 \rangle \end{pmatrix} = \begin{pmatrix} \langle \varphi_0, \varphi_0 \rangle & \langle \varphi_0, \varphi_1 \rangle \\ \langle \varphi_1, \varphi_0 \rangle & \langle \varphi_1, \varphi_1 \rangle \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix}$$

so,

$$\begin{pmatrix} \langle x^3, 1 \rangle \\ \langle x^3, x \rangle \end{pmatrix} = \begin{pmatrix} \langle 1, 1 \rangle & \langle 1, x \rangle \\ \langle x, 1 \rangle & \langle x, x \rangle \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix}$$

so that,

$$\langle 1, 1 \rangle = \int_{-1}^{1} (1)(1) dx = 2,$$

$$\langle 1, x \rangle = \langle x, 1 \rangle = \int_{-1}^{1} (1)(x) dx = 0,$$

$$\langle x, x \rangle = \int_{-1}^{1} (x)(x) dx = \frac{2}{3},$$

$$\langle x^3, 1 \rangle = \int_{-1}^{1} (x^3)(1) dx = 0,$$

$$\langle x^3, x \rangle = \int_{-1}^{1} (x^3)(x) dx = \int_{-1}^{1} x^4 dx = \frac{2}{5}.$$

this implies that

$$\begin{pmatrix} 2 & 0 \\ 0 & \frac{2}{3} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{2}{5} \end{pmatrix}$$

we get $a_0^* = 0$, $a_1^* = \frac{3}{5}$ *hence* $p_1(x) = \frac{3}{5}x$.

8.4 Application to the discrete case

Let f define at $N+1$ points of $[a,b]$ then $(f(x_0), f(x_1), \cdots, f(x_N)) \in \mathbb{R}^{N+1}$, we define the scalar product

$$\langle f, g\rangle = \sum_{i=0}^{N} w(x_i) f(x_i) g(x_i),$$

and the associated standard norm

$$\| f \| = \sum_{0}^{N} w(x_i) f^2(x_i).$$

Let E_{n+1} be a vector space of dimension $n+1$ of $C([a,b])$ with $n < N$.

Definition 8.6 *We define the best approximation of* $\varphi^* \in E_{n+1}$ *by least-squares points of* f *as follows*

$$\sum_{i=0}^{N} w(x_i)[f(x_i) - \varphi^*(x_i))]^2 = \min_{\varphi \in E_{n+1}} \sum_{i=0}^{N} w(x_i)[f(x_i) - \varphi(x_i))]^2.$$

with φ^* *is unique.*

We have

$$\langle f - \varphi^*, \varphi\rangle = 0, \ \forall \varphi \in E_{n+1},$$

as E_{n+1} generated by $\varphi_0, \varphi_1, \cdots, \varphi_n$ then

$$\langle f - \varphi^*, \varphi_j\rangle = 0, \quad j = 0, \cdots, n$$

hence

$$\langle f, \varphi_j\rangle = \sum_{k=0}^{n} a_k^* \langle \varphi_k, \varphi_j\rangle, \quad j = 0, \cdots, n$$

so that,

$$\varphi^* = \sum_{k=0}^{n} a_k^* \varphi_k,$$

finally, we have

$$\sum_{i=0}^{N} w(x_i) f(x_i) \varphi_j(x_i) = \sum_{i=0}^{N} w(x_i) \varphi_k(x_i) \varphi_j(x_i) \sum_{k=0}^{n} a_k. \quad j = 1, \cdots, n$$

Example 8.4.1 *Let* f *be defined by*

x_i	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1
$f(x_i)$	$-\frac{3}{2}$	0	$\frac{1}{4}$	0	0

Find $p_2 \in \mathbb{P}_2$ which satisfies the best least squares pointwise approximation on the interval $[-1,1]$ such that $w(x) = 1$, we have

$$\begin{aligned} p_2^*(x) &= a_0^*\varphi_0 + a_1^*\varphi_1 + a_2^*\varphi_2 \\ &= a_0^* + a_1^*x + a_2^*x^2, \end{aligned}$$

according to theorem (8.1), we get

$$\begin{pmatrix} \langle\varphi_0,\varphi_0\rangle & \langle\varphi_0,\varphi_1\rangle & \langle\varphi_0,\varphi_2\rangle \\ \langle\varphi_1,\varphi_0\rangle & \langle\varphi_1,\varphi_1\rangle & \langle\varphi_1,\varphi_2\rangle \\ \langle\varphi_2,\varphi_0\rangle & \langle\varphi_2,\varphi_1\rangle & \langle\varphi_2,\varphi_2\rangle \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \\ a_2^* \end{pmatrix} \begin{pmatrix} \langle f,\varphi_0\rangle \\ \langle f,\varphi_1\rangle \\ \langle f,\varphi_2\rangle \end{pmatrix}$$

then

$$\begin{pmatrix} \langle 1,1\rangle & \langle 1,x\rangle & \langle 1,x^2\rangle \\ \langle x,1\rangle & \langle x,x\rangle & \langle x,x^2\rangle \\ \langle x^2,1\rangle & \langle x^2,x\rangle & \langle x^2,x^2\rangle \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \\ a_2^* \end{pmatrix} \begin{pmatrix} \langle f,1\rangle \\ \langle f,x\rangle \\ \langle f,x^2\rangle \end{pmatrix}.$$

We have

$$\begin{aligned} &\langle 1,1\rangle = \textstyle\sum_{i=0}^4 1 = 5,\ \langle 1,x\rangle = \langle x,1\rangle = \sum_{i=0}^4 x_i = 0, \\ &\langle x,x\rangle = \langle 1,x^2\rangle = \langle x^2,1\rangle = \textstyle\sum_{i=0}^4 x_i^2 = \dfrac{5}{2}, \\ &\langle x,x^2\rangle = \langle x^2,x\rangle = \textstyle\sum_{i=0}^4 x_i^3 = 0,\ \langle x^2,x^2\rangle = \sum_{i=0}^4 x_i^4 = \dfrac{17}{8} \end{aligned}$$

and

$$\langle f,1\rangle = \textstyle\sum_{i=0}^4 f(x_i) = -\dfrac{5}{4},\ \langle f,x\rangle = \sum_{i=0}^4 f(x_i)x_i = \dfrac{3}{2},\ \langle f,x^2\rangle = \sum_{i=0}^4 f(x_i)x_i^2 = -\dfrac{3}{2},$$

so that,

$$\begin{pmatrix} 5 & 0 & \frac{5}{2} \\ 0 & \frac{5}{2} & 0 \\ \frac{5}{2} & 0 & \frac{17}{8} \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \\ a_2^* \end{pmatrix} \begin{pmatrix} -\frac{5}{4} \\ \frac{3}{2} \\ -\frac{3}{2} \end{pmatrix}.$$

this implies that

$$(a_0^*, a_1^*, a_2^*) = \left(\frac{3}{5}, -1, \frac{1}{4}\right)$$

finally,

$$p_2(x) = \frac{1}{4} + \frac{3}{5}x - x^2.$$

8.5 Application to the matrix case

It is often difficult to obtain an exact solution to an applied mathematics problem. However, it is generally equally useful for finding arbitrarily close approximations to a solution. In particular, finding 'linear approximations' is a powerful technique in applied mathematics. A base case is the situation where a system of linear equations has no solution, and it is desirable to find a "best approximation" to a solution to the system. In this section the best approximations are defined and a method to find them is described. The result is then applied to the "least squares" approximation of the data. For more details, see [9] (https://math.libretexts.org).

We suppose that A is a matrix $m \times n$ and that b is a column of $\mathbb{R}^m$, and we consider

$$A\mathbf{x} = \mathbf{b},$$

the system of m linear equations with n variables. This doesn't need to have a solution. However, given any column $\mathbf{z} \in \mathbb{R}^n$, the distance $\| \mathbf{b} - A\mathbf{z} \|$ is a measure of the distance between $A\mathbf{z}$ and $\mathbf{b}$.

It is therefore natural to ask whether there is a column $\mathbf{z}$ in $\mathbb{R}^n$ as close as possible to a solution in the sense that

$$\| \mathbf{b} - A\mathbf{z} \|,$$

is the minimum value of $\| \mathbf{b} - A\mathbf{x} \|$ when $\mathbf{x}$ spans all columns of $\mathbb{R}^n$.

The answer is "yes", we define

$$U = \{A\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\},$$

This is a subspace of $\mathbb{R}^n$ (check) and we want a vector $A\mathbf{z}$ in U as close as possible to $\mathbf{b}$. That there is such a vector is clear geometrically if $n = 3$ by the diagram. Moreover, the projection theorem gives a simple way to calculate $\mathbf{z}$ because it also shows that the vector $\mathbf{b} - A\mathbf{z}$ is orthogonal to any vector $A\mathbf{x}$ in U. Thus, for all $\mathbf{x}$ in $\mathbb{R}^n$,

$$0 = (A\mathbf{x}) \cdot (\mathbf{b} - A\mathbf{z}) = (A\mathbf{x})^t \cdot (\mathbf{b} - A\mathbf{z}) = \mathbf{x}^t A^t \cdot (\mathbf{b} - A\mathbf{z}) = \mathbf{x}^t \cdot [A^t \cdot (\mathbf{b} - A\mathbf{z})].$$

In other words, the vector $A^t(\mathbf{b} - A\mathbf{z})$ in $\mathbb{R}^n$ is orthogonal to any vector in $\mathbb{R}^n$ and therefore must be zero (being orthogonal to itself). So $\mathbf{z}$ satisfies

Definition 8.7 *(Normal equations). This is a system of linear equations called the* ***normal equations*** *for* $\mathbf{z}$.

The matrix A^tA of $n \times n$ is invertible if (and only if) the columns of A are linearly independent; so in this case $\mathbf{z}$ is uniquely determined and is given explicitly by $\mathbf{z} = (A^tA)^{-1}A^t\mathbf{b}$. However, the most efficient way to find $\mathbf{z}$ is to apply Gaussian elimination to the normal equations.

This discussion is summarized in the following **best approximation** theorem.

Theorem 8.2 *Let A be a matrix of $m \times n$, let $\mathbf{b}$ be any column of $\mathbb{R}^m$, and consider the system*

$$A\mathbf{x} = \mathbf{b},$$

of m equations with n variables.

1. *Any solution $\mathbf{z}$ to the normal equations*

   $$(A^tA)\mathbf{z} = A^t\mathbf{b}$$

   is a better approximation to a solution of $A\mathbf{x} = \mathbf{b}$ in the sense that $\| \mathbf{b} - A\mathbf{z} \|$ is the minimum value of $\| \mathbf{b} - A\mathbf{x} \|$ when $\mathbf{x}$ spans all columns of $\mathbb{R}^n$.

2. *If the columnss of A are linearly independent, then A^tA is invertible and $\mathbf{z}$ is given uniquely by $\mathbf{z} = (A^tA)^{-1}A^t\mathbf{b}$.*

We note in passing that if A is $n \times n$ and invertible, then

$$\mathbf{z} = (A^tA)^{-1}A^t\mathbf{b} = A^{-1}\mathbf{b}$$

is the solution of the system of equations, and $\| \mathbf{b} - A\mathbf{z} \| = 0$. So if A has independent columns, then $(A^tA)^{-1}A^t$ acts as the inverse of the non-square matrix A. The matrix $A^t(AA^t)^{-1}$ plays a similar role when the rows of A are linearly independent. These are two particular cases of the inverse generalization of a matrix A, here.

Example 8.5.1 *The system of linear equations*

$$\begin{aligned} 3x - y &= 4 \\ x + 2y &= 0 \\ 2x + y &= 1 \end{aligned}$$

has no solution. Find the vector $\mathbf{z} = (x_0, y_0)^t$ that comes close as to a solution.

Correction 8.5.1 *In this case,*

$$\begin{pmatrix} 3 & -1 \\ 1 & 2 \\ 2 & 1 \end{pmatrix}, \text{ so } A^tA = \begin{pmatrix} 3 & 1 & 2 \\ -1 & 2 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ 1 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 14 & 1 \\ 1 & 6 \end{pmatrix}$$

is reversible. The normal equations $(A^tA)\mathbf{z} = A^t\mathbf{b}$ *are*

$$\begin{pmatrix} 14 & 1 \\ 1 & 6 \end{pmatrix} \mathbf{z} = \begin{pmatrix} 14 \\ -3 \end{pmatrix}, \text{ so } \mathbf{z} = \frac{1}{83} \begin{pmatrix} 87 \\ -56 \end{pmatrix}.$$

Thus $x_0 = \dfrac{87}{83}$ *and* $y_0 = -\dfrac{56}{83}$. *With these values of* x *and* y, *the left sides of the equations are, approximately*

$$\begin{aligned} 3x_0 - y_0 &= \frac{317}{83} = 3.82 \\ x_0 + 2y_0 &= -\frac{25}{83} = -0.30 \\ 2x_0 + y_0 &= \frac{118}{83} = 1.42 \end{aligned}$$

This is as close, as to a solution as possible.

Example 8.5.2 *The average number* g *of goals scored per game by a hockey player seems to be linearly related to two factors, the number* x_1 *of years of experience and the number* x_2 *of goals in previous* 10 *matches. The data on the following was collected from four players. Find the linear function* $g = a_0 + a_1x_1 + a_2x_2$ *that best fits this data.*

g	x_1	x_2
0.8	5	3
0.8	3	4
0.6	1	5
0.4	2	1

Correction 8.5.2 *If the relation is given by* $g = r_0 + r_1x_1 + r_2x_2$, *then the data can be described as follows:*

$$\begin{pmatrix} 1 & 5 & 3 \\ 1 & 3 & 4 \\ 1 & 1 & 5 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} r_0 \\ r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 0.8 \\ 0.6 \\ 0.4 \end{pmatrix}.$$

Using the theorem notation (8.2), we get

$$\mathbf{z} = (A^tA)^{-1}A^t\mathbf{b} = \frac{1}{42} \begin{pmatrix} 119 & -17 & -19 \\ -17 & 5 & 1 \\ -19 & 1 & 5 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 5 & 3 & 1 & 2 \\ 3 & 4 & 5 & 1 \end{pmatrix} \begin{pmatrix} 0.8 \\ 0.8 \\ 0.6 \\ 0.4 \end{pmatrix} = \begin{pmatrix} 0.14 \\ 0.08 \\ 0.06 \end{pmatrix}.$$

The best-fitting function is therefore $g = 0.14 + 0.09x_1 + 0.08x_2$. *The amount of computation would have been reduced if the normal equations had been constructed and then solved by Gaussian elimination.*

8.5.1 Least-squares approximation

In many scientific researches, data are collected that relates two variables. For example, if x is the number of dollars spent on advertising by a manufacturer and y is the value of sales in the region in question, the manufacturer could generate data by spending $x_1, x_2, \cdots, x_n$ dollars at different times and measuring the corresponding sales values $y_1, y_2, \cdots, y_n$.

Suppose we know that there is a linear relation between the variables x and y, that is, $y = a+bx$ for some constants a and b. If the data is plotted, the points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ may appear to lie on a straight line and the estimate of a and b requires finding the "best fitting" line through these data points. For example, if five data points occur as shown in the chart, the 1 line is clearly a better fit than the 2 line. In general, the problem is to find the values of the constants a and b such that the line $y = a + bx$ best matches the data in question. Note that an exact fit would be obtained if a and b were such that $y_i = a + bx_i$ were true for each data point (x_i, y_i).

Experimental measurement errors are inevitable, so the choice of a and b must be made in such a way that the errors between the observed values y_i and the corresponding fitted values $a + bx_i$ are somehow so minimized. Least squares approximation is one way to do this. The first thing we need to do is explain exactly what we mean by the best fit of a line $y = a + bx$ to an observed set of data points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. For convenience, write the linear function $r_0 + r_1 x$ as

$$f(x) = r_0 + r_1 x,$$

so that the fitted points (on the line) have coordinates $(x_1, f(x_1)), \cdots, (x_n, f(x_n))$.

The second diagram is a sketch of what the line $y = f(x)$ might look like as. For each i, the observed data point (x_i, y_i) and the fitted point $(x_i, f(x_i))$ need not be the same, and the distance d_i between them measures how far the line misses the observed point. For this reason, d_i is often called the error at x_i, and a natural measure of how close the line $y = f(x)$ is to the observed data points is the sum $d_1 + d_2 + \cdots + d_n$ of all these errors. However, it turns out to be better to use the sum of squares

$$S = d_1^2 + d_2^2 + \cdots + d_n^2$$

as a measure of error, and the row $y = f(x)$ should be chosen so as to make this sum as small as possible. This line is said to be the least squares approximation line for the data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. The squared error d_i is given by $d_i^2 = [y_i - f(x_i)]^2$ for each i, so the quantity S to be minimized is the sum

$$S = [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \cdots + [y_n - f(x_n)]^2.$$

Note that all numbers x_i and y_i are given here; what is required is that the function f be chosen so as to minimize S. As $f(x) = r_0 + r_1 x$, this amounts to choosing r_0 and r_1 to minimize S. This problem can be solved using theorem 8.2. The following notation is convenient.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \text{and} f(\mathbf{x}) = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} r_0 + r_1 x_1 \\ r_0 + r_1 x_2 \\ \vdots \\ r_0 + r_1 x_n \end{bmatrix}.$$

Then the problem takes the following form: Choose r_0 and r_1 such that

$$S = [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \cdots + [y_n - f(x_n)]^2 = \| \mathbf{y} - f(\mathbf{x}) \|^2$$

is as small as possible. Now, we write

$$M = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \text{and} \, \mathbf{r} = \begin{bmatrix} r_0 \\ r_1 \end{bmatrix}$$

Then $M\mathbf{r} = f(\mathbf{x})$, so we are looking for a column $\mathbf{r}^t = [r_0, r_1]$ such that $\| \mathbf{t} - M\mathbf{r} \|^2$ be as small as possible. In other words, we are looking for a better approximation $\mathbf{z}$ of the system $M\mathbf{r} = \mathbf{y}$. Theorem 8.2 therefore applies directly, and we have

Theorem 8.3 *[2] Suppose n data points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ are given, where at least two of $x_1, x_2, \cdots, x_n$ are distinct . We set*

$$\mathbf{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \text{and} \, M = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},$$

Then the least squares approximation line for these data points has the equation

$$y = z_0 + z_1 x,$$

where $\mathbf{z}^t = [z_0, z_1]$ *is found by Gaussian elimination of normal equations*

$$M^T M\mathbf{z} = M^T\mathbf{y}.$$

The condition that at least two of $x_1, x_2, \cdots, x_n$ *are distinct ensures that* $M^T M$ *is an invertible matrix, so* $\mathbf{z}$ *is unique:*

$$\mathbf{z} = (M^T M)^{-1} M^T\mathbf{y}.$$

Example 8.5.3 *Let the data points* $(x_1, y_1), (x_2, y_2), \cdots, (x_5, y_5)$ *be given as in the table. Find the least squares approximation line for this data.*

x_i	y_i
1	1
3	2
4	3
6	4
7	5

Correction 8.5.3 *In this case, we have*

$$\begin{aligned} M^T M &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \end{bmatrix} \\ &= \begin{bmatrix} 5 & x_1 + x_2 + x_3 + x_4 + x_5 \\ x_1 + x_2 + x_3 + x_4 + x_5 & x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 \end{bmatrix} = \begin{bmatrix} 5 & 21 \\ 21 & 111 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} M^T\mathbf{y} &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} \\ &= \begin{bmatrix} y_1 + y_2 + y_3 + y_4 + y_5 \\ x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5 \end{bmatrix} = \begin{bmatrix} 15 \\ 78 \end{bmatrix} \end{aligned}$$

so the normal equations $(M^T M)\mathbf{z} = M^T\mathbf{y}$ *for* $\mathbf{z}^T = [z_0, z_1]$*, it comes*

$$\begin{bmatrix} 5 & 21 \\ 21 & 111 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 15 \\ 78 \end{bmatrix}$$

The solution (using Gaussian elimination) is $\mathbf{z}^T = [z_1, z_2] = [0.24, 0.66]$ *to two decimal places, so the least squares line of approximation for these data is* $y = 0.24 + 0.66x$*. Note that* $M^T M$ *is indeed invertible here (the determinant is* 114*), and the exact solution is*

$$\mathbf{z} = (M^T M)^{-1} M^T\mathbf{y} = \frac{1}{114} \begin{bmatrix} 111 & -21 \\ -21 & 5 \end{bmatrix} \begin{bmatrix} 15 \\ 78 \end{bmatrix} = \frac{1}{114} \begin{bmatrix} 27 \\ 75 \end{bmatrix} = \frac{1}{38} \begin{bmatrix} 9 \\ 25 \end{bmatrix}.$$

8.5.2 Least-squares approximation polynomials

Suppose now that, rather than a straight line, we will find a polynomial

$$y = f(x) = r_0 + r_1 x + r_2 x^2 + \cdots + r_m x^m$$

of degree m which comes closest to the data pairs (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n). As before, we write

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \text{ and } \quad f(\mathbf{x}) = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix}.$$

For each x_i, we have two values of the variable $\mathbf{y}$, the observed value y_i and the calculation value $f(x_i)$. The problem is to choose $f(\mathbf{x})$, i.e. to choose $r_0, r_1, \cdots, r_m$ such that the $f(x_i)$ are as close as possible to the y_i. Once again, we define "as close as possible" by the least squares condition, we choose the r_i such that

$$\| \mathbf{y} - f(\mathbf{x}) \|^2 = [y_0 - f(x_0)]^2 + [y_1 - f(x_1)]^2 + \cdots + [y_n - f(x_n)]^2$$

is as small as possible.

Definition 8.8 *(Least squares approximation) A polynomial $f(\mathbf{x})$ satisfying this condition is called* ***a polynomial approximated by least squares*** *of degree m for the given pairs of data.*

If we write

$$M = \begin{bmatrix} 1 & x_1 & \cdots & x_1^m \\ 1 & x_2 & \cdots & x_2^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^m \end{bmatrix}, \quad \text{and} \quad \mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}.$$

we see that $f(\mathbf{x}) = M\mathbf{r}$. We therefore want to find $\mathbf{r}$ such that $\| \mathbf{y} - M\mathbf{r} \|^2$ is the smallest possible; that is, we want a best approximation $\mathbf{z}$ of the system $M\mathbf{r} = \mathbf{y}$. Theorem 8.2 gives the first part of theorem 8.4.

Theorem 8.4 *Let n data pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be given, and write*

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, M = \begin{bmatrix} 1 & x_1 & \cdots & x_1^m \\ 1 & x_2 & \cdots & x_2^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^m \end{bmatrix}, \quad \textit{and} \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}.$$

1. *If $\mathbf{z}$ is any solution to normal equations*

$$(M^T M)\mathbf{z} = M^T \mathbf{y}$$

then the polynomial

$$z_0 + z_1 x + z_2 x^2 + + z_m x^m$$

is a polynomial approximated by least squares of degree m for the given data pairs.

2. *If at least $m+1$ of the numbers $x_1, x_2, \cdots, x_n$ are distinct (therefore $n \geq m+1$), the matrix $M^T M$ is invertible and $\mathbf{z}$ is uniquely determined by*

$$\mathbf{z} = (M^T M)^{-1} M^T \mathbf{y}.$$

Proof. It remains to prove (2), and for this we show that the columns of M are linearly independent. It is assumed that a comlinear combination of columns disappears:

$$r_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + r_1 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \cdots + r_m \begin{bmatrix} x_1^m \\ x_2^m \\ \vdots \\ x_n^m \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

If we write $q(x) = r_0 + r_1 x + \cdots + r_m x^m$, the equation of the coefficients shows that

$$q(x_1) = q(x_2) = \cdots = q(x_n) = 0,$$

then $q(\mathbf{x})$ is a polynomial of degree m with at least $m+1$ distinct roots, so $q(\mathbf{x})$ must be the zero polynomial. Thus $r_0 = r_1 = \cdots = r_m = 0$ as required. ■

Example 8.5.4 *Find the approximating least squares of quadratically $y = z_0 + z_1x + z_2x^2$ for the following data points.*

$$(-3,3),(-1,1),(0,1),(1,2),(3,4)$$

Correction 8.5.4 *This is an instance of theorem 8.4 with $m = 2$. Here*

$$\mathbf{y} = \begin{bmatrix} 3 \\ 1 \\ 1 \\ 2 \\ 4 \end{bmatrix} \qquad M = \begin{bmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{bmatrix}.$$

so,

$$M^TM = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -3 & -1 & 0 & 1 & 3 \\ 9 & 1 & 0 & 1 & 9 \end{bmatrix} \begin{bmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 20 \\ 0 & 20 & 0 \\ 20 & 0 & 164 \end{bmatrix}$$

and

$$M^T\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -3 & -1 & 0 & 1 & 3 \\ 9 & 1 & 0 & 1 & 9 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 11 \\ 4 \\ 66 \end{bmatrix}$$

the normal equations for $\mathbf{z}$ *are*

$$\begin{bmatrix} 5 & 0 & 20 \\ 0 & 20 & 0 \\ 20 & 0 & 164 \end{bmatrix} \mathbf{z} = \begin{bmatrix} 11 \\ 4 \\ 66 \end{bmatrix} \quad \textit{hence} \quad \mathbf{z} = \begin{bmatrix} 1.15 \\ 0.20 \\ 0.26 \end{bmatrix}$$

This means that the least squares approximating the quadratic for this data are

$$y = 1.15 + 0.20x + 0.26x^2.$$

8.5.3 Other functions

There is an extension of theorem 8.4 that should be mentioned. Given the data pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, this theorem shows how to find a polynomial

$$f(x) = r_0 + r_1x + \cdots + r_mx^m,$$

such that $\| \mathbf{y} - f(\mathbf{x}) \|^2$ is as small as possible, where $\mathbf{x}$ and $f(\mathbf{x})$ are like before. Choosing the appropriate polynomial $f(\mathbf{x})$ amounts to choosing the coefficients $r_0, r_1, \cdots, r_m$, and theorem 8.4 gives a formula for the optimal choices. Here $f(\mathbf{x})$ is a linear combination of the functions $1, x, x^2, \cdots, x^m$ where the r_i are the coefficients, which suggests applying the method to other functions. If $f_0(x), f_1(x), \cdots, f_m(x)$ are given functions, we write

$$f(x) = r_0f_0(x) + r_1f_1(x) + \cdots + r_mf_m(x),$$

where the r_i, $(i = 0, ..., m)$ are real numbers. Then the more general question is whether $r_0, r_1, \cdots, r_m$ can be found such that $\| \mathbf{y} - f(\mathbf{x}) \|^2$ is that small as possible where

$$f(\mathbf{x}) = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{bmatrix}$$

Such a function $f(\mathbf{x})$ is called **best least squares approximation** for these data pairs of the form $r_0f_0(x) + r_1f_1(x) + \cdots + r_mf_m(x)$, r_i in $\mathbb{R}$. The proof of The theorem 8.4 requires proving

Theorem 8.5 *[2] Let n data pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be given, and suppose that $m+1$ functions $f_0(x), f_1(x), \cdots, f_m(x)$ are specified. We write*

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, M = \begin{bmatrix} 1 & f_0(x_1) & \cdots & f_m(x_1) \\ 1 & f_0(x_2) & \cdots & f_m(x_2) \\ \vdots & \vdots & & \vdots \\ 1 & f_0(x_n) & \cdots & f_m(x_n) \end{bmatrix}, \quad \text{and} \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}.$$

1. *If $\mathbf{z}$ is any solution to normal equations*

$$(M^T M)\mathbf{z} = M^T \mathbf{y}$$

then the polynomial

$$z_0 f_0(x) + z_1 f_1(x) + z_2 f_2(x) + + z_m f_m(x)$$

is the best approximation of this data among all functions of the form $r_0 f_0(x) + r_1 f_1(x) + \cdots + r_m f_m(x)$ where the r_i are in $\mathbb{R}$.

2. *If $M^T M$ is invertible (that is, if $\operatorname{rg}(M) = m+1$), then $\mathbf{z}$ is uniquely determined; actually,*

$$\mathbf{z} = (M^T M)^{-1} M^T \mathbf{y}.$$

Heis clear that theorem 8.5 contains theorem 8.4 as a special case, but there is no simple test in general to know if $M^T M$ is invertible. The conditions for this to be valid depend on the choice of functions $f_0(x), f_1(x), \cdots, f_m(x)$.

Example 8.5.5 *Given the data pairs $(-1, 0)$, $(0, 1)$ and $(1, 4)$, find the least squares approximation function of the form $r_0 x + r_1 2^x$.*

Correction 8.5.5 *The functions are $f_0(x) = x$ and $f_1(x) = 2^x$, so the matrix M is*

$$M = \begin{bmatrix} f_0(x_1) & f_1(x_1) \\ f_0(x_2) & f_1(x_2) \\ f_0(x_3) & f_1(x_3) \end{bmatrix} = \begin{bmatrix} -1 & 2^{-1} \\ 0 & 2^0 \\ 1 & 2^1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -2 & 1 \\ 0 & 2 \\ 2 & 4 \end{bmatrix}.$$

In this case

$$M^T M = \frac{1}{4} \begin{bmatrix} 8 & 6 \\ 6 & 21 \end{bmatrix}$$

$M^T M$ is invertible, so the normal equations

$$\frac{1}{4} \begin{bmatrix} 8 & 6 \\ 6 & 21 \end{bmatrix} \mathbf{z} = \begin{bmatrix} 4 \\ 9 \end{bmatrix}$$

this system admits a unique solution

$$\mathbf{z} = \frac{1}{11} \begin{bmatrix} 10 \\ 16 \end{bmatrix}.$$

Therefore, the best-fit function of the form $r_0 x + r_1 2^x$ is $\overline{f}(x) = \frac{10}{11} x + \frac{16}{11} 2^x$.
We take note that

$$\overline{f}(\mathbf{x}) = \begin{bmatrix} \overline{f}(-1) \\ \overline{f}(0) \\ \overline{f}(1) \end{bmatrix}$$

compared to

$$\mathbf{y} = \begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix}$$

8.6 Exercises with solutions

Exercise 8.1 *Fit a second-order polynomial to the following data*

i	1	2	3	4	5	6
x	0	0.5	1	1.5	2	2.5
y	0	0.25	1	2.25	4	6.25

Exercise 8.2 *Find the least squares parabola that corresponds to the following data set*

x	0	1	2	3	4	5
y	2.1	7.7	13.6	27.2	40.9	61.1

Exercise 8.3 *The following data was created from the equation* $y = 5 + 4x_1 - 3x_2$

x_1	x_2	y
0	0	5
2	1	10
2.5	2	9
1	3	0
4	6	3
7	2	27

Use multiple linear regression to fit this data.

Exercise 8.4 *Use multiple linear regression to fit*

x	0	1	1	2	2	3	3	4	4
y	0	1	2	1	2	1	2	1	2
z	15	18	12.8	25.7	20.6	35	29.8	45.5	40.3

8.6.1 Solutions

Solution 8.1 *Another alternative is to fit the polynomials to the data using* ***polynomial regression****. The least squares procedure can be easily extended to fit the data to a higher order polynomial. For example, suppose we are fitting a polynomial or a second-order quadratic*

$$y = a_0 + a_1 x + a_2 x^2.$$

The normal equations for finding a least squares parabola are

$$\begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^2 y_i \end{pmatrix}$$

since the order is 2*, the matrix form to solve is*

$$n = 6, \quad \sum_{i=1}^6 x_i = 7.5, \quad \sum_{i=1}^6 x_i^2 = 13.75, \quad \sum_{i=1}^6 x_i^3 = 28.125, \quad \sum_{i=1}^6 x_i^4 = 61.1875$$
$$\sum_{i=1}^6 y_i = 13.75, \quad \sum_{i=1}^6 x_i y_i = 28.125, \quad \sum_{i=1}^6 x_i^2 y_i = 61.1875$$

then

$$\begin{pmatrix} 6 & 7.5 & 13.75 \\ 7.5 & 13.75 & 28.125 \\ 13.75 & 28.125 & 61.1875 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 13.75 \\ 28.125 \\ 61.1875 \end{pmatrix}$$

using the inversion method

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \text{inv} \begin{pmatrix} 6 & 7.5 & 13.75 \\ 7.5 & 13.75 & 28.125 \\ 13.75 & 28.125 & 61.1875 \end{pmatrix} \begin{pmatrix} 13.75 \\ 28.125 \\ 61.1875 \end{pmatrix}$$

or uses the elimination of Gauss, one gives the solution to the coefficients

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

implies that $f(x) = 0 + 0x + 1x^2$. *This matches the data exactly. In other words,* $f(x) = y$ *since* $y = x^2$.

Solution 8.2 *Since the order is* 2, *we have*

$$n = 6, \quad \sum_{i=1}^{6} x_i = 15, \quad \sum_{i=1}^{6} x_i^2 = 55, \quad \sum_{i=1}^{6} x_i^3 = 225, \quad \sum_{i=1}^{6} x_i^4 = 979$$
$$\sum_{i=1}^{6} y_i = 152.6, \quad \sum_{i=1}^{6} x_i y_i = 585.6, \quad \sum_{i=1}^{6} x_i^2 y_i = 2488.6$$

so that

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2.479 \\ 2.359 \\ 1.861 \end{pmatrix}$$

hence

$$y = 2.479 + 2.359x + 1.861x^2.$$

Solution 8.3 *Multiple linear regression is used when* y *is a linear function of* 2 *or more independent variables. We have*

$$y = a_0 + a_1 x_1 + a_2 x_2$$

with

$$\begin{pmatrix} n & \sum_{i=1}^n x_{1,i} & \sum_{i=1}^n x_{2,i} \\ \sum_{i=1}^n x_{1,i} & \sum_{i=1}^n x_{1,i}^2 & \sum_{i=1}^n x_{2,i}^2 \\ \sum_{i=1}^n x_{2,i} & \sum_{i=1}^n x_{1,i}x_{2,i} & \sum_{i=1}^n x_{1,i}x_{2,i} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{1,i}y_i \\ \sum_{i=1}^n x_{2,i}y_i \end{pmatrix}$$

The calculations needed to develop the normal equations for the above data

	y	x_1	x_2	x_1^2	x_2^2	x_1x_2	x_1y	x_2y
	5	0	0	0	0	0	0	0
	10	2	1	4	1	2	20	10
	9	2.5	2	6.25	4	5	22.5	18
	0	1	3	1	9	3	0	0
	3	4	6	16	36	24	12	18
	27	7	2	49	4	14	189	54
$\sum$	54	16.5	14	76.25	54	48	243.5	100

so that

$$\begin{pmatrix} 6 & 16.5 & 14 \\ 16.5 & 76.25 & 48 \\ 14 & 48 & 54 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 54 \\ 243.5 \\ 100 \end{pmatrix}.$$

Which can be solved for $a_0 = 5$, $a_1 = 4$, $a_2 = -3$, *which is consistent with the original equation from which the data was derived.*

Solution 8.4 *We have*

$$n = 9, \quad \sum_{i=1}^{9} x_i = 20, \quad \sum_{i=1}^{9} x_i^2 = 60, \quad \sum_{i=1}^{9} y_i = 12, \quad \sum_{i=1}^{9} y_i^2 = 20$$
$$\sum_{i=1}^{9} x_i y_i = 30, \quad \sum_{i=1}^{9} z_i = 242.7, \quad \sum_{i=1}^{9} x_i z_i = 661, \quad \sum_{i=1}^{9} y_i z_i = 331.2$$

then

$$\begin{pmatrix} 9 & 20 & 12 \\ 20 & 60 & 30 \\ 12 & 30 & 20 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 242.7 \\ 661 \\ 331.2 \end{pmatrix}$$

we get $a_0 = 14.40$, $a_1 = 9.03$, $a_2 = -5.62$, *then* $z = 14.4 + 9.03x - 5.62y$.

8.7 Exercises without solutions

Exercise 8.1 *1. Let the points be* $x_0 = -1$, $x_1 = -\frac{1}{2}$, $x_2 = 0$, $x_3 = \frac{1}{2}$, $x_4 = 1$.
Find the polynomial $P^* \in \mathbb{P}_2$ *which realizes the following minimum*

$$\min_{P \in \mathbb{P}_2} \sum_{i=0}^{4} \frac{1}{1 + x_i^2} (|x_i| - P(x_i)),$$

where $\mathbb{P}_2$ *is the set of polynomials of degree* ≤ 2.

Exercise 8.2 *Let* $f(x) = |x|$ *and* $< g, h >= \sum_{i=0}^{4} g(x_i)h(x_i)$, $\forall g, h \in C([-1, 1], \mathbb{R})$ *where*

$$x_0 = -1, \; x_1 = -\frac{1}{2}, \; x_2 = 0, \; x_3 = \frac{1}{2}, \; x_4 = 1.$$

1. *Determine the polynomial* P *of* $\mathbb{P}$ *which achieves the best approximation of* f *in the sense of least squares.*
2. *Verify that* $Q(x) = \frac{7}{2}x^2 - \frac{4}{3}x^4$ *is the interpolation polynomial of* f *at points* x_i, $i = 0, ..., 4$.
3. *In the same frame, draw the graphs of* f, P *and* Q *on* $[-1, 1]$. *Comment.*

Exercise 8.3 *We consider the set* $E = C([-1, 1])$ *endowed with the scalar product*

$$\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)dx.$$

Let $\mathbb{P}_1$ *be the set of polynomials of degree* ≤ 1 *and* $f \in E$.

1. *Find as a function of* f, *the polynomial* $P_1(x) = a + bx$ *which realizes the best approximation in the least squares sense of* f *in* P_1.
2. *Show that* P_1 *can be put in the form:*

$$P_1(x) = \int_{-1}^{1} (1 + 3xt)f(t)dt.$$

3. *Find* P_1 *pour* $f(x) = -x^3 - 2x$. *Draw the graphs of* f *and* P_1.
4. *Evaluate the error and make its graphical representation.*

Exercise 8.4 *We are looking for the polynomial* $P_2(x) = a_0 + a_1x + a_2x^2$ *which minimizes the expression*

$$\int_{-1}^{1} (f(x) - P_2(x))^2 \frac{dx}{\sqrt{1 - x^2}},$$

1. *Consider the orthogonal polynomials, given by the relation* $T_0(x) = 1$, $T_1(x) = x$, $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$, $n \leq 1$. *Calculate* T_2, T_3, T_4.
2. *We take* $f(x) = 2x^3 + x^4$. *Determine* P_2 *in the base* $\{T_1, T_2, T_3\}$ *and deduce the values of* a_0, a_1 *and* a_2.
3. *Evaluate the error made.*

Exercise 8.5 *Find the best approximation to a solution of each of the following systems of equations.*

$$\begin{aligned} x + y - z &= 5 \\ 2x - y + 6z &= 1 \\ 3x + 2y - z &= 6 \\ -x + 4y + z &= 0 \end{aligned}$$

and

$$3x + y + z = 62x + 3y - z = 12x - y + z = 03x - 3y + 3z = 8$$

Exercise 8.6 *Find the least squares approximations of the line* $y = z_0 + z_1x$ *for each of the following sets of data points.*

a. $(1, 1)$, $(3, 2)$, $(4, 3)$, $(6, 4)$

b. $(2, 4)$, $(4, 3)$, $(7, 2)$, $(8, 1)$

c. $(-1, -1)$, $(0, 1)$, $(1, 2)$, $(2, 4)$, $(3, 6)$

d. $(-2, 3)$, $(-1, 1)$, $(0, 0)$, $(1, -2)$, $(2, -4)$

Exercise 8.7 *Find the least squares approximations of quadratic* $y = z_0 + z_1x + z_2x^2$ *for each of the following sets of data points.*

a. $(0, 1)$, $(2, 2)$, $(3, 3)$, $(4, 5)$

b. $(-2, 1)$, $(0, 0)$, $(3, 2)$, $(4, 3)$

Exercise 8.8 *Find an approximate least-squares function of the form* $r_0x + r_1x^2 + r_2 2^x$ *for each of the following sets of data pairs.*

a. $(-1, 1)$, $(0, 3)$, $(1, 1)$, $(2, 0)$

b. $(0, 1)$, $(1, 1)$, $(2, 5)$, $(3, 10)$

Exercise 8.9 *Find the least-squares approximation function of the form* $r_0 + r_1x^2 + r_2 \sin(\frac{\pi}{2}x)$ *for each of the following sets of data pairs.*

a. $(0, 3)$, $(1, 0)$, $(1, -1)$, $(-1, 2)$

b. $(-1, \frac{1}{2})$, $(0, 1)$, $(2, 5)$, $(3, 9)$

Exercise 8.10 *Newton's laws of motion imply that an object released from rest at a height of* 100 *meters will be at a height* $s = 100 - \frac{1}{2}gt^2$ *meters* t *seconds later, where* g *is a constant called acceleration due to gravity.*
The values of s *and* t *given in the table are respected. Write down* $x = t^2$, *find the least squares approximating the line* $s = a + bx$ *for these data, and use* b *to estimate* g. *Next, find the least squares quadratic approximant* $s = a_0 + a_1t + a_2t^2$ *and use the value of* a_2 *to estimate* g.

t	1	2	3
s	95	80	56

Exercise 8.11 *A naturalist measured the heights y_i (in meters) of several spruces with trunk diameter x_i (in centimeters). The data is as indicated in the table. Find the least squares approximation line for this data and use to estimate the height of a spruce tree with a trunk* $10cm$ *in diameter.*

x_i	5	7	8	12	13	16
y_i	2	3.3	4	7.3	7.9	10.1

Exercise 8.12 *Wheat yield y in bushels per acre appears to be a linear function of the number of days x_1 of sunshine, the number of inches x_2 of rain, and the number of pounds x_3 of fertilizer applied per acre . Find the best fit to the table data by an equation of the form $y = r_0 + r_1x_1 + r_2x_2 + r_3x_3$. (If a calculator for inverting A^TA is not available, the inverse is given in the answer).*

y	x_1	x_2	x_3
28	50	18	10
30	40	20	16
21	35	14	10
23	40	12	12
23	30	16	14

Appendix A

Numerical simulation

A.1 Introduction

Matlab is a numerical calculation software produced by MathWorks (see the website *http* : *//www.mathworks.com/*). It is available on multiple platforms. Matlab is a simple and very efficient language, optimized for processing matrices and for digital calculation. Matlab is short for "matrix laboratory", all objects in play are matrices, including scalars (1×1 matrices). This software is specially designed for scientific computing and manipulation of vectors and matrices. MATLAB is both a programming language and a development environment developed and marketed by the American company MathWorks. MATLAB is used by scientists and engineers, for different applications (signal and image processing, system control, statistics, data analysis and processing, modelling, etc.).

MATLAB is an interpreter: instructions are interpreted and executed line by line. MATLAB works in several environments such as Windows, Macintosh, UNIX, Linux.

There are two modes of operation:

1. Interactive mode: MATLAB executes instructions as they are given by the user.
2. Executive mode: MATLAB executes an 'M file' line by line (program in MATLAB language).

A.2 Initiation to Matlab

We propose a non-exhaustive introduction to useful commands in scientific computing. First use of Matlab, the online help Once Matlab is launched, the instructions for Matlab must be typed in the command window the acronym >> means a line command, For example:

```
>>1+2
ans =
3
>>t=1+2
t =
3
>>t=1+1;
>>t
t=
2
>>u=sin(t)
u =
0.9093
>>v=exp(u)
v=
2.4826
>>long format
```

```
>>v
v=
2.48257772801500
>>short size
>>v
v=
2.4826
>>who
Your variables are:
years u v
>>whos
Name Size Bytes Class
years 1x1 8 dual array
t 1x1 8 double array
u 1x1 8 double array
v 1x1 8 dual-array
>> clear
>> v
```

We note that:

1. by default, any calculation is assigned to the variable **ans**
2. the command **format** allows to modify the display of the format of the different variables
3. the **clear** command clears the contents of all used variables.

A.2.1 Main Commands

help
For more information on a command, you can use the help command followed by the name of the command requested: **help <command>**
Example: help help
quit
This command is used to quit MATLAB, at the end of our work.
clc
To clear the window.
clear / clear all
It resets the environment (the "workspace") by destroying all the active variables in memory.
whos / who
All active variables can be consulted using the whos and who commands.

A.2.2 The command

```
>> y=sin(0.15*pi);
```

the calculation of y has been performed but it is not displayed on the screen. To display it, just type y

```
>> y
y= 0.4540
```

***The command line**

```
>> x=50, y=9, z=x+y,
```

You can define several variables in the same command line. The separation between variable is with comma $\ll$, $\gg$.
The semicolon ";"
You must end the operation with a semicolon ";" otherwise, all the steps of the calculation will be displayed on the screen. Example: In the command window, type:

```
>>a = 4*5;
>> x=5; y=5; z=x+y
z=
10
```

A.2.3 Edit Window

The "Edit Window" can be seen as a text editor where:

- Comments are written in green and start with "%"
- Variables and equations appear in black
- Characters appear in red
- Keywords in Matlab like loops appear in blue

A.2.4 Special Variables

pi, *i*, *realmin*, *realmax*, *eps*, *ans* ... are predefined variables.

```
>> eps
y =
2.2204e-16
>> 1.+eps
1.0000
>> realmax
ans =
1.7977e+308
>> realmin
ans =
2.2251e-308
>> v=1.e+400
v=
whos
>> pi
```

Machine zero eps is the largest positive real such that $1 + eps <= 1$!!!!!.

```
>>y = 0.444444444444444e-30
>>x = 0.222222222222222e-20
>> x+y
0.222222222666667e-20
>>u = 0.444444444444444e+30
>>v = 0.222222222222222e+20
>> u+v
0.444444444666666e+30
```

The calculation is done by reducing to the highest power. In the first case, the next way

$$x + y = 10^{-20}(0.222222222222222 + 0.000000000444444)$$

and in the second

$$u + v = 10^{30}(0.444444444444444 + 0.000000000222222).$$

```
>> c1 = 1-2i
c1 =
1.0000 - 2.0000i
>> c2 = 3*(2-sqrt(-1)*3)
c2 =
```

```
6.0000 - 9.0000i
>>c3= conj(c2)
>> real(c1)
>> imag(c2)
>> abs(c2)
>> angle(c1)
ans =
-1.1071
```

A.2.5 Display

FORMAT Set output format. All computations in MATLAB are done in double precision. FORMAT may be used to switch between different output display formats as follows: FORMAT SHORT (default) Scaled fixed point format with 5 digits. LONG FORMAT Scaled fixed point format with 15 digits.

```
>> pi
ans =
3.1416
>> long format
>> pi
ans =
3.141592653589
```

A.2.6 Comments

It is important to be able to write comments when developing a program. For that, on one line everything after the % symbol is unread.

```
>> pi % to know the value of pi
```

A.2.7 Vectors - Matrices

Arrays (vectors, matrices, ...) are used in the matlab command window or in programs. The indices of vectors and matrices always start at 1. The index 0 does not exist in matlab. The ith component of a vector x is called x(i); the coefficient of the ith row, jth column of a matrix A is A(i, j). There is no need to declare the arrays, but it is however recommended to reserve a memory space for the matrices by initializing them to 0. (see below).

A.2.7.1 Creating vectors.

−− A row vector is defined

(i) either by a relation of the form

$$x = [x1x2xn],$$

or

$$x = [x1, x2, , xn],$$

where xi elements are separated by spaces or commas and enclosed in [ ].
Examples:

```
>> x1 = [1.1 2.3 3.5 -4. -8.]% (vector of 5 components)
x1 =
1.1000 2.3000 3.5000 -4.0000 -8.0000
>> x2 = [3 -5.2 2*sqrt(3)] % (vectors of 3 components).
x2 =
3.0000 -5.2000 3.4641
>>x2(3)
```

(ii) or by a subdivision of a given interval (a, b) into sub-intervals of given length h,

$x = a : h : b$ go from a to b in steps of h,
$x = a : b$ (default $h = 1$).

```
>> x = 5:-1:1
x = 5 4 3 2 1
>> x=1:1.1:5
x = 1.0000 2.1000 3.2000 4.3000
>> x=5:1
x =
Empty matrix: 1-by-0
>> x = 0: pi/2: 2 * pi
x =
0 1.5708 3.1416 4.7124 6.2832
```

(iii) either by the instruction

x = linspace(a, b, n) defines a vector of n components $x(i) = a + (i-1)\dfrac{b-a}{n-1}$, for $i = 1 \cdots n$. Default: $n = 100$.

```
>>y= linspace(0,pi,9)
y =
0 0.3927 0.7854 1.1781 1.5708 1.9635 2.3562 2.7489 3.1416
```

(iv) or by using a function. For example, if x is a row vector, then sin(x) and cos(x) (for example) are row vectors with the same number of components.

A.2.7.2 Transposed vector.

−− A column vector can be defined as the transpose of a row vector, x′ denotes the transposed vector of vector x.

```
>> x = (1:4)'; y = [3 -4.5 2.1]' % two column vectors x
                                  % and y of components
```

ii) or directly, using ; instead of ,.

$x = [x1; x2; \quad ; xn]$

```
>> z = [3.1452; -3; 4.; 5.256] % defines a column vector
                               % z of components
```

iii) using a function. If x is a column vector, $\sin(x)$ and $\cos(x)$, for example, are column vectors of the same length as x.

```
>> x = (0:0.2:1); % x is a column vector
>> y = exp(x); % y is a column vector
```

A.2.7.3 Vector operations.

−− Some vector operations:

+ *	:	addition of two vectors u and v of the same length.
dot(u, v) or u'* v	:	scalar product of two vectors u and v.
.*	:	multiplies two component vectors by components.
./	:	ditargets two by two the components of two vectors.
.	:	raises the components of one vector to the power of the components of the second.
sum(u)	:	sum of the components of a vector u.
mean(u)	:	mean of the components of a vector u.
length(u)	:	gives the length of a vector u.
min(u)	:	gives the smallest component of a vector u.
max(u)	:	gives the largest component of a vector u.

```
>> u = (1:4) , v = (2:5)
u =
1 2 3 4
v=
2 3 4 5
>> z = u .* v % z(i) = u(i)*v(i)
z =
2 6 12 20
>> w = v./u % w(i) = v(i) = u(i)
w =
2.0000 1.5000 1.3333 1.2500
>> t = v .^u % t(i) = v(i) ^ u(i)
t =
2 9 64 625
>> x = (0:0.2:1)
x =
0 0.2000 0.4000 0.6000 0.8000 1.0000
>> z = exp(x).* cos(x)
z =
1.0000 1.1971 1.3741 1.5039 1.5505 1.4687
%z is a vector whose components are the elements
%z(i) = exp(x(i)) cos(x(i)) for i = 1; ; 6 ,
>> disp('mean(z) = '), mean(z)
>>disp('length(z) = '), length(z)
>>disp('min(z) = '), min(z)
>>disp('max(z) = '), max(z)
>>disp('sum(z) = '), sum(z)
>>disp('max(abs(z)) = '), max(abs(z))
```

disp('phrase to display'): displays on the screen the sentence "phrase to display"

A.2.8 Matrix operations

A.2.8.1 Creating matrices

i) A matrix of n rows and p columns can be defined by a relation of the form:

$$A = [a_{11}\, a_{12}\, a_{1p};\ a_{21}\, a_{22}\, a_{2p};\ \cdots\ a_{n1}\, a_{n2}\, a_{np}]$$

```
>> A = [1 2 3 4 ; 5 6 7 8; 9 10 11 12] % A is a matrix of
                                        %3 rows and 4 columns
A=
1 2 3 4
5 6 7 8
9 10 11 12
```

A' designates the transpose of A if A is real, the adjoint of A if A is complex. $A(i, j)$ designates the element of row i and column j.

ii) A matrix can also be constructed from vectors or smaller matrices **concatenation**: Example

```
>> A = [1 2 3; 4 5 6]
A=
1 2 3
4 5 6
>> B = [A; 7 8 9]
B=
1 2 3
4 5 6
7 8 9
```

A.2.8.2 Creating matrices $n \times p$.

−− Initializing matrices

A=zeros(n,p)	:	initialization at 0 of a matrix n rows and p columns
A=eye(n)	:	order identity matrix n
A=ones(n, p)	:	matrix n rows, p columns, consisting of 1
A= rand(n, p)	:	random number array $\in]0;1[$.

Example.

```
>> x = zeros(1, 5) % defines the zero vector of 5 components
x = 0 0 0 0 0
>> x = zeros(5 , 1) %defines a column vector of 5 components equal to 0.
>> x = ones(1, 5) %define the row vector of 5 components
x = 1 1 1 1 1
```

Some possible operations

$+ *$	:	addition, multiplication of two compatible matrices.
$x = A(1,:)$	:	first line of A.
$y = A(:, 2:3)$	:	second and third columns of$A$.
$w = A(3, i:j)$	:	from i-th to j-th elements of line 3.
$u = A(:, 2)$	:	second column of A.
$z = A(:)$	:	format a column.
$c * A$	:	multiply all elements of A by the scalar c.
$A.m$	:	raising to the power m of each element of the matrix A.
A^m	:	exponentiation m of matrix$A$.
$size(A)$	:	gives the dimensions of the matrix $A([m, n] = size(A))$
$eig(A)$	:	vector giving the eigenvalues of the matrix A
$det(A)$	:	gives the determinant of the matrix A.
$rank(A)$	:	rank of matrix A
$trace(A)$	:	trace of matrix A
$spy(A)$	:	graphical representation of the matrix A (in the case of large matrices).

Example:

```
>>A= rand(6,6)
>>DetA=det(A)
>>PropVal = eig(A)
>>A(1,:)=1
>>A(4,3)=-3.3333
>>rankA= rank(A)
>>B = rand(20,30);
>>B(3:6,10:18)=0.;
>>spy(B)
>>[n,m]=size(B)
>>NbLineB = size(B,1)
>>NbColumnB= size(B,2)
```

A.2.9 M-Files or scripts

A script (or M-file) is a file (premiertp.m for example) containing Matlab instructions.
All matlab files must end with the suffix .m
Here is an example script or matlab program:

```
% my first matlab program
%firsttp.m calculates det, eigenvalues of a matrix
%
clear all
```

```
%%% Matrices
n=input('Give the dimension of the matrix n = ')
10.10. FUNCTIONS 119
A= rand(n,n)
DetA=det(A)
PropVal = eig(A)
A(1,:)=1
A(4,3)=-3.3333
rankA= rank(A)
nbl=input('Give the number of rows (> 10) of the matrix B = ')
nbc=input('Give the number of columns (>20) of the matrix = ')
B = rand(nbl,nbc);
B(4:8,10:18)=0.;
spy(B)
[n,m]=size(B)
NbLineB = size(B,1)
NbColumnB= size(B,2)
```

Matlab offers you an editor to write and debug your M-files:
To **run** the program just click on run in the editor menu,
Or
in the command window, we execute an M-file using the script name as the command:

```
>> premiertp
```

The M-files (or program) are **executed sequentially** in the "workspace", that is to say that they can access the variables which are already there, modify them, create others etc.

A.2.10 Functions

A function is a script admitting input variables and output variables, for example example $f(x, y) = x^2 + y^2$ admits x and y as input variables and the result $z = x^2 + y^2$ as an output argument. Here is a "fonc" function defined in a file $fonc.m$

```
function [z,t] = func(x,y,m)
z = x^2+y^2;
t=x-y+m;
end
```

The variables x, y and m are the input variables and they must not be modified inside of the function. The variables z and t are the output arguments, the function must assign a value for z and t.
Using this function. A function is called from a script (a main program) or in a command window.

```
>> [z,t] = func(1., 0., -4.)
>>[y1,y2]= func(-1., sin(1.), sqrt(2.) )
```

A.2.11 Help

All matlab features are illustrated from the **help** menu.

```
   >>help function
>> help for
>> help switch
>> help if
```

A.2.12 Graphics

```
>> help 2-D Plots
```

put the following instructions in a *courbes.m* script

```
% curves.m
% Examples of 2D graphics
% graph 1
face
x=0:0.05:5;
y=sin(x.^2);
plot(x,y);
xlabel('Time')
ylabel('Amplitude')
title('my first curve')
legend('toto')

%graph 2
face
z = -2.9:0.2:2.9;
bar(z,exp(-z.*z));

%graph 3
face
subplot(2,1, 1)
plot(x,y);
subplot(2,1,2)
bar(z,exp(-z.*z));

% graph 4
face
Y=[0:0.05:1];Z1=sin(2*pi*Y);Z2=cos(2*pi*Y);
plot(Y,Z1,':b',Y,Z2,'+k');
title('Example of curves');
xlabel('Y');ylabel('Z');
caption('sin','cos');
```

Comment on each instruction in this program.

Exercise. Define the function *gaussian.m*,

```
function [g] = Gaussian(x,xc,s)
g=exp( - (x-xc)^2/s^2)
end
```

plot the function for different values of xc and s on the same curve (make a script $tracer-gauss.m$). In the Gaussian function the argument x is a scalar, write a function "$gaussianv(x, xc, s)$" with x a vector. Plot the function for different value of xc and s on the same curve.

A.2.13 Tick tock

Calculate the CPU execution time of a program. Just do the *tic* instruction at the beginning of the script and then to do *toc* at the end of the script to have the execution time of the program.

A.2.14 Math functions

```
sin, cos, tan, sinh, cosh, tanh, ...
asin, acos, atan, asinh, acosh, atanh, ...
exp, log, log10, sqrt, ...
```

fix(x)	:	gives the smallest integer less than or equal to the real x
floor(x)	:	gives the integer part of x
ceil(x)	:	nearest integer greater than or equal to x
round(x)	:	
mod(x)	:	division remainder
sign	:	
other functions	:	factor isprime primes $gcd(gcd)$, $lcm(ppcm)$.

A.3 Programming with MATLAB

Matlab can be used as an advanced programming language. We can write

- scripts (command files)
- Where to define new functions

These scripts and functions can use Matlab's built-in functions.

A.3.1 Scripts

To execute a script, it is necessary to mention its file name in the line of Matlab command. The script instructions execute one after the other as if they were typed on the Matlab command line. The variables defined in the scripts remain in the memory of Matlab after the execution of the script.

Example A.3.1 *To create a script, just type the command edit*

```
>> edit
```

or go to menu Home ⟶ New ⟶ Script

Enter the following instructions, and save the file under the name ***script.m***

To run a script, we return to the Matlab window

- *either by typing it with the keyboard*

```
>> script

D =

     5     4     8
```

- *or by clicking on the **Run** icon from the edit window.*

A.3.2 Loops and control

A.3.2.1 Logical comparison operators.

```
< smaller
> bigger
<= less than or equal
>= greater than or equal
== equal (compare whether two numbers are equal or not)
~= different or not equal
/&  logical and
|  logical or
~ not
```

A.3.3 Conditions and Loops

- for loop

  ```
  \textbf{for} variable = expression
  Instructions
  \textbf{end}
  ```

  Example. Write the matrix A such that $a_{ij} = i + j^2$ with $1 \leq i \leq 4$ and $1 \leq j \leq 5$.

  ```
  A = zeros(4.5)
  for i=1:4
  for j=5:-1:1
  A(i,j) = i+j^2;
  end
  ```

- While Loop

  ```
  \textbf{while} condition
  Instructions
  \textbf{end}
  ```

  Example. Use a while loop to calculate factorial (10).

  ```
  n = 10;
  f = n;
  while n > 1
      n = n-1;
      f = f*n;
  end
  disp('n! = ' num2str)
  ```

- In a script, you can use the conditional statements

  ```
  \textbf{If} condition
  instructions
  \textbf{else}
  instructions
  \textbf{end}
  ```

Example

```
a= log(2.)
b= sqrt(2.)/2.
if (a > b)
disp('a is greater than b')
else
disp('a is smaller than b')
end
```

Or

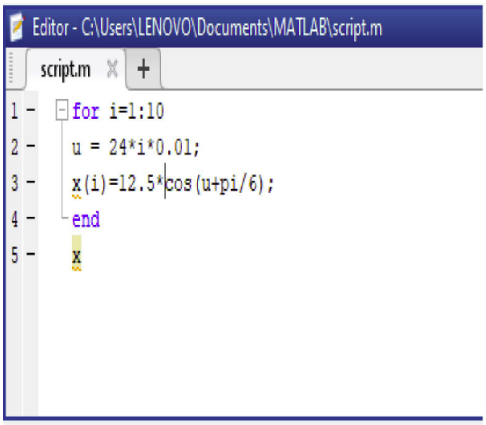
```
\textbf{switch} expression
\textbf{case} value1
instructions
\textbf{case} value2
instructions
otherwise
instructions
\textbf{end}
```

Example

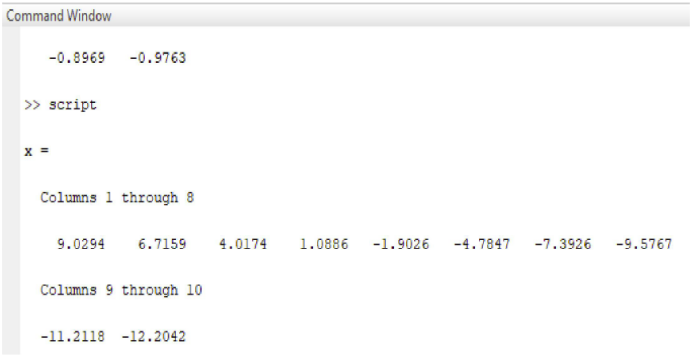
```
switch m % do according to the value of m
case 0 % if m=0
x=0
case {1, 2, 9} % if m=1 or m=2 or m=9
x= m^2
case {-3,-1, 99} % if m=-3 or m=-1 or m= 99
x= m+ m^2
otherwise % else
x=-9999999
end
```

Example A.3.2 *For loop example*

Script:



Running the script gives



A.3.4 Functions

A function is a script that

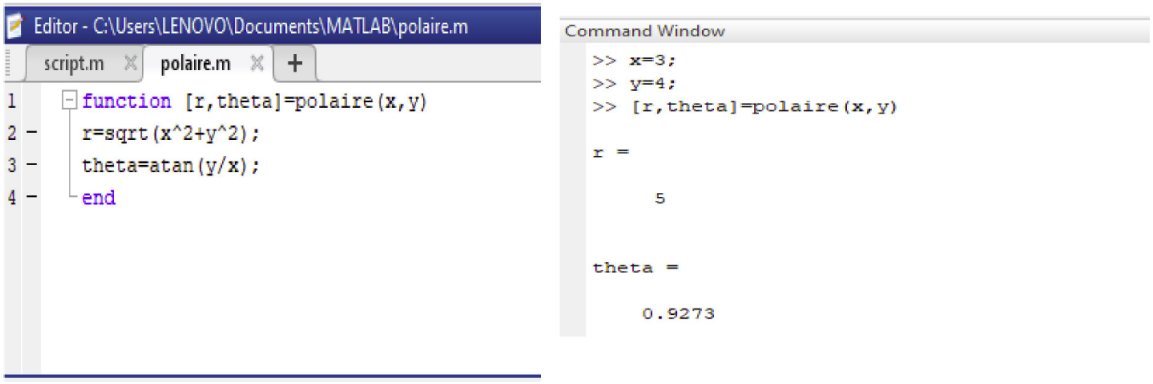
- receives input arguments
- returns results
- locally defined variables will be cleared after execution of the function

```
function [y_1,...,y_m]=function_name(x_1,....,x_n)
```

Or name_function is the name of the function, $x_1, ..., x_n$, the n arguments input and $y_1, ..., y_m$, the m output arguments.

The function must be saved in a file with the name of the function and the extension **.m** (name_function.m).

Example A.3.3 *Function example (Function)*



Bibliography

[1] **G. Allaire, S.M. Kaber,** *Algèbre linéaire numérique ellipse,* Mathématiques 2e cycle édition, 2002.

[2] **R.L. Burden, J.D Faires,** *Numerical analysis,* 9th édition, Brook/cole, Pacific Grove, 2007.

[3] **O.G. Ciarlet,** *Introduction et analyse numérique matricielle et à l'optimisation,* Dunod.

[4] **M. Crouzeix, AL Mignot** *Analyse numérique des équations différentielles,* collec. Math. Appli. pour la maitrise. Masson, 1984.

[5] **J.P. Demailly,** *Analyse numérique et équations différentielles,* collection Grenoble Sciences.

[6] **P. Lascaux, R. Théodor.** *Analyse numérique matricielle appliquée à l'art de l'ingénieur,* Tomes 1 et 2 Masson 1986.

[7] **J. Solomon,** *Numerical algorithms: methods for computer vision, machine learning, and graphics,* CRC Press, Taylor & Francis Group, 2015.

[8] **S. Saha Ray,** *Numerical analysis with algorithms and programmings,* CRC Press, Taylor & Francis Group, 2016.

[9] https://math.libretexts.org.