Applied Mathematics
Class Notes

Master 1 - Mathématiques Appliquées, Statistics
Parcours Data Science and Parcours Computational Mathematical Biology
Aix-Marseille Université

FLORENCE HUBERT AND FRÉDÉRIC RICHARD

December 15, 2024

Year 2024 — 2025

Table of Contents

Bibliography

[Cia95] P.G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, 1995.

[Ros67] Robert Rosen, *Optimality principles in biology*, Butterworths, London, 1967.

[Ser02] D. Serre, *Matrices, theory and applications*, Springer, 2002.

Introduction

Advanced mathematical models and methods are more and more used within most areas of science, industry, medicine, economics and social sciences. Two important issues arise:

- Parameters involved in the models have to be calibrated/estimated from the real data.
- Models are too complex to be solved exactly.

For that we need statisticis, approximation tools and now machine learning, that are based on linear algebra and optimization.

Here are some optimization problems that arise in various context in biology:

- Level of protein expression. Different proteins have different expression levels. Evolutionary theories suggest that protein expression levels may maximize fitness.
- Branching structure of the vascular tree
- Optimized protocol in medecine
- Optimal Leaf Size in Relation to Environment
- ....

For more examples, see [Ros67].

Chapter 1 Basics in linear algebra

We consider in this chapter a finite-dimensional vector space E over a field $\mathbb{K}$. Denote by n its dimension and $\mathcal{L}(E)$ the set of linear map from E to E. This chapter is primarily composed of concepts that you are expected to have already acquired. We assume that the following notions of matrix calculus are known: matrix multiplications, determinant, trace, ...

1.1 Vector space, linear map, matrix representation

1.1.1 Vector space and bases

Definition 1.1: Vector space

A vector space over a field $\mathbb{K}$ is a non-empty set V together with

- a binary operation called *addition* that assigns to any two vectors $\mathbf{x}$ and $\mathbf{y}$ in V a third vector $\mathbf{x}+\mathbf{y} \in V$,
- a binary function called *scalar multiplication* that assigns to any scalar $\lambda \in \mathbb{K}$ and any vector $\mathbf{x} \in V$ another vector $\lambda\mathbf{x} \in V$

with the following axioms

1. **Associativity** of the vector addition: $\mathbf{x}+(\mathbf{y}+\mathbf{z}) = (\mathbf{x}+\mathbf{y})+\mathbf{z}$
2. **Commutativity** of the vector addition: $\mathbf{x}+\mathbf{y} = \mathbf{y}+\mathbf{x}$
3. **Identity element** of the vector addition: there exists an element $\mathbf{0} \in V$ such that $\mathbf{x}+\mathbf{0} = \mathbf{x}$
4. Inverse elements of the vector addition: for every $\mathbf{x} \in V$, there exists an element $-\mathbf{x} \in V$ such that
$$\mathbf{x}+(-\mathbf{x}) = \mathbf{0}$$
5. **Compatibility** of scalar multiplication with field multiplication: $\lambda(\mu\mathbf{x}) = (\lambda\mu)\mathbf{x}$
6. **Identity element** of the scalar multiplication: there exists an element $1 \in \mathbb{K}$ such that $1\mathbf{x} = \mathbf{x}$
7. **Distributivity** of scalar multiplication with respect to vector addition: $\lambda(\mathbf{x}+\mathbf{y}) = \lambda\mathbf{x}+\lambda\mathbf{y}$
8. **Distributivity** of scalar multiplication with respect to field addition: $(\lambda+\mu)\mathbf{x} = \lambda\mathbf{x}+\mu\mathbf{x}$

for $\mathbf{x},\mathbf{y},\mathbf{z}$ in V and λ,μ in $\mathbb{K}$.

Definition 1.2: Linear vector subspace

A vector subspace W of a $\mathbb{K}$−vector space V is a non-empty subset of V that is a vector space for the induced addition and scalar multiplication.

Proposition 1.1

A subspace W of a $\mathbb{K}$−vector space V that is a vector subspace of V if and only if

1. W is a non-empty subset
2. for all $\mathbf{x}, \mathbf{y} \in W$, $\lambda \in \mathbb{K}$, $\mathbf{x} + \lambda \mathbf{y} \in W$

Definition 1.3: Linear combination

Given a set G of elements of a $\mathbb{K}$-vector space V, a **linear combination** of elements of G is an element of V of the form

$$a_1\mathbf{g}_1 + a_2\mathbf{g}_2 + \cdots + a_k\mathbf{g}_k,$$

where $a_1, \ldots, a_k \in \mathbb{K}$ and $\mathbf{g}_1, \ldots, \mathbf{g}_k \in G$. The scalars $a_1, \ldots, a_k$ are called the **coefficients** of the linear combination.

Definition 1.4: Linear independence

The elements of a subset G of a $\mathbb{K}$-vector space V are said to be **linearly independent** if and only if for any family $\mathbf{g}_1, \ldots, \mathbf{g}_k$ of G,

$$a_1\mathbf{g}_1 + a_2\mathbf{g}_2 + \cdots + a_k\mathbf{g}_k = 0 \text{ implies } a_1 = a_2 = \cdots = a_k = 0.$$

Definition 1.5: Linear span

Given a subset G of a $\mathbb{K}$−vector space V, the **linear span** or simply the span of G is the set of all linear combinations of elements of G. If W is the span of G, one says that G spans or generates W, and that G is a generating set of W.

Definition 1.6: Basis

A subset of a vector space is a basis if its elements are **linearly independent** and span the vector space. It implies that every $\mathbf{v} \in V$ may be written $\mathbf{v} = a_1\mathbf{b}_1 + \cdots + a_n\mathbf{b}_n$, with $a_1, \ldots, a_n \in \mathbb{K}$, and that this decomposition is unique. The scalars $a_1, \ldots, a_n$ are called the coordinates of v on the basis. They are also said to be the coefficients of the decomposition of v on the basis.

Definition 1.7: Dimension

Assume that a $\mathbb{K}$−vector space V admits a basis $\mathcal{B}$, then any other basis of V has the same cardinality called the **dimension**. In particular, if $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$, the cardinality of $\mathcal{B}$ is equal to n and V is said to be of **finite dimension** equal to n and is denoted by $\dim V$.

Example 1.1

Let $\mathbf{x} \in V$ a non-zero vector, the span generated by $\mathbf{x}$ has a dimension 1. It is a line also denoted by $\mathbb{K}\mathbf{x}$.

Example 1.2

Let $\mathbf{x}, \mathbf{y} \in V$ two vectors linearly independant. The span generated by $\mathbf{x}$ and $\mathbf{y}$ is a plane of dimension 2.

Example 1.3

For a field $\mathbb{K}$, the space $\mathbb{K}^n = \{\mathbf{x} = (x_1, \cdots, x_n),\ \text{for } x_1, \cdots, x_n \in \mathbb{K}\}$ is a vector space over the field $\mathbb{K}$. Its dimension is equal to n. We call canonical basis, the family

$$\mathbf{e}_1 = (1, 0, \cdots, 0), \cdots, \mathbf{e}_n = (0, \cdots, 0, 1).$$

Definition 1.8: Line, plane, hyperplane

Let V be a vector space over a field $\mathbb{K}$.

- A **line** is a subset of V of dimension 1.
- A **plane** of V is a subset of V of dimension 2.
- A **hyperplane** of a a vector space of finite dimension n is a subspace of dimention $n-1$.

Remark.

The one-to-one correspondence between vectors and their coordinate vectors maps vector addition to vector addition and scalar multiplication to scalar multiplication. It is thus a vector space isomorphism, which allows translating reasonings and computations on vectors into reasonings and computations on their coordinates.

1.1.2 Linear map and its matrix representation

Definitions

Definition 1.9: Linear map

Let V and W be vector spaces over the same field $\mathbb{K}$. A function $f : V \to W$ is said to be a **linear map** if for any two vectors $\mathbf{x}, \mathbf{y} \in V$ and any scalar $\lambda \in \mathbb{K}$ the following two conditions are satisfied:

1. **Additivity**
$$f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$$
2. **Homogeneity of degree 1**
$$f(\lambda\mathbf{x}) = \lambda f(\mathbf{x})$$

Proposition 1.2

If $f : V \to W$ is a linear map, then we have the following properties:

- f preserves linear combinations, for any vectors $\mathbf{x}_1, \dots, \mathbf{x}_n \in V$ and scalars $\lambda_1, \dots, \lambda_n \in \mathbb{K}$, we have
$$f(\lambda_1\mathbf{x}_1 + \cdots + \lambda_n\mathbf{x}_n) = \lambda_1 f(\mathbf{x}_1) + \cdots + \lambda_n f(\mathbf{x}_n).$$
- Denoting the zero elements of the vector spaces V and W, $\mathbf{0}_V$ and $\mathbf{0}_W$ respectively, we have
$$f(\mathbf{0}_V) = f(0\mathbf{x}) = 0f(\mathbf{x}) = \mathbf{0}_W \text{ for all } \mathbf{x} \in V.$$

Definition 1.10: Linear functional

A linear map $V \to \mathbb{K}$ with $\mathbb{K}$ viewed as a one-dimensional vector space over itself is called a linear functional.

Definition 1.11: Matrix representation

Let V and W be two finite-dimensional vector spaces endowed respectively with bases $\mathcal{B}_V = (\mathbf{v}_1, \cdots, \mathbf{v}_n)$ and $\mathcal{B}_W = (\mathbf{w}_1, \cdots, \mathbf{w}_m)$. If $f : V \longrightarrow W$ is a linear map, the **representation matrix** of f in the bases $\mathcal{B}_V, \mathcal{B}_W$ is a matrix $M_{\mathcal{B}_W, \mathcal{B}_V}(f)$ denoted by $A = (a_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}}$ for short where $(a_{ij})_{1 \le i \le m}$ are the coordinates of the vectors $f(\mathbf{v}_j)$ in the basis $\mathcal{B}_W$:

$$f(\mathbf{v}_j) = \sum_{i=1}^{m} a_{ij} w_i = a_{ij}\mathbf{w}_1 + \cdots, a_{mj}\mathbf{w}_m$$

For any vector $\mathbf{v} = \lambda_1 \mathbf{v}_1 + \cdots + \lambda_n \mathbf{v}_n \in V$, the vector $f(\mathbf{v})$ is then given by

$$f(\mathbf{v}) = \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{m} a_{ij} \mathbf{w}_i$$

Proposition 1.3: Addition operation

Let V and W be three finite-dimensional vector spaces endowed respectively with bases $\mathcal{B}_V = (\mathbf{v}_1, \cdots, \mathbf{v}_n)$ and $\mathcal{B}_W = (\mathbf{w}_1, \cdots, \mathbf{w}_m)$. If $f : V \longrightarrow W$ and $g : V \longrightarrow W$ are two linear maps, then $f + g$ is also a linear map. Moreover,

$$M_{\mathcal{B}_W, \mathcal{B}_V}(f + g) = M_{\mathcal{B}_W, \mathcal{B}_V}(f) + M_{\mathcal{B}_W, \mathcal{B}_V}(g).$$

Proposition 1.4: Composition operation

Let V, Y and W be three finite-dimensional vector spaces endowed respectively with bases $\mathcal{B}_V = (\mathbf{v}_1, \cdots, \mathbf{v}_n)$, $\mathcal{B}_Y = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$ and $\mathcal{B}_W = (\mathbf{w}_1, \cdots, \mathbf{w}_m)$. If $f : V \longrightarrow Y$, and $g : Y \longrightarrow W$ are two linear maps, then $g \circ f : V \longrightarrow W$ is a linear map. Moreover,

$$M_{\mathcal{B}_W, \mathcal{B}_V}(g \circ f) = M_{\mathcal{B}_W, \mathcal{B}_Y}(g) M_{\mathcal{B}_Y, \mathcal{B}_V}(f).$$

Notion of range and kernel

Definition 1.12

If $f : V \longrightarrow W$ is a linear map, we define the kernel and the image or range of f by

$$\begin{aligned} \mathrm{Ker}(f) &= \{\mathbf{x} \in V : f(\mathbf{x}) = \mathbf{0}\} \\ \mathrm{Im}(f) &= \{\mathbf{w} \in W : \mathbf{w} = f(\mathbf{x}), \mathbf{x} \in V\} \end{aligned}$$

Theorem 1.5: Rank-nullity theorem

The set $\mathrm{Ker}(f)$ is a subspace of V and $\mathrm{Im}(f)$ is a subspace of W. Assume that V and W have a finite dimension, we have

$$\dim(\mathrm{Ker}(f)) + \dim(\mathrm{Im}(f)) = \dim(V).$$

The number $\dim(\mathrm{Im}(f))$ is also called the rank of f and written as $\mathrm{Rank}(f)$; the number $\dim(\mathrm{Ker}(f))$ is called the nullity of f and written as (f).

Remark.

If V and W are finite-dimensional, bases have been chosen and f is represented by the matrix A, then the rank and nullity of f are equal to the rank and nullity of the matrix A respectively.

Notion of isomorphism, endomorphism and automorphism

Proposition 1.6: Injectivity, surjectivity, bijectivity

Let V and W denote vector spaces over a field $\mathbb{K}$ and let $f : V \to W$ be a linear map. We have

- f is injective if and only if $\ker(f) = \{\mathbf{0}_V\}$.
- f is surjective if and only if $\mathrm{Im}(f) = W$.
- f is bijective if and only if f is bijective if and only if $\ker(f) = \{\mathbf{0}_V\}$ and $\mathrm{Im}(f) = W$.

Definition 1.13: Isomorphism, endomorphism, automorphism

- An isomorphism is a bijective linear map.
- A linear map f from V to V where V is a $\mathbb{K}-$ vector space is an **endomorphism**.
- An automorphism is a bijective endomorphism.

Proposition 1.7

Let V and W denote vector spaces over a field $\mathbb{K}$.

- The set $\mathcal{L}(V, W)$ of the linear maps from V to W forms a vector space over $\mathbb{K}$.
- The set $\mathcal{L}(V)$ of the endomorphisms of V together with the addition and the composition operations of linear maps forms an associative algebra.
- The set $\mathcal{GL}(V)$ of the automorphisms of V together with the composition operation forms a group.

Proposition 1.8

Let V and W be two finite-dimensional vector spaces endowed respectively with a basis $\mathcal{B}_V = (\mathbf{v}_1, \cdots, \mathbf{v}_n)$ and $\mathcal{B}_W = (\mathbf{w}_1, \cdots, \mathbf{w}_m)$.

- There exists an isomorphism between $\mathcal{L}(V, W)$ and $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$ that associates to f the linear map $\widetilde{f} : \mathbb{K}^n \to \mathbb{K}^m$ defined by $\widetilde{f}(\mathbf{x}) = A\mathbf{x}$ where A is the representation matrix of f in $\mathcal{B}_V, \mathcal{B}_W$.
- There exists an isomorphism between $\mathcal{L}(V, W)$ and $\mathcal{M}_{m,n}(\mathbb{K})$ (set of matrices of size $m \times n$) that associates to f the representation matrix A of f in $\mathcal{B}_V, \mathcal{B}_W$.
- There exists an isomorphism between $\mathcal{GL}(V)$ and $\mathcal{GL}_n(\mathbb{K})$ (set of invertible matrices of size $n \times n$) that associates to f the representation matrix A of f in $\mathcal{B}_V, \mathcal{B}_V$.

Take home message.

In finite dimension, working with linear maps is equivalent to working with matrices.

1.1.3 Some important examples

Elementary matrix

Definition 1.14: Elementary matrix

Let i, j two distinct integers from $\{1, \cdots, m\} \times \{1, \cdots, n\}$ $(i \neq j)$ and define the matrix $E_{ij} \in \mathcal{M}_{m,n}(\mathbb{K})$ whose coefficients are all equal to 0 except the coefficient i, j that takes 1 as value.

1.1.4 Identity

Proposition 1.9

Consider the endomorphism $f : V \longrightarrow V$ such that $f(\mathbf{x}) = \mathbf{x}$ for each $\mathbf{x} \in V$. In any basis $\mathcal{B}$ of V, the matrix representation of f is the identity matrix

$$Id = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

We will also denote by Id this endomorphism.

Dilatation

Definition 1.15: Dilatation map

We say that $f \in \mathcal{L}(V)$ is a dilatation map if $\mathrm{Ker}(f - Id)$ is an hyperplane of V and f is invertible.

Definition 1.16: Dilatation matrix

Let $i \in \{1, \cdots, n\}$ and define for $a \in \mathbb{K}$, $a \neq 0$ the diagonal matrix $D_i(a) \in \mathcal{M}_n(\mathbb{K})$ whose diagonal coefficients are all equal to 1 except the i^{th} that takes a as value:

$$D_i(a) = \left(\begin{array}{ccccccc} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & & \vdots \\ \vdots & & \ddots & a & \ddots & & \vdots \\ \vdots & & & \ddots & 1 & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{array}\right) \begin{array}{l} \\ \\ \\ - i^{th} \text{ row} \\ \\ \\ \\ \end{array}$$

If f is a dilatation map, there exists $a \neq 0$ and a basis $\mathcal{B}$ in which the representation matrix of f is the matrix $D_i(a)$. In the case $a = -1$, we say that f is a reflexion.

Proposition 1.10

Let $A \in \mathcal{M}_{m,n}(\mathbb{K})$.

- The matrix $D_i(a)A$ is the matrix obtained from A by multiplying by a its i^{th} row by a.
- The matrix $AD_i(a)$ is the matrix obtained from A by multiplying by a its i^{th} column by a.

Example 1.4

Let $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$. Then

$$D_2(a)A = \begin{pmatrix} 1 & 2 & 3 \\ 4a & 5a & 6a \\ 7 & 8 & 9 \end{pmatrix}, AD_2(a) = \begin{pmatrix} 1 & 2a & 3 \\ 4 & 5a & 6 \\ 7 & 8a & 9 \end{pmatrix}.$$

Transvection

Definition 1.17: Transvection map

We say that f is a transvection map if f is not the identity and if there exists a hyperplane $\mathcal{H}$ of V such that $f_{|\mathcal{H}} = Id_{|\mathcal{H}}$.

Definition 1.18: Transvection matrix

Let i, j two distinct integers from $\{1, \cdots, n\}$ $(i \neq j)$, and $\lambda \in \mathbb{K}$. We define the matrix $T_{ij}(\lambda) = I_n + \lambda E_{ij}$ or

$$T_{ij}(\lambda) = \begin{array}{l}\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \lambda & 0 & 1 & 0 & & 0 \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ -i^{th}\text{ row} \\ \\ \\ \\ \end{matrix} \\ \qquad\quad | \\ \qquad\quad j^{th}\text{ column} \end{array}$$

If f is a transvection map, there exists $\lambda \neq 0$ and a basis $\mathcal{B}$ in which the representation matrix of f is the matrix $T_{ij}(\lambda)$. A tranvection matrix is a triangular matrix (upper if $i < j$, lower else). It is an invertible matrix, with $T_{ij}(\lambda)^{-1} = T_{ij}(-\lambda)$.

Proposition 1.11

Let $A \in \mathcal{M}_{m,n}(\mathbb{K})$.

- The matrix $T_{ij}(\lambda)A$ is the matrix obtained from A by adding to the i^{th} row of A, λ times the j^{th} row.
- The matrix $AT_{ij}(\lambda)$ is the matrix obtained from A by adding to the j^{th} column of A, λ times the i^{th} column.

Example 1.5

Let $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$. Then $T_{21}(\lambda) = \begin{pmatrix} 1 & 0 & 0 \\ \lambda & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and

$$T_{21}(\lambda)A = \begin{pmatrix} 1 & 2 & 3 \\ 4+\lambda & 5+2\lambda & 6+3\lambda \\ 7 & 8 & 9 \end{pmatrix}, \quad AT_{21}(\lambda) = \begin{pmatrix} 1+2\lambda & 2 & 3 \\ 4+5\lambda & 5 & 6 \\ 7+8\lambda & 8 & 9 \end{pmatrix}$$

Permutation matrix

Definition 1.19

Consider a permutation σ of $\{1, \cdots, n\}$. The permutation matrix P_σ associated to σ is the matrix defined by

$$P_{ij} = \begin{cases} 1 \text{ if } j = \sigma(i) \\ 0 \text{ else} \end{cases}$$

Note that the permutation matrix are invertible.

Proposition 1.12

Let $A \in \mathcal{M}_{m,n}(\mathbb{K})$.

- The matrix $P_\sigma A$ is the matrix obtained from A by exchanging the rows of A, following the permutation σ.
- The matrix AP_σ is the matrix obtained from A by exchanging the columns of A following the permutation σ.

Example 1.6

Let $\sigma = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{bmatrix}$ and $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$. Then $P_\sigma = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ and

$$P_\sigma A = \begin{pmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 1 & 2 & 3 \end{pmatrix}, AP_\sigma = \begin{pmatrix} 2 & 5 & 8 \\ 3 & 6 & 9 \\ 1 & 4 & 7 \end{pmatrix}$$

.

Circulant matrix

Define P the permutation matrix associated to the permutation σ defined by $\sigma(i) = i + 1 mod(n)$ for all $i = 1, \cdots, n$:

$$P = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & \ddots & \ddots & \vdots & 0 \\ 0 & \ddots & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

It is easy to check that

$$P^2 = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 \\ 0 & \ddots & & \ddots & 1 \\ 1 & \ddots & \ddots & & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \cdots, P^{n-1} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & \ddots & 1 \\ 1 & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

Definition 1.20

A circulant matrix is a matrix that takes the form:

$$A = \sum_{i=0}^{n-1} a_i P^i = \begin{pmatrix} a_0 & a_{n-1} & \cdots & a_2 & a_1 \\ a_1 & \ddots & \ddots & & a_2 \\ a_2 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{n-1} \\ a_{n-1} & \cdots & a_2 & a_1 & a_0 \end{pmatrix}$$

where $a_i \in \mathbb{K}$ for $i = 0, \cdots, n-1$.

Rank 1 endomorphism

Proposition 1.13

Let V be a vector space on a field $\mathbb{K}$. The endomorphism f is of rank 1 if and only if there exists a vector $\mathbf{u}$ such that

$$\text{for each } \mathbf{x} \in V, f(\mathbf{x}) = \lambda_{\mathbf{x}} \mathbf{u} \text{ where } \mathbf{x} \mapsto \lambda_{\mathbf{x}} \in \mathcal{L}(V, \mathbb{K}).$$

Remark.

If V is endowed with a scalar product $(.,.)$, then f is a rank 1 endomorphism if and only if there exists a vector $\mathbf{u}$ such that

$$f(\mathbf{x}) = (\mathbf{x}, \mathbf{u})\mathbf{u} \text{ for all } \mathbf{x} \in V.$$

Proposition 1.14

Let $\mathbf{u} \in \mathbb{K}^n$ and $\mathbf{v} \in \mathbb{K}^n$. The matrix $\mathbf{u}\mathbf{v}^t$ is a rank 1 -matrix and $\mathbb{Im}(\mathbf{u}\mathbf{v}^t) = \mathbb{K}\mathbf{u}$.

Projection

Definition 1.21

Let V be a vector space over a field $\mathbb{K}$. We said that $p \in \mathcal{L}(V)$ is a projection if $p \circ p = p$.

Proposition 1.15

Let $p \in \mathcal{L}(V)$ be a projection. We have

- $p_{|\mathrm{Im}(p)} = Id_{|\mathrm{Im}(p)}$
- $\mathrm{Im}(p)$ and $\mathrm{Ker}(p)$ are complementary (*i.e.* $V = \mathrm{Im}(p) + \mathrm{Ker}(p)$ and $\mathrm{Im}(p) \cap \mathrm{Ker}(p) = \{0\}$ or simply $V = \mathrm{Im}(p) \oplus \mathrm{Ker}(p)$). There exists a basis $\mathcal{B}$ in which the representation matrix of p can be written by block as
$$M_{\mathcal{B}}(p) = \begin{pmatrix} Id & 0 \\ 0 & 0 \end{pmatrix}$$

Proof. We just need to write that for each $\mathbf{x} \in V$,

$$\mathbf{x} = \underbrace{p(\mathbf{x})}_{\in \mathrm{Im}(p)} + \underbrace{\mathbf{x} - p(\mathbf{x})}_{\in \mathrm{Ker}(p)}.$$

and to construct a basis $\mathcal{B}$ adapted to the decomposition $V = \mathrm{Im}(p) \oplus \mathrm{Ker}(p)$. □

Frobenius compagnion matrix

Definition 1.22

Let $P(x) = x^n + a_{n-1}X^{n-1} + \cdots + a_1 X + a_0$ a monic polynomial. We call Frobenius compagnion matrix associated to P, the matrix

$$C(P) = \begin{pmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & \ddots & \ddots & \vdots & -a_1 \\ 0 & \ddots & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -a_{n-1} \end{pmatrix}$$

Matrix of the discretization of the Laplace operator

Important examples of matrice come from the discretization of differential operators. The most famous one is the Laplace operator that associates to a a function u defined on the interval $[0,1]$ its derivative $-u''$. This operator is usually endowed with what is call boundary conditions. Classical examples are :

- **Homogeneous Dirichlet boundary conditions.** $u(0) = u(1) = 0$.
- **Periodic boundary conditions.** $u(0) = u(1)$.

The finite difference discretization process consists in approximating the operator at n points ih for $i=1,\cdots,n$ with $h=\frac{1}{n+1}$. Using taylor expansion, we see that

$$-u''(ih) = -\frac{1}{h^2}\left(u((i+1)h) - 2u(ih) + u((i-1)h)\right) + \mathcal{O}(h^2).$$

Approximating the solution of the problem $-u''=f$ then leads to the resolution of a linear system with a matrix of the form

$$A_{dir} = (n+1)^2 \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \text{ for Dirichlet BC,} \quad A_{per} = (n+1)^2 \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & -1 \\ -1 & \ddots & \ddots & & \ddots & 0 \\ 0 & \ddots & & \ddots & & \vdots \\ \vdots & & \ddots & & \ddots & 0 \\ 0 & \ddots & & \ddots & \ddots & -1 \\ -1 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \text{ for Periodic BC}$$

1.1.5 Exercises

Exercise 1.1

Given $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, we introduce $A \in \mathcal{M}_{m,n}(\mathbb{R})$ the matrix defined by $A = uv^t$. The vector u is supposed to be normed.

1. Prove that $\mathrm{Rk}(A) = 1$.
2. Give the matrix of the orthogonal projection on $\mathrm{Im}(A)$.

---

1. It is easy to see that for all $x \in \mathbb{R}^n$
$$Ax = u \underbrace{v^t x}_{\in\mathbb{R}} \in \mathrm{Vect}(u)$$
and thus $\mathrm{Rk}A = 1$.
2. Let $B = uu^t$ and $y \in \mathrm{Im}(A)$, we have $y = \lambda u$ and thus $By = uu^t(\lambda u) = \lambda u \underbrace{u^t u}_{=1} = y$. Now, if $z \in (\mathrm{Im}A)^{\perp}$ then $u^t z = 0$, so $Bz = uu^t z = 0$. We deduce that $B = uu^t$ is the matrix of the orthogonal projection on $\mathrm{Im}(A)$.

Exercise 1.2

Given two families of vectors $(u_i)_{i=1\cdots,l} \in \mathbb{R}^m$ and $(v_i)_{i=1,\cdots,l} \in \mathbb{R}^n$, we introduce $A \in \mathcal{M}_{m,n}(\mathbb{R})$ the matrix given by $A = \sum_{i=1}^{l} u_i v_i^t$. We assume that the two families $(u_i)_{i=1\cdots,l}$ and $(v_i)_{i=1\cdots,l}$ are orthonormal.

1. Determine $\mathrm{Rk}(A)$.
2. Give the matrix of the orthogonal projection on $\mathrm{Im}(A)$.

---

1. It is easy to see that for all $x \in \mathbb{R}^n$
$$Ax = \sum_{i=1}^{l} u_i \underbrace{v_i^t x}_{\in\mathbb{R}} \in \mathrm{Vect}(u_1,\cdots,u_l)$$
and thus $\mathrm{Rk}A \le l$. Now, remarking that $Av_i = u_i$, we deduce that $\mathrm{Im}(A) = \mathrm{Vect}(u_i)$ and thus $\mathrm{Rk}(A) = l$.
2. Let $B = \sum_{i=1}^{l} u_i u_i^t$. We have for $i = 1,\cdots,l$, $Bu_i = u_i$ so that $B_{|\mathrm{Im}(A)} = Id$ and if $z \in \mathrm{Im}(A)0erp$, for all $i = 1,\cdots,l$, $u_i^t z = 0$ and thus $Bz = 0$.

Exercise 1.3

Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$ a matrix of rank $r \leq p = \min(m,n)$, let $[U, \Sigma, V]$ be its SVD decomposition. Denote by $(u_1, \cdots, u_m)$ the columns vectors of U and $(v_1, \cdots, v_n)$ those of V.

1. Prove that $\text{Im}(A) = \text{Vect}(u_1, \cdots, u_r)$.
2. Prove that $\text{Ker}(A) = \text{Vect}(v_{r+1}, \cdots, v_n)$.
3. Prove that $\text{Im}(A^t) = \text{Vect}(v_1, \cdots, v_r)$.
4. Prove that $\text{Ker}(A^t) = \text{Vect}(u_{r+1}, \cdots, u_m)$.
5. Determine the matrices of the orthogonal projection on $\text{Im}(A)$, $\text{Ker}(A)$, $\text{Im}(A^t)$, $\text{Ker}(A^t)$ thanks to U and V.
6. Find the SVD decomposition of $A = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ -1 & 1 \end{pmatrix}$ and the matrices of the orthogonal projections $\text{Im}(A)$, $\text{Ker}(A)$, $\text{Im}(A^t)$, $\text{Ker}(A^t)$.

We first recall that if $A = U\Sigma V^t$ then equivalently we have

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^t.$$

1. Consider $y \in \text{Im}(A)$, then

$$y = Ax = \sum_{i=1}^{r} \sigma_i u_i v_i^t x = \sum_{i=1}^{r} \underbrace{\sigma_i v_i^t x}_{\in\mathbb{R}} u_i \in \text{Vect}(u_1, \cdots, u_r)$$

As the dimension of ImA $= r = dim(\text{Vect}(u_1, \cdots, u_r))$, we conclude the proof of the first point.

2. For $l = r+1, \cdots, n$, we have

$$Av_l = \sum_{i=1}^{r} \sigma_i u_i \underbrace{v_i^t v_l}_{=0}$$

Therefore, $\text{Vect}(v_{r+1}, \cdots, v_n) \subset \text{Ker}(A)$. We conclude by using an dimension argument.

3. We remark that

$$A^t = \sum_{i=1}^{r} \sigma_i v_i u_i^t.$$

Consider $y \in \text{Im}(A^t)$, then

$$y = A^t x = \sum_{i=1}^{r} \sigma_i v_i u_i^t x = \sum_{i=1}^{r} \underbrace{\sigma_i u_i^t x}_{\in\mathbb{R}} v_i \in \text{Vect}(v_1, \cdots, v_r)$$

We conclude by using an dimension argument.

4. For $l = r+1, \cdots, n$, we have

$$A^t u_l = \sum_{i=1}^{r} \sigma_i v_i \underbrace{u_i^t u_l}_{=0}$$

Therefore, $\text{Vect}(u_{r+1}, \cdots, u_n) \subset \text{Ker}(A^t)$. We conclude by using an dimension argument.

5. We recall that the orthogonal projection P on a subset F is such that $Px = x$ for all $x \in F$ and $Px = 0$ for all $x \in F^\perp$. Note that $Id - P$ is the orthogonal projection on $F^\perp$.

   - The orthogonal projection on $\text{Im}(A)$ is thus $P = \sum_{i=1}^{r} u_i u_i^t$. Indeed, for $l = 1, \cdots, r$ using question 1

$$Pu_l = \sum_{i=1}^{r} u_i \underbrace{u_i^t u_l}_{=\delta_{il}} = u_l$$

and as $\text{Im}(A)^\perp = \ker(A^t)$, using question 4, we check that for $l \geq r+1$ $Pu_l = 0$.

- ▶ Similarly, the orthogonal projection on $\mathrm{Im}(A^t)$ is thus $P = \sum_{i=1}^{r} v_i v_i^t$.
- ▶ Using the fact that $\mathrm{Im}(A)^\perp = \mathrm{Ker}(A^t)$, we deduce that the orthogonal projection on $\mathrm{Ker}(A^t)$ is given by

$$Id - \sum_{i=1}^{r} u_i u_i^t$$

- ▶ Using the fact that $\mathrm{Im}(A^t)^\perp = \mathrm{Ker}(A)$, we deduce that the orthogonal projection on $\mathrm{Ker}(A)$ is given by

$$Id - \sum_{i=1}^{r} v_i v_i^t$$

6. We remark that

$$A^t A = \begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix}$$

Its eigenvalues are $\sigma_1^2 = 7$ and $\sigma_2^2 = 2$ with eigenvectors $v_1 = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $v_2 = \frac{1}{\sqrt{5}}\begin{pmatrix} -1 \\ 2 \end{pmatrix}$. We deduce

$$u_1 = \frac{1}{\sigma_1} A v_1 = \frac{1}{\sqrt{5}}\frac{1}{\sqrt{7}}\begin{pmatrix} 3 \\ 5 \\ -1 \end{pmatrix}, \; u_2 = \frac{1}{\sigma_2} A v_2 = \frac{1}{\sqrt{5}}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}$$

We now choose $u_3 = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$ orthogonal to u_1 and u_2:

$$3\alpha + 5\beta - \gamma = 0, \; \alpha + 3\gamma = 0, \; \alpha^2 + \beta^2 + \gamma^2 = 1$$

We deduce

$$\alpha = -3\gamma, \; \beta = \frac{10}{5}\gamma, \; \gamma^2\left(9 + \frac{100}{25} + 1\right) = 1 \text{ or } \gamma = \frac{5}{\sqrt{350}} = \frac{1}{\sqrt{14}}$$

Or $u_3 = \frac{1}{\sqrt{14}}\begin{pmatrix} -3 \\ 2 \\ 1 \end{pmatrix}$.

We have

$$u_1 u_1^t = \frac{1}{35}\begin{pmatrix} 9 & 15 & -3 \\ 15 & 25 & -5 \\ -3 & -5 & 1 \end{pmatrix}, \; u_2 u_2^t = \frac{1}{10}\begin{pmatrix} 1 & 0 & 3 \\ 0 & 0 & 0 \\ 3 & 0 & 9 \end{pmatrix}, \; u_3 u_3^t = \frac{1}{14}\begin{pmatrix} 9 & -6 & 3 \\ -6 & 4 & 2 \\ -3 & 2 & 1 \end{pmatrix}$$

and

$$v_1 v_1^t = \frac{1}{5}\begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}, \; v_2 v_2^t = \frac{1}{5}\begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix}$$

Therefore,

- ▶ the matrix of the projection on $\mathrm{Im}(A)$ is given by

$$P_{\mathrm{Im}(A)} = u_1 u_1^t + u_2 u_2^t = \frac{1}{70}\begin{pmatrix} 25 & 30 & 15 \\ 30 & 50 & -10 \\ 15 & -10 & 65 \end{pmatrix} = \frac{1}{14}\begin{pmatrix} 5 & 6 & 3 \\ 6 & 10 & -3 \\ 3 & -2 & 13 \end{pmatrix}$$

- ▶ the matrix of the projection on $\mathrm{Im}(A^t)$ is given by

$$P_{\mathrm{Im}(A^t)} = v_1 v_1^t + v_2 v_2^t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- ▶ $\mathrm{Ker}(A)$ is reduced to 0.
- ▶ the matrix of the projection on $\mathrm{Ker}(A^t)$ is given by

$$P_{\mathrm{Ker}(A)} = u_3 u_3^t = \frac{1}{14}\begin{pmatrix} 9 & -6 & 3 \\ -6 & 4 & 2 \\ -3 & 2 & 1 \end{pmatrix}$$

1.2 Polynomial of endomorphism

Let V be a vector space of a field $\mathbb{K}$ and $\mathcal{L}(V)$ the set of endomorphisms defined on V. For any $f \in \mathcal{L}(V)$ and any polynomial $P \in \mathbb{K}[X]$, we can define the endomorphism $P(f)$:

$$\text{if } P(X) = \sum_{i=0}^{k} a_i X^i \text{ then } P(f) = \sum_{i=0}^{k} a_i f^i \text{ where } f^i = \underbrace{f \circ \cdots \circ f}_{i \text{ times}}.$$

Let I_f be the set of annihilating polynomials of f:

$$I_f = \{P \in \mathbb{K}[X] \text{ such that } P(f) = 0\}.$$

1.2.1 Minimal and characteristic polynomials

Proposition 1.16

The set of annihilating polynomials I_f is a non zero ideal of $\mathbb{K}[X]$, generated by a unique monic polynomial μ_f called the **minimal polynomial**.

We recall that

- a monic polynomial $P(X) = \sum_{i=0}^{k} a_i X^i$ is polynomial such that $a_k = 1$.
- an ideal I is a group for the operation addition and for the operation multiplication, it satisfies:

$$\text{if } p \in I \text{ then for all } q \in \mathbb{K}[X],\ pq \in I.$$

Remark.

The degree of the minimal polynomial is at least equal to 1. Any annihilating polynomial of f is thus a multiple of μ_f.

Example 1.7

- Let $A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}$. We have $P(A) = \begin{pmatrix} P(2) & 0 & 0 \\ 0 & P(3) & 0 \\ 0 & 0 & P(4) \end{pmatrix}$. Thefore $P(A) = 0$ if and only if $P(X) = (X-2)(X-3)(X-4)Q(X)$. We deduce that

$$\mu_A(X) = (X-2)(X-3)(X-4).$$

- Let $A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}$. We have $P(A) = \begin{pmatrix} P(2) & 0 & 0 \\ 0 & P(2) & 0 \\ 0 & 0 & P(4) \end{pmatrix}$. Thefore, $P(A) = 0$ if and only if $P(X) = (X-2)(X-4)Q(X)$. We deduce that

$$\mu_A(X) = (X-2)(X-4).$$

- Let $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. We have $A - I = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq 0$, but $(A-I)^2 = 0$. Consequently,

$$\mu_A(X) = (X-1)^2.$$

Be carefull, the minimal polynomial is not always irreductible.

Definition 1.23

The **characteritic polynomial** of f is the polynomial defined by

$$\chi_f(X) = \det(f - Xid) = (-1)^n \left(X^n - \mathrm{Tr}(f) X^{n-1} + \cdots + (-1)^n \det(f)\right).$$

Definition 1.24

A scalar $\lambda \in \mathbb{K}$ is an **eigenvalue** of f if and only if $f - \lambda id$ is not invertible. In particular, if λ is an eigenvalue then there exists a non-zero vector $\mathbf{x} \neq 0$, called **eigenvector** such that $f(\mathbf{x}) = \lambda \mathbf{x}$.

Remark.

- The eigenvalues are the roots of the characteristic polynomials.
- If $\mathbf{x}$ is an eigenvector of f, the vector subspace $\mathbb{K}\mathbf{x}$ is stable by f $(f(\mathbb{R}\mathbf{x}) \subset \mathbb{R}\mathbf{x})$.

1.2.2 Properties of the minimal and characteristic polynomial

Proposition 1.17

- If $A = \begin{pmatrix} A' & B \\ 0 & A'' \end{pmatrix}$ then $\chi_A(X) = \chi_{A'}(X)\chi_{A''}(X)$.
- If $B = \begin{pmatrix} B' & 0 \\ 0 & B'' \end{pmatrix}$ then $\mu_B(X) = LCM(\mu_{B'}(X), \mu_{B''}(X))$.

Proof. We just need to prove that $P(B) = \begin{pmatrix} P(B') & 0 \\ 0 & P(B'') \end{pmatrix}$ and thus $P(B) = 0$ if and only if $P(B') = P(B'') = 0$. □

Proposition 1.18: Minimal and characteristic polynomial of a restriction

Let $f \in \mathcal{L}(E)$ and F a subset **stable** by f (*i.e.* $f(F) \subset F$). Then $\chi_{f_{|F}}$ divide χ_f and $\mu_{f_{|F}}$ divide μ_f.

Proof. Let $\mathcal{B}'$ be a basis of F and $\mathcal{B}''$ a basis of it supplementary. Then the representation matrix of f in $\mathcal{B} = (\mathcal{B}', \mathcal{B}'')$ is given by:

$$\begin{pmatrix} A' & B \\ 0 & A'' \end{pmatrix}.$$

We deduce immediatly the result for the characteristic polynomial and for the minimal polynomial we remark that for all $\mathbf{x} \in F$, we have $\mu_f(\mathbf{x}) = 0$ so that μ_f is an annihilating polynomial for $f_{|F}$ therefore it is divided by $\mu_{f_{|F}}$. □

1.2.3 The Cayley-Hamilton theorem

We propose to prove the famous Cayley-Hamilton theorem by using the cyclic endomorphisms.

Theorem 1.19: Cayley-Hamilton theorem

The characteristic polynomial of u satisfies

$$\chi_u(u) = 0.$$

Therefore, μ_u divide χ_u.

Definition 1.25: Cyclic subspace

Let $u \in \mathcal{L}(E)$. A subspace F of E is **cyclic** if there exists $\mathbf{x} \in E$ such that

$$F = \mathrm{Span}\left(\mathbf{x}, u(\mathbf{x}), u^2(\mathbf{x}), \cdots, u^k(\mathbf{x}), \cdots\right).$$

We will denote by $E_{\mathbf{x}}$ or $\mathbb{K}[u](\mathbf{x})$ this subspace.

Definition 1.26

An endomorphism u is said to be **cyclic** if there exists $\mathbf{x} \in E$ such that $E = E_{\mathbf{x}}$. This element $\mathbf{x}$ is called a **cyclic vector**.

Theorem 1.20

Let $u \in \mathcal{L}(E)$ a cyclic endomorphism.

1. If $\mathbf{x}$ is a cyclic vector for u, then $\mathcal{B} = (\mathbf{x}, u(\mathbf{x}), \cdots, u^{n-1}(\mathbf{x}))$ forms a basis of Ein which
$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} 0 & & & & b_0 \\ 1 & \ddots & & (0) & b_1 \\ & \ddots & \ddots & & \vdots \\ & (0) & \ddots & 0 & \vdots \\ & & & 1 & b_{n-1} \end{pmatrix}$$
2. The endomorphisms $id, u, \cdots, u^{n-1}$ are linearly independant on $\mathcal{L}(E)$.
3. An endomorphism v commutes with u if and only if $v \in \mathbb{K}[u]$.

Proof.

1. Let us introduce
$$k = \max\{j \text{ tel que } (\mathbf{x}, u(\mathbf{x}), \cdots, u^{j-1}(\mathbf{x})) \text{ is linearly independant}\}.$$
By assumption, $k \leq n$. Moreover, by definition of k, we have
$$u^k(\mathbf{x}) \in \mathrm{Span}\left(\mathbf{x}, u(\mathbf{x}), \cdots, u^{k-1}(\mathbf{x})\right)$$
We deduce that
$$u^{k+1}(\mathbf{x}) \in \mathrm{Span}\left(u(\mathbf{x}), u^2(\mathbf{x}), \cdots, u^k(\mathbf{x})\right) \subset \mathrm{Span}\left(\mathbf{x}, u(\mathbf{x}), \cdots, u^{k-1}(\mathbf{x})\right)$$
and then by induction for all $j \in \mathbb{N}$
$$u^j(\mathbf{x}) \in \mathrm{Span}\left(\mathbf{x}, u(\mathbf{x}), \cdots, u^{k-1}(\mathbf{x})\right)$$
As $\mathbf{x}$ is a cyclic vector, we have $E \subset \left\langle \mathbf{x}, u(\mathbf{x}), \cdots, u^{k-1}(\mathbf{x})\right\rangle$. We conclude that $k = n$ and $\mathcal{B} = (\mathbf{x}, u(\mathbf{x}), \cdots, u^{n-1}(\mathbf{x}))$ is a basis of E. We denote $u^n(\mathbf{x}) = b_0\mathbf{x} + b_1 u(\mathbf{x}) + \cdots + b_{n-1}u^{n-1}(\mathbf{x})$. The matrix of u in $\mathcal{B}$ can be deduced immediatly.
2. Assume the existence of $\lambda_0, \cdots, \lambda_{n-1}$ such that
$$\sum_{i=0}^{n-1} \lambda_i u^i = 0.$$
In particular, we have $\sum_{i=0}^{n-1} \lambda_i u^i(\mathbf{x}) = 0$ and as $\mathcal{B}$ a family of linearly independant vectors, we deduce that $\lambda_i = 0$ for all i.
3. Let $v \in \mathbb{K}[u]$ then $v \circ u = u \circ v$. Conversely, assume that $v \circ u = u \circ v$. As $\mathcal{B}$ is a basis of E and as $v(x) \in E$, there exists $\lambda_0, \cdots, \lambda_{n-1}$ such that
$$v(\mathbf{x}) = \sum_{i=0}^{n-1} \lambda_i u^{i-1}(\mathbf{x}).$$
Consider the endomorphism w defined by $w = \sum_{i=0}^{n-1} \lambda_i u^{i-1}$. By assumption, $\mathbf{x} \in \mathrm{Ker}(w - v)$. As u and $w - v$ commute, we have $(w - v)(u(\mathbf{x})) = u((w - v)(\mathbf{x})) = 0$, therefore we can deduce that $u(\mathbf{x}) \in \mathrm{Ker}(w - v)$ and then by induction that $u^j(\mathbf{x}) \in \mathrm{Ker}(w - v)$ for each integer j. We deduce that $v = w$ and thus $v \in \mathbb{K}[u]$.

□

We have used in the previous result the following lemma

Lemma 1.21

If two endomorphisms u and v commute, then each eigenspace of u is stable by v and reciprocally.

Proposition 1.22

Let u be a cyclic endomorphism of a vector space E of dimension n, then $\mu_u = (-1)^n \chi_u$. In particular, $\chi_u(u) = 0$.

Proof. From the previous proposition, the family $id, \cdots, u^{n-1}$ is linearly independant, so that the minimal polynomial of u has a degree greater that n. Moreover, in the basis $(\mathbf{x}, u(\mathbf{x}), \cdots, u^{n-1}(\mathbf{x}))$ of E (where $\mathbf{x}$ is a cyclic vector of u), the matrix of the endomorphism u is given by

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} 0 & & & & b_0 \\ 1 & \ddots & & (0) & b_1 \\ & \ddots & \ddots & & \vdots \\ & (0) & \ddots & 0 & \vdots \\ & & & 1 & b_{n-1} \end{pmatrix}$$

This matrix is a **Frobenius compagnion matrix**.

We associate to this matrix the polynomial $P(X) = X^n - \sum_{i=0}^{n-1} b_i X^i$. By assumption, $P(u)(\mathbf{x}) = 0$ and $u^k \circ P(u)(\mathbf{x}) = P(u)(u^k(\mathbf{x})) = 0$ donc comme $(\mathbf{x}, u(\mathbf{x}), \cdots, u^{n-1}(\mathbf{x}))$ defines a basis of E, we can deduce that $P(u) = 0$ and thus that $\mu_u = P$.

It is then easy to check that $\chi_u = (-1)^n P$. Indeed,

$$\chi_u(X) = \underbrace{\begin{vmatrix} -X & & & & b_0 \\ 1 & \ddots & & (0) & b_1 \\ & \ddots & \ddots & & \vdots \\ & (0) & \ddots & -X & \vdots \\ & & & 1 & b_{n-1} - X \end{vmatrix}}_{Det(b_0, \cdots, b_{n-1})} \overset{\text{Dev \% } L_1}{=} -X \underbrace{\begin{vmatrix} -X & & & & b_1 \\ 1 & \ddots & & (0) & b_2 \\ & \ddots & \ddots & & \vdots \\ & (0) & \ddots & -X & \vdots \\ & & & 1 & b_{n-1} - X \end{vmatrix}}_{Det(b_1, \cdots, b_{n-1})} + (-1)^n b_0$$

We deduce by induction that n.

We thefore proved that for any cyclic endomorphism, we have $\mu_u = (-1)^n \chi_u$. □

Proof of the Cayley-Hamilton theorem We want to prove that $\chi_u(u) = 0$. Let us fix $x \in E \setminus \{0\}$ and define $E_x = \left\langle x, u(x), \cdots, u^k(x), \cdots \right\rangle$. The space E_x is a vectorial subspace of E of dimension $p \leq n$. This subspace E_x is stable by u. Let us introduce $v = u_{|E_x}$. The vector x is a cyclic vector for v. Consequently, v is a cyclic endomorphism on E_x and from the previous proposition, $\mu_v = (-1)^p \chi_v$.

We also know that as E_x is stable by u, χ_v divide χ_u. We deduce that

$$\chi_u(u)(x) = \chi_v(u) \circ P(u)(x) = P(u)(\chi_v(u)(x)) = 0$$

We thus have proved that for any $x \neq 0$, $\chi_u(u)(x) = 0$, and thuspar $\chi_u(u) = 0$.

1.2.4 Link between minimal and characteristic polynomial

Proposition 1.23

Let $u \in \mathcal{L}(E)$, then

1. For all $P \in \mathbb{K}[X]$,
$$P(u) \text{ invertible } \Leftrightarrow P \wedge \mu_u = 1 \Leftrightarrow P \wedge \chi_u = 1.$$
2. The polynomial μ_u and χ_u share the same irreductible factors.

Proof.

1. Assume that $P \wedge \mu_u = 1$, from Bezout theorem, there exists two polynomials $A, B \in \mathbb{K}[X]$ such that $AP + B\mu_u = 1$. In particular, $P(u)A(u) = id$ and thus $P(u)$ is invertible.

   Conversely, if $P(u)$ is invertible, we know that $P(u)$ is a polynomial in u denoted by $Q(u)$. We deduce that $(PQ-1)(u) = 0$ and thus $(PQ-1)$ is an annihilating polynomial of u. We deduce that μ_u divide $PQ-1$ and thus that $PQ + R\mu_u = 1$, or equivalently $P \wedge \mu_u = 1$.

   Factoring P in $\widehat{\mathbb{K}}$ leads to
$$P(X) = \prod_{i=1}^{n}(X - \xi_i), \quad \xi_i \text{ not necessarily distincts.}$$

   We then have $P(u) = \prod_{i=1}^{n}(u - \xi_i id)$ and
$$\det(P(u)) = \prod_{i=1}^{n}\det(u - \xi_i\, id) = \prod_{i=1}^{n}\chi_u(\xi_i).$$

   We then see that
$$\begin{array}{lcl} P(u) \text{ is invertible} & \Leftrightarrow & \det(P(u)) \neq 0 \\ & \Leftrightarrow & \chi_u(\xi_i) \neq 0 \text{ for all } i = 1, \cdots, n \\ & \Leftrightarrow & \text{none of the irreductible factor (on } \widehat{\mathbb{K}}\text{) of } P \text{ is a factor of } \chi_u \\ & \Leftrightarrow & P \wedge \chi_u = 1 \end{array}$$

2. Now, let Δ be an irreductible factor of χ_u, we have $\chi_u \wedge \Delta = \Delta \neq 1$, consequently $\Delta(u)$ is not invertible so that $\Delta \wedge \mu_u \neq 1$. As Δ is irreductible Δ divide μ_u. The proof is the same for the irreductible factors of μ_u.

□

Corollary 1.24

The minimal and characteristic polynomila can be decomposed in the form:
$$\begin{array}{rcl} \mu_u & = & \Delta_1^{m_1} \cdots \Delta_s^{m_s} \\ \chi_u & = & (-1)^n \Delta_1^{c_1} \cdots \Delta_s^{c_s} \end{array}$$

the the Δ_i are irreductible on $\mathbb{K}$ and $m_i \leq c_i$. In particular, if one of the polynomial μ_u and χ_u is the product of linear factors
$$\begin{array}{rcl} \mu_u & = & \prod_{i=1}^{s}(X - \lambda_i)^{m_i} \\ \chi_u & = & (-1)^n \prod_{i=1}^{s}(X - \lambda_i)^{c_i} \end{array}$$

1.2.5 Exercises

Exercise 1.4

Determine the characteristic and the minimal plynomial of the following matrices:

1. $A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}$.

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

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

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

---

1. $A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}$. On a alors $P(A) = \begin{pmatrix} P(2) & 0 & 0 \\ 0 & P(3) & 0 \\ 0 & 0 & P(4) \end{pmatrix}$. Par conséquent, $P \in \mathrm{Ker}(\phi_u)$ si et seulement si $P(X) = (X-2)(X-3)(X-4)Q(X)$. On en déduit que
$$\mu_A(X) = (X-2)(X-3)(X-4) = \chi'_A X).$$

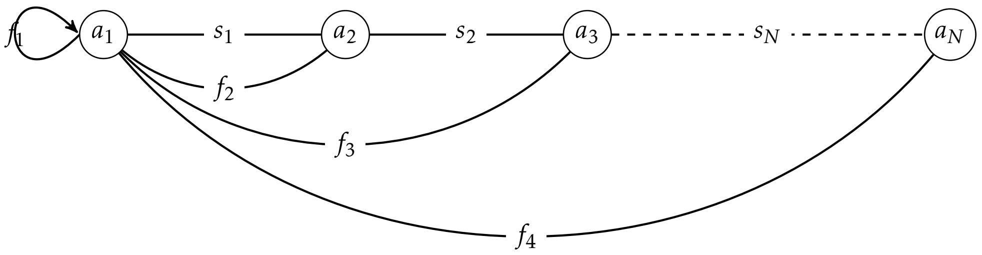
2. Soit $A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}$. On a alors $P(A) = \begin{pmatrix} P(2) & 0 & 0 \\ 0 & P(2) & 0 \\ 0 & 0 & P(4) \end{pmatrix}$. Par conséquent, $P \in \mathrm{Ker}(\phi_u)$ si et seulement si $P(X) = (X-2)(X-4)Q(X)$. On en déduit que
$$\mu_A(X) = (X-2)(X-4), \chi_A(X) = (X-2)^2(X-4).$$

3. Soit $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. On a $A - I = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq 0$, par contre $(A-I)^2 = 0$. Par conséquent,
$$\mu_A(X) = (X-1)^2 = \chi_A(X).$$

4.
$$\chi_A(X) = (X+1)^4(X-2)^9(X-3)^5, \mu_A(X) = (X+1)(X-2)^4(X-3)^5$$

Exercise 1.5: Leslie matrices

Consider a population structured in age. Let x_i^n be the number of females of age a_i in the population at a time $t_n = ndt$, where $i \in \{1, \cdots, N\}$, with N the number of classes of age and $n \in \{1, \cdots, N_t\}$.



We assume that the duration of the time step dt is equal to the duration of a class of age da, for example a year. The model is written in the following way taking into account the aging of individuals and reproduction:

$$x_1^{n+1} = \sum_{i=2}^{N} f_i x_i^n, \quad x_i^{n+1} = s_i x_{i-1}^n, \text{ for } i = 2, \cdots, N,$$

where $f_i \geq 0$ denotes the birth rate of the population of the age-class i and $s_i > 0$ denotes the growth rate which is the proportion of individuals leaving age-class i and entering the age-class $i+1$ at each time step. It is assumed that if an individual in age-class i does not successfully in age-class $i+1$ it dies. We obtain the linear recurrence:

$$X^{n+1} = LX^n \text{ where } X^n = (x_i^n)_{i=1,\cdots,N} \text{ and } L = \begin{pmatrix} f_1 & f_2 & \cdots & \cdots & f_n \\ s_1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & s_{n-1} & 0 \end{pmatrix}.$$

The matrix L is classicaly called a **Leslie matrix**.

1. Prove that the characteristical polynomial associated to L is given by

$$\det(\lambda Id - L) = \lambda^N - f_1 \lambda^{N-1} - \sum_{i=2}^{N} \left(f_i \prod_{j=1}^{i-1} s_i \right) \lambda^{n-i}$$

   and deduce that L admits a unique positive eigenvalue that is necessarily $\rho(L)$. (You can use the fact that all the coefficients of that polynomial except the coefficient of λ^N are non positive and study its variation. This result is a particular case of the famous **Descartes rule**).

   Note that the eigenvalue $\rho(L)$ has an important biological significance: it represents usually the asymptotic growth rate of the population. The components of the eigenvector v also have an important biological significance: they represent the proportions of individuals in the different age groups in the asymptotic regime.

2. Explicit the eigenvector v knowing $\rho(L)$.

1.3 Diagonalizibility

1.3.1 Kernel Lemma

Theorem 1.25: Kernel Theorem

Let $u \in \mathcal{L}(E)$ and $Q_1, \cdots, Q_s \in \mathbb{K}[X]$, pairwise coprime. Then:

1. The vector subspaces $E_i = \operatorname{Ker}(Q_i(u))$ are in direct sum.
2. $$\bigoplus_{i=1}^{s} \operatorname{Ker}(Q_i(u)) = \operatorname{Ker}\left(\prod_{i=1}^{s} Q_i(u)\right).$$
3. The projectors associated with this direct sum are all in $\mathbb{K}[u]$.

Proof.

- Let's show that $\bigoplus_{i=1}^{s} \operatorname{Ker}(Q_i(u)) = \operatorname{Ker}\left(\prod_{i=1}^{s} Q_i(u)\right)$.

  Let $P_k = \prod_{i \neq k} Q_i(u)$. The polynomials P_i are coprime in their entirety ($P_1 \wedge \cdots \wedge P_s = 1$). Indeed, if Δ divides the P_i's, then Δ divides the product $Q_2 \cdots Q_s$. According to Gauss' theorem, there exists some $i_0 > 1$ such that Δ divides Q_{i_0}. But Δ also divides P_{i_0}, and by Gauss' theorem again, Δ divides Q_i for $i \neq i_0$, which contradicts the fact that the Q_i's are pairwise coprime.

  By applying Bézout's theorem to the P_k's, we find that there exist A_i for $i = 1, \cdots, s$ such that
  $$A_1 P_1 + \cdots + A_s P_s = 1.$$
  Now let $x \in \operatorname{Ker}\left(\prod_{i=1}^{s} Q_i(u)\right)$. According to the preceding, we can write
  $$x = \underbrace{A_1(u) \circ P_1(u)(x)}_{x_1} + \cdots + \underbrace{A_s(u) \circ P_s(u)(x)}_{x_s}.$$
  We will now show that $x_k \in \operatorname{Ker}(Q_k(u))$. Indeed, we have
  $$Q_k(u) \circ A_k(u) \circ P_k(u)(x) = A_k(u) \circ \underbrace{Q_k(u) \circ P_k(u)}_{\prod_{i=1}^{s} Q_i(u)}(x) = 0.$$
  Hence, we have shown that $\operatorname{Ker}\left(\prod_{i=1}^{s} Q_i(u)\right) \subseteq \bigoplus_{i=1}^{s} \operatorname{Ker}(Q_i(u))$.

  Conversely, if $x \in \operatorname{Ker}(Q_i(u))$, we have
  $$Q_1(u) \cdots Q_s(u)(x) = P_i(u) \circ Q_i(u)(x) = 0.$$
  Therefore,
  $$\operatorname{Ker}\left(\prod_{i=1}^{s} Q_i(u)\right) = \bigoplus_{i=1}^{s} \operatorname{Ker}(Q_i(u)).$$

- Now, let's show that the sum is direct. Assume that
  $$w_1 + \cdots + w_s = 0 \text{ where } w_i \in E_i = \operatorname{Ker}(Q_i(u)).$$
  We apply $A_i P_i(u)$. For $j \neq i$, $w_j \in \operatorname{Ker}(P_i)$, and hence $A_i(u) \circ P_i(u)(w_j) = 0$. Thus, $A_i(u) \circ P_i(u)(w_i) = 0$.

  By summing over all j, we get
  $$\sum_{j=1}^{s} A_j(u) \circ P_j(u)(w_i) = 0 \quad \forall i = 1, \cdots, s.$$
  Since $\sum_{i=1}^{s} A_j P_j = 1$, it follows that $w_i = 0$ for all i.

  Therefore, we have shown points 1 and 2.

- The projection of $x \in \operatorname{Ker}\left(\prod_{i=1}^{s} Q_i(u)\right)$ onto $\operatorname{Ker}(Q_j(u))$ parallel to $\bigoplus_{i \neq j} \operatorname{Ker}(Q_i(u))$ is given by $A_j \circ P_j(u)(x)$. This projection is therefore an element of $\mathbb{K}[u]$.

□

1.3.2 Definition

Definition/Proposition 1.1: Diagonalizibility of an endomorphism

Let $u \in \mathcal{L}(E)$, the endomorphism u is diagonalizable if one of these equivalent properties is satisfied:

1. The space E can be decomposed as a direct sum of the eigenspaces of u:

$$E = \bigoplus_{\lambda \in \mathrm{Sp}(u)} E_\lambda \text{ where } E_\lambda = \mathrm{Ker}(u - \lambda\, id).$$

2. There exists a basis consisting of eigenvectors of u.

Proof.

- To verify that 1 implies 2, it is enough to gather the bases of the spaces E_λ.

- Conversely, we define $E'_\lambda = \mathrm{Vect}\langle e_i \,|\, e_i \in \mathcal{B} \text{ such that } u(e_i) = \lambda e_i \rangle$. We know that $E'_\lambda \subset E_\lambda$. The family E_λ is a direct sum (this is always true, by using the kernel lemma), and since $E = \bigoplus_{\lambda \in \mathrm{Sp}(u)} E'_\lambda$, we conclude that $E_\lambda = E'_\lambda$.

□

Definition 1.27: Diagonalizibility of a matrix

A matrix $A \in \mathcal{M}_n(\mathbb{K})$ is diagonalizable if it is similar to a diagonal matrix, or equivalently, if it represents a diagonalizable endomorphism in a certain basis.

1.3.3 Link with annihilating polynomials

Theorem 1.26: Necessary and Sufficient Conditions for Diagonalization

Let $u \in \mathcal{L}(E)$, then the following three propositions are equivalent:

1. u is diagonalizable.
2. The minimal polynomial μ_u is split with simple roots

$$\mu_u(X) = \prod_{i=1}^{s} (X - \lambda_i)^{m_i}, \quad m_i = 1.$$

3. There exists an annihilating polynomial of u that is split with simple roots.
4. The characteristic polynomial χ_u is split, and for each eigenvalue, $\dim(E_\lambda) = c_\lambda$

$$\chi_u(X) = \prod_{i=1}^{s} (X - \lambda_i)^{c_i}, \quad c_i = \dim(E_{\lambda_i}).$$

Proof. Assume 1. There exists a basis in which u can be written as

$$\begin{array}{c}
\begin{array}{cccccccccc} & \overset{\dim(E_{\lambda_1})}{\longleftrightarrow} & & & \overset{\dim(E_{\lambda_2})}{\longleftrightarrow} & & & & \overset{\dim(E_{\lambda_s})}{\longleftrightarrow} & \end{array} \\
\begin{pmatrix}
\lambda_1 & & & & & & & & & \\
& \ddots & & & & & & & & \\
& & \lambda_1 & & & & & & & \\
& & & \lambda_2 & & & & & & \\
& & & & \ddots & & & & & \\
& & & & & \lambda_2 & & & & \\
& & & & & & \ddots & & & \\
& & & & & & & \lambda_s & & \\
& & & & & & & & \ddots & \\
& & & & & & & & & \lambda_s
\end{pmatrix}
\end{array}$$

Therefore,

$$\chi_u(X) = \begin{vmatrix}
\lambda_1 - X & & & & & & & & & \\
& \ddots & & & & & & & & \\
& & \lambda_1 - X & & & & & & & \\
& & & \lambda_2 - X & & & & & & \\
& & & & \ddots & & & & & \\
& & & & & \lambda_2 - X & & & & \\
& & & & & & \ddots & & & \\
& & & & & & & \lambda_s - X & & \\
& & & & & & & & \ddots & \\
& & & & & & & & & \lambda_s - X
\end{vmatrix} = \prod_{i=1}^{s} (\lambda_i - X)^{\dim(E_{\lambda_i})}.$$

Similarly, for any polynomial $P \in \mathbb{K}[X]$, we have

$$P(X) = \begin{pmatrix}
P(\lambda_1) & & & & & & & & & \\
& \ddots & & & & & & & & \\
& & P(\lambda_1) & & & & & & & \\
& & & P(\lambda_2) & & & & & & \\
& & & & \ddots & & & & & \\
& & & & & P(\lambda_2) & & & & \\
& & & & & & \ddots & & & \\
& & & & & & & P(\lambda_s) & & \\
& & & & & & & & \ddots & \\
& & & & & & & & & P(\lambda_s)
\end{pmatrix}$$

Thus, $P(u) = 0$ if and only if $P(\lambda_i) = 0$ for all $i = 1, \ldots, s$. We deduce that

$$\mu_u(X) = \prod_{i=1}^{s} (X - \lambda_i).$$

We deduce that Point 1 induces Points 2 and 4.

Point 2 clearly implies point 3.

Assume Point 3 there exists $P(X) = \prod_{i=1}^{k}(X - \lambda_i)$ such that $P(u) = 0$. Applying the kernel lemma to the polynomial P, we have:

$$E = \bigoplus_{i=1}^{k} \underbrace{\mathrm{Ker}(u - \lambda_i \, id)}_{E_{\lambda_i}}.$$

The non-zero subspaces E_{λ_i} are the eigenspaces of u, and we can therefore conclude.

Assume now Point 4, we just need to remark that the assumption $c_i = \dim E_{\lambda_i}$ induces that $n = \dim E = \sum_i c_i = \dim \bigoplus_{i=1}^{k} \underbrace{\mathrm{Ker}(u - \lambda_i \, id)}_{E_{\lambda_i}}$ and thus $E = \bigoplus_{i=1}^{k} \underbrace{\mathrm{Ker}(u - \lambda_i \, id)}_{E_{\lambda_i}}$. Th erefore u is diagonalizable. □

Definition 1.28: Eigenspace and characteristic subspace

The space $\mathrm{Ker}(u-\lambda_i)$ is called the **eigenspace** associated to the eigenvalue λ_i and the space $\mathrm{Ker}(u-\lambda_i)^{c_i}$ is called the **characteristic space**.

Corollary 1.27

The restriction of an endomorphism to a vector subspace stable under that endomorphism is diagonalizable.

Proof. Let F be a vector subspace of E. Since F is stable under u, we can use the fact that $\mu_{u_{|F}}$ divides μ_u and is therefore split with simple roots. □

1.3.4 Simultaneous Diagonalization

Theorem 1.28: Simultaneous Diagonalization

Let $(u_i)_{i\in I}$ be a family of diagonalizable endomorphisms in $\mathcal{L}(E)$ (not necessarily finite). Then the following propositions are equivalent:

1. There exists a basis $\mathcal{B}$ of E in which the endomorphisms u_i are diagonalizable.
2. The endomorphisms u_i commute pairwise.

Proof. The implication 1 implies 2 is immediate: we take the basis that diagonalizes the endomorphisms u_i. The associated matrices are all diagonal, so they obviously commute.

For the converse, we proceed by induction on $n = \dim(E)$. If $n = 1$, the result is evident. If $n > 1$, and if all the endomorphisms u_i are homotheties, their matrices in any basis are diagonal, so they commute.

Now, suppose that one of the u_{i_0} is not a homothety. Let E_{λ_i}, $i = 1,\dots,s$, be the eigenspaces of u_{i_0}. We have $E = \bigoplus_{i=1}^{s} E_{\lambda_i}$, and since $u_j \circ u_{i_0} = u_{i_0} \circ u_j$, we have $u_j(E_{\lambda_i}) \subset E_{\lambda_i}$. The restrictions $(u_j)_{|E_{\lambda_i}}$ are diagonalizable and commute. Since u_{i_0} is not a homothety, $\dim(E_{\lambda_i}) < n$, so we can apply the induction hypothesis. There exists a basis $\mathcal{B}_i$ of E_{λ_i} that diagonalizes all the $(u_j)_{|E_{\lambda_i}}$. Let $\mathcal{B} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_s$; this is a basis that diagonalizes all the u_i. □

1.3.5 Exercises

Exercise 1.6

Let $A = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{pmatrix}$. Determine its minimal polynomial and discuss its diagonalizability.

$$A^2 = \begin{pmatrix} 7 & 12 & 6 \\ 6 & 13 & 6 \\ 6 & 12 & 7 \end{pmatrix} = Id + 6\underbrace{\begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix}}_{=A-Id} = 6A - 5Id$$

We deduce that

$$\mu_A(X) = X^2 - 6X + 5 = (X-1)(X-5)$$

We deduce that A is diagonalizable.

Exercise 1.7

Discuss the diagonalizability of

1. a projection

2. a transvection
3. a permutation
4. a cyclic matrix

Exercise 1.8

Compute the eigenvalues of the matrices

$$A_{dir} = (n+1)^2 \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}, \quad A_{per} = (n+1)^2 \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & -1 \\ -1 & \ddots & \ddots & & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & & \ddots & \ddots & -1 \\ -1 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

and deduce that the two matrix are diagonalizable.

1.4 Triangularizability

1.4.1 Definition

Definition 1.29

An endomorphism $u \in \mathcal{L}(E)$ is said to be **triangularizable** if there exists a basis $\mathcal{B}$ in which the matrix of u is upper triangular.

Remark.

The diagonal coefficients of $\mathrm{Mat}_{\mathcal{B}}(u)$ are then the eigenvalues of u. "If we want a result involving only the eigenvalues of u, it suffices to triangularize u."

1.4.2 Necessary and/or Sufficient Conditions for Triangularization

Theorem 1.29: Necessary and Sufficient Conditions for Triangularization

An endomorphism $u \in \mathcal{L}(E)$ is triangularizable if and only if the characteristic polynomial χ_u is split.

Proof. The implication 1 implies 2 is evident: we compute the characteristic polynomial in the basis $\mathcal{B}$ that triangularizes u.

For the converse, we proceed by induction on $n = \dim(E)$. The result is evident for $n = 1$. Assume the result is true for $n-1$. We use a **duality argument**. For n, we consider u^t. Since $\chi_u = \chi_{u^t}$ is split, u^t has an eigenvector $\phi \in E^*$ associated with the eigenvalue λ: $u^t \circ \phi = \lambda\phi = \phi \circ u$. It follows that $\mathrm{Ker}(\phi)$ is stable under u. Moreover, $\mathrm{Ker}(\phi)$ is a hyperplane of dimension $n-1$. We apply the induction hypothesis to $v = u_{|\mathrm{Ker}(\phi)}$. Since χ_v divides χ_u and χ_u is split, we deduce that χ_v is split, and thus there exists a basis of $\mathrm{Ker}(\phi)$ in which $\mathrm{Mat}_{\mathcal{B}_\phi}(v)$ is upper triangular. We extend $\mathcal{B}_\phi$ to a basis $\mathcal{B}$, and in this basis $\mathrm{Mat}_{\mathcal{B}}(v)$ is also upper triangular. □

Remark.

Geometrically, an eigenvector for u^t corresponds to a stable hyperplane for u. The converse is also true: if there is a stable hyperplane F for u, there exists a linear form whose kernel is equal to F.

Corollary 1.30

Any endomorphism $u \in \mathcal{L}(E)$ for E a vector space on the field $\mathbb{K} = \mathbb{C}$ is triangularizable.

1.4.3 Simultaneous Triangularization

Theorem 1.31: Simultaneous Triangularization

Let $(u_i)_{i \in I}$ be a family of triangularizable endomorphisms (not necessarily finite) that commute pairwise. Then the endomorphisms $(u_i)_{i \in I}$ are simultaneously triangularizable.

Remark.

Note: two triangular matrices do not always commute:

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}.$$

Proof. We again proceed by induction on $n = \dim(E)$. The result is true for $n = 1$. Assume the result holds for dimension $n-1$. We distinguish between two cases:

- **First case:** All the endomorphisms $(u_i)_{i \in I}$ are homotheties: $u_i = \lambda_i\, id$. Any basis is suitable.
- **Second case:** Fix u_{i_0}, which is not a homothety. There exists F, a subspace of dimension strictly less than n, which is an eigenspace of u_{i_0}. The subspace F is stable under all the endomorphisms u_i since it is an eigenspace. Let G be a complement of F: $E = F \oplus G$. The linear maps $u_i|_F$ are triangularizable by induction.

□

1.5 Hilbertian analysis

1.5.1 Inner-Product Spaces

Definition 1.30

Let V be a linear space on a field $\mathbb{K}$ that is either $\mathbb{C}$ or $\mathbb{R}$. An inner-product on V is a function

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{K}$$

which satisfies the following:

1. $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Rightarrow x = 0 \quad \forall x \in V$
2. $\langle y, x \rangle = \overline{\langle x, y \rangle} \quad \forall x, y \in V$
3. $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle \quad \forall x, y, z \in V$
4. $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle \quad \forall x, y \in V, \forall \lambda \in \mathbb{K}$

Example 1.8

On $\mathbb{R}^n$, $\langle x, y \rangle = y^t x = \sum_{i=1}^n x_i y_i$ is an inner product.

Example 1.9

On $\mathbb{C}^n$, $\langle x, y \rangle = \bar{y}^t x = \sum_{i=1}^n x_i \bar{y}_i$ is an inner product.

Remark.

1. If $\mathbb{K} = \mathbb{R}$, an inner product is also what we call a *symmetric positive definite bilinear form*.
2. If $\mathbb{K} = \mathbb{C}$, an inner product is also what we call a *hermitian positive definite bilinear form*.

Definition 1.31: Euclidian versus Hermitian space

A linear space equipped with an inner-product operation is referred to as an **inner-product space**. In finite dimension,

1. if $\mathbb{K} = \mathbb{R}$, it is an **euclidean space**.
2. if $\mathbb{K} = \mathbb{C}$, it is an **hermitian space**.

Proposition 1.32: Inner product and norm

Let $\langle \cdot, \cdot \rangle$ an inner product on V, then

$$x \mapsto \sqrt{(x,x)}$$

defines a norm on V.

Lemma 1.33: The Cauchy-Schwarz Inequality

Let V be an inner-product space. Then

$$|(x,y)| \leq ||x||||y||$$

Lemma 1.34: Parallelogram law

Let V be an inner-product space and $||.||$ the associated norm. For any $x, y \in V$ we have

$$||x+y||^2 + ||x-y||^2 = 2||x||^2 + 2||y||^2$$

Lemma 1.35: Polarisation identity

Let V be an inner-product space and $||.||$ the associated norm. For any $x, y \in V$ we have

$$||x+y||^2 - ||x-y||^2 = 4\mathcal{R}e(x,y)$$

Definition 1.32: Hilbert space

A Hilbert space is a complete inner-product space; that is, an inner product space in which every Cauchy sequence is convergent.

1.5.2 Orthogonal and Orthonormal Families

Let V be an inner-product space.

Definition 1.33

Two vectors $x, y \in V$ are said to be **orthogonal** if $\langle x, y \rangle = 0$. We denote this as $x \perp y$.

Definition 1.34

A family $(x_i)_{i \in I}$ is said to be orthogonal if the vectors composing it are pairwise orthogonal.

An orthogonal family made up of non-zero vectors is a linearly independent family.

Proposition 1.36

1. If $(x_1,\ldots,x_p)$ is an orthogonal family, the following Pythagorean relation holds:
$$\left\|\sum_{k=1}^{p} x_k\right\|^2 = \sum_{k=1}^{p} \|x_k\|^2.$$
2. If X is a subset of V, the orthogonal complement of X, denoted by $X^{\perp}$, is defined as:
$$X^{\perp} = \{y \in E\,;\, \forall x \in X, \langle x,y\rangle = 0\}.$$
The orthogonal complement $X^{\perp}$ is always a subspace of V and $V = V \oplus V^{\perp}$.

Definition 1.35

1. A vector $x \in V$ is said to be unitary or normalized if $\|x\| = 1$.
2. A family (x_i) is orthonormal if it is an orthogonal family and all its vectors are normalized.

Theorem 1.37: Gram-Schmidt orthonormalization process

If $(x_1,\ldots,x_p)$ is a linearly independent family in V, there exists a unique orthonormal family $(e_1,\ldots,e_p)$ satisfying the following two conditions:

- For all $k \in \{1,\ldots,p\}$, $\text{vect}(e_1,\ldots,e_k) = \text{vect}(x_1,\ldots,x_k)$;
- For all $k \in \{1,\ldots,p\}$, $\langle e_k, x_k\rangle > 0$.

Definition 1.36: Orthonormal basis

We assume in this section that V is an n-dimensional Euclidean space.
An orthonormal basis of V is an orthogonal family $(e_1,\ldots,e_n)$ where all vectors are unit vectors. Specifically, it is a basis of V.
V has orthonormal bases. If $(e_1,\ldots,e_n)$ is an orthonormal basis of V, then every $x \in V$ can be uniquely written as:
$$x = \sum_{i=1}^{n} \langle x, e_i\rangle e_i$$
The real number $\langle x, e_i\rangle$ is called the coordinate of x with respect to e_i in the basis $(e_1,\ldots,e_n)$.

If $(e_1,\ldots,e_n)$ is an orthonormal basis of V, and if $x = \sum_{i=1}^{n} x_i e_i$ and $y = \sum_{i=1}^{n} y_i e_i$, then the scalar product and the norm can be calculated using the following formulas:

$$\langle x,y\rangle = \sum_{i=1}^{n} x_i y_i$$

Proposition 1.38: Matrix of an orthogonal/orthonormal familly - case $\mathbb{K} = \mathbb{R}$

Consider $\mathcal{B} = (e_1,\cdots,e_n)$ the canonic basis of $\mathbb{R}^n$

1. If $\mathcal{B}' = (\varepsilon_1,\cdots,\epsilon_n)$ is an orthogonal basis of $\mathbb{R}^n$, then the matrix P of $\mathcal{B}'$ in $\mathcal{B}$ is normal *i.e.* $P^tP = PP^t$.
2. If $\mathcal{B}' = (\varepsilon_1,\cdots,\epsilon_n)$ is an orthonormal basis of $\mathbb{R}^n$, then the matrix P of $\mathcal{B}'$ in $\mathcal{B}$ is orthogonal *i.e.* $P^tP = PP^t = Id$.

where if $P = (P_{ij})_{ij}$, $P^* = P^t = (P_{ji})_{ij}$.

Proposition 1.39: Matrix of an orthogonal/orthonormal familly - case $\mathbb{K} = \mathbb{C}$

Consider $\mathcal{B} = (e_1, \cdots, e_n)$ the canonic basis of $\mathbb{C}^n$.

1. If $\mathcal{B}' = (\varepsilon_1, \cdots, \epsilon_n)$ is an orthogonal basis of $\mathbb{C}^n$, then the matrix P of $\mathcal{B}'$ in $\mathcal{B}$ is normal *i.e.* $P^*P = PP^*$.
2. If $\mathcal{B}' = (\varepsilon_1, \cdots, \epsilon_n)$ is an orthonormal basis of $\mathbb{C}^n$, then the matrix P of $\mathcal{B}'$ in $\mathcal{B}$ is an unitary matrix *i.e.* $P^*P = PP^* = Id$.

where if $P = (P_{ij})_{ij}$, $P^* = (\overline{P_{ji}})_{ij}$.

Proposition 1.40: Orthogonal/Unitary group

The set of orthogonal matrices and the set of unitary matrices form two groups :

$$\mathcal{O}(n) = \{P \in \mathcal{M}_n(\mathbb{R}) \text{ such that } PP^t = P^tP = Id\},\ U(n)\{P \in \mathcal{M}_n(\mathbb{C}) \text{ such that } PP^* = P^*P = Id\}$$

1.5.3 Adjoint of an endomorphism

Proposition 1.41: Adjoint

If $u \in \mathcal{L}(V)$, where V is a Euclidean or an Hermitian space, there exists an unique endomorphism u^* called the adjoint of u, defined by:

$$\forall x, y \in V, \quad \langle u(x), y\rangle = \langle x, u^*(y)\rangle$$

Proposition 1.42: Matrix of the adjoint

Let A be the matrix of an endomorphism u in a basis $\mathcal{B}$.

1. If $\mathbb{K} = \mathbb{R}$, the matrix of the adjoint is the transpose of A *i.e.* $A^* = A^t$.
2. If $\mathbb{K} = \mathbb{C}$, the matrix of the adjoint is the conjugate transpose matrix of A *i.e.* $A^* = \overline{A}^t$.

Proposition 1.43

If $\in \mathcal{L}(V)$,

$$(\mathbb{Im}(u))^\perp = \text{Ker}(u^*), (\text{Ker}(u)^\perp = \mathbb{Im}(u^*)$$

1.5.4 Symmetric endomorphism on an Euclidean space/ hermitian endomorphism on a Hermitian space

Definition 1.37

Let V be an Euclidean space (resp. Hermitian) equipped with a scalar product $\langle\cdot,\cdot\rangle$, and let $u \in \mathcal{L}(V)$.
We say that u is a symmetric (resp. Hermitian) endomorphism if:

$$\forall x, y \in V, \quad \langle u(x), y\rangle = \langle x, u(y)\rangle$$

or equivalently $u^* = u$.

Proposition 1.44

Let $u \in \mathcal{L}(V)$ where $(V, \langle \cdot, \cdot \rangle)$ is a Euclidean (resp. Hermitian) space. The following three statements are equivalent:

i) The endomorphism u is a symmetric (resp. hermitian) endomorphism of V.

ii) There exists an orthonormal basis $\mathcal{B}$ of V such that the matrix $\mathrm{Mat}_{\mathcal{B}}(u)$ is a symmetric (resp. hermitian) matrix.

iii) For all orthonormal bases $\mathcal{B}$ of V, the matrix $\mathrm{Mat}_{\mathcal{B}}(u)$ is a symmetric (resp. hermitian) matrix.

Lemma 1.45

Let u be a symmetric (resp. hermitian) endomorphism of a Euclidean (resp. hermitian) space $(V, \langle \cdot, \cdot \rangle)$, and let F be a subspace of V. If F is stable under u, then $F^{\perp}$ is also stable under u.

Theorem 1.46: Reduction

Let u be an endomorphism of a Euclidean (resp. Hermitian) space $(E, \langle \cdot, \cdot \rangle)$. If u is a symmetric (resp. Hermitian) endomorphism, then:

- u is diagonalizable.
- There exists an orthonormal basis of V consisting of eigenvectors of u.

Theorem 1.47: Reduction of a symmetric (resp. hermitian) matrix

If A is a real symmetric (resp. complex hermitian) matrix, then:

- The matrix A is diagonalizable.
- There exists an orthogonal matrix P and a real diagonal matrix D such that:

$$D = P^{-1}AP = P^{*}AP$$

Definition 1.38: Symmetric matrix

Let A be a symmetric real matrix. We say that

- A is a semi-definite positive matrix if $(Ax, x) \geq 0$ for all $x \in \mathbb{R}^n$.
- A is a definite positive matrix if $(Ax, x) > 0$ for all $x \neq 0$.

Proposition 1.48: Caracterization of a (semi-)positive definite matrix

Let A be a symmetric real matrix. The following properties are equivalent

1. A is definite positive (resp. semi-definite positive)
2. All the eigenvalues of A are positive (resp. non-negative)
3. All the leading principal minors of A are positive (resp. non negative)

1.5.5 Reduction of a normal endomorphism

Definition 1.39: Normal endomorphism

We say that $u \in \mathcal{L}(V)$ is *normal* if u^* and u commute, i.e.,

$$u^*u = uu^*$$

Lemma 1.49

Let $u \in \mathcal{L}(V)$ be a normal operator and let F be a subspace of V that is stable under u. Then:

- F is stable under u^*.
- $F^\perp$ is stable under both u and u^*.

Moreover, if u is normal, then for any subspace F stable under u, the restriction $u|_F$ is also normal.

Theorem 1.50

Let V be a Euclidean space and $u \in \mathcal{L}(V)$ a normal operator. Then there exists an orthonormal basis $\mathcal{B}$ of V such that:

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} \lambda_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \lambda_r & \ddots & & \vdots \\ \vdots & & \ddots & \tau_1 & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \tau_s \end{pmatrix}$$

where for all i, $\lambda_i \in \mathbb{R}$ and for all j,

$$\tau_j = \begin{pmatrix} a_j & -b_j \\ b_j & a_j \end{pmatrix} \in \mathcal{M}_2(\mathbb{R}).$$

Theorem 1.51

Let V be a Hermitian space and $u \in \mathcal{L}(V)$ a normal operator. Then there exists an orthonormal basis $\mathcal{B}$ of V such that:

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_r \end{pmatrix}$$

where for all i, $\lambda_i \in \mathbb{C}$. In other words, u can be diagonalized in an orthonormal basis.

Remark.

A normal operator on an Euclidean space can be seen as a normal operator on a hermitian space and thus can be diagonalized on $\mathbb{C}$ in a orthonormal basis of $\mathbb{C}$.

1.5.6 Hilbert space and linear forms

Theorem 1.52: Riesz theorem

Let H be a Hilbert space and $L \in \mathcal{L}(H, \mathbb{R})$ a continuous linear form on H. Then there exists a unique $a \in H$ such that

$$L(x) = \langle a, x \rangle, \text{ for all } x \in H.$$

Note that this result is true in both finite and infinite dimension. The assumption L continuous is automatically satisfied in finite dimension.

A classical application of the Riesz theorem, that is frequently used to solve elliptic Partial Differential Problem is the following result.

Theorem 1.53: Lax-Milgram theorem

Let H be a Hilbert space, $a : H \times H \to \mathbb{R}$ a bilinear form, and $L : H \to \mathbb{R}$ a linear form. We assume that:

- The bilinear form a is continuous:

$$\exists C > 0 \text{ such that } |a(x,y)| \leq C\|x\|\|y\|, \forall x, y \in H$$

- The bilinear form a is coercive:

$$\exists \alpha > 0 \text{ such that } a(x,x) \geq \alpha\|x\|^2, \forall x \in H$$

- The linear form L is continuous:

$$\exists C > 0 \text{ such that } |L(x)| \leq C\|x\|, \forall x \in H$$

Then, there exists a unique solution $u \in H$ to the following problem:

$$a(u,v) = L(v), \quad \forall v \in H.$$

Proof. In the case where a is symmetric, the proof of this theorem is an immediat application of the Riesz theorem remarking that H endowed with the inner product a is still a Hilbert space and that L is also continuous for this inner product.

In the general case, we apply the Riesz theorem to the linear maps

$$\begin{cases} y \mapsto a(x,y) \text{ for } x \in H \text{ fixed} \\ y \mapsto L(y) \end{cases}$$

and deduce the existence for all $x \in H$, $A_x \in H$ such that

$$a(x,y) = \langle A_x, y\rangle$$

and $b \in H$ such that

$$L(y) = \langle b, y\rangle$$

It is then easy to see that the map $A : x \mapsto A_x$ is linear and thus that the problem is reduced to the resolution of the problem: finding $x \in H$ such that $Ax = b$. The coercivity of the bilinear form a induces the injectivity of A. Indeed, if $Ax = 0$ then

$$0 = \langle Ax, x\rangle = a(x,x) \geq \alpha\|x\|^2 \implies x = 0.$$

In finite dimension, that ends the proof. In infinite dimension, we still need to prove that $\mathrm{Im}(A)$ is closed. For that, we prove that $\mathrm{Im}(A)^\perp = \{0\}$ using again the coercivity property of a. □

1.5.7 Exercises

Exercise 1.9

Let φ be defined on $M_3(\mathbb{R}) \times M_3(\mathbb{R})$ by $\varphi(A,B) = \mathrm{Tr}(A^t B)$.

1. Show that φ is an inner product.
2. Show that $\psi : M_3(\mathbb{R}) \to M_3(\mathbb{R})$, defined by $\psi(M) = M^\top$, is an orthogonal and symmetric endomorphism.
3. Find the eigenvalues and eigenvectors of ψ.

---

1. This is very standard. The form is well-defined because the trace takes real values. It is linear on the left due to the linearity of the trace. It is symmetric because $\mathrm{Tr}(A^\top B) = \mathrm{Tr}((A^\top B)^\top) = \mathrm{Tr}(B^\top A)$. For positive definiteness, we calculate:

$$(A^\top A)_{i,j} = \sum_{k=1}^{3} a_{k,i} a_{k,j}$$

Hence,

$$\mathrm{Tr}(A^\top A) = \sum_{i=1}^{3}\sum_{k=1}^{3} a_{i,k}^2$$

which is a sum of squared real numbers, thus non-negative. If it equals zero, all the coefficients of A are zero, implying that A is the zero matrix.

2. We have:

$$\varphi(\psi(A),\psi(B)) = \varphi(A^\top, B^\top) = \mathrm{Tr}(AB^\top) = \mathrm{Tr}((AB^\top)^\top) = \mathrm{Tr}(B^\top A) = \varphi(A,B)$$

so ψ is orthogonal. Additionally:

$$\varphi(\psi(A),B) = \varphi(A^\top, B) = \mathrm{Tr}(AB) = \mathrm{Tr}(BA) = \mathrm{Tr}((BA)^\top) = \mathrm{Tr}(A^\top B^\top) = \varphi(A,\psi(B))$$

thus ψ is symmetric.

3. Since $\psi \circ \psi = \mathrm{Id}$, ψ is an involution. Therefore, its eigenvalues must be among $\{-1,1\}$. The eigenspace corresponding to the eigenvalue 1, denoted E_1, consists of symmetric matrices (of dimension 6), and the eigenspace corresponding to the eigenvalue -1, denoted E_{-1}, consists of antisymmetric matrices (of dimension 3).

Exercise 1.10

Let $A \in \mathcal{M}_n(\mathbb{R})$ a symmetric matrix with eigenvalues $\lambda_1,\dots,\lambda_n$ such that $\lambda_1 \leq \cdots \leq \lambda_n$. We endow $\mathbb{R}^n$ with the Euclidean scalar product and its associated norm.

1. Show that for all $X \in \mathbb{R}^n$,
$$\lambda_1 \|X\|^2 \leq X^\top AX \leq \lambda_n \|X\|^2.$$

2. Deduce that
$$\lambda_1 = \min_{X\in\mathbb{R}^n\setminus\{0\}} \frac{X^\top AX}{\|X\|^2}$$
and
$$\lambda_n = \max_{X\in\mathbb{R}^n\setminus\{0\}} \frac{X^\top AX}{\|X\|^2}.$$

Exercise 1.11

Consider $C \in \mathcal{M}_{mn}(\mathbb{R})$ and A the matrices defined by $A = CC^t$ and $B = C^tC$. Prove that the matrices A and B are semi-definite positive. At what condition are they definite positive?

Exercise 1.12

Prove that the matrices

$$A_{dir} = (n+1)^2 \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}, \quad A_{per} = (n+1)^2 \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & -1 \\ -1 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & -1 \\ -1 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

are respectively definite positive and semi-definite positive.

Exercise 1.13: Orthogonal projection

Let F be a subspace of V an euclidean space. Let p be the orthogonal projection on F.

1. Give the matrix of p in an adapted basis of V.

2. Prove that $p(x) = \min_{y\in F} \|x-y\|$.

3. Evaluate $\langle p(x)-x, p(x)-y\rangle$ for all $y \in F$.

1. $V = F \oplus F^{\perp}$. In an orthonormal basis $\mathcal{B} = (\varepsilon_1, \cdots, \varepsilon_p, \varepsilon_{p+1}, \cdots, \varepsilon_n)$ adapted to this decomposition, $M_{\mathcal{B}}(p) = \begin{pmatrix} Id & 0 \\ 0 & 0 \end{pmatrix}$. The family $\varepsilon_1, \cdots, \varepsilon_p$ is a basis of F and $\varepsilon_{p+1}, \cdots, \varepsilon_n$ a basis of $F^{\perp}$.
2. Let $x \in V$ and $y \in F$, we have
$$x = \sum_{i=1}^{n} x_i \varepsilon_i,\ y = \sum_{i=1}^{p} y_i \varepsilon_i \text{ and } p(x) = \sum_{i=1}^{p} x_i \varepsilon_i$$
Therefore,
$$\|x-y\|^2 = \sum_{i=1}^{p}(x_i - y_i)^2 + \sum_{i=p+1}^{n} x_i^2 \geq \sum_{i=p+1}^{n} x_i^2 = \|x - p(x)\|^2.$$
3. We have
$$p(x) - x = \sum_{i=p+1}^{n} x_i \varepsilon_i,\ p(x) - y = \sum_{i=1}^{p}(x_i - y_i)\varepsilon_i$$
Thefore,
$$\langle p(x) - x, p(x) - y \rangle = 0$$

Exercise 1.14

Let H be a Hilbert space, $a : H \times H \to \mathbb{R}$ a bilinear symmetric form, and $L : H \to \mathbb{R}$ a linear form. We assume that:

- ▶ The bilinear form a is continuous:
$$\exists C > 0 \text{ such that } |a(x,y)| \leq C\|x\|\|y\|,\ \forall x, y \in H$$
- ▶ The bilinear form a is coercive:
$$\exists \alpha > 0 \text{ such that } a(x,x) \geq \alpha\|x\|^2,\ \forall x \in H$$
- ▶ The linear form L is continuous:
$$\exists C > 0 \text{ such that } |L(x)| \leq C\|x\|,\ \forall x \in H$$

Prove that the unique solution $x \in H$ of the problem $a(x,y) = L(y) \forall y \in H$ is the solution of the minimizing problem: finding $x \in H$ such that
$$J(x) = \min_{y \in H} J(y) \text{ where } J(y) = \frac{1}{2}a(y,y) - L(y).$$

Chapter 2 Classical matrix decompositions

2.1 Dunford Decomposition and Applications

2.1.1 Nilpotent Endomorphisms

Definition 2.1

We say that $u \in \mathcal{L}(E)$ is nilpotent if there exists $k > 0$ such that $u^k = 0$. We say that k is the index of nilpotency of u if, moreover, $u^{k-1} \neq 0$.

Proposition 2.1

An endomorphism $u \in \mathcal{L}(E)$ is nilpotent if and only if $\chi_u(X) = (-1)^n X^n$.

Proof. If $\chi_u(X) = (-1)^n X^n$, then Cayley-Hamilton's theorem ensures that $(-1)^n u^n = 0$. Conversely, if $u^k = 0$, then the minimal polynomial of u divides X^k. Since μ_u and χ_u have the same irreducible factor, we deduce that $\chi_u(X) = (-1)^n X^n$. □

In particular, we have $\mu_u(X) = X^k$ for $1 \leq k \leq n$.

Corollary 2.2

An endomorphism $u \in \mathcal{L}(E)$ is nilpotent if and only if its spectral radius (on $\widehat{\mathbb{K}}$) is $\rho(u) = 0$.

Proof. If u is nilpotent and if λ is an eigenvalue of u, we have $u^k(x) = \lambda^k x = 0$ for $x \neq 0$, thus $\lambda = 0$. Conversely, if all eigenvalues of u are zero, there exists a triangularization basis in which u can be written as:

$$A = \begin{pmatrix} 0 & * & \cdots & \cdots & * \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & * \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

We then verify that:

$$A^2 = \begin{pmatrix} 0 & 0 & * & \cdots & * \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & * \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

and by recurrence that $A^n = 0$. Therefore, u is indeed nilpotent. □

Theorem 2.3: Reduction of a nilpotent

Let $u \in \mathcal{L}(E)$ be a nilpotent endomorphism. There exists a basis $\mathcal{B}$ in which

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} N_{q_1} & | & & & \\ -- & & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & -- \\ & & & | & N_{q_k} \end{pmatrix}$$

The N_{q_i} are unique up to ordering. The nilpotency index α of u is equal to the maximum value of the q_i.

Proof.

- **Existence**

  We have $\mu_u(X) = X^\alpha$. Let $F_i = \mathrm{Ker}(u^i)$. We know from exercise 2.1.3 that

  $$\{0\} \subsetneq F_1 \subsetneq \cdots \subsetneq F_{\alpha-1} \subsetneq F_\alpha = E.$$

  Let $(e_{1,1}, \cdots, e_{1,i_1})$ be a basis of a supplementary space $G_{\alpha-1}$ of $F_{\alpha-1}$ in E:

  $$E = F_{\alpha-1} \oplus \underbrace{G_{\alpha-1}}_{\text{basis } (e_{1,1},\cdots,e_{1,i_1})}$$

  – The family $(u(e_{1,1}), \cdots, u(e_{1,i_1}))$ is linearly independent in $F_{\alpha-1}$.

  Indeed, suppose that $\sum_{i=1}^{i_1} \lambda_i u(e_{1,i}) = 0$. Applying $u^{\alpha-2}$, we obtain

  $$u^{\alpha-1}\left(\sum_{i=1}^{i_1} \lambda_i e_{1,i}\right) = 0,$$

  which means $\sum_{i=1}^{i_1} \lambda_i e_{1,i} \in F_{\alpha-1} \cap G$, so $\lambda_j = 0$ for all $j = 1, \cdots, i_1$.

  – The vector space generated by $(u(e_{1,1}), \cdots, u(e_{1,i_1}))$ is in direct sum with $F_{\alpha-2}$.

  Indeed, if $\sum_{i=1}^{i_1} \lambda_i u(e_{1,i}) \in F_{\alpha-2}$, this implies

  $$u^{\alpha-2}\left(\sum_{i=1}^{i_1} \lambda_i u(e_{1,i})\right) = u^{\alpha-1}\left(\sum_{i=1}^{i_1} \lambda_i e_{1,i}\right) = 0,$$

  so again $\lambda_j = 0$ for all $j = 1, \cdots, i_1$.

  – We complete $(u(e_{1,1}), \cdots, u(e_{1,i_1}))$ by $(e_{2,1}, \cdots, e_{2,i_2})$ into a basis of a supplementary space $G_{\alpha-2}$ of $F_{\alpha-2}$ in $F_{\alpha-1}$:

  $$E = F_{\alpha-2} \oplus \underbrace{G_{\alpha-2}}_{\text{basis } (u(e_{1,1}),\cdots,u(e_{1,i_1}),e_{2,1},\cdots,e_{2,i_2})} \oplus \underbrace{G_{\alpha-1}}_{\text{basis } (e_{1,1},\cdots,e_{1,i_1})}$$

  The family $(u^2(e_{1,1}), \cdots, u^2(e_{1,i_1}), u(e_{2,1}), \cdots, u(e_{2,i_2}))$ is a linearly independent set in $F_{\alpha-1}$, in direct sum with $F_{\alpha-3}$. We repeat this process...

  – We obtain a basis of E that can be reordered in a clever way:

  $$\mathcal{B} = \Big(\begin{array}{l} u^{\alpha-1}(e_{11}), \cdots, e_{11}, u^{\alpha-1}(e_{12}), \cdots, e_{12}, \cdots, u^{\alpha-1}(e_{1i_1}), \cdots, e_{1i_1}, \\ u^{\alpha-2}(e_{21}), \cdots, e_{21}, u^{\alpha-2}(e_{22}), \cdots, e_{22}, \cdots, u^{\alpha-2}(e_{2i_2}), \cdots, e_{2i_2}, \\ \vdots \\ e_{\alpha 1}, \cdots e_{\alpha i_\alpha} \Big) \end{array}$$

The matrix of u in this basis is indeed of the desired form. We have i_1 blocks N_α, ... i_α zero blocks.

- **Uniqueness**

  Notice that

$$\begin{aligned} i_1 &= \dim(E) - \dim(F_{\alpha-1}) = \dim(F_\alpha) - \dim(F_{\alpha-1}) \\ i_1 + i_2 &= \dim(F_{\alpha-1}) - \dim(F_{\alpha-2}) \\ &\vdots \\ i_1 + i_2 + \cdots + i_\alpha &= \dim(F_1) - \dim(F_0) \end{aligned}$$

  The number of blocks of a given type depends only on the dimensions of the spaces F_i and not on the specific construction we chose.

□

2.1.2 Dunford Reduction

Theorem 2.4

Let $u \in \mathcal{L}(E)$ such that χ_u or μ_u is split over $\mathbb{K}$.
Then:

- There exist two endomorphisms d and n such that $u = d + n$, with d diagonalizable, n nilpotent, and $d \circ n = n \circ d$.
- Moreover, the endomorphisms are polynomials in u, $d, n \in \mathbb{K}[u]$ and they are unique.

Proof.

- **Existence**

  We decompose E into a direct sum of the characteristic subspaces of u:

$$E = \bigoplus_{i=1}^{s} F_i,\ F_i = \mathrm{Ker}(u - \lambda_i id)^{m_i}.$$

  Note that to define d and n, it is sufficient to define their restriction to the spaces F_i. Set $d_{|F_i} = \lambda_i id$ and $n_{|F_i} = u - \lambda_i id$. Then clearly d is diagonalizable and n is nilpotent with index m_i on F_i.

  We note that if we denote $P_1, \cdots P_s$ as the projectors onto F_i parallel to $\bigoplus_{j \neq i} F_j$, the kernel theorem assures us that $P_i \in \mathbb{K}[u]$. Thus, $d = \lambda_1 P_1 + \cdots + \lambda_s P_s$ is also an element of $\mathbb{K}[u]$ and therefore $n = u - d \in \mathbb{K}[u]$. We conclude that d and n commute. We could also have shown that they commute by observing that they commute on each subspace F_i.

- **Uniqueness**

  Suppose we have another decomposition (d', n') of u such that d' and n' commute. We note that d' commutes with $d' + n' = u$ and thus with any polynomial in u, in particular with d. The endomorphisms d and d' commute and are diagonalizable, hence they are simultaneously diagonalizable. In particular, we deduce that their difference $d - d' = n' - n$ is both diagonalizable and nilpotent (the sum of two nilpotents is nilpotent). We deduce that all the eigenvalues of $d - d'$ are zero, so $d = d'$ from which $n = n'$ follows.

□

Remark.

Note that if $A = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} = \underbrace{\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}}_{D} + \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{N}$, we do not have a Dunford decomposition because D and N do not commute. Here, the Dunford decomposition of A is given by $D = A$ and $N = 0$.

2.1.3 Exercises

Exercise 2.1

Let $u \in \mathcal{L}(E)$ where $\dim E = n$. If $\mu_u(X) = X^n$, prove that there exists a basis $\mathcal{B}$ in which

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

Since $u^n = 0$ and $u^{n-1} \neq 0$, we know there exists $x \neq 0$ such that $u^{n-1}(x) \neq 0$. We will show that $(x, u(x), \cdots, u^{n-1}(x))$ is a basis of E. It suffices to show that this set is linearly independent, as it has n elements in E, a vector space of dimension n.
Let $(\lambda_0, \cdots, \lambda_{n-1})$ be such that

$$\lambda_0 x + \cdots + \lambda_{n-1} u^{n-1}(x) = 0.$$

Applying u^{n-1} gives us $\lambda_0 u^{n-1}(x) = 0$, thus $\lambda_0 = 0$. Then applying u^{n-2} yields $\lambda_1 u^{n-1}(x) = 0$, and so on, which implies all λ_i are zero. The matrix of u in this basis is clearly of the form

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

Exercise 2.2

Let N be the nilpotent matrix

$$N = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

Show that

$$\mathbb{K}[N] = \left\{ \begin{pmatrix} a_0 & a_1 & & a_{n-2} & a_{n-1} \\ \vdots & \ddots & \ddots & \ddots & a_{n-2} \\ \vdots & & \ddots & \ddots & \\ \vdots & & & \ddots & a_1 \\ 0 & \cdots & \cdots & \cdots & a_0 \end{pmatrix}, a_0, \cdots, a_{n-1} \in \mathbb{K} \right\}.$$

Show that the invertible elements of $\mathbb{K}[N]$ are those elements such that $a_0 \neq 0$ and compute their inverse.

We have already shown that since $\mu_N(X) = X^n$, the set $\{Id, N, \cdots, N^{n-1}\}$ is a basis for $\mathbb{K}[N]$. Now,

$$N^k = \begin{pmatrix} 0 & 0 & & 1 & & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & \ddots & 1 \\ \vdots & & & \ddots & \ddots & \\ \vdots & & & & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \end{pmatrix} \quad \text{(the } (k+1)^{th} \text{ row)}$$

Let $B \in \mathbb{K}[N]$, $B = a_0 Id + \cdots + a_{n-1}N^{n-1}$. If $a_0 \neq 0$, we can perform the decomposition of 1 by the polynomial $a_0 + \cdots + a_{n-1}X^{n-1}$ of order $n-1$. We have

$$1 = (a_0 + \cdots + a_{n-1}X^{n-1})(\alpha_0 + \cdots + \alpha_{n-1}X^{n-1}) + X^n R(X).$$

We deduce that

$$Id = B\underbrace{(\alpha_0 Id + \cdots + \alpha_{n-1}N^{n-1})}_{B^{-1}}.$$

Exercise 2.3

Let $u \in \mathcal{L}(V)$ such that $\mu_u(X) = \displaystyle\prod_{i=1}^{s}(X-\lambda_i)^{m_i}$. Show that

$$\{0\} \subsetneq \operatorname{Ker}(u-\lambda_i id) \subsetneq \operatorname{Ker}(u-\lambda_i id)^2 \cdots \subsetneq \operatorname{Ker}(u-\lambda_i id)^{m_i} = \operatorname{Ker}(u-\lambda_i id)^k, \forall k \geq m_i$$

Let $N_j = \operatorname{Ker}(u - \lambda_i\, id)^j$. It is clear that

$$N_1 \subset N_2 \subset \cdots \subset N_k \subset N_{k+1} \subset \cdots.$$

Let us show that $N^{m_i-1} \subsetneq N^{m_i}$. We proceed by contradiction: if $N^{m_i-1} = N^{m_i}$, then by the kernel lemma, we have

$$\begin{aligned} E &= \bigoplus_{j\neq i}^{s} \operatorname{Ker}(u-\lambda_j id)^{m_j} \bigoplus \operatorname{Ker}(u-\lambda_i id)^{m_i} \\ &= \bigoplus_{j\neq i}^{s} \operatorname{Ker}(u-\lambda_j id)^{m_j} \bigoplus \operatorname{Ker}(u-\lambda_i id)^{m_i-1} \\ &= \operatorname{Ker}\left((u-\lambda_i)^{m_i-1}\prod_{j\neq i}(u-\lambda_j id)^{m_j}\right) \end{aligned}$$

We deduce that the polynomial $(X-\lambda_i)^{m_i-1}\prod_{j\neq i}(X-\lambda_j id)^{m_j}$ is an annihilating polynomial for u, which contradicts the fact that μ_u is the minimal polynomial.

Now, let $x \in N_{m_i} \setminus N_{m_{i-1}}$. We have $(u-\lambda_i id)(x) \in N_{m_{i-1}} \setminus N_{m_{i-2}}$, and by induction, $(u-\lambda_i id)^k(x) \in N_{m_{i-k}} \setminus N_{m_{i-k-1}}$. Therefore, all the previous inclusions are strict.

Now, let us show that $N_{m_i} = N_{m_i+1}$. We apply the kernel lemma to $\mu_u(X)(X-\lambda_i)$. We have

$$E = \bigoplus_{j\neq i}^{s} \operatorname{Ker}(u-\lambda_j id)^{m_j} \bigoplus \operatorname{Ker}(u-\lambda_i id)^{m_i+1}$$

We deduce that $\dim(N_{m_i}) = \dim(N_{m_{i+1}})$, and since $N_{m_i} \subset N_{m_{i+1}}$, we conclude the result. Moreover, we know that once we have an equality, all subsequent inclusions are equalities.

2.1.4 Application to Matrix Power Calculations

We consider $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$.

Exercise 2.4

Let n be a nilpotent endomorphism with nilpotency index r and $q \in \mathbb{N}$.
Show that the sequence $((\lambda id + n)^q)_{q\in\mathbb{N}}$ is bounded in $\mathcal{M}_p(\mathbb{K})$ if and only if

- $|\lambda| < 1$

  or

- $|\lambda| = 1$ and $r = 1$.

Since homotheties commute with all endomorphisms:

$$(\lambda id + n)^q = \sum_{i=0}^{q} C_q^i \lambda^{q-i} n^i = \sum_{i=0}^{r-1} C_q^i \lambda^{q-i} n^i \text{ if } q \geq r.$$

The family $(id, n, \cdots, n^{r-1})$ being linearly independent in $\mathcal{M}_p(\mathbb{K})$, the sequence $((\lambda id{+}n)^q)_q$ is bounded in $\mathcal{M}_p(\mathbb{K})$ if all the coefficients (in a basis $\mathcal{B}$ whose first elements are $(id, n, \cdots, n^{r-1})$) $\lambda^q, \cdots, C_q^{r-1}\lambda^{r-1}$ are bounded in $\mathbb{K}$. We deduce that a necessary condition is that $|\lambda| \leq 1$. If $|\lambda| < 1$ since C_q^i is a polynomial in q, all sequences are bounded. If $|\lambda| = 1$, then we need $r = 1$, in other words, $n = 0$.

Exercise 2.5

Let n be a nilpotent endomorphism with index r. Suppose $\lambda \neq 0$. Under what condition on n and on λ is the sequence $((\lambda id + n)^q)_{q\in\mathbb{Z}}$ bounded in $\mathcal{M}_p(\mathbb{K})$?

We verify that

$$(\lambda id + n)^{-1} = \lambda^{-1}(id + \frac{n}{\lambda})^{-1} = \lambda^{-1}\left(id - \frac{n}{\lambda} + \cdots + (-1)^{r-1}\frac{n^{r-1}}{\lambda^{r-1}}\right) = \lambda^{-1} id + n'.$$

The endomorphism n' is a sum of nilpotent endomorphisms that commute, thus it is a nilpotent endomorphism. According to the previous exercise, the sequence will only be bounded over $\mathbb{Z}$ if $|\lambda| = 1$ and $r = 1$, in other words $|\lambda| = 1$ and $n = 0$.

Exercise 2.6

Give a necessary and sufficient condition for the sequence A^q ($A \in \mathcal{M}_n(\mathbb{C})$) to be bounded

1. for $q \in \mathbb{N}$.
2. for $q \in \mathbb{Z}$.

Let $\lambda_1, \cdots, \lambda_s$ be the eigenvalues of A and F_i the characteristic subspaces $\text{Ker}(u - \lambda_i Id)^{m_i}$, with u being the endomorphism associated with A. We know (Kernel Theorem) that $E = \bigoplus_{i=1}^{s} F_i$ and that $u_i = u_{|F_i} = \lambda_i id + n_i$ where n_i is nilpotent. We then have the equivalence between

- the sequence $(A^q)_q$ is bounded
- the sequence $(u^q)_q$ is bounded
- the sequences $(u_i^q)_q$ are bounded for all $i = 1, \cdots, s$.

From what precedes, the sequence $(A^q)_{q\in\mathbb{N}}$ is bounded if and only if for all $i = 1, \cdots, s$, $|\lambda_i| < 1$ or $|\lambda_i| = 1$ and $n_i = 0$ ($\Leftrightarrow u_i = \lambda_i id \Leftrightarrow F_i = E_i = \text{Ker}(u - \lambda_i)$).
From what precedes, the sequence $(A^q)_{q\in\mathbb{Z}}$ is bounded if and only if A is diagonalizable and its spectrum is included in the unit disk.

Exercise 2.7: Variant

Show that the sequence A^q converges to 0 as $q \to \infty$ if and only if for every $\lambda \in \mathrm{Sp}(A)$, $|\lambda| < 1$.

2.2 Jordan Reduction

We define the Jordan blocks of size q as follows:

$$J_{\lambda,q} = \lambda Id + N_q, \text{ with } N_q = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \in \mathcal{M}_q(\mathbb{R})$$

in other words

$$J_{\lambda,q} = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda \end{pmatrix}.$$

The corresponding Jordan form of a matrix A is the direct sum of its Jordan blocks, each corresponding to an eigenvalue of A.

We denote $\mathcal{J}$ the space of upper triangular matrices. The Jordan form can be characterized as a unique upper triangular matrix J such that:

$$A \sim J \iff \exists P \in \mathcal{GL}_n(\mathbb{R}), J = P^{-1}AP \text{ (conjugation)} \iff A \text{ is similar to } J.$$

2.2.1 Jordan Reduction

Theorem 2.5

Let $u \in \mathcal{L}(E)$ be an endomorphism of $\mathcal{L}(E)$ such that μ_u is split. There exists a basis $\mathcal{B}$ in which

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} J_{\lambda_1,q_1} & | & & & \\ -- & & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & -- \\ & & & | & J_{\lambda_k,q_k} \end{pmatrix}$$

The J_{λ_i,q_i} are unique up to order.

This is referred to as the block diagonalization of the endomorphism u.

Proof. The proof of this result follows from Dunford's decomposition: we decompose $E = \bigoplus F_i$, with F_i being the characteristic subspaces. On each stable subspace F_i, we have $u_{|F_i} = \lambda_i id + n_i$. We then use the reduction of nilpotents we have just seen. □

2.2.2 Application to the Solution of Linear Recurrence Sequences

Let **E** be the set of solutions to the linear recurrence

$$u_{n+p} = a_1 u_{n+p-1} + \cdots + a_p u_n.$$

Theorem 2.6

Let $\mathbf{E}$ be the set of solutions for $a_p \neq 0$ of

$$u_{n+p} = a_1 u_{n+p-1} + \cdots + a_p u_n.$$

Suppose that $r_1, \cdots, r_s$ are the roots of multiplicities $m_1, \cdots, m_s$ of the equation

$$r^p = a_1 u^{p-1} + \cdots + a_p.$$

Then the sequences $(r_1^n), (nr_1^n), \cdots, (n^{m_1-1} r_1^n), \cdots, (r_s^n), (nr_s^n), \cdots, (n^{m_s-1} r_s^n)$ form a basis for $\mathbf{E}$.

Proof. We can define a mapping

$$\phi : \begin{array}{ccc} \mathbf{E} & \rightarrow & \mathbb{K}^p \\ (u_n)_n & \mapsto & (u_0, \cdots, u_p) \end{array}$$

This mapping is clearly linear, injective, and surjective. We deduce that $\mathbf{E}$ is a vector space of dimension p. Now, let $X_n = \begin{pmatrix} u_n \\ \vdots \\ u_{n+p-1} \end{pmatrix}$, we have

$$X_n = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ a_p & \cdots & \cdots & \cdots & a_1 \end{pmatrix} X_{n-1} = AX_{n-1}.$$

Thus, $X_n = A^n X_0$.

Let

$$f : \begin{array}{ccc} \mathbb{K}^{\mathbb{N}} & \rightarrow & \mathbb{K}^{\mathbb{N}} \\ (v_0, v_1, \cdots) & \mapsto & (v_1, v_2, \cdots) \end{array},$$

we verify that

$$\mathbf{E} = \operatorname{Ker}(f^p - a_1 f^{p-1} - \cdots - a_p id).$$

By hypothesis,

$$X^p - a_1 X^{p-1} - \cdots - a_p = (X - r_1)^{m_1} \cdots (X - r_s)^{m_s}.$$

We apply the kernel theorem to f, which can be done because it does not require finite dimensionality. We obtain

$$\mathbf{E} = \bigoplus_{i=1}^{s} \underbrace{\operatorname{Ker}(f - r_i id)^{m_i}}_{F_i}.$$

We have seen that $\dim(F_i) = m_i$, and the elements of F_i are the solutions of the recurrence $(f - r_i id)^{m_i} = 0$. We have shown that the set $\mathbf{E}_i$ of solutions of this recurrence is a vector space of dimension m_i, the length of the recurrence. It suffices to show that for every $j = 1, \cdots, m_i$, $n^{j-1} r_i^n \in \mathbf{E}_i$ and that this family is linearly independent in $\mathbf{E}_i$.

First, we show that $n^{j-1} r_i^n \in \mathbf{E}_i$ by showing that $(f - r_i id)(n^{j-1} r_i^n) = 0$. Indeed,

$$\begin{array}{rcl} (f - r_i id)(n^{j-1} r_i^n) & = & \Big((n+1)^{j-1} - n^{j-1}\Big) r_i^{n+1} \\ & \in & \operatorname{Vect}\Big\{r_i^n, \cdots, n^{j-2} r_i^n\Big\} \\ (f - r_i id)^2(n^{j-1} r_i^n) & = & \Big((n+2)^{j-1} - 2(n+1)^{j-1} + n^{j-1}\Big) r_i^{n+2} \\ & \in & \operatorname{Vect}\Big\{r_i^n, \cdots, n^{j-3} r_i^n\Big\} \\ & \cdots & \\ (f - r_i id)^{j-1}(n^{j-1} r_i^n) & \in & \operatorname{Vect}\Big\{r_i^n\Big\} \end{array}$$

Thus, $(f - r_i id)^j (n^{j-1} r_i^n) = 0$, and since $j \leq m_i$, we have for all $j = 1, \cdots, m_i$, $(n^{j-1} r_i^n)_n \in F_i$.

It remains to show the independence of these elements. Suppose $\lambda_1, \cdots, \lambda_i$ such that

$$\sum_{j=1}^{m_i} \lambda_j n^{j-1} r_i^n = 0 \Rightarrow r_i^n \left(\sum_{j=1}^{m_i} \lambda_j n^{j-1} \right) = 0.$$

Since $a_p \neq 0$, we have $r_i \neq 0$, and thus $\sum_{j=1}^{m_i} \lambda_j n^{j-1} = 0$. The polynomial $Q(X) = \sum_{j=1}^{m_i} \lambda_j X^{j-1}$ vanishes on $\mathbb{N}$, hence it must be identically zero, which means all its coefficients λ_j are zero. □

2.2.3 Application to the exponential Matrix calculations

Theorem 2.7

Let $u \in \mathcal{L}(E)$ be an endomorphism of $\mathcal{L}(E)$ such that μ_u is split. Consider a basis $\mathcal{B}$ in which

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} J_{\lambda_1,q_1} & | & & & \\ -- & & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & -- \\ & & & | & J_{\lambda_k,q_k} \end{pmatrix}$$

Then for $t \in \mathbb{K}$

$$\mathrm{Mat}_{\mathcal{B}}(\exp(tu)) = \begin{pmatrix} \exp(tJ_{\lambda_1,q_1}) & | & & & \\ -- & & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & -- \\ & & & | & \exp(tJ_{\lambda_k,q_k}) \end{pmatrix}$$

with

$$\exp(tJ_{\lambda,q}) = e^{t\lambda_i} \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{q-1}}{q!} \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{t2}{2!} \\ \vdots & & \ddots & \ddots & t \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}.$$

2.2.4 Exercises

Exercise 2.8

Determine the sequences that satisfies

$$u_{n+1} = 2u_n - u_{n-1} \, \forall n \geq 1$$

What happen if we impose one of these additionnal condition:

1. $u_0 = u_{N+1} = 0$,
2. $u_{n+N} = u_n$ for all $n \in \mathbb{N}$.

Exercise 2.9

Determine e^{tA} for $A = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

2.3 SVD decomposition

2.3.1 The decomposition

The SVD decomposition is due to Carl ECKART and Gale YOUNG in 1936. The work has been published in 1939 and can be find in : http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.bams/1183501633

For more details on matrice decomposition see also [Ser02].

Theorem 2.8: Existence and uniqueness

For all $A \in \mathcal{M}_{m,n}(\mathbb{C})$, there exists $U \in \mathcal{M}_m(\mathbb{C})$, $\Sigma \in \mathcal{M}_{m,n}(\mathbb{C})$, $V \in \mathcal{M}_n(\mathbb{C})$ such that

$$A = U\Sigma V^*$$

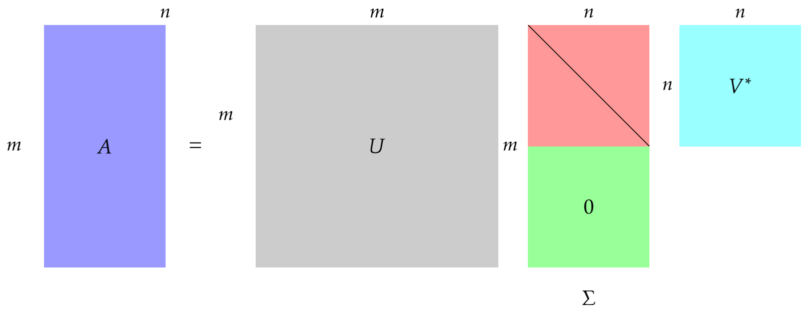
with U and V two unitary matrices ($UU^* = U^*U = Id$, $VV^* = V^*V = Id$), Σ a diagonal matrix with non negative $(\sigma_i)_{i=1,\cdots,p}$ coefficients, with $p = \min(m,n)$ and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p.$$

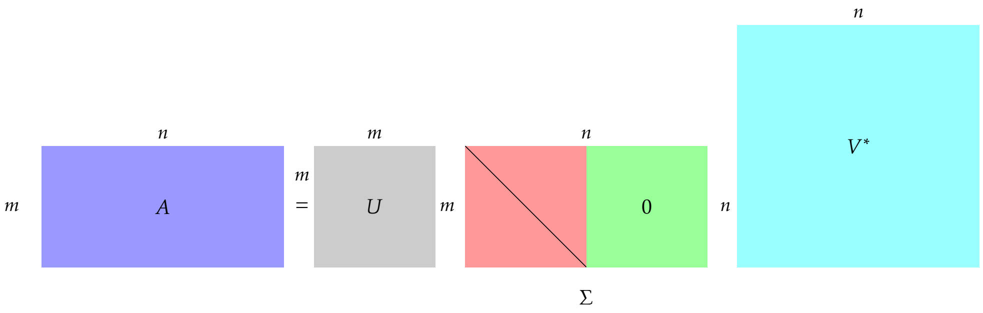
The values $(\sigma_i)_{i=1,\cdots,p}$ are called singular values of A. The matrix Σ is unique, but not U and V.

Proof. Let us visualize the shape of the matrices according to the values of m and n.

- Case $m > n$.



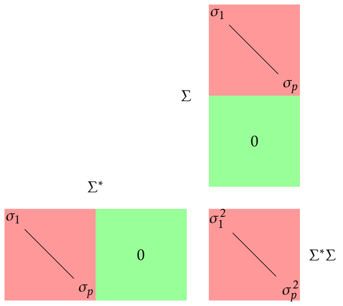
- Case $m < n$.



Assume that such a decomposition exists $A = U\Sigma V^*$, then

$$A^*A = (U\Sigma V^*)^*(U\Sigma V^*) = V\Sigma^*\Sigma V^*$$

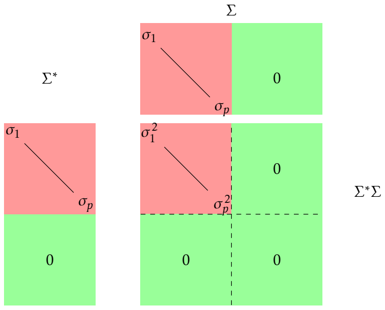
We deduce that A^*A and $\Sigma^*\Sigma$ are two symmetric positive $n \times n$ matrices that are similar and thus they share the same eigenvalues. In the case $m > n$,



so that

$$\Sigma^*\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_p^2).$$

In the case $n > m$,



so that

$$\Sigma^*\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_p^2, 0, \cdots, 0).$$

Consequently, $\sigma_1^2, \cdots, \sigma_p^2$ are nothing but eigenvalues of A^*A and the columns of V correspond to the eigenvectors. As A^*A is and hermitian matrix, it is always possible to chose a unitary matrix V . We can emphasized that $\sigma_i = 0$ for all $i > r = \mathrm{Rk}(A)$ with $r \leq p$.

Let us now construct the matrix U. Consider first the case $\mathrm{Ker}(A) = \{0\}$, $n = m$. We have for all $i = 1, \cdots, p$, $\sigma_i > 0$ and thus Σ is invertible. Therefore, we can define $U = AV\Sigma^{-1}$ that is a unitary matrix because

$$U^*U = (AV\Sigma^{-1})^*(AV\Sigma^{-1}) = (\Sigma^{-1})^*V^*A^*AV\Sigma^{-1} = (\Sigma^{-1})^*\Sigma^*\Sigma\Sigma^{-1} = Id$$

That conclude the proof of the theorem in the simplest case $\mathrm{Ker}(A) = \{0\}$, $n = m$.

Let us comme back to the general case. Introduce $u_1, \cdots, u_m$ the column vector of the matrix U and $v_1, \cdots, v_n$ the column vector of the matrix V. We look for a unitary matrix U such that $U\Sigma = AV$. The matrix U will be unitary if and only if $u_i^*u_j = \delta_{ij}$. Remarking that

$$\begin{aligned} U\Sigma &= \begin{pmatrix} \sigma_1 u_1 & \sigma_2 u_2 & \cdots & \sigma_p u_p & 0 & \cdots & 0 \end{pmatrix} \text{if } n > m \quad i.e. \quad p = m \\ &= \begin{pmatrix} \sigma_1 u_1 & \sigma_2 u_2 & \cdots & \sigma_p u_p \end{pmatrix} \text{if } n \leq m \quad i.e. \quad p = n \end{aligned}$$

and

$$AV = \begin{pmatrix} Av_1 & Av_2 & \cdots & Av_n \end{pmatrix}$$

Note that $Av_i = 0$ for all $i > r = \mathrm{Rk}(A)$ and $\sigma_i = 0 \Leftrightarrow i > r$. The problem $U\Sigma = AV$ can be reduced to

$$u_i = \frac{1}{\sigma_i} Av_i \, \forall i \leq r$$

Remark that

$$u_j^*u_i = \frac{1}{\sigma_i \sigma_j} v_j^* A^* A v_i = \delta_{ij}.$$

We conclude by completing the family $(u_i)_{i=1,\cdots,r}$ into an orthogonal basis.

□

Remark.

If $A \in \mathcal{M}_{m,n}(\mathbb{R})$ has real coefficients, then matrices U and V can be choosen with real coefficients and are consequently orthogonals: $U^tU = UU^t = Id$ and $V^tV = VV^t = Id$.

Remark.

The SVD decomposition of a matrix $A \in \mathcal{M}_{m,n}(\mathbb{C})$ can also be written as

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^*$$

or if $A \in \mathcal{M}_{m,n}(\mathbb{R})$

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^t$$

Indeed, let $B = \sum_{i=1}^{r} \sigma_i u_i v_i^*$ et evaluate Bv_j. We have

$$Bv_j = \sum_{i=1}^{r} \underbrace{\sigma_i u_i}_{=Av_i} \underbrace{v_i^* v_j}_{=\delta_{ij}} = Av_i$$

The images of the base (v_j) by the two matrix coincide, therefore $A = B$.

2.3.2 Application to the least square method

Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^m$ and $J : \mathbb{R}^n \to \mathbb{R}$, let us define $J(X) = \|AX - b\|^2$. We will use SVD to solve the normal equation associated to this optimization problem

$$0 = \nabla J(X) = A^t AX - A^t b = V\Sigma^t \Sigma V^t X - V\Sigma U^t b \Leftrightarrow \Sigma^t \Sigma V^t X = \Sigma U^t b \tag{2.1}$$

Introduce $Z = V^t X = \left(Z_i\right)_{i=1,\cdots,n}$ where Z_i are the coordinates of Z and $U^t b = \left(\hat{b}_i\right)_{i=1,\cdots,m}$. The system (2.1) reads

$$z_i = \frac{1}{\sigma_i}\hat{b}_i \,\forall i = 1,\cdots,r \text{ and } z_i \text{ chosen anyway } \forall i > r.$$

The set $\mathcal{S}$ of the solution of the least square problem is thus given by

$$\mathcal{S} = \left\{X \in \mathbb{R}^n \text{ such that } X = V\left(\frac{1}{\sigma_1}\hat{b}_1,\cdots,\frac{1}{\sigma_r}\hat{b}_r, z_{r+1},\cdots,z_n\right)^t, z_{r+1},\cdots,z_n \text{chosen anyway}\right\}.$$

SLet $X \in \mathcal{S}$, as V is orthogonal

$$\|X\|^2 = \|VX\|^2 = \underbrace{\sum_{i=1}^{r}\left(\frac{1}{\sigma_i}\hat{b}_i\right)^2}_{\text{fixed}} + \underbrace{\sum_{i \geq r+1} (z_i)^2}_{\geq 0}.$$

We deduce the existence of a unique vector $\bar{X} \in \mathcal{S}$ with a minimum norm given by

$$\bar{X} = V\left(\frac{1}{\sigma_1}\hat{b}_1,\cdots,\frac{1}{\sigma_r}\hat{b}_r, 0,\cdots,0\right)^t$$

2.3.3 Application to the low rank approximation

Theorem 2.9: Low rank approximation

Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and its SVD decomposition: $U \in \mathcal{O}(m)$, $V \in \mathcal{O}(n)$, $\Sigma \in \mathcal{M}_{m,n}(\mathbb{R})$ diagonal such that $A = U\Sigma V^t$, $p = \min(n, m)$. Introduce $u_1,\cdots,u_m$ the columns of U, $v_1,\cdots,v_n$ the columns of V et $\sigma_1 \geq \cdots \geq \sigma_p$ the diagonal coefficients of Σ. The matrix

$$A_k = \sum_{l=1}^{k} \sigma_l u_l v_l^t \in \mathcal{M}_{m,n}(\mathbb{R})$$

satisfies

$$\|A_k - A\| = \sigma_{k+1}$$

where $\|.\|$ denotes the matrix norm induced by the 2-norm on $\mathbb{R}^m$.

Proof. The first step consists in remarking that

$$A = \sum_{l=1}^{p} \sigma_l u_l v_l^t.$$

So that $\delta A = A - A_k = \sum_{l=k+1}^{p} \sigma_l u_l v_l^t$. Recall that $||\delta A||_2 = \sqrt{\rho(\delta A^t \delta A)}$. Using the orthogonality of U and V we obtain

$$\begin{aligned} \delta A^t \delta A &= \left(\sum_{l=k+1}^{p} \sigma_l u_l v_l^t\right)^t \left(\sum_{l=k+1}^{p} \sigma_l u_l v_l^t\right) \\ &= \sum_{l=k+1}^{p} \sigma_l^2 v_l v_l^t \end{aligned}$$

For all $i = 1, \cdots, n$, we deduce

$$\delta A^t \delta A v_i = \begin{cases} 0 \text{ if } i \le k \\ \sigma_i^2 v_i \text{ if } i > k \end{cases}$$

The family (v_i) is a basis constituted of eigenvectors of $\delta A^t \delta A$ with eigenvalues $(0, \cdots, 0, \sigma_{k+1}^2, \cdots, \sigma_p^2)$. We conclude easily that $\rho(\delta A^t \delta A) = \sigma_{k+1}^2$. □

Remark.

If the singular values of A decrease quikly, we can approximate the matrix A by A_k for small k and only store the informations necessary to define A_k, that is $k * (n + m + 1)$ coefficients instead $n * m$. This approximation is use in image compressing or in the Principal Composant Approximation (PCA), one of the machine learning strategy.

Remark.

Note that if $\sigma_1, \cdots, \sigma_p$ design the singular values of A, then

$$\sigma_{k+1} = \min\{||A - B||_2 \text{ such that } \mathrm{Rk}(B) \le k\}.$$

Indeed, we just have seen that $\sigma_{k+1} = ||A - A_k||_2$, therefore

$$\sigma_{k+1} \ge \min\{||A - B||_2 \text{ such that } \mathrm{Rk}(B) \le k\}.$$

Reciprocally, if a matrix B a a rank smaller than k, its kernel a dimension bigger than $n-k$ and thus it intersects the space $\mathrm{Vect}(v_1, \cdots, v_{k+1})$, subspace of dimension $k+1$. So its exists $x \in \mathbb{R}^n$ such that $||x|| = 1$ and $x \in Ker(B) \cap \mathrm{Vect}(v_1, \cdots, v_{k+1})$. Therefore

$$||(A - B)x||_2 = ||Ax||_2 = ||U\Sigma V^t x||_2 = ||\Sigma V^t x||_2 \ge \sigma_{k+1}||V^t x||_2 \ge \sigma_{k+1}$$

We deduce that

$$\sigma_{k+1} \le ||A - B||_2 \, \forall B \text{ such that } \mathrm{Rk}(B) \le k.$$

2.3.4 Application to image compression

Any image can be represented by a matrix (for a black and white image) or a set of three matrices (for a coloured image: R red G green B blue). Each pixel of the image is a point on a cartesian grid and each element of the matrix represents the intensity of the color of the image on the associated grid point.

The most famous example is certainly the photo of Lena :

Lena

Gene Howard Golub (1932-2007)

William Morton Kahan (1933-.)

This image can be represented by a $2^9 = 512$ square matrix A, or by $2^{18} = 262144$ coeficcients (pixels). Using the SVD decomposition we will see how to compress this image.

The algorithm was proposed in 1965 by Gene GOLUB (US) and William KAHAN (canada), based on the theorem 2.3.3. It just consists in approximating A by the matrices A_k. The figure 2.1 shows the efficency of the method. For a matrix with mn coefficients, the gain $1 - \frac{k(n+m+1)}{nm}$, so in this particular case, it gives a gain of 93.75% if $k = 4$ and 75% if $k = 6$.

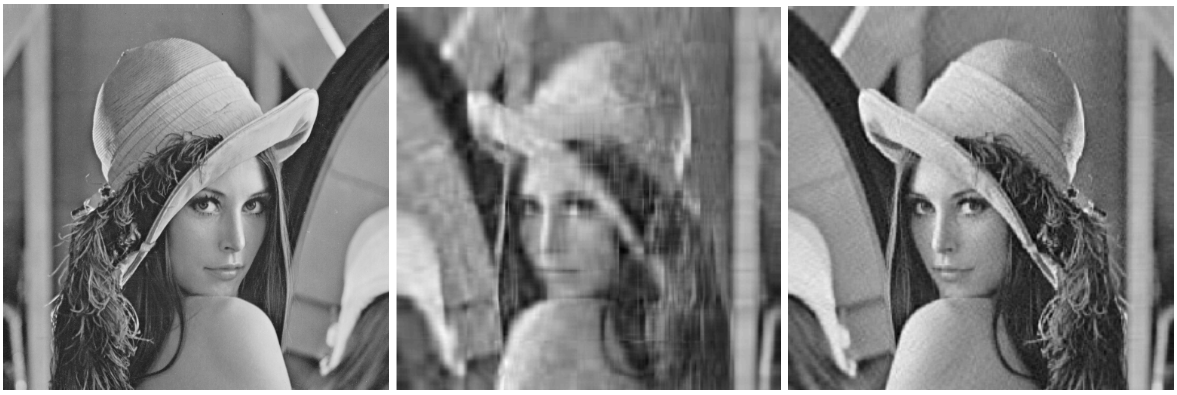


Figure 2.1. Lena. (Left) Original picture (Center) $k = 2^4$. (Right) $k = 2^6$.

2.3.5 Exercises

Exercise 2.10

Consider $A = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$.

1. Determine the singular values of A.
2. Give the svd decompsition of A.

We have $A^tA = \begin{pmatrix} 4 & 0 \\ 0 & 9 \end{pmatrix}$, thus $\sigma_1 = 3$ and $\sigma_2 = 2$,

$$V = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}, U = AV\Sigma^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Exercise 2.11

Consider $A = \begin{pmatrix} 0 & 3 \\ -2 & 0 \\ 0 & 0 \end{pmatrix}$.

1. Determine the singular values of A.
2. Give the svd decompsition of A.

We have

$$A^tA = \begin{pmatrix} 0 & -2 & 0 \\ 3 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 3 \\ -2 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 4 & 0 \\ 0 & 9 \end{pmatrix}$$

We deduce that $\sigma_1 = 3$ and $\sigma_2 = 2$,

$$V = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 2 \\ 0 & 0 \end{pmatrix}, u_1 = \frac{1}{\sigma_1}(AV)_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, u_2 = \frac{1}{\sigma_2}(AV)_2 = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}$$

We can complete the matrix U the vector $u_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, therefore

$$U = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Exercise 2.12

Consider $A = \begin{pmatrix} 0 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix}$.

1. Determine the singular values of A.
2. Give the svd decompsition of A.

We have

$$A^tA = \begin{pmatrix} 0 & 1 \\ 3 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 9 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

and thus the singular values of A are $\sigma_1 = 3$, $\sigma_2 = \sqrt{2}$, $\sigma_3 = 0$ with

$$V = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}, AV = \begin{pmatrix} 3 & 0 & 0 \\ 0 & \frac{2}{\sqrt{2}} & 0 \end{pmatrix}$$

We deduce that $U = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Exercise 2.13

Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$ a matrix of rank $r \leq p = \min(m,n)$, let $[U,\Sigma,V]$ be its SVD decomposition. Denote by $(u_1,\cdots,u_m)$ the columns vectors of U and $(v_1,\cdots,v_n)$ those of V.

1. Prove that $\mathrm{Im}(A) = \mathrm{Vect}(u_1,\cdots,u_r)$.
2. Prove that $\mathrm{Ker}(A) = \mathrm{Vect}(v_{r+1},\cdots,v_n)$.
3. Prove that $\mathrm{Im}(A^t) = \mathrm{Vect}(v_1,\cdots,v_r)$.
4. Prove that $\mathrm{Ker}(A^t) = \mathrm{Vect}(u_{r+1},\cdots,u_m)$.
5. Determine the matrices of the orthogonal projection on $\mathrm{Im}(A)$, $\mathrm{Ker}(A)$, $\mathrm{Im}(A^t)$, $\mathrm{Ker}(A^t)$ thanks to U and V.
6. Find the SVD decomposition of $A = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ -1 & 1 \end{pmatrix}$ and the matrices of the orthogonal projections $\mathrm{Im}(A)$, $\mathrm{Ker}(A)$, $\mathrm{Im}(A^t)$, $\mathrm{Ker}(A^t)$.

We first recall that if $A = U\Sigma V^t$ then equivalently we have

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^t.$$

1. Consider $y \in \mathrm{Im}(A)$, then

$$y = Ax = \sum_{i=1}^{r} \sigma_i u_i v_i^t x = \sum_{i=1}^{r} \underbrace{\sigma_i v_i^t x}_{\in\mathbb{R}} u_i \in \mathrm{Vect}(u_1,\cdots,u_r)$$

As the dimension of $\mathrm{ImA} = r = dim(\mathrm{Vect}(u_1,\cdots,u_r))$, we conclude the proof of the first point.

2. For $l = r+1, \cdots, n$, we have
$$Av_l = \sum_{i=1}^{r} \sigma_i u_i \underbrace{v_i^t v_l}_{=0}$$
Therefore, $\mathrm{Vect}(v_{r+1}, \cdots, v_n) \subset \mathrm{Ker}(A)$. We conclude by using an dimension argument.

3. We remark that
$$A^t = \sum_{i=1}^{r} \sigma_i v_i u_i^t.$$
Consider $y \in \mathrm{Im}(A^t)$, then
$$y = A^t x = \sum_{i=1}^{r} \sigma_i v_i u_i^t x = \sum_{i=1}^{r} \underbrace{\sigma_i u_i^t x}_{\in \mathbb{R}} v_i \in \mathrm{Vect}(v_1, \cdots, v_r)$$
We conclude by using an dimension argument.

4. For $l = r+1, \cdots, n$, we have
$$A^t u_l = \sum_{i=1}^{r} \sigma_i v_i \underbrace{u_i^t u_l}_{=0}$$
Therefore, $\mathrm{Vect}(u_{r+1}, \cdots, u_n) \subset \mathrm{Ker}(A^t)$. We conclude by using an dimension argument.

5. We recall that the orthogonal projection P on a subset F is such that $Px = x$ for all $x \in F$ and $Px = 0$ for all $x \in F^\perp$. Note that $Id - P$ is the orthogonal projection on $F^\perp$.
   - ▶ The orthogonal projection on $\mathrm{Im}(A)$ is thus $P = \sum_{i=1}^{r} u_i u_i^t$. Indeed, for $l = 1, \cdots, r$ using question 1
   $$Pu_l = \sum_{i=1}^{r} u_i \underbrace{u_i^t u_l}_{=\delta_{il}} = u_l$$
   and as $\mathrm{Im}(A)^\perp = \ker(A^t)$, using question 4, we check that for $l \geq r+1$ $Pu_l = 0$.
   - ▶ Similarly, the orthogonal projection on $\mathrm{Im}(A^t)$ is thus $P = \sum_{i=1}^{r} v_i v_i^t$.
   - ▶ Using the fact that $\mathrm{Im}(A)^\perp = \mathrm{Ker}(A^t)$, we deduce that the orthogonal projection on $\mathrm{Ker}(A^t)$ is given by
   $$Id - \sum_{i=1}^{r} u_i u_i^t$$
   - ▶ Using the fact that $\mathrm{Im}(A^t)^\perp = \mathrm{Ker}(A)$, we deduce that the orthogonal projection on $\mathrm{Ker}(A)$ is given by
   $$Id - \sum_{i=1}^{r} v_i v_i^t$$

6. We remark that
$$A^t A = \begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix}$$
Its eigenvalues are $\sigma_1^2 = 7$ and $\sigma_2^2 = 2$ with eigenvectors $v_1 = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $v_2 = \frac{1}{\sqrt{5}}\begin{pmatrix} -1 \\ 2 \end{pmatrix}$. We deduce
$$u_1 = \frac{1}{\sigma_1} A v_1 = \frac{1}{\sqrt{5}} \frac{1}{\sqrt{7}} \begin{pmatrix} 3 \\ 5 \\ -1 \end{pmatrix},\ u_2 = \frac{1}{\sigma_2} A v_2 = \frac{1}{\sqrt{5}} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}$$
We now choose $u_3 = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$ orthogonal to u_1 and u_2:
$$3\alpha + 5\beta - \gamma = 0,\ \alpha + 3\gamma = 0,\ \alpha^2 + \beta^2 + \gamma^2 = 1$$
We deduce
$$\alpha = -3\gamma,\ \beta = \frac{10}{5}\gamma,\ \gamma^2 (9 + \frac{100}{25} + 1) = 1 \text{ or } \gamma = \frac{5}{\sqrt{350}} = \frac{1}{\sqrt{14}}$$
Or $u_3 = \frac{1}{\sqrt{14}} \begin{pmatrix} -3 \\ 2 \\ 1 \end{pmatrix}$.

We have
$$u_1 u_1^t = \frac{1}{35} \begin{pmatrix} 9 & 15 & -3 \\ 15 & 25 & -5 \\ -3 & -5 & 1 \end{pmatrix},\ u_2 u_2^t = \frac{1}{10} \begin{pmatrix} 1 & 0 & 3 \\ 0 & 0 & 0 \\ 3 & 0 & 9 \end{pmatrix},\ u_3 u_3^t = \frac{1}{14} \begin{pmatrix} 9 & -6 & 3 \\ -6 & 4 & 2 \\ -3 & 2 & 1 \end{pmatrix}$$

and

$$v_1 v_1^t = \frac{1}{5}\begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}, \; v_2 v_2^t = \frac{1}{5}\begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix}$$

Therefore,

- the matrix of the projection on $\mathrm{Im}(A)$ is given by

$$P_{\mathrm{Im}(A)} = u_1 u_1^t + u_2 u_2^t = \frac{1}{70}\begin{pmatrix} 25 & 30 & 15 \\ 30 & 50 & -10 \\ 15 & -10 & 65 \end{pmatrix} = \frac{1}{14}\begin{pmatrix} 5 & 6 & 3 \\ 6 & 10 & -3 \\ 3 & -2 & 13 \end{pmatrix}$$

- the matrix of the projection on $\mathrm{Im}(A^t)$ is given by

$$P_{\mathrm{Im}(A^t)} = v_1 v_1^t + v_2 v_2^t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- $\mathrm{Ker}(A)$ is reduced to 0.
- the matrix of the projection on $\mathrm{Ker}(A^t)$ is given by

$$P_{\mathrm{Ker}(A)} = u_3 u_3^t = \frac{1}{14}\begin{pmatrix} 9 & -6 & 3 \\ -6 & 4 & 2 \\ -3 & 2 & 1 \end{pmatrix}$$

Exercise 2.14

Consider $A \in \mathcal{M}_n(\mathbb{R})$, we denote by $A = U\Sigma V^t$ the SVD decomposition of A and σ_j for $j = 1, \cdots, n$ its singular values.

1. Prove that the eigenvalues of the matrix$M = \begin{pmatrix} 0 & A^t \\ A & 0 \end{pmatrix}$ are given by $\pm\sigma_j$ for $j = 1, \cdots, n$ and are associated to the eigenvectors $\begin{pmatrix} v_j \\ \pm u_j \end{pmatrix}$.
2. Evaluate $M^t M$. Deduce the singular values of the matrix M.
3. In the case $n = 2$ and A invertible, explicit the SVD decomposition of M.
4. In the case $n = 2$ and A of rank 1, explicit the SVD decomposition of the matrix M.
5. How can we generalize the result in dimension n ?

---

1. On rappelle que $AV = U\Sigma$ et que $A^t U = V\Sigma^t$. Par conséquent , pour $\epsilon = 1$ ou $\epsilon = -1$

$$M\begin{pmatrix} v_j \\ \epsilon u_j \end{pmatrix} = \begin{pmatrix} \epsilon A^t u_j \\ A v_j \end{pmatrix} = \epsilon\sigma_j \begin{pmatrix} v_j \\ \epsilon u_j \end{pmatrix}$$

2. On vérifie quelconque

$$M^t M = \begin{pmatrix} 0 & A^t \\ A & 0 \end{pmatrix}\begin{pmatrix} 0 & A^t \\ A & 0 \end{pmatrix} = \begin{pmatrix} A^t A & 0 \\ 0 & AA^t \end{pmatrix}$$

Les valeurs singulières de M sont donc σ_j pour $j = 1, \cdots, n$ comptées deux fois.

3. Dans le cas $n = 2$ et A inversible, les valeurs singulières de A sont donc $\sigma_1 \geq \sigma_2 > 0$. Les vecteurs propres de $A^t A$ sont les v_j, ceux de AA^t les u_j. Par conséquent, si $M = \bar U \bar\Sigma \bar V$, alors

$$\bar V = \frac{1}{\sqrt{2}}\begin{pmatrix} v_1 & v_1 & v_2 & v_2 \\ u_1 & -u_1 & u_2 & -u_2 \end{pmatrix}, \; \bar\Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_1 & 0 & 0 \\ 0 & 0 & \sigma_2 & 0 \\ 0 & 0 & 0 & \sigma_2 \end{pmatrix}, \; \bar U = \frac{1}{\sqrt{2}}\begin{pmatrix} v_1 & -v_1 & v_2 & -v_2 \\ u_1 & u_1 & u_2 & u_2 \end{pmatrix}$$

avec $\bar U_j = \frac{1}{\bar\Sigma_{jj}} M \bar V_j = \epsilon V_j$ en remarquant que $\bar V_j = \begin{pmatrix} v_i \\ \epsilon u_i \end{pmatrix}$ avec $\epsilon = 1$ ou -1 et $\bar\Sigma_{jj} = \sigma_i$.

4. Même réponse avec $\sigma_2 = 0$.

5. Par conséquent, si $M = \bar{U}\bar{\Sigma}\bar{V}$, alors

$$\bar{V} = \frac{1}{\sqrt{2}}\begin{pmatrix} v_1 & v_1 & v_2 & v_2 & \dots & v_n & v_n \\ u_1 & -u_1 & u_2 & -u_2 & \dots & u_n & -u_n \end{pmatrix}, \bar{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \sigma_2 & \dots & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & \dots & \sigma_n & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & \sigma_n \end{pmatrix},$$

$$\bar{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} v_1 & -v_1 & v_2 & -v_2 & \dots & v_n & -v_n \\ u_1 & u_1 & u_2 & u_2 & \dots & u_n & u_n \end{pmatrix}$$

2.4 Row reduced echelon form, LU and Choleski decomposition

2.4.1 Row-reduced echelon form matrix

Definition 2.2: Partial row-echelon form

A matrix is in *partial row-echelon form* if:

1. Any rows of zeros are at the bottom of the matrix.
2. The first nonzero entry of each nonzero row is 1. This entry 1 is called a *pivot*.
3. The pivot of a lower row is to the right of the pivot of a higher row.

Definition 2.3: Row-echelon form

A matrix is in *row-echelon form* if:

1. it is in partial row-reduced echelon form
2. The other entries of a column containing a pivot are all zero. (Such a column is called a *pivotal column*.

Example 2.1

The matrices

$$A_1 = \begin{bmatrix} 0 & 1 & 3 & 1 & 4 & -1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 1 & 1 & 1 & 0 & 4 & -1 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 4 & -5 & 0 \end{bmatrix}$$

are in partial row-echelon form.
The matrices

$$A_3 = \begin{bmatrix} 0 & 1 & 3 & 0 & 4 & -1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad A_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 4 & -1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 4 & -5 & 0 \end{bmatrix}$$

are in row-echelon form. Note that they differ from respectively A_1 and A_2 only in their pivotal columns.
The matrices

$$A_5 = \begin{bmatrix} 0 & 1 & 3 & 0 & 4 & -1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad A_6 = \begin{bmatrix} 1 & 0 & 0 & 0 & 4 & -1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 4 & -5 & 0 \end{bmatrix}$$

are not. The second and fourth columns of A_1 above are pivotal columns, as are the first three third columns of A_2. The other columns of A_1 and A_2 are *nonpivotal*.

Note that if a matrix is in row-echelon form, then all the submatrices formed from its first k rows and first l columns are also in row-echelon form.

Theorem 2.10

Every matrix A can be transformed into

1. a matrix $\overline{A}$ in partial row-echelon form by a succession of row operations. That is there exists a matrix $\overline{P}$ product of transvection and dilatation such that $\overline{A} = \overline{P}A$ is in partial row-reduced echelon form.

2. a matrix $\widetilde{A}$ in row-echelon form by a succession of row operations. That is there exists a matrix $\widetilde{P}$ product of transvection and dilatation such that $\widetilde{A} = \widetilde{P}A$ is in row-reduced echelon form.

The matrix $\widetilde{A}$ in row-echelon form is unique.

Example 2.2

The pivotal 1's are in bold.

$$\begin{bmatrix} 1 & 2 & 3 & 1 \\ -1 & 1 & 0 & 2 \\ 1 & 0 & 1 & 2 \end{bmatrix} \xrightarrow[\substack{T_{21}(1)\\ T_{31}(-1)}]{} \begin{bmatrix} \mathbf{1} & 2 & 3 & 1 \\ 0 & 3 & 3 & 3 \\ 0 & -2 & -2 & 1 \end{bmatrix} \xrightarrow[D_2\left(\frac{1}{3}\right)]{} \begin{bmatrix} \mathbf{1} & 2 & 3 & 1 \\ 0 & \mathbf{1} & 1 & 1 \\ 0 & -2 & -2 & 1 \end{bmatrix} \xrightarrow[\substack{T_{32}(2)\\ D_3\left(\frac{1}{3}\right)}]{} \begin{bmatrix} \mathbf{1} & 2 & 3 & 1 \\ 0 & \mathbf{1} & 1 & 1 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix} := \overline{A}$$

$$\xrightarrow[T_{12}(-2)]{} \begin{bmatrix} \mathbf{1} & 0 & 1 & -1 \\ 0 & \mathbf{1} & 1 & 1 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix} \xrightarrow[\substack{T_{13}(1)\\ T_{23}(-1)}]{} \begin{bmatrix} \mathbf{1} & 0 & 1 & 0 \\ 0 & \mathbf{1} & 1 & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix} := \widetilde{A}$$

Note that we have obtained a matrix in a partial row-reduced echelon form after the first line's operations. The operations that we made are all associated in that case to lower triangular matrix:

$$\overline{P} = D_3\left(\frac{1}{3}\right)T_{32}(2)D_2\left(\frac{1}{3}\right)T_{31}(-1)T_{21}(1).$$

It is due to the fact that we did not requiered any permutation of lines to obtain the pivotal entries of the matrix.

Example 2.3

In that case, the first coefficient of the first column is 0, we need to use the second row of the matrix to create a non zero pivotal entry:

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 2 & 0 \end{bmatrix} \xrightarrow[T_{12}(1)]{} \begin{bmatrix} \mathbf{1} & 3 & 1 \\ 1 & 2 & 0 \end{bmatrix} \xrightarrow[T_{21}(-1)]{} \begin{bmatrix} \mathbf{1} & 3 & 1 \\ 0 & -1 & -1 \end{bmatrix} \xrightarrow[D_2(-1)]{} \begin{bmatrix} \mathbf{1} & 3 & 1 \\ 0 & \mathbf{1} & 1 \end{bmatrix} := \overline{A}$$

$$\xrightarrow[T_{12}(-3)]{} \begin{bmatrix} \mathbf{1} & 0 & -2 \\ 0 & \mathbf{1} & 1 \end{bmatrix} := \widetilde{A}$$

The matrix $\overline{P} = D_2(-1)T_{21}(-1)T_{12}(1)$ is not a triangular matrix.

2.4.2 LU decomposition of a matrix

Note that what is usually called the Gaussian pivot corresponds to a slightly different algorithm from that of row-reduction. In this algorithm, we are allowed to exchange columns of the system (which corresponds to right multiplication by transvection and dilation matrices), and we do not require the pivots of the algorithm to be equal to 1, thus only using transvections. In the case where all the leading principal minors of the initial matrix are non-zero, these operations on the columns of the matrix are not necessary. We then obtain what is called an LU decomposition of the matrix. The algorithm can be expressed as follows:

Given a square matrix $A = (a_{n,n})$ of size N, we define $A^{(0)} := A$ and the iterations are performed for $n = 1, \dots, N-1$ as follows.

On the n^{th} column of $A^{(n-1)}$, we eliminate the elements below the diagonal by performing transvections $T_{in}(l_{i,n})$ where

$$l_{i,n} := -\frac{a_{i,n}^{(n-1)}}{a_{n,n}^{(n-1)}}.$$

The coefficients $a_{n,n}^{(n-1)}$ are called the pivots of the algorithm. We obtain a matrix PA that is upper triangular, while P, the product of lower triangular transvections, is a lower triangular matrix with 1s on the diagonal.

Theorem 2.11

1. Let A be an invertible matrix. The matrix A can be expressed as $A = PLU$ where P is a permutation matrix, L is a lower triangular matrix with 1s on the diagonal, and U is an upper triangular matrix.
2. Now suppose that all the leading principal minors of the matrix A are non-zero; then there exists a unique lower triangular matrix L with 1s on the diagonal and a unique upper triangular matrix U such that $A = LU$.

Definition 2.4: Bandwidth of a matrix

Consider a sparse matrix A such that

$$a_{ij} = 0 \text{ for all } (i,j) \text{ satisfying } |i-j| > k. \tag{2.2}$$

The bandwidth of the matrix is the smaller value of k such that (2.2) is satisfied.

Remark.

The LU decomposition preserves the bandwidth of a matrix.

2.4.3 Choleski decomposition of a matrix

Theorem 2.12: Choleski

Let $A \in \mathcal{M}_n(\mathbb{R})$ with $n \geq 1$. We assume that A is symmetric and positive definite. Then, there exists a unique matrix $C \in \mathcal{M}_n(\mathbb{R})$, $C = (c_{i,j})_{i,j=1}^n$, such that:

1. C is lower triangular (i.e. $c_{i,j} = 0$ if $j > i$),
2. $c_{i,i} > 0$, for all $i \in \{1, \ldots, n\}$,
3. $A = CC^t$.

Proof. The result can be obtain either by induction on the matrix size or as a direct consequence of the LU decomposition of the matrix. □

Remark.

The Choleski decomposition preserves also the bandwidth odf a matrix.

2.4.4 Exercises

Exercise 2.15

Are the following statements true or false?

1. The matrix $B = \begin{pmatrix} 1 & -2 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$ is symmetric positive definite.
2. The matrix B above has an LU decomposition.
3. The matrix $\begin{pmatrix} 1 & -1 \\ 1 & 3 \end{pmatrix}$ can be written as C^tC.
4. The matrix $A = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}$ admits a Cholesky decomposition $A = C^tC$ with $C = \begin{pmatrix} -1 & -1 \\ 0 & -2 \end{pmatrix}$.

Exercise 2.16

Find the LU and the Choleski decomposition of the matrix $A_{dir} = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$ and observe that the bandwidth of the matrix is preserved.

Exercise 2.17

show that the matrix A, defined by the following coefficients:

$$a_{ij} = \begin{cases} -1 & \text{if } i > j, \\ 1 & \text{if } i = j \text{ or } j = n, \\ 0 & \text{otherwise,} \end{cases}$$

admits an LU decomposition, where L is a lower triangular matrix and U is an upper triangular matrix, we proceed in three steps:

2.5 QR decomposition

2.5.1 The decomposition

Theorem 2.13

Let $A \in \mathcal{M}_n(\mathbb{R})$. Then there exist an orthogonal matrix Q and an upper triangular matrix R with non-negative diagonal entries such that $A = QR$. If the matrix A is invertible, then this decomposition is unique.

Proof. In the general case, the proof relies on the use of Householder matrices, defined later. The proof is done by induction on n, with the first step consisting of transforming the first column vector of the matrix using a Householder matrix into a vector of the form $\begin{pmatrix} a \\ 0 \\ \vdots \\ 0 \end{pmatrix}$. One can refer to [Cia95] for a complete demonstration of the result.

In the case of an invertible matrix, note that if a_k and q_k denote respectively the column vectors of the matrices A and Q, then

$$\begin{aligned} a_1 &= r_{11} q_1, \\ a_2 &= r_{21} q_1 + r_{22} q_2, \\ &\vdots \\ a_n &= r_{n1} q_1 + \cdots + r_{nn} q_n \end{aligned}$$

with q_k being an orthonormal family. In other words, the QR decomposition reduces to orthonormalizing the vectors a_k. To achieve this, we use the Gram-Schmidt orthonormalization process. □

2.5.2 Application to the determination of the eigenvalues

Starting from an invertible matrix $A = A_0$, the QR method consists of the following steps:

1. We use Gram-Schmidt to find Q_0 and R_0 such that $A_0 = Q_0 R_0$, with Q_0 orthogonal and R_0 upper triangular.
2. We multiply the factors in reverse order: we set $A_1 = R_0 Q_0$.

Then we repeat this process, writing $A_1 = Q_1 R_1$ and setting $A_2 = R_1 Q_1$, and so on.

Remark.

The equation $A_0 = Q_0R_0$ implies that $Q_0^\top A = R_0$, from which it follows that $A_1 = Q_0^\top AQ_0 = Q_0^{-1}AQ_0$. In particular, $A_0, A_1, \cdots$ are all conjugate and thus have the same eigenvalues.

Example 2.4

In the following example, we pushed the iterations until the coefficients of A_k below the diagonal were less than 10^{-4}.

For the matrix

$$\begin{pmatrix} 1 & 3 & 7 \\ 2 & -1 & 2 \\ 2 & -3 & 4 \end{pmatrix},$$

the successive steps are as follows:

	Q	R	RQ
1	$\begin{bmatrix} -0.333 & 0.882 & 0.331 \\ -0.666 & 0.027 & -0.744 \\ -0.666 & -0.469 & 0.579 \end{bmatrix}$	$\begin{bmatrix} -3.000 & 1.666 & -6.333 \\ 0 & 4.027 & 4.358 \\ 0 & 0 & 3.144 \end{bmatrix}$	$\begin{bmatrix} 4.111 & 0.367 & -5.903 \\ -5.590 & -1.933 & -0.474 \\ -2.096 & -1.474 & 1.821 \end{bmatrix}$
2	$\begin{bmatrix} -0.567 & -0.677 & -0.468 \\ 0.771 & -0.237 & -0.590 \\ 0.289 & -0.695 & 0.657 \end{bmatrix}$	$\begin{bmatrix} -7.249 & -2.125 & 3.508 \\ 0 & 1.236 & 2.846 \\ 0 & 0 & 4.240 \end{bmatrix}$	$\begin{bmatrix} 3.486 & 2.977 & 6.954 \\ 1.776 & -2.274 & 1.141 \\ 1.226 & -2.951 & 2.787 \end{bmatrix}$
3	$\begin{bmatrix} -0.850 & 0.511 & 0.126 \\ -0.433 & -0.542 & -0.719 \\ -0.299 & -0.666 & 0.682 \end{bmatrix}$	$\begin{bmatrix} -4.100 & -0.663 & -7.241 \\ 0 & 4.722 & 1.075 \\ 0 & 0 & 1.962 \end{bmatrix}$	$\begin{bmatrix} 5.939 & 3.093 & -4.984 \\ -2.367 & -3.278 & -2.665 \\ -0.586 & -1.308 & 1.339 \end{bmatrix}$
4	$\begin{bmatrix} -0.925 & -0.369 & -0.086 \\ 0.368 & -0.819 & -0.438 \\ 0.091 & -0.437 & 0.894 \end{bmatrix}$	$\begin{bmatrix} -6.420 & -4.190 & 3.750 \\ 0 & 2.115 & 3.442 \\ 0 & 0 & 2.797 \end{bmatrix}$	$\begin{bmatrix} 4.737 & 4.168 & 5.747 \\ 1.094 & -3.239 & 2.151 \\ 0.255 & -1.224 & 2.502 \end{bmatrix}$
8	$\begin{bmatrix} -0.998 & -0.061 & -0.001 \\ 0.061 & -0.996 & -0.061 \\ 0.002 & -0.061 & 0.998 \end{bmatrix}$	$\begin{bmatrix} -5.631 & -4.298 & 4.838 \\ 0 & 3.265 & 2.910 \\ 0 & 0 & 2.066 \end{bmatrix}$	$\begin{bmatrix} 5.367 & 4.333 & 5.099 \\ 0.207 & -3.430 & 2.706 \\ 0.004 & -0.126 & 2.062 \end{bmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
12	$\begin{bmatrix} -1.000 & -0.009 & -0.000 \\ 0.009 & -0.999 & -0.006 \\ 0.000 & -0.006 & 1.000 \end{bmatrix}$	$\begin{bmatrix} -5.497 & -4.244 & 5.034 \\ 0 & 3.444 & 2.743 \\ 0 & 0 & 2.006 \end{bmatrix}$	$\begin{bmatrix} 5.455 & 4.264 & 5.063 \\ 0.034 & -3.462 & 2.719 \\ 0.0001 & -0.013 & 2.006 \end{bmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
20	$\begin{bmatrix} -1.000 & -0.0003 & -0.0000 \\ 0.0003 & -1.0000 & -0.0001 \\ 0.000 & -0.0001 & 1.0000 \end{bmatrix}$	$\begin{bmatrix} -5.472 & -4.242 & 5.062 \\ 0 & 3.471 & 2.714 \\ 0 & 0 & 2.000 \end{bmatrix}$	$\begin{bmatrix} 5.471 & 4.243 & 5.062 \\ 0.0009 & -3.471 & 2.714 \\ 0.0000 & -0.0002 & 2.000 \end{bmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
28	$\begin{bmatrix} -1.000 & -0.000 & -0.000 \\ 0.000 & -1.000 & -0.000 \\ 0.000 & -0.000 & 1.000 \end{bmatrix}$	$\begin{bmatrix} -5.472 & -4.242 & 5.063 \\ 0 & 3.472 & 2.713 \\ 0 & 0 & 2.000 \end{bmatrix}$	$\begin{bmatrix} 5.472 & 4.242 & 5.063 \\ 0.000 & -3.472 & 2.713 \\ 0.000 & -0.000 & 2.000 \end{bmatrix}$

After 28 iterations, the absolute values of the terms below the diagonal are strictly less than 10^{-4}; at this precision, the eigenvalues have converged. In fact, the eigenvalues are 2 and $1 \pm 2\sqrt{5}$; since $1 + 2\sqrt{5} = 5.4721359\cdots$ and $1 - 2\sqrt{5} = -3.472135954\cdots$, we see that the algorithm has indeed found these numbers with the required precision.

Thanks to the command *qr* in Python, it is extremely easy to implement this method. The following program does just that:

```python
import numpy as np

def methodQR(A,n):
    for i in range(n):
        Q,R = np.qr(A)
        A=R.dot(Q)
    return A
```

Theorem 2.14: Convergence of the QR algorithm

Let A be a matrix in $\mathcal{M}_n(\mathbb{C})$ with n distinct eigenvalues $\lambda_1,\ldots,\lambda_n$ and n associated eigenvectors $(\mathbf{v}_1,\ldots,\mathbf{v}_n)$. Assume that

$$|\lambda_1| > \cdots > |\lambda_n| > 0. \tag{2.3}$$

Let $(\mathbf{e}_1,\ldots,\mathbf{e}_n)$ be the standard basis of $\mathbb{C}^n$ and assume that for each $i = 1,\ldots,n$ the vectors

$$\mathbf{e}_1,\ldots,\mathbf{e}_i,\mathbf{v}_{i+1},\ldots,\mathbf{v}_n \tag{2.4}$$

form a basis of $\mathbb{C}^n$. Let A_k be the sequence defined by

$$A_0 = A,\; A_{k+1} = R_k Q_k \text{ where } A_k = Q_k R_k.$$

Then the diagonal entry $(A_k)_{i,i}$ of A_k converges to λ_i and each entry of A_k below the diagonal converges to 0.

2.6 Polar decomposition

Theorem 2.15

1. **Polar decomposition in $\mathcal{GL}_n(\mathbb{R})$.**

   Let $\mathcal{S}_n$ denote the set of matrices in $\mathcal{M}_n(\mathbb{R})$ that are symmetric, and $\mathcal{S}_n^+$ denote the set of matrices in $\mathcal{M}_n(\mathbb{R})$ that are symmetric and positive definite. Let $\mathcal{O}_n$ denote the set of the orthogonal matrices.

   (a) The map

   $$\begin{array}{rcl}\Psi : \mathcal{O}_n(\mathbb{R}) \times \mathcal{S}_n^+ & \to & \mathcal{GL}_n(\mathbb{R}) \\ (O,S) & \mapsto & OS\end{array}$$

   is a homeomorphism.

   (b) The map

   $$\begin{array}{rcl}\mathcal{O}_n(\mathbb{R}) \times \mathcal{S}_n & \to & \mathcal{M}_n(\mathbb{R}) \\ (O,S) & \mapsto & OS\end{array}$$

   is surjective but generally not injective.

2. **Polar decomposition in $\mathcal{GL}_n(\mathbb{C})$.**

   Let $\mathcal{H}_n$ denote the set of matrices in $\mathcal{M}_n(\mathbb{C})$ that are Hermitian, and $\mathcal{H}_n^+$ denote the set of matrices in $\mathcal{M}_n(\mathbb{C})$ that are Hermitian and positive definite. Let $\mathcal{U}_n$ denote the set of the unitary matrices.

   (a) The map

   $$\begin{array}{rcl}\mathcal{U}_n(\mathbb{R}) \times \mathcal{H}_n^+ & \to & \mathcal{GL}_n(\mathbb{C}) \\ (O,S) & \mapsto & OS\end{array}$$

   is a homeomorphism.

   (b) The map

   $$\begin{array}{rcl}\mathcal{U}_n(\mathbb{R}) \times \mathcal{H}_n & \to & \mathcal{M}_n(\mathbb{C}) \\ (O,S) & \mapsto & OS\end{array}$$

   is surjective but generally not injective.

Proof.

1. (a) ▶ Surjectivity. Let $A \in \mathcal{GL}_n(\mathbb{R})$. If $A = OS$, then $A^tA = S^2$. The matrix A^tA is symmetric and positive definite, thus diagonalizable in an orthonormal basis:

$$A^tA = P^t \operatorname{diag}(\lambda_1,\cdots,\lambda_n)P,\; \lambda_i > 0.$$

Set $S = P^t \mathrm{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_n})P$. We verify that $S \in \mathcal{S}_n^+$ and $S^2 = A^tA$. Moreover, S is a polynomial in A^tA, we just need to consider the Lagrange interpolation polynomial which satisfies $Q(\lambda_i) = \sqrt{\lambda_i}$ and we verify that $S = Q(A^tA)$. By setting $O = AS^{-1}$ we indeed have $OO^t = I_n$.

- ► Injectivity. Suppose that $O_1S_1 = O_2S_2$. The matrix $M = O_2^{-1}O_1 = S_2S_1^{-1}$ is a matrix that is both orthogonal and positive definite symmetric, so that $M = Id$. In fact, if $S \in \mathcal{S}_n^+$, then $S = P^tDP$ with D diagonal and $P \in \mathcal{O}_n$. Thus, D is a diagonal matrix with positive real coefficients that is orthogonal, meaning it must be the identity matrix.

- ► Homeomorphism. The map Ψ is clearly continuous. It remains to show the continuity of Ψ^{-1}. Let $M_p = O_pS_p$ be a sequence of matrices in $\mathcal{GL}_n(\mathbb{R})$ converging to $M = OS$ in $\mathcal{GL}_n(\mathbb{R})$. Since the group $\mathcal{O}_n(\mathbb{R})$ is compact, we can extract a subsequence of O_p that converges to a matrix $O' \in \mathcal{O}_n(\mathbb{R})$. Set $S' = O'^{-1}M$. We know that $O_{\phi(p)}^{-1}M_{\phi(p)}$ converges to $O'^{-1}M$, in other words, the sequence $S_{\phi(p)}$ converges to a matrix S' which is symmetric and positive (both symmetry and positivity are closed properties), and $\det(S') \neq 0$, hence $S' \in \mathcal{S}_n^+$. By the uniqueness of the decomposition, we have $O' = O$ and $S' = S$. Therefore, the sequence O_p has only one limit point in $\mathcal{O}_n$, so it converges to O, and consequently, S_p converges to S. We deduce that the map Ψ^{-1} is continuous.

(b) We simply need to repeat the proof of surjectivity from the previous question. Non-injectivity can be observed by noting that if $S = 0$, $OS = 0$ for any matrix $O \in \mathcal{O}_n$.

2. The proofs are analogous.

□

Chapter

3 Topology on matrices

The aim of this chapter is to master the notion of matrix norms and their link with spectral radius, to know the link between the speed of convergence of a recurrent sequence and spectral radius.

3.1 Matrix Norms

3.1.1 Induced Norms or Operator Norms

Let E be a finite-dimensional vector space of dimension n over a field $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. We equip E with a norm $\|\cdot\|$.

Proposition 3.1

Every endomorphism $u \in \mathcal{L}(E)$ is continuous, and

$$|||u||| = \sup_{x \neq 0} \frac{\|u(x)\|}{\|x\|} = \sup_{\|x\|=1} \|u(x)\| = \inf\{k > 0 \text{ such that for all } x \in E, \|u(x)\| \leq k\|x\|\}$$

defines a norm on $\mathcal{L}(E)$. Such a norm is called an **induced norm** or **operator norm**.

Proof.

- It is clear that $|||u||| \geq 0$ for all $u \in \mathcal{L}(E)$.
- If $|||u||| = 0$, then $0 = \sup_{\|x\|=1} \|u(x)\|$. Therefore, for all $x \neq 0$, $u\left(\frac{x}{\|x\|}\right) = 0$ and thus $u(x) = 0$.
- We have
$$|||u+v||| = \sup_{\|x\|=1} \|u(x)+v(x)\| \leq \sup_{\|x\|=1} \|u(x)\| + \sup_{\|x\|=1} \|v(x)\| = |||u||| + |||v|||.$$

□

Remark.
We have $|||\lambda u||| = |\lambda|\,|||u|||$ for all $\lambda \in \mathbb{K}$.

Proposition 3.2

If $|||\cdot|||$ is an induced norm associated with a norm $\|\cdot\|$ on E, then we have:

$$\|u(x)\| \leq |||u|||\,\|x\| \text{ for all } x \in E, \tag{3.1}$$

$$|||u \circ v||| \leq |||u|||\,|||v||| \text{ for all } u, v \in \mathcal{L}(E). \tag{3.2}$$

Proof.

- We use the fact that if $x \neq 0$, $u(x) = \|x\| u\left(\frac{x}{\|x\|}\right)$. Therefore,

$$\|u(x)\| = \|x\| \left\| u\left(\frac{x}{\|x\|}\right) \right\| \leq \|x\| \sup_{\|y\|=1} \left\| u(y) \right\| \leq ||| u ||| \, \|x\|.$$

- We also have

$$\|u \circ v(x)\| \leq ||| u ||| \, \|v(x)\| \leq ||| u ||| \, ||| v ||| \, \|x\|.$$

Thus, $||| u \circ v ||| \leq ||| u ||| \, ||| v |||$.

□

Definition 3.1

Any norm that satisfies (3.2) is called an **algebra norm**.

Remark.

If $||| \cdot |||$ is an induced norm, then $||| \mathrm{id} ||| = 1$.

Proposition 3.3: Common Induced Norms

We show that on $\mathbb{K}^n$:

1. The induced norm associated with the infinity norm, $\|X\|_\infty = \sup_{1 \leq i \leq n} |x_i|$, is given by:

$$||| A |||_\infty = \sup_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

2. The induced norm associated with the 1-norm, $\|X\|_1 = \sum_{1 \leq i \leq n} |x_i|$, is given by:

$$||| A |||_1 = \sup_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|.$$

3. The induced norm associated with the 2-norm, $\|X\|_2 = \left(\sum_{1 \leq i \leq n} |x_i|^2\right)^{\frac{1}{2}}$, is given by:

$$||| A |||_2 = \sqrt{\rho(A^t A)}.$$

It is noteworthy that this norm is invariant under orthonormal basis changes.

3.1.2 Norm and Spectral Radius

Proposition 3.4

Given a norm $\| \cdot \|$ on $\mathcal{M}_p(\mathbb{C})$, if $A \in \mathcal{M}_p(\mathbb{C})$, we have:

$$\rho(A) = \lim_{n \to \infty} \|A^n\|^{\frac{1}{n}}.$$

Proof. We note that it is sufficient to show this for a well-chosen norm, since by the equivalence of norms, it will hold for all norms. Indeed, if $\| \cdot \|_*$ denotes a norm equivalent to $\| \cdot \|$, for any matrix $B \in \mathcal{M}_p(\mathbb{C})$, we have:

$$C_1 \|B\|_* \leq \|B\| \leq C_2 \|B\|_*.$$

Thus,

$$C_1^{\frac{1}{n}} \|A^n\|_*^{\frac{1}{n}} \leq \|A^n\|^{\frac{1}{n}} \leq C_2^{\frac{1}{n}} \|A^n\|_*^{\frac{1}{n}}.$$

If we have shown that

$$\lim_{n \to \infty} \|A^n\|_*^{\frac{1}{n}} = \rho(A),$$

since $\lim_{n\to\infty} C_i^{\frac{1}{n}} = 1$, it follows that:

$$\lim_{n\to\infty} \|A^n\|^{\frac{1}{n}} = \rho(A).$$

Note also that if the two norms are equivalent, we have:

$$\overline{\lim_{n\to\infty}}\, \|A^n\|_*^{\frac{1}{n}} = \overline{\lim_{n\to\infty}}\, \|A^n\|^{\frac{1}{n}} \text{ and } \underline{\lim_{n\to\infty}}\, \|A^n\|_*^{\frac{1}{n}} = \underline{\lim_{n\to\infty}}\, \|A^n\|^{\frac{1}{n}}. \tag{3.3}$$

Now, suppose that $\|\cdot\|$ is an induced norm, and let $X \in \mathbb{C}^p, X \neq 0$ be such that $AX = \lambda X$ with $|\lambda| = \rho(A)$. We then have:

$$\|A^n X\| = \rho(A)^n \|X\| \leq \|A^n\| \|X\|,$$

Thus,

$$\rho(A) \leq \|A^n\|^{\frac{1}{n}} \text{ for all } n \in \mathbb{N},$$

and consequently,

$$\rho(A) \leq \underline{\lim} \|A^n\|^{\frac{1}{n}}.$$

This holds for any induced norm, and therefore for all norms, by (3.3).

Conversely, let's first assume that $\rho(A) < 1$. We will show that $A^n \to 0$ as $n \to \infty$. Using the Dunford decomposition, $A = D + N$, where D is a diagonalizable matrix and N is a nilpotent operator such that $DN = ND$. We have seen that this decomposition can be written more precisely in a basis adapted to the decomposition into characteristic subspaces of A as follows:

$$A = P^{-1} \begin{pmatrix} \lambda_1 I_{q_1} + N_{q_1} & | & & & \\ ----- & & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_k I_{q_k} + N_{q_k} \end{pmatrix} P, \quad |\lambda_i| < 1.$$

We observe that:

$$(\lambda_i I_{q_i} + N_{q_i})^n = \lambda_i^n I_{q_i} + C_n^1 \lambda_i^{n-1} N_i + \cdots + C_n^{q_i - 1} \lambda_i^{n-q_i-1} N_{q_i}^{q_i-1} \underset{n\to 0}{\longrightarrow} 0.$$

Hence, $A^n \to 0$ as $n \to \infty$.

Now, without assuming $\rho(A) < 1$, we define $B = \frac{A}{\rho(A)+\epsilon}$ for some fixed $\epsilon > 0$. From the above, since $\rho(B) = \frac{\rho(A)}{\rho(A)+\epsilon} < 1$, we have $B^n \to 0$. Therefore, there exists n_0 such that for all $n \geq n_0$,

$$\|B^n\| < 1,$$

and hence, for all $n \geq n_0$,

$$\|A^n\| \leq (\rho(A) + \epsilon)^n.$$

We deduce that:

$$\overline{\lim} \|A^n\|^{\frac{1}{n}} \leq \rho(A) + \epsilon, \forall \epsilon > 0.$$

Thus,

$$\overline{\lim} \|A^n\|^{\frac{1}{n}} \leq \rho(A).$$

□

Remark.

Another approach is to show that for every $\epsilon > 0$, there exists a change of basis (depending on ϵ) such that:

$$A = P_\epsilon^{-1} \begin{pmatrix} \lambda_1 I_{q_1} + \epsilon N_{q_1} & | & & & \\ ----- & & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_k I_{q_k} + \epsilon N_{q_k} \end{pmatrix} P_\epsilon = P_\epsilon^{-1} A^\epsilon P_\epsilon$$

where N_{q_i} is the $q_i \times q_i$ matrix with coefficients $n_{ij} = \delta_{j,i+1}$.
We define $\|X\|_\epsilon = \|P_\epsilon X\|_\infty$ and $|||\cdot|||_\epsilon$ as the associated induced norm. Then,

$$\|AX\|_\epsilon = \|A^\epsilon P^\epsilon X\|_\infty,$$

so that $|||A|||_\epsilon = |||A^\epsilon|||_\infty \leq \rho(A)+\epsilon$.
Using the fact that $|||\cdot|||_\epsilon$ is an induced norm, we have:

$$|||A^n|||_\epsilon^{\frac{1}{n}} \leq |||A|||_\epsilon \leq \rho(A)+\epsilon.$$

Thus,

$$\overline{\lim_{n\to\infty}} |||A^n|||_\epsilon^{\frac{1}{n}} \leq \rho(A)+\epsilon.$$

Since all norms are equivalent, it follows that for any $\epsilon > 0$, we have:

$$\rho(A) \leq \underline{\lim_{n\to\infty}} |||A^n|||_\infty^{\frac{1}{n}} \leq \overline{\lim_{n\to\infty}} |||A^n|||_\infty^{\frac{1}{n}} \leq \rho(A)+\epsilon.$$

We conclude that:

$$\lim_{n\to\infty} |||A^n|||_\infty^{\frac{1}{n}} = \rho(A),$$

and thus, for any norm $\|\cdot\|$ on $\mathcal{M}_p(\mathbb{C})$, we have $\lim_{n\to\infty} \|A^n\|_\infty^{\frac{1}{n}} = \rho(A)$.

3.1.3 Application to the convergence of the recurrent sequences

3.1.4 Exercises

Chapter 4 Solving linear systems

The aim of this chapter is to state the main properties of classical methods for solving linear systems: cost of the Gauss algorithm or LU methods, interest of iterative methods.

4.1 Direct solvers

Let $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and $b \in \mathbb{R}^m$. We want to solve the linear system

$$\text{Find } X \in \mathbb{R}^n \text{ such that } AX = b.$$

Let us define

$$\mathcal{S} = \{X \text{ such that } AX = b\}.$$

We know that $\mathcal{S} \neq \emptyset$ if and only if $b \in \mathrm{Im}(A)$ and in that case

$$\mathcal{S} = X_0 + \mathrm{Ker}(A)$$

where X_0 is a particular solution.

4.1.1 Case where A is in a row reduced echelon form

Theorem 4.1

Let $A \in \mathcal{M}_{m,n}(\mathbb{K})$ be a totally echelon matrix of A and let i_{max} be the number of non-zero rows of A and $J = \{j_i, i \leq i_{max}$ such that $a_{ij_i} = 1\}$. Then

- $\mathcal{S} = \emptyset$ if $(b)_i \neq 0$ for some $i > i_{max}$.
- $\mathcal{S} \neq \emptyset$ if $(b)_i = 0$ for all $i > i_{max}$. In this case, $\mathcal{S} = X^0 + \ker(A)$ where

$$X^0 = (x_j^0)_j, \text{ with } \begin{cases} x_{j_i}^0 = (b)_i, \, i \leq i_{max} \\ x_j^0 = 0, \, \forall j \notin J \end{cases}$$

Example 4.1

Let $A = \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ and $b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$. Then $\mathcal{S} \neq \emptyset$ if and only if $b_3 = 0$. We see that $J = \{1, 3\}$, $X_0 = \begin{pmatrix} b_1 \\ 0 \\ b_2 \\ 0 \end{pmatrix}$ and we obtain also easily

$$\ker(A) = \left\{ \begin{pmatrix} -2x_2 \\ x_2 \\ -x_4 \\ x_4 \end{pmatrix}, x_2, x_4 \in \mathbb{R} \right\} = \text{Span}\left(\begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix} \right).$$

Remark.

As the cost of the obtention of the row reduced echelon form of a matrix is $\mathcal{O}(\max(n,m)^3)$, the cost of the resolution of a general linear system is also $\mathcal{O}(\max(n,m)^3)$. If the matrix A is already triangular, the cost of the resolution of the linear system becomes $\mathcal{O}(\max(n,m)^2)$.

4.1.2 Case where we should use LU or Choleski decomposition

Note that if we know the LU decomposition of the matrix A, solving the system $AX = b$ consists in solving two triangular systems

$$\begin{cases} LY = b \\ UX = Y \end{cases}$$

Similarly, if we know the choleski decomposition of a matrix A solving the system $AX = b$ consists in solving two triangular systems

$$\begin{cases} C^tY = b \\ CX = Y \end{cases}$$

In both case, the main cost is due to the decomposition the cost of the resolution of the two linear systems being in $\mathcal{O}(\max(n,m)^2)$. When we have several systems to solve with the same matrix A: $AX_i = b_i$ it is worthly to use the LU or the Choleski decomposition.

4.1.3 Case where we should use least square method

When $\mathcal{S} = \emptyset$, we can try to find a vector $X \in \mathbb{R}^n$ such that

$$||AX - b|| = \min_{Y \in \mathbb{R}^n} ||AY - b||$$

where $||.||$ denotes the euclidian norm. We say that we solve the system by the least square method. You will see in Chapter **??** that the problem admits at least a solution that solve the linear system called **the normal equations**

$$A^tAX = A^tb.$$

4.2 Approximate Resolution of Solutions of a Linear System

4.2.1 The principle

For simplicity, we will assume in this section that A is a matrix in $\mathcal{GL}_n(\mathbb{R})$ and $b \in \mathbb{R}^n$. We are looking for the solution X of the system

$$(S): \quad AX = b,$$

as the limit of a sequence defined recursively:

$$X^{k+1} = BX^k + C, \quad X^0 = X_0 \tag{4.1}$$

There exists two main families of method. In the case of symmetric positive definite matrix, we have seen that solving $AX = b$ is equivalent to solve the minimizing problem $J(X) = min_{Y \in \mathbb{R}^n} J(Y)$ with $J(Y) = \frac{1}{2}(AY, Y) - (b, Y)$. The first family of methods is a gradient method that you will see in the last chapter.

For the second family of method, we suppose that the matrix A can be decomposed as $A = M - N$, where M is a matrix that is "easy" to invert (diagonal or triangular, for example). The system $AX = b$ can then be rewritten equivalently as $MX = NX + b$, or as $X = M^{-1}NX + M^{-1}b$. By setting $B = M^{-1}N$ and $C = M^{-1}b$, we see that if the sequence X^k defined by (4.1) converges, then it converges to the unique solution of the system $AX = b$ and it converges if $\rho(B) < 1$ as state in proposition **??**.

4.2.2 Classical Methods

We decompose A as follows: $A = D - E - F$ where D is the diagonal of A, E is lower triangular, and F is upper triangular.

- **Jacobi Method** $B = J = D^{-1}(E+F)$ and $C = D^{-1}b$.
- **Gauss-Seidel Method** $B = G = (D-E)^{-1}F$ and $C = (D-E)^{-1}b$.
- **Relaxation Method** $B = S_\omega = (\frac{1}{\omega}D - E)^{-1}((\frac{1}{\omega} - 1)D + F)$ and $C = (\frac{1}{\omega}D - E)^{-1}b$.

4.2.3 Convergence Results

Proposition 4.2

If A is symmetric and positive definite, M is invertible, and $M^t + N$ is positive definite, then the iterative method converges.

Proof. Since A is a symmetric positive definite matrix, the mapping $(x, y) \mapsto (Ax, y)$ defines an inner product on $\mathbb{R}^n$, denoted $||\cdot||_A$ for the associated norm, and $||\cdot||_A$ for the induced matrix norm. To show that the iterative method converges, it is sufficient to show that $||M^{-1}N||_A < 1$, which will imply $\rho(M^{-1}N) < 1$ and hence the result. We must therefore show that for all $x \in \mathbb{R}^n$ such that $||x||_A = 1$,

$$||M^{-1}Nx||_A < 1,$$

or equivalently,

$$||x - M^{-1}Ax||_A \leq 1.$$

Let $y = M^{-1}Ax$, then $Ax = My$ and $x^tAy^tM^t$, hence

$$\begin{aligned} ||x-y||_A^2 &= \underbrace{||x||_A}_{=1} - x^tAy - y^tAx + y^tAy \\ &= 1 - y^tM^ty - y^tMy + y^tAy \\ &= 1 - y^t(M^t + M - A)y. \end{aligned}$$

We use the assumption that $M^t + N$ is positive definite to conclude. □

Consequence: If A is symmetric and positive definite, the Gauss-Seidel and relaxation methods converge if $\omega \in]0, 2[$. Indeed, if A is symmetric, $E^t = F$, and therefore $M^t + N = \frac{1}{\omega}D - E^t + \left(\frac{1}{\omega} - 1\right)D + F = \left(\frac{2}{\omega} - 1\right)D$. This matrix is symmetric and positive definite if $\omega \in]0, 2[$.

Note that this result does not generally provide information about the Jacobi method.

Proposition 4.3

If A is tridiagonal, then the Jacobi and Gauss-Seidel methods are of the same nature. The Gauss-Seidel method converges twice as fast.

Proof. We write $A = D - (E + E^t)$, $J = D^{-1}(E + E^t)$ and $G = (D - E)^{-1}E^t$. We note that since A is positive definite, the matrix D is invertible. Let λ be an eigenvalue of J, $\det(D^{-1}(E + E^t) - \lambda I) = 0$, or equivalently, $\det(\varepsilon D_\varepsilon^{-1}D(D^{-1}E + E^t - \lambda I)D_\varepsilon) = \det\left(E + \varepsilon^2E^* - \lambda\varepsilon D\right) = 0$ for any $\varepsilon \neq 0$ and $D_\varepsilon = \text{diag}(\varepsilon^{i-1})$. Choosing $\varepsilon = \lambda$, we see that if λ is an eigenvalue of the Jacobi matrix J, then λ^2 is an eigenvalue of the Gauss-Seidel matrix. Conversely, if μ is an eigenvalue of G, $\det((D - E)^{-1}E^t - \mu I) = 0$, or equivalently, $\det(\varepsilon D_\varepsilon^{-1}(D - E)((D - E)^{-1}E^t - \mu I)D_\varepsilon) = \det(\mu\varepsilon^2E + E^t - \mu\varepsilon D) = 0$. If we choose ε such that $\mu\varepsilon^2 = 1$, then $\varepsilon\mu = \sqrt{\mu}$ is an eigenvalue of J. We conclude that $\rho(G) = \rho(J)^2 < \rho(J)$. □

Proposition 4.4

If A has an invertible diagonal, then the relaxation method diverges for $\omega \notin]0,2[$. If A is tridiagonal and symmetric positive definite, then the relaxation method converges for $\omega \in]0,2[$.

Proof. Let λ_i be the eigenvalues of the matrix S_ω. We recall that

$$\prod_{i=1}^{n} \lambda_i = \frac{\det\left(\left(\frac{1}{\omega}-1\right)D+E^t\right)}{\det\left(\frac{1}{\omega}D-E\right)} = \frac{\prod_{i=1}^{n}\left(\frac{1}{\omega}-1\right)d_i}{\prod_{i=1}^{n}\frac{1}{\omega}d_i} = (1-\omega)^n.$$

We deduce that $\rho(S_\omega)^n \geq \prod_{i=1}^{n}|\lambda_i| = |\omega-1|^n$. Consequently, if $\omega \notin (0,2)$, then $\rho(S_\omega) > 1$.

If $\omega \in (0,2)$, we have already seen that the relaxation method converges, in other words $\rho(S_\omega) < 1$. Let λ be an eigenvalue of S_ω satisfying

$$\det\left(\left(\frac{1}{\omega}-1\right)D+E^t-\lambda\left(\frac{1}{\omega}D-E\right)\right) = \det\left(E^t+\lambda E-\left((\lambda-1)\frac{1}{\omega}+1\right)D\right) = 0$$

If the matrix is tridiagonal, we obtain

$$\det\left(E^t+\lambda\epsilon^2 E-\epsilon\left((\lambda-1)\frac{1}{\omega}+1\right)D\right) = 0$$

If $\lambda\epsilon^2 = 1$, we see that λ being an eigenvalue of S_ω implies that $\mu^{\pm}(\lambda,\omega) = \pm\frac{1}{\sqrt{\lambda}}\left((\lambda-1)\frac{1}{\omega}+1\right)$ is an eigenvalue of J. We deduce that if $\mu = \mu^{\pm}(\lambda,\omega)$,

$$\lambda\omega^2\mu^2 = (\lambda+\omega-1)^2 \qquad \text{or } \lambda^2+\left(2(\omega-1)-\omega^2\mu^2\right)\lambda+(\omega-1)^2$$

or equivalently

$$\lambda^{\pm}(\omega,\mu) \quad = \quad \frac{1}{2}\left(\omega^2\mu^2-2(\omega-1)\right) \pm \omega\mu\sqrt{(\omega^2\mu^2-4\omega+4)}.$$

In other words,

$$\rho(S_\omega) = \max_{\mu\in Sp(J)}(\lambda^{\pm}(\omega,\mu)).$$

We can then verify that $\rho(J)$ being fixed, there exists a value $\omega_0 = \frac{2}{1+\sqrt{1-\rho(J))^2}}$ such that

$$\rho(S_{\omega_0}) \leq \rho(S_\omega), \forall\omega.$$

See the books [?], [Ser02] for more details. □

Remark.

Iterative methods are often used for large sparse matrices, resulting in significant storage savings.

4.3 Exercises

Exercise 4.1

Consider the matrix $A_n \in \mathcal{M}_n(\mathbb{R})$ whose coefficients are $(A_n)_{ij} = \min(i,j)$.

1. Show that A_n admits a Choleski decomposition that should be determined.

2. Solve the system $A_3X = \begin{pmatrix}3\\5\\6\end{pmatrix}$.

Exercise 4.2

Etudier la convergence de la méthode de Jacobi pour la matrice $A_{dir} + \alpha Id$ pour $\alpha \geq 0$ et

$$A_{dir} = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}.$$

Exercise 4.3

Let $a \in \mathbb{R}$ and

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

Show that A is symmetric positive definite if and only if $-1/2 < a < 1$, and that the Jacobi method converges if and only if $-1/2 < a < 1/2$.

1. We note that A is clearly symmetric and has a simple eigenvalue $1 + 2a$ and a double eigenvalue $1 - a$. Its eigenvalues are therefore strictly positive if $-1/2 < a < 1$.
2. The associated Jacobi method is written as $M = Id$,

   $$X_{k+1} = BX_k + b \text{ with } B = \begin{pmatrix} 0 & -a & -a \\ -a & 0 & -a \\ -a & -a & 0 \end{pmatrix}$$

   The eigenvalues of B are a (double) and $-2a$. Therefore, $\rho(B) = 2|a|$ and the method converges if and only if $-1/2 < a < 1/2$.

Thus, there exist SPD matrices for which the Jacobi method does not converge.

Exercise 4.4: Matrix conditionning and linear systems

Let $A \in \mathcal{M}_N(\mathbb{R})$ be an invertible matrix, and $b \in \mathbb{R}^N$, $b \neq 0$. We equip $\mathbb{R}^N$ with a norm $|\cdot|$, and $\mathcal{M}_N(\mathbb{R})$ with the induced norm. Let $\delta_A \in \mathcal{M}_N(\mathbb{R})$ and $\delta_b \in \mathbb{R}^N$. We assume that $||\delta_A|| < \dfrac{1}{||A^{-1}||}$.

1. Show that the matrix $(Id + A^{-1}\delta_A)$ is invertible and verify that $||Id + A^{-1}\delta A|| \leq \frac{1}{1-||A^{-1}||||\delta A||}$.
2. Deduce that the matrix $A + \delta A$ is invertible.
3. Let x be the solution of $Ax = b$ and $\bar{x} = x + \delta_x$ the solution of $(A + \delta A)\bar{x} = b + \delta b$. Show that

   $$\frac{||\delta_x||}{||x||} \leq \frac{\operatorname{cond}(A)}{1 - ||A^{-1}||||\delta_A||}\left(\frac{||\delta_b||}{||b||} + \frac{||\delta_A||}{||A||}\right)$$

   where $\operatorname{cond}(A) = ||A||||A^{-1}||$.

1. Recall that the matrix $Id + B$ is invertible if $||B|| < 1$, and its inverse is then given by $\sum_{n\geq 0} B^n$. We then observe that the matrix $A + \delta A$ is invertible if and only if the matrix $A^{-1}(A + \delta A) = Id + A^{-1}\delta A$ is invertible. Now, by hypothesis, we have

   $$||A^{-1}\delta A|| \leq ||A^{-1}||||\delta A|| < 1,$$

   hence, the matrix $A + \delta A$ is invertible.

2. We have

$$(A + \delta A)(x + \delta x) = b + \delta b = Ax + \delta b,$$

which implies

$$\delta x = A^{-1}(Id + A^{-1}\delta A)^{-1}(\delta b - \delta A x).$$

We observe that $||(Id + A^{-1}\delta A)^{-1}|| \leq \frac{1}{1-||A^{-1}||||\delta A||}$ and that $||b|| = ||Ax|| \leq ||A||||x||$. Consequently, we have

$$||\delta x|| \leq \frac{||A^{-1}||}{1 - ||A^{-1}||||\delta A||}(||\delta b|| + ||\delta A||||x||) \leq \frac{||A^{-1}||}{1 - ||A^{-1}||||\delta A||}\left(\frac{||\delta b||}{||b||} + \frac{||\delta A||}{||A||}\right)||A||||x||.$$

We conclude by recalling that $\mathrm{cond}(A) = ||A||||A^{-1}||$.

Chapter 5 Introduction to optimization in data science

5.1 Statistical Inference

5.1.1 Statistical Model

Statistical inference. Statistical inference is an approach in statistics aimed at obtaining general information about a subject of study based on partial observations of that object. The observations are data that can take various forms (variables, vectors, signals, images, etc.) and can be collected in different ways (surveys, cohort studies in medicine, experimental or numerical simulations in science, web scraping, data collection via smartphone apps, etc.). In inferential statistics, these data are generally viewed as realizations of random objects: We define the space $\mathcal{X}$ of the values that the data from a study can take and equip it with a sigma-algebra $\mathcal{F}$. On this sigma-algebra, we define a family $\mathcal{P}$ of probability laws, which aim to describe the random phenomena that produced the data.

Definition 5.1: Statistical Model

The triplet $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ is called a statistical model.

The law governing the random phenomena that produced the data is unknown, or at least partially known. The objective of statistical inference is to determine this law or some of its characteristics by using the data.

Parametric estimation. In a common approach, the family of probability laws $\mathcal{P}$ is parameterized by a set of parameters θ belonging to a set Θ:

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\},$$

where P_θ is a probability law on $\mathcal{F}$.

Statistical inference then consists of specifying the parameters of the model that best describe the data, known as parametric estimation.

i.i.d. samples In some cases, observations consist of an indexed set $x_1, \cdots, x_n$ of values related to the observation of the same phenomenon under similar conditions. We can then assume that each of these data points x_i is the realization of a random variable X_i (or other random objects) distributed according to the same law P. Furthermore, when conditions permit, we can sometimes assume that the variables X_i are independent. This leads to a particular statistical model, which is the independent and identically distributed (i.i.d.) sample model of law P. Formally, this model is described as follows.

Definition 5.2: i.i.d. Sample Statistical Model

Let $\mathcal{X}$ be an observation space and $\mathcal{P}$ a family of probability laws defined on a sigma-algebra $\mathcal{F}$ on $\mathcal{X}$. For $n \in \mathbb{N}^*$, the statistical model defined by the triplet $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, \{P^{\otimes n}, P \in \mathcal{P}\})$ is an i.i.d. sample model of law P.

In a sample model, the joint law of the sample $(X_1, \cdots, X_n)$ is expressed, for any n-tuple $(A_1, \cdots, A_n)$ of $\mathcal{F}^{\otimes n}$, as follows:

$$\mathbb{P}(X_1 \in A_1, \cdots, X_n \in A_n) = \prod_{i=1}^{n} P(A_i).$$

5.1.2 Likelihood

We now focus on parameterized statistical models that have a likelihood function.

Definition 5.3: Likelihood

Let $(\mathcal{X}, \mathcal{F}, \{P_\theta, \theta \in \Theta\})$ be a parameterized statistical model. If there exists a positive σ-finite measure ν on $\mathcal{F}$ and a function $\mathcal{L}$ from $\Theta \times \mathcal{X}$ to $\mathbb{R}^+$ such that

$$\forall A \in \mathcal{F},\ P_\theta(A) = \int_A \mathcal{L}(\theta; x) d\nu(x),$$

then the statistical model is said to be dominated by the measure ν. The function $\mathcal{L}$ is called the likelihood of the model.

i.i.d sample with a density. Consider now an i.i.d. sample $X_1, \cdots, X_n$ i.i.d. a parameterized law P_θ. Suppose that P_θ is a continuous probability law with density f_θ with respect to the Lebesgue measure on $\mathcal{X} = \mathbb{R}$. For all $A \in \mathcal{F}$, we have:

$$P_\theta(A) = \int_A f_\theta(x) dx.$$

Thus, for any n-tuple $(A_1, \cdots, A_n)$ of $\mathcal{F}^{\otimes n}$, the joint law of the sample $(X_1, \cdots, X_n)$ is given by:

$$\mathbb{P}(X_1 \in A_1, \cdots, X_n \in A_n) = \int_{A_1 \times \cdots \times A_n} \prod_{i=1}^{n} f_\theta(x_i) dx_1 \cdots dx_n.$$

Proposition 5.1

A i.i.d. sample model with density f_θ is dominated by the Lebesgue measure on $\mathbb{R}^n$ and has the likelihood function:

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

i.i.d sample with a discrete distribution. Let us now assume that P_θ is a discrete distribution on a countable space $\mathcal{X}$. In this case, for $A \in \sigma(\mathcal{X})$,

$$P_\theta(A) = \sum_{x \in \mathcal{X}} \delta_x(A) P_\theta(x), \tag{5.1}$$

where δ_x is a measure on $\sigma(\mathcal{X})$ such that $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. Let us define $\nu = \sum_{x \in \mathcal{X}} \delta_x$, the counting measure on $\sigma(\mathcal{X})$. We can then write

$$P_\theta(A) = \int_A P_\theta(x) d\nu(x). \tag{5.2}$$

Thus, for any n-tuple $(A_1, \cdots, A_n)$ of $\sigma(\mathcal{X})^{\otimes n}$, the joint distribution of the sample $(X_1, \cdots, X_n)$ is of the form

$$\mathbb{P}(X_1 \in A_1, \cdots, X_n \in A_n) = \prod_{i=1}^{n} P_\theta(A_i) = \prod_{i=1}^{n} \int_{A_i} P_\theta(x_i) d\nu(x_i) = \int_{A_1 \times \cdots \times A_n} \prod_{i=1}^{n} P_\theta(x_i) d\nu(x_1) \cdots d\nu(x_n). \tag{5.3}$$

We obtain the following result.

Proposition 5.2

An i.i.d. sample model with discrete law P_θ is dominated by the counting measure on $\sigma(\mathcal{X})^{\otimes n}$ and has the likelihood function

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} P_\theta(x_i). \tag{5.4}$$

5.2 Maximum Likelihood Estimation

Let us consider a parametric statistical model $(\mathcal{X}, \sigma(\mathcal{X}), \{P_\theta, \theta \in \Theta\})$. In inferential statistics, the parameters of the model are estimated by functions that depend only on the observations or known parameters. Formally, estimators are measurable functions from $(\mathcal{X}, \sigma(\mathcal{X}))$ to Θ equipped with a sigma-algebra.

When the statistical model is dominated and has a likelihood $\mathcal{L}$, there exists a systematic method to construct estimators for the parameters. This method is known as maximum likelihood estimation, which is defined as follows.

Definition 5.4: Maximum Likelihood Estimator (MLE)

Let $(\mathcal{X}, \sigma(\mathcal{X}), \{P_\theta, \theta \in \Theta\})$ be a dominated statistical model with likelihood $\mathcal{L}$. Suppose that, for any fixed x in $\mathcal{X}$, the function $\theta \mapsto \mathcal{L}(\theta; x)$ has a maximum on Θ and that this maximum is reached at a unique point $\hat{\theta}(x)$ in Θ. Then, the random function $\hat{\theta}(X)$ is the maximum likelihood estimator (MLE) of θ.

In the case of an i.i.d. sample model with discrete law or density, the likelihood takes the form

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} \mathcal{L}_0(\theta; x_i). \tag{5.5}$$

Over the set of points $(x_1, \cdots, x_n)$ where the likelihood is strictly positive, we often reduce the problem of maximizing $\mathcal{L}$ to minimizing the function

$$\ell(\theta; x_1, \cdots, x_n) = -\log_e(\mathcal{L}(\theta; x_1, \cdots, x_n)), \tag{5.6}$$

where $\log_e(\mathcal{L}(\theta; x_1, \cdots, x_n))$ is often referred to as the (negative) log-likelihood. Indeed, since the function $t \mapsto -\log_e(t)$ is strictly decreasing, the points where $\mathcal{L}$ is maximal are those where ℓ is minimal. The advantage of using log-likelihood is that it transforms the product of functions into a sum of functions:

$$\ell(\theta; x_1, \cdots, x_n) = \sum_{i=1}^{n} \ell_0(\theta; x_i), \tag{5.7}$$

with $\ell_0(\theta; x_i) = -\log_e(\mathcal{L}_0(\theta; x_i))$.

Optimization problems related to finding the MLE are quite varied. Let's see a few examples with sample models.

Example 5.1: Bernoulli Sample

Here, $\theta = p$ and $\Theta = (0, 1)$. The likelihood of this model is given by

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}, \tag{5.8}$$

and its log-likelihood (inverse) is

$$\ell(\theta; x_1, \cdots, x_n) = -\sum_{i=1}^{n} x_i \log(\theta) - (n - \sum_{i=1}^{n} x_i)\log(1-\theta). \tag{5.9}$$

In this model, the maximum likelihood estimator exists and is unique. It is given by $\hat{\theta} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. This will be justified in an exercise.

Example 5.2: Normal Sample

Here, $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times \mathbb{R}^+$. The likelihood of this model is given by

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right), \tag{5.10}$$

and its negative log-likelihood is

$$\ell(\theta; x_1, \cdots, x_n) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{5.11}$$

The maximum likelihood estimator is $\hat{\theta} = (\bar{X}_n, S_n^2)$ with $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. This will be justified in an exercise.

Example 5.3: Gamma Sample

The parameters of this distribution satisfy $a > 0$ and $b > 0$. The density of the distribution is given by

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \mathbb{1}_{x \geq 0}, \tag{5.12}$$

where $\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x}\, dx$. In this case, the maximum likelihood estimator exists but does not have an analytical expression. Numerical methods are used to approximate it.

Example 5.4: Sample of Gaussian Vectors

Let's consider a sample of vectors of size p from the distribution $\mathcal{N}(\mu, \Sigma)$. In this case, $\theta = (\mu, \Sigma)$ and $\Theta = \mathbb{R}^p \times \mathcal{M}_p^+$, where $\mathcal{M}_p^+$ denotes the space of symmetric positive definite matrices of size $p \times p$. The likelihood of this model can be expressed as

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \frac{1}{(2\pi)^{np/2} \det \Sigma^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right).$$

It achieves a maximum at $\hat{\theta} = (\hat{\mu}, \widehat{\Sigma})$ with $\hat{\mu} = \bar{X}_n$ and

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T.$$

Remark.

A maximum likelihood estimator is constructed by solving an optimization problem. For the estimator to be well-defined, the problem must have a unique solution. The solution to the problem is not always explicit. Numerical methods are most often used to approximate it.

5.3 Bayesian Inference

Bayesian inference is a statistical approach to estimation in which the parameters to be estimated are considered as random variables. The classical statistical model $(\mathcal{X}, \sigma(\mathcal{X}), \{P_\theta, \theta \in \Theta\})$ is complemented by a distribution model on the parameters θ: we assume that the parameters θ are realizations of random variables T taking values in a space Θ and that they are distributed according to a fixed distribution P_T, called the prior distribution. This distribution may have a density f_T with respect to a measure μ.

Remark.

The Bayesian approach can be used to incorporate known information about the parameters of a model. It is also useful for addressing high-dimensional problems where the number of parameters is greater than the number of observations.

In a Bayesian model, the probability law of the observations is interpreted conditionally on the values of the parameters:

$$\forall\, A \in \sigma(\mathcal{X}), \quad P_\theta(A) = \mathbb{P}(X \in A \mid T = \theta).$$

If the statistical model is dominated, we additionally have

$$P_\theta(A) = \int_A \mathcal{L}(\theta, x)\, d\nu(x).$$

The marginal law of the observations is then obtained through integration:

$$\mathbb{P}(X \in A) = \int_\Theta \mathbb{P}(X \in A \mid T = \theta)\, f_T(\theta)\, d\mu(\theta).$$

We can verify that

$$\mathbb{P}(X \in A) = \int_A f_X(x)\, d\nu(x),$$

where f_X is a function called the marginal likelihood, defined by

$$f_X(x) = \int_\Theta \mathcal{L}(\theta; x)\, f_T(\theta)\, d\mu(\theta).$$

Finally, the law of the variable X given T, also called the posterior law, is characterized by a function $f_{T|X=x}$, known as the posterior likelihood. This function can be expressed as

$$f_{T|X=x}(\theta) = \frac{f_{X,T}(x,\theta)}{f_X(x)} = \frac{\mathcal{L}(\theta; x)\, f_T(\theta)}{f_X(x)}$$

using Bayes' theorem.

In this framework, one way to estimate the parameters is to maximize the *posterior* likelihood. The resulting estimator takes into account both the observations and the prior information provided by the distribution law on the parameters.

Definition 5.5: Maximum A Posteriori Estimator (MAP)

Let us consider a Bayesian model defined by the statistical model $(\mathcal{X}, \sigma(\mathcal{X}), \{P_\theta, \theta \in \Theta\})$ and a distribution model on the parameters θ. Suppose that for every fixed x in $\mathcal{X}$, the posterior likelihood $\theta \to f_{T|X=x}(\theta)$ has a maximum over Θ and that this maximum is attained at a unique point $\hat{\theta}(x)$ in Θ. Then, the random function $\hat{\theta}(X)$ is the maximum a posteriori (MAP) estimator of θ.

Just like for the maximum likelihood estimator (MLE), the search for the MAP leads to a wide variety of optimization problems.

Example 5.5

See the reference to `exo-map`.

5.4 Regression Problems

5.4.1 Generalities

A regression problem consists of establishing a link between a variable y, called the response variable, and a set of variables $x = (x^1, \cdots, x^p)$, known as explanatory variables, based on a set of observations $(x_i^1, \cdots, x_i^p, y_i)_{i=1}^n$ of these variables. In this section, we will assume that the variables y_i are realizations of random variables Y_i and that the x_i are fixed variables.

When the response variable is quantitative, a common model of the relationship between Y and x is the additive model

$$Y_i = g(x_i^1, \cdots, x_i^p) + \epsilon_i, \quad i = 1, \cdots, n,$$

where g is a function and ϵ_i is a random variable that accounts for the randomness in the relationship between the variables.

To tackle a regression problem, we define a criterion J to measure the discrepancies between the observations y_i and the model $g(x_i^1, \cdots, x_i^p)$. In the context of an additive model involving independent and identically distributed (i.i.d.) Gaussian random variations ϵ_i, a suitable criterion is the least squares criterion

$$J(g) = \sum_{i=1}^{n} (y_i - g(x_i))^2.$$

We also choose a function space $\mathcal{J}$. We seek in this space a function $\hat{g}$ that minimizes the criterion J over $\mathcal{J}$.

In the following, we will present in more detail two particular cases of regression: linear regression and logistic regression.

5.4.2 Linear Regression

Linear regression applies when the variables y and x are both quantitative (however, categorical explanatory variables can be integrated into linear regression by transforming them into quantitative variables).

The linear regression model is an additive model where the function that relates y to x is assumed to be affine. It can be expressed as

$$Y_i = \theta_0 + \sum_{j=1}^{p} \theta_j x_i^j + \epsilon_i, \quad i = 1, \cdots, n,$$

where $\theta = (\theta_j)_{j=0}^p$ groups a set of real parameters, called regression parameters. The function

$$g_\theta(x_i) = \theta_0 + \sum_{j=1}^{p} \theta_j x_i^j$$

is a parameterized function model that explains Y in terms of x. By taking the convention $x_i^0 = 1$ and setting $x_i = (x_i^j)_{j=0}^p$ and $\theta = (\theta_j)_{j=0}^p$, we can write

$$g_\theta(x_i) = \theta^T x_i = \langle \theta, x_i \rangle$$

with the inner product $\langle \cdot, \cdot \rangle$ of $\mathbb{R}^{p+1}$.

The linear regression model is accompanied by assumptions about the random variations ϵ_i. In its simplest version, it is assumed that the variations ϵ_i are centered, with a common variance σ^2 and are uncorrelated. If we add the assumption that the ϵ_i are normally distributed, the observations Y_i are then independent (but not identically distributed) with respective distributions

$$\mathcal{N}\left(\langle \theta, x_i \rangle, \sigma^2\right).$$

Furthermore, the joint probability of the observations $Y = (Y_1, \cdots, Y_n)^T$ can be written, for any A in $\sigma(\mathbb{R}^n)$, as

$$\mathbb{P}(Y \in A) = \int_A \mathcal{L}(\theta, \sigma^2; y_1, \cdots, y_n)\, dy_1 \cdots dy_n,$$

with

$$\mathcal{L}(\theta, \sigma^2; y_1, \cdots, y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle \theta, x_i \rangle)^2 \right). \tag{5.13}$$

Thus, the observations Y are described by a statistical model with parameters (θ, σ^2) in $\mathbb{R}^p \times \mathbb{R}_*^+$ and likelihood $\mathcal{L}$. In this model, the maximum likelihood estimator $\hat{\theta}$ of θ corresponds to the values that minimize the least squares criterion:

$$J_0(\theta) = \sum_{i=1}^{n} (y_i - \langle \theta, x_i \rangle)^2.$$

It is also a solution of the linear system

$$X^T X \theta = X^T Y,$$

where X is a matrix of size $n \times (p+1)$ (design matrix) with entries $X_{ij} = x_i^j$ for $j = 0, \cdots, p$ and $i = 1, \cdots, n$. This system has a unique solution as long as $\det X^T X \neq 0$. Therefore, the search for the maximum likelihood estimator of the regression parameters θ in a linear model reduces to the numerical solution of a linear system.

5.4.3 Ridge Linear Regression

Ridge regression is a Bayesian regression model where it is assumed that the parameters θ_p of the model are random, independent, and identically distributed according to a normal distribution centered with variance $\frac{\sigma^2}{\lambda}$, where $\lambda > 0$ is a hyperparameter that must be fixed.

Remark.

Ridge regression helps to address problems with solving the linear system that arise when the determinant $\det X^T X$ is zero or close to zero. These problems occur when the explanatory variables are highly correlated (the multicollinearity problem) or in high-dimensional situations when the number of variables p exceeds the number of observations n ($p > n$).

The posterior likelihood of this Bayesian model is proportional to

$$\mathcal{L}(\theta, \sigma^2; y)\ \exp\left(-\frac{\lambda}{2\sigma^2}\sum_{i=0}^{p}(\theta_p)^2\right),$$

where the likelihood $\mathcal{L}$ is the function from equation (5.13). Searching for a maximum of the posterior likelihood corresponds to finding a minimum of a penalized least squares criterion:

$$J_\lambda(\theta) = J_0(\theta) + \lambda|\theta|^2, \tag{5.14}$$

where J_0 is defined by equation (5.4.2). Again, the solution to this minimization problem is the solution of a linear system:

$$(X^T X + \lambda I)\theta = X^T Y,$$

where X is the design matrix of the linear model and I is the identity matrix of size $(p+1)\times(p+1)$.

Other distribution models can be used to describe the prior law of the regression parameters. This is the case in LASSO regression, where by assuming that the coefficients follow exponential distributions, one is led to minimize the criterion

$$\tilde{J}_\lambda(\theta) = J_0(\theta) + \lambda|\theta|_1, \tag{5.15}$$

with a penalty $\lambda|\theta|_1 = \sum_{j=0}^{p}|\theta_j|$ formed by the L_1 norm of the coefficients.

5.4.4 Generalized Linear Regression

In certain situations, the observations Y_i are correlated, and it is unacceptable to assume that the variations ϵ_i are uncorrelated. We can then assume that the observations $Y = (Y_1, \cdots, Y_n)^T$ form a Gaussian vector with distribution $\mathcal{N}(X\theta, \Sigma)$, where $X\theta$ is a vector in $\mathbb{R}^n$ describing the expectation of Y and Σ is a covariance matrix with terms $\Sigma_{ik} = \mathrm{Cov}(\epsilon_i, \epsilon_k)$ for $i, k = 1, \cdots, n$. This model has the likelihood

$$\widetilde{\mathcal{L}}(\theta, \Sigma; y) = \frac{1}{(2\pi)^{n/2}}\frac{1}{\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}(y - X\theta)^T\Sigma^{-1}(y - X\theta)\right)$$

with respect to the Lebesgue measure on $\mathbb{R}^n$. The classical linear model is retrieved when $\Sigma = \sigma^2 I$, where I is the identity matrix of size $n \times n$. With Σ fixed, maximizing the likelihood of a generalized linear model corresponds to minimizing a generalized least squares criterion:

$$\tilde{J}(\theta) = (y - X\theta)^T\Sigma^{-1}(y - X\theta).$$

The solution to this optimization problem is the solution of the linear system

$$X^T\Sigma^{-1}X\theta = X^T\Sigma^{-1}Y.$$

5.4.5 Linear Regression and Optimization

In all the linear regression models we have seen, the search for the regression parameters reduces to solving a linear system of the form $A\theta = b$. This type of problem can be solved in several ways.

A first approach is to directly solve the linear system. To do this, we can factor the matrix $A = LU$ into the product of a lower triangular matrix L and an upper triangular matrix U. The solution of the system $A\theta = b$ then becomes equivalent to the simpler solution of two systems:

$$Lz = b \text{ and } U\theta = z.$$

The second approach relies on minimizing the quadratic functional

$$J(\theta) = \frac{1}{2}\langle A\theta, \theta\rangle - \langle b, \theta\rangle.$$

It can be shown that J achieves a global minimum at $\hat{\theta}$ if and only if $\hat{\theta}$ is a solution to the system $A\theta = b$. Thus, minimization methods can be applied to J to solve the system. These minimization methods typically take the form of iterative descent algorithms. At each iteration t of these algorithms, we update the current estimated value $\theta^{(t)}$ of θ:

$$\theta^{(t+1)} = \theta^{(t)} + \rho_t u^{(t)},$$

taking a step $\rho_t > 0$ and a direction $u^{(t)}$ ensuring that $J(\theta^{(t+1)}) < J(\theta^{(t)})$. This approach is less costly than a factorization. It can be particularly interesting when the number of variables p is large.

5.4.6 Logistic Regression

Logistic regression is another form of regression that applies when the response variable is qualitative. Suppose the variables Y_i can take two states: $\{0,1\}$. The variables Y_i are not directly linked to the explanatory variables x_i^j but rather through a parameterization of their probability distribution. We set:

$$\mathbb{P}(Y_i = 1) = \sigma\left(\theta_0 + \sum_{j=1}^{p} \theta_j x_i^j\right) = \sigma(\langle\theta, x_i\rangle),$$

where σ is a function from $\mathbb{R}$ to $]0,1[$, called the logistic function, defined by

$$\sigma(t) = \frac{1}{1+e^{-t}}. \tag{5.16}$$

This implies that the variables Y_i follow a Bernoulli distribution with parameter $p_i(\theta) = \sigma(\langle\theta, x_i\rangle)$. We also assume that the variables Y_i are independent. This defines a statistical model whose parameters are θ and whose likelihood is given by

$$\mathcal{L}(\theta; y_1, \cdots, y_n) = \prod_{i=1}^{n} \sigma(\langle\theta, x_i\rangle)^{y_i}(1 - \sigma(\langle\theta, x_i\rangle))^{1-y_i}$$

and the log-likelihood (negative) is

$$\ell(\theta, y_1, \cdots, y_n) = -\sum_{i=1}^{n} [y_i \log(\sigma(\langle\theta, x_i\rangle)) + (1-y_i)\log(1 - \sigma(\langle\theta, x_i\rangle))].$$

The latter function is called the binary cross entropy. The maximum likelihood estimator (MLE) of θ for this model is computed by minimizing this function. Unlike the linear model, this optimization problem does not reduce to solving a linear system. It does not have an analytic solution. Its resolution is typically performed using a Newton-Raphson type descent method.

5.5 Exercises

Exercise 5.1: Maximum Likelihood Estimator

Let P_θ be a probability distribution with density

$$f_\theta(x) = \begin{cases} \theta(1-x)^{\theta-1} & \text{if } x \in [0,1], \\ 0 & \text{otherwise,} \end{cases}$$

defined for $\theta > 1$ with respect to the Lebesgue measure. Consider an n-sample $(X_1, \cdots, X_n)$ from the distribution P_θ.

1. Write the statistical model associated with the sample.
2. Specify the likelihood of the model.
3. Determine the maximum likelihood estimator (MLE) $\hat{\theta}$ of θ.

Exercise 5.2: Maximum A Posteriori Estimator

Let $X = (X_1, \cdots, X_n)$ be an i.i.d. sample from a Bernoulli distribution $B(p)$. Assume that p is a random parameter distributed according to a Beta distribution $\text{Beta}(\alpha, \beta)$ with density

$$g(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}\mathbf{1}_{[0,1]}(p),$$

where $B(\alpha,\beta) = \int_0^1 q^{\alpha-1}(1-q)^{\beta-1}dq$.

1. Show that the posterior distribution of p is a Beta distribution, and determine its parameters.
2. Determine the maximum a posteriori (MAP) estimator of p. Compare it with the maximum likelihood estimator.

Exercise 5.3: Quadratic Functional

Let $A \in M_{n,p}$ be a matrix of size $n \times p$ and $b \in \mathbb{R}^p$ a vector of size p. Consider the real-valued function defined for all $\theta \in \mathbb{R}^p$ by

$$F(\theta) = \frac{1}{2}|A\theta|^2 - \langle b, \theta \rangle, \tag{5.17}$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product on $\mathbb{R}^p$ and $|\cdot|$ is its associated norm.

1. Expand $F(\theta + h) - F(\theta)$ in terms of A, b, and h.
2. Show that if a vector θ^* in $\mathbb{R}^p$ satisfies $A^T A\theta = b$, then it also satisfies $F(\theta^*) \leq F(\theta), \forall \theta \in \mathbb{R}^p$.
3. Prove the converse and conclude.

Exercise 5.4: Generalized Linear Regression

Consider a linear model

$$Y_i = \theta_0 + \sum_{j=1}^{p} x_i^j \theta_j + \epsilon_i, \quad i = 1, \cdots, n, \tag{5.18}$$

where the x_i^j are fixed variables, the θ_j are unknown parameters, and the ϵ_i are random variables. Assume that $\epsilon = (\epsilon_i)_{i=1}^n$ is a centered Gaussian vector with covariance matrix Σ.

1. Write the model in matrix form.
2. Specify the likelihood of the model.

3. Show that finding the MLE of θ reduces to a minimization problem of a function of the form of equation (5.17).
4. Deduce that the MLE is the solution of a linear system that you will specify.
5. Specify this system when $\Sigma = \sigma^2 I$ with $\sigma > 0$.

Exercise 5.5: Ridge Regression

We consider again the model in equation (5.18) in the case where $\Sigma = \sigma^2 I$. Furthermore, we assume that the parameters θ_j are i.i.d. with a normal distribution centered at zero and variance $\frac{\sigma^2}{\lambda}$ (with $\lambda > 0$ fixed).

1. Write the posterior likelihood of θ.
2. Show that finding the MAP estimator of θ reduces to a minimization problem of a function of the form of equation (5.17). Connect this to ridge regression.
3. Deduce that the MAP estimator is the solution of a linear system that you will specify.

Exercise 5.6: Rayleigh Quotient

Let A be a real symmetric matrix of size $p \times p$. We recall that A is diagonalizable in an orthonormal basis $\mathcal{U} = \{u_1, \cdots, u_p\}$ of eigenvectors and that its eigenvalues $\lambda_1 \leq \cdots \leq \lambda_p$ are real.
For any non-zero $v \in \mathbb{R}^p$, the Rayleigh quotient is defined as

$$R_A(v) = \frac{\langle Av, v\rangle}{|v|^2}.$$

We aim to find the maximum of R_A.

1. Let $\alpha = (\alpha_1, \cdots, \alpha_p)$ be the coordinates of v in $\mathcal{U}$. Show that

$$R_A(v) = \frac{\sum_{j=1}^{p} \lambda_j \alpha_j^2}{\sum_{m=1}^{p} \alpha_m^2}.$$

2. Deduce that R_A is bounded above by λ_p.
3. Show that this upper bound is achieved and specify at which v^* in $\mathbb{R}^p$.

Exercise 5.7: Principal Component Analysis

We observe the realizations $x_1, \cdots, x_n$ of random vectors following the same distribution as a random vector X of dimension p with covariance matrix Σ. We assume that $n > p$ and estimate Σ using the empirical covariance matrix

$$S_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)(x_i - \bar{x}_n)^T,$$

where $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$.
Let $v \in \mathbb{R}^p$ such that $|v| = 1$. We consider the coefficient X_v of the orthogonal projection of X onto the vector subspace F_v of $\mathbb{R}^p$ spanned by v.

1. Specify the variance of the random variable X_v in terms of S_n and v.
2. Using the result from Exercise 5.5, find a vector v for which the variance of X_v is maximal and specify this maximum value.

Chapter 6 Basics in differential calculus

6.1 Differential

6.1.1 Definition and Examples

We assume that $(E, |\cdot|_E)$ and $(F, |\cdot|_F)$ are normed vector spaces.

Definition 6.1: Differential

Let U be an open set in E and $f : U \to F$. We say that f is differentiable at a point a in U if there exists a continuous linear map $L \in \mathcal{L}(E, F)$ such that

$$f(a+h) = f(a) + L(h) + \epsilon(h), \tag{6.1}$$

where ϵ is a function from E into F satisfying

$$\lim_{h \to 0} \frac{|\epsilon(h)|_F}{|h|_E} = 0.$$

If this condition holds, we call L the differential of f at a. When it exists, the differential is unique.

There are several notations for the differential of f at a. The simplest is $f'(a)$, but its use requires familiarity with the concept of differential to avoid confusion with the derivative. In this chapter, we will temporarily denote the differential as df_a to distinguish it from the derivative or partial derivatives.

Remark.

The notion of differential generalizes the derivative of a function from $\mathbb{R}$ into itself. Indeed, the derivative of a function f at x can be defined as the unique value $f'(x)$ in $\mathbb{R}$ such that

$$f(x+h) = f(x) + f'(x)h + o(h) \text{ as } h \to 0.$$

Equation (6.1), which characterizes the differential, extends this development to any mapping from E into F. Furthermore, this development shows that if the function has a derivative $f'(x)$ at x, it admits a differential at x which is the linear map df_x defined for all $h \in \mathbb{R}$ by

$$df_x(h) = f'(x)h.$$

This establishes the connection between the derivative and the differential.

Example 6.1: Differential of a constant function

If f is a constant function, then its differential is zero at every point x, i.e., $df_x \equiv 0$.

Example 6.2: Differential of a linear map

If f is a linear map, then its differential at every point is itself. Indeed, in this case,

$$f(x+h)-f(x)=f(h),$$

which directly identifies the differential in the development of f around x.

Example 6.3: Differential of a bilinear map

Let $f \in \mathcal{L}_2(E,F)$ be a continuous bilinear map from E^2 to F. We endow the product space E^2 with the norm defined, for all $h=(h_1,h_2)\in E^2$, by

$$|h|_{E^2}=\sqrt{|h_1|_E^2+|h_2|_E^2}.$$

The map f is differentiable at any $x \in E^2$. Its differential at x is the map df_x in $\mathcal{L}(E^2,F)$ defined for all $h\in E^2$ by

$$df_x(h_1,h_2)=f(x_1,h_2)+f(h_1,x_2),\ \forall h\in E^2.$$

6.1.2 Link with Partial Derivatives

Consider the case where E is of finite dimension p. We equip E with a basis $(e_1,\cdots,e_p)$.

Definition 6.2: Partial Derivatives

Let f be a function from an open set U of E to $\mathbb{R}$. For a fixed a in E, the function

$$g_j:\mathbb{R}\to\mathbb{R},\quad h\mapsto g_j(h)=f(a+he_j)$$

is the partial function of f at a with respect to the jth variable. When the derivative of g_j exists at 0, this derivative is called the partial derivative of f at a with respect to the jth variable, denoted by $\partial_j f(a)$. In other words,

$$\partial_j f(a)=\lim_{h\to 0}\frac{g_j(h)-g_j(0)}{h}=\lim_{h\to 0}\frac{f(a+he_j)-f(a)}{h}=g_j'(0).$$

Proposition 6.1

Let f be a function from an open set U of E to $\mathbb{R}$. If f is differentiable at $a\in U$, then its partial derivatives exist at a and

$$\partial_j f(a)=df_a(e_j),\forall j=1,\cdots,p.$$

In this case, we also have, for any $h\in E$,

$$df_a(h)=\sum_{j=1}^{p}\partial_j f(a)h_j,$$

where the h_j are the coordinates of h in the basis of E.

Proof. See the exercise in Section 6.4. □

The converse of this proposition is not true in general. There are functions that have partial derivatives but are not differentiable.

Now, consider functions f defined on an open set U of E taking values in $\mathbb{R}^q$. These functions can be described by their components, i.e., the q functions $f_1,\cdots,f_q$ from E to $\mathbb{R}$ such that

$$f(h)=\begin{pmatrix}f_1(h)\\ \vdots\\ f_q(h)\end{pmatrix},\ \forall\, h\in E.$$

The notion of partial derivative can be applied to each of these components. When all the components have partial derivatives, they are assembled into a matrix called the Jacobian matrix of f.

Definition 6.3: Jacobian Matrix

Let f be a function from an open set U of E to $\mathbb{R}^q$ whose components have partial derivatives at a point a of U. The Jacobian matrix of f at a, denoted by $Df(a)$, is the matrix of size $q \times p$, defined by

$$Df(a) = \begin{pmatrix} \partial_1 f_1(a) & \cdots & \partial_p f_1(a) \\ \vdots & \ddots & \vdots \\ \partial_1 f_q(a) & \cdots & \partial_p f_q(a) \end{pmatrix}.$$

When $p = q$, the determinant of $Df(a)$ is called the Jacobian of f at a.

Proposition 6.2

If a function f from an open set U of E to $\mathbb{R}^q$ is differentiable at $a \in U$, then

$$df_a(h) = Df(a)h = \begin{pmatrix} \sum_{j=1}^p h_j \partial_j f_1(a) \\ \vdots \\ \sum_{j=1}^p h_j \partial_j f_q(a) \end{pmatrix},$$

where, by abuse of notation, $h = (h_j)_{j=1}^p$ denotes the vector in $\mathbb{R}^p$ formed by the coordinates of h in the basis of E.

6.1.3 Gradient

Let $(E, \langle \cdot, \cdot \rangle)$ be a Hilbert space and f be a function from an open set U of E to $\mathbb{R}$ differentiable at $a \in U$.

Definition 6.4: Gradient

The gradient of f at a, denoted by $\nabla f(a)$, is the element of E such that

$$df_a(h) = \langle \nabla f(a), h \rangle, \ \forall\, h \in E.$$

When f is differentiable at a, we can write:

$$f(a+h) = f(a) + \langle \nabla f(a), h \rangle + o(h) \text{ as } h \to 0.$$

Proposition 6.3

Let E be a finite-dimensional Hilbert space of dimension p with a basis $(e_1, \cdots, e_p)$. The gradient of f at a has coordinates $\partial_j f(a)$ in the basis. By abuse of notation, we write

$$\nabla f(a) = \begin{pmatrix} \partial_1 f(a) \\ \vdots \\ \partial_p f(a) \end{pmatrix}.$$

With this notation, we also have

$$\nabla f(a) = Df(a)^T.$$

6.2 Higher-Order Differentials

6.2.1 Definition and Examples

Definition 6.5: Continuously differentiable

Let f be a function defined on an open set $U \subset E$ with values in F. If f is differentiable at every point of U, then we say that f is differentiable on U. In this case, we can define a function

$$\begin{array}{cccc} df & : & U \subset E & \longrightarrow & \mathcal{L}(E,F) \\ & & a & \longrightarrow & df_a \end{array},$$

called the differential of f. If, furthermore, df is continuous on U, we say that f is continuously differentiable on U. We denote by $\mathcal{C}^1(U,F)$ the set of continuously differentiable functions on U with values in F.

Definition 6.6: Differentiable of order 2

Let f be a function from U to F with a differential df on U. If this differential is itself differentiable on U, then we say that f is twice differentiable on U. We denote the differential of the differential of f by $d^2f = d(df)$. This function, called the second-order differential of f, takes values d^2f_a in the space $\mathcal{L}(E,\mathcal{L}(E,F))$, which we denote as $\mathcal{L}^2(E,F)$.

Remark.

We can define higher-order differentials by induction: we assume that f has a differential d^nf of order $n \geq 1$. This function is defined on U, and its values d^nf_a belong to the space $\mathcal{L}^n(E,F)$. If d^nf is differentiable on U, we say that f is $n+1$-times differentiable on U and define the function $d^{(n+1)}f = d(d^nf)$ on U, with values $d^{(n+1)}f_a$ in $\mathcal{L}^{n+1}(E,F) := \mathcal{L}(E,\mathcal{L}^n(E,F))$.

Definition 6.7: Class $\mathcal{C}^n$

We say that a function f from $U \subset E$ to F is of class $\mathcal{C}^n$ on U if it is n-times differentiable on U and if its differential d^nf of order n is continuous on U. We denote by $\mathcal{C}^n(U,F)$ the space of functions of class $\mathcal{C}^n$ on U.

Example 6.4: Differential of order 2 of a bilinear function

A continuous bilinear function f defined on E^2 with values in F is twice differentiable. Its differential at $x \in E^2$ is the function d^2f_x in $\mathcal{L}^2(E,F)$ defined for all h and k in E^2 by

$$d^2f_x(h)(k) = f(h_1,k_2) + f(k_1,h_2).$$

6.2.2 Second-Order Differential and Bilinear Function

Let f be a function from $U \subset E$ to F that admits a second-order differential on U. As seen, d^2f_a belongs to $\mathcal{L}^2(E,F) = \mathcal{L}(E,\mathcal{L}(E,F))$. Thus, for $h,k \in E$, $d^2f_a(h) \in \mathcal{L}(E,F)$ and $d^2f_a(h)(k) \in F$.

Furthermore, the functions

$$\begin{array}{ccc} k \in E & \rightarrow & d^2f_a(h)(k) \in F \\ h \in E & \rightarrow & d^2f_a(h)(k) \in F \end{array}$$

are both linear. Like a bilinear function, the second differential is thus linear with respect to each of its arguments. Fundamentally, the space $\mathcal{L}^2(E,\mathcal{L}(E,F))$ is isomorphic to the space of bilinear functions $\mathcal{L}_2(E,F)$. Therefore, every second-order differential can be identified with a bilinear function. Referring to this bilinear function, we often write $d^2f_a(h,k)$ to denote the values $d^2f(h)(k)$.

The second-order differential has another important property stated in the following theorem.

Theorem 6.4: Schwarz

Let f be a function from $U \subset E$ open in F that admits a second-order differential at $a \in U$. Then d^2f_a is a symmetric bilinear function: for all $h,k \in E$,

$$d^2f_a(h)(k) = d^2f_a(h,k) = d^2f_a(k,h).$$

6.2.3 Link with Partial Derivatives

We are considering the case where E is a normed vector space of finite dimension p with a basis $(e_1, \cdots, e_p)$.

Definition 6.8: Partial derivatives of order 2

Let f be a mapping from E to $\mathbb{R}$ with partial derivatives at every point of an open set U. For $i = 1, \cdots, p$, we consider the mapping $\partial_i f$ which associates each $a \in U$ with $\partial_i f(a)$. If this function has a partial derivative with respect to a j-th variable, it is called the second-order partial derivative of f with respect to the variables i and j. It is denoted as

$$\partial^2_{ji} f(a) = \partial_j(\partial_i f)(a).$$

Remark.

Higher-order partial derivatives are defined recursively, analogously to differentials. The recurrence formula for obtaining the partial derivative $\partial^{n+1}_{i_{n+1}\cdots i_1} f(a)$ of f at a of order $n+1$ with respect to the indices $i_1, \cdots, i_{n+1}$ is:

$$\partial^{n+1}_{i_{n+1}\cdots i_1} f(a) = \partial_{i_{n+1}}(\partial^n_{i_n\cdots i_1} f)(a).$$

Now, suppose a function f from E to $\mathbb{R}$ is twice differentiable on an open set U of E. By definition,

$$df_{a+h} - df_a = d^2 f_a(h) + \epsilon(h),$$

with

$$\lim_{h\to 0} \frac{|\epsilon(h)|}{|h|} = 0.$$

For all $k \in E$, we can also write

$$df_{a+h}(k) - df_a(k) = d^2 f_a(h,k) + \epsilon(h)(k).$$

In particular, for $h = te_j$ and $k = e_i$ with $t \in \mathbb{R}^*$, we have

$$df_{a+te_j}(e_i) - df_a(e_i) = td^2 f_a(e_i, e_j) + \epsilon(te_j)(e_i).$$

Since we know that $\partial_i f(a + te_j) = df_{a+te_j}(e_i)$ and $\partial_i f(a) = df_a(e_i)$, it follows that

$$\frac{\partial_i f(a+te_j) - \partial_i f(a)}{t} = d^2 f_a(e_i, e_j) + \frac{\epsilon(te_j)(e_i)}{t}.$$

Consequently,

$$\partial^2_{ji} f(a) = \lim_{t\to 0} \frac{\partial_i f(a+te_j) - \partial_i f(a)}{t} = d^2 f_a(e_i, e_j).$$

Thus, we establish a connection between differentials and second-order partial derivatives.
Additionally, since $d^2 f_a(e_i, e_j) = d^2 f_a(e_j, e_i)$, we have

$$\partial^2_{ji} f(a) = \partial^2_{ij} f(a).$$

This result provides a formulation of the link between differentials and second-order partial derivatives.

Proposition 6.5: Link between differential and partial derivatives of order 2

Let f be a mapping from $U \subset E$ (an open set) to $\mathbb{R}$, differentiable to the second order on U. Then, f has second-order partial derivatives on U, and

$$\partial^2_{ji} f(a) = d^2 f_a(e_i, e_j).$$

Moreover, for all $h, k \in E$,

$$d^2 f_a(h,k) = \sum_{i=1}^{p} \sum_{j=1}^{p} h_i k_j \partial^2_{ij} f(a),$$

where $(h_i)_i$ and $(k_j)_j$ are the coordinates of h and k in the basis of E, respectively.

Definition 6.9: Hessian Matrix

Let f be a mapping from $U \subset E$ (an open set) to $\mathbb{R}$, differentiable to the second order on U. At every point a in U, we define the matrix

$$\nabla^2 f(a) = \begin{pmatrix} \partial^2_{11} f(a) & \cdots & \partial^2_{1p} f(a) \\ \vdots & \ddots & \vdots \\ \partial^2_{p1} f(a) & \cdots & \partial^2_{pp} f(a) \end{pmatrix}.$$

This matrix is called the Hessian matrix of f at a. It is symmetric.

Proposition 6.6: Matricial expression of the second order differential

Using the Hessian matrix, we can write

$$d^2 f_a(h,k) = h^T \nabla^2 f(a) k = \left\langle \nabla^2 f(a) h, k \right\rangle,$$

where, as a slight abuse of notation, we denote by h and k the column vectors formed by the coordinates of h and k, respectively.

Remark.

For partial derivatives of order higher than 2, we have the more general formula

$$d^n f(a)(h^{(1)}, \cdots, h^{(n)}) = \sum_{i_1=1}^{p} \cdots \sum_{i_n=1}^{p} h^{(1)}_{i_1} \cdots h^{(n)}_{i_n} \partial^n_{i_1 \cdots i_n} f(a),$$

for all $h^{(1)}, \cdots, h^{(n)} \in E$.

6.2.4 Second-order Differential of Composed Functions

Theorem 6.7: second-order differential of composed functions

Let E, F, and G be three normed vector spaces. Consider
- a function f defined on an open set U in E, valued in F and twice differentiable on U, - a function g defined on an open set V in F containing $f(U)$, valued in G and twice differentiable on V.
Then the composite function $h = g \circ f$ is twice differentiable at any point a in U and has a differential

$$d^2 h_a = dg_{f(a)}(d^2 f_a) + d^2 g_{f(a)}(df_a, df_a),$$

that is, for all $u, v \in E$,

$$d^2 h_a(u,v) = dg_{f(a)}(d^2 f_a(u,v)) + d^2 g_{f(a)}(df_a(u), df_a(v)).$$

6.3 Fundamental theorems

6.3.1 Mean Value Theorem

Theorem 6.8: Mean Value Theorem

Let f be a function from $\mathbb{R}$ to $\mathbb{R}$ continuous on an interval $[a,b]$ and differentiable on $]a,b[$. Then, there exists $c \in]a,b[$ such that

$$f'(c) = \frac{f(b)-f(a)}{b-a}.$$

Proof: The proof of this theorem is a consequence of Rolle's theorem.

Theorem 6.9: Rolle's Theorem

Let g be a function from $\mathbb{R}$ to $\mathbb{R}$, continuous on an interval $[a,b]$ and differentiable on $]a,b[$, such that $g(a) = g(b)$. Then, there exists $c \in]a,b[$ with $g'(c) = 0$.

It suffices to apply this theorem to the function

$$g(t) = f(t) - \frac{f(b)-f(a)}{b-a}t.$$

□

Applied to a function f, the mean value theorem implies that

$$|f(b)-f(a)| \leq K|a-b|.$$

This property is generalized in the following version of the mean value theorem.

Theorem 6.10: Mean Value Theorem (Generalization)

Let E and F be two Banach spaces, and U an open convex subset of E. Consider a function f defined on U, valued in F. Suppose that f is differentiable on U and that there exists $K > 0$ such that

$$|df_x| \leq K, \quad \forall x \in U.$$

Then, for all $x, y \in U$,

$$|f(x)-f(y)|_F \leq K\,|x-y|_E.$$

In other words, f is K-Lipschitz.

Corollary 6.11

Under the conditions of the generalized mean value theorem, if $K = 0$ then f is constant on U.

Corollary 6.12

Under the conditions of the generalized mean value theorem, if E has finite dimension and $F = \mathbb{R}$, then $f \in \mathcal{C}^1(U, \mathbb{R})$ if and only if its partial derivatives exist and are continuous.

6.3.2 Some Taylor's Formulas

To conclude this chapter, we present some formulas essential for optimization.

Theorem 6.13: Taylor-Young Formula

Let f be a function from $\mathbb{R}$ to $\mathbb{R}$, $n-1$-differentiable on an open set U in $\mathbb{R}$ and admitting an n-th order derivative at a point a of U. Then,

$$f(a+h) = f(a) + \sum_{k=1}^{n} \frac{f^{(k)}(a)}{k!} h^k + o_{h\to 0}(h^n).$$

We now state a generalization of the Taylor-Young formula for a function f from a Banach space E to $\mathbb{R}$.

Theorem 6.14: Taylor-Young Formula (Generalization)

If f is $n-1$ times differentiable on an open set U of E and admits an n-th order differential at a point a of U, then for any h such that $a+h \in U$,

$$f(a+h) = P[f]_a^n(h) + |h|^n \epsilon(h),$$

where $P[f]^n$ is the function from E to $\mathbb{R}$ defined for any h in E by

$$P[f]_a^n(h) = f(a) + \sum_{k=1}^{n} \frac{1}{k!} d^{(k)} f_a(h, \cdots, h),$$

and ϵ is a function from E to $\mathbb{R}$ satisfying

$$\lim_{h\to 0} \epsilon(h) = 0.$$

We also provide a finite-dimensional version of this theorem, often used in optimization.

Corollary 6.15: Taylor-Young Formula (Finite dimension)

Let E be a finite-dimensional space with an inner product $\langle \cdot, \cdot \rangle$ and a basis. Under the conditions of the Taylor-Young theorem at order $n = 2$,

$$f(a+h) = f(a) + \langle \nabla f(a), h \rangle + \frac{1}{2} \left\langle \nabla^2 f(a) h, h \right\rangle + |h|^2 \epsilon(h).$$

6.4 Exercises

Exercise 6.1: First and Second Differentials of a Bilinear Map

Let f be a continuous bilinear map from E^2 to F. We equip the product space E^2 with the norm

$$|(h,k)|_{E^2} = \sqrt{|h|_E^2 + |k|_E^2}, \quad \forall (h,k) \in E^2.$$

1. Show that f is differentiable on E^2 and compute its differential.
2. Show that f is twice differentiable and compute its second differential.
3. Suppose E is finite-dimensional and $F = \mathbb{R}$. Let $(e_1, \ldots, e_p)$ be an orthonormal basis for E. We equip E with the inner product

$$\langle x, y \rangle = \sum_{i=1}^{p} x_i y_i,$$

where $(x_i)_{i=1}^p$ and $(y_i)_{i=1}^p$ are the coordinates of x and y in this basis, respectively.

   (a) Specify the gradient and Hessian of f in this basis of E.

   (b) Write the first and second differentials in terms of the gradient and Hessian, respectively.

Exercise 6.2: Connection between Differential and Partial Derivatives

Let E be a finite-dimensional space and let f be a function from an open set $U \subset E$ to $\mathbb{R}$. We equip E with a basis $(e_1, \dots, e_p)$.

1. Show that if f is differentiable at $a \in U$, then its partial derivatives exist at a and

$$\partial_j f(a) = df_a(e_j), \quad \forall j = 1, \dots, p.$$

2. Furthermore, show that for any $h \in E$,

$$df_a(h) = \sum_{j=1}^{p} \partial_j f(a) h_j,$$

where h_j are the coordinates of h in the basis of E.

Exercise 6.3: Gamma Distribution Sample

Let $(Y_1, \dots, Y_n)$ be a sample from a Gamma distribution with positive parameters a and b, with the density function defined by

$$f(x) = \frac{b^a}{\Gamma(a)} e^{-bx} x^{a-1} \mathbf{1}_{(0,+\infty)}(x),$$

where $\Gamma(a)$ is the following integral:

$$\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1}\, dx.$$

1. Write the likelihood $\mathcal{L}(a, b; y_1, \dots, y_n)$ of the model.
2. For fixed y_i, set $\ell(a,b) = -\log(\mathcal{L}(a, b; y_1, \dots, y_n))$. Show that ℓ is differentiable and compute its partial derivatives in terms of Γ and its derivative.

Exercise 6.4: Linear Model

Consider the generalized linear model

$$Y = X\theta + \epsilon,$$

where the design matrix X is of size $n \times p$ and rank p, the parameters θ are in $\mathbb{R}^p$, and the errors ϵ are distributed according to $\mathcal{N}(0, \Sigma)$ with a positive definite covariance matrix Σ.
We define the generalized least squares criterion used to estimate

$$J(\theta) = (Y - X\theta)^T \, \Sigma^{-1} \, (Y - X\theta).$$

1. Justify that this criterion is of class $\mathcal{C}^\infty$.
2. Determine the first and second-order differentials of this criterion.
3. Equip $\mathbb{R}^p$ with the usual inner product. Specify the gradient and the Hessian matrix of J in the canonical basis of $\mathbb{R}^p$.
4. Verify that the Hessian matrix of J is positive definite.

Exercise 6.5: Logistic Regression

Consider a sample consisting of n independent random variables Y_i following a Bernoulli distribution with parameter p_i. We assume that for each i, the parameter p_i depends on explanatory variables $x_i \in \mathbb{R}^p$ and unknown parameters $\theta \in \mathbb{R}^p$ through the relation $p_i = \sigma(x_i^T \theta)$, where $\sigma(t) = \frac{1}{1+e^{-t}}$.

1. Show that

$$\sigma(t) + \sigma(-t) = 1$$

and that

$$(\log_e \circ \sigma)'(t) = \sigma(-t).$$

2. Determine the likelihood $\mathcal{L}(\theta; y_1, \dots, y_n)$ of the model.

3. For fixed y_i, set $\ell(\theta) = -\log_e(\mathcal{L}(\theta; y_1, \dots, y_n))$, the logistic loss function. Show that ℓ is of class $\mathcal{C}^\infty$.
4. Show that the gradient $\nabla\ell(\theta)$ and the Hessian matrix $\nabla^2\ell(\theta)$ of ℓ are given by:

$$\nabla\ell(\theta) = -\sum_{i=1}^{n} x_i(y_i - \sigma(x_i^T\theta)) \quad \text{and} \quad \nabla^2\ell(\theta) = \sum_{i=1}^{n} \sigma(x_i^T\theta)(1-\sigma(x_i^T\theta))x_i x_i^T.$$

5. Under what condition(s) is $\nabla^2\ell(\theta)$ a positive definite matrix?

Exercise 6.6: Neural Networks

We consider a neural network that relates a variable $y \in \mathbb{R}$ to a vector $x \in \mathbb{R}^p$. The output variable z is obtained from the input vector x in the following manner:

$$z = g(\langle x, \theta \rangle),$$

where g is a $\mathcal{C}^\infty$ function from $\mathbb{R}$ to $\mathbb{R}$ and $\theta \in \mathbb{R}^p$ represents a set of parameters. For training the network, we have a dataset $(x_i, y_i)_{i=1}^n$. We adjust the network parameters by minimizing the least squares criterion with respect to θ:

$$J(\theta) = \sum_{i=1}^{n} (y_i - g(\langle x_i, \theta \rangle))^2.$$

1. Justify that the criterion J is of class $\mathcal{C}^\infty$.
2. Calculate its differential in terms of the derivatives of g. Specify its gradient.
3. Calculate its second differential and specify its Hessian.

Chapter 7 Basics in unconstrained optimisation

7.1 Introduction

7.1.1 Definitions and notations

Let f be a function defined on a set V of a space E with values in $\mathbb{R}$. Optimizing f over V involves finding the points x^* in V that achieve a minimum (resp. a maximum) of f, i.e., points satisfying:

$$f(x^*) \leq f(x), \forall x \in V \quad (\text{resp. } f(x^*) \geq f(x), \forall x \in V). \tag{7.1}$$

The point x^* satisfying equation (7.1) is a **minimizer** (resp. maximizer) of the function f over V, while the value $f(x^*)$ is a **minimum** (resp. maximum) of f on V. The minimum (resp. maximum) value of f on V is denoted by:

$$\min_{x \in V} f(x) \quad \left(\text{resp. } \max_{x \in V} f(x)\right).$$

Example 7.1: Linear model

In data science, the problem of linear regression is translated into a minimization problem. Starting from a linear relationship model

$$Y = X\theta + \epsilon$$

between an observation vector Y and regression variables X, we seek the regression coefficients θ that best describe this relationship. This problem is treated by minimizing a least-squares criterion

$$\min_{\theta \in \mathbb{R}^{p+1}} \|Y - X\theta\|^2.$$

The set of all minimizers (resp. maximizers) of a function f over V is denoted by:

$$\arg\min_{x \in V} f(x) \quad \left(\text{resp. } \arg\max_{x \in V} f(x)\right).$$

If x^* is a minimizer, we write $x^* \in \arg\min_{x \in V} f(x)$. When this minimizer is unique, we may abuse notation and write $x^* = \arg\min_{x \in V} f(x)$, but it is important to remember that $\arg\min_{x \in V} f(x)$ is a subset of V.

Example 7.2: Linear model

In linear regression, we have

$$\arg\min_{\theta \in \mathbb{R}^{p+1}} \|Y - X\theta\|^2 = \left\{\theta \in \mathbb{R}^{p+1} \mid X^\top X\theta = X^\top Y\right\}.$$

Finding the maximum of a function f is equivalent to finding the minimum of the function $-f$. Without loss of generality, we will focus on minimization problems.

Minimization problems are **constrained** when V is strictly included in the space E; in this case, V generally represents a set of mathematical constraints that limit the search for the minimum. When the minimum is sought over the entire E, the minimization problem is **unconstrained**.

Example 7.3: Linear model

In its simpliest form, the linear regression problem is unconstrained. However, for modeling purposes, constraints are sometimes added to the regression coefficients. For example, one might require that these coefficients are all positive, leading to the minimization of a least-squares criterion over the subset of $\mathbb{R}^{p+1}$:

$$V = \left\{\theta \in \mathbb{R}^{p+1} \mid \theta_j \geq 0, \forall j\right\}.$$

The minimum of a function f over the entire set E is called the **global minimum** of f. Sometimes, a weaker notion of minimum is used, the **local minimum**: f has a local minimum at x^* if there exists a neighborhood $\mathcal{N}$ of x such that:

$$f(x^*) \leq f(x), \forall x \in \mathcal{N}.$$

Moreover, we distinguish discrete minimization problems from continuous minimization problems based on whether the space E is countable or not.

In this chapter, we will study unconstrained continuous minimization problems within the classical framework where the functions f are defined on Hilbert spaces $(E, \langle\cdot,\cdot\rangle)$ (e.g., $\mathbb{R}^d$ with the dot product). We will also assume that the functions f are differentiable on E.

Remark.

The fundamental examples in linear and logistic regressions fit within this chapter framework. This is also true for ridge and generalized linear regressions. However, the chapter does not cover the LASSO model, whose criterion is not differentiable.

7.1.2 Classical Questions in Minimization Problems

The classical questions in a minimization problem concern the existence of a minimizer and its uniqueness. Indeed, several scenarios can arise:

- The function f is not bounded below. In this case, the infimum of f over E (i.e., the greatest of its lower bounds) is
$$\inf_{x\in E} f(x) = -\infty.$$
- The function f is bounded below, but its infimum is not reached. In this case, $\inf_{x\in E} f(x)$ is finite, but there does not exist a minimum of f, and
$$\arg\min_{x\in E} f(x) = \emptyset.$$
- The function f is bounded below, and the infimum is reached. In other words, there exists a minimizer of f, which can also be expressed as
$$\arg\min_{x\in E} f(x) \neq \emptyset.$$
- The function f is bounded below, and the infimum is reached at a unique point x^*. In other words, there exists a unique minimizer of f, which can also be expressed as
$$\arg\min_{x\in E} f(x) = \{x^*\}.$$

Thus, the task is to determine under what conditions a minimizer of a function exists and whether that minimizer is unique. Another important question is the characterization of minimizers when they exist.

Example 7.4: Linear model

In linear regression, two situations may arise depending on the rank of the matrix X. If X is not full rank, then $\arg\min_{\theta\in\mathbb{R}^{p+1}} \|Y - X\theta\|^2$ contains an infinite number of elements corresponding to the solutions of the linear system $X^\top X\theta = X^\top Y$. If X is full rank, $\arg\min$ reduces to a single element, corresponding to the unique solution of the linear system.

7.2 Existence of Minima

7.2.1 Conditions for Existence

We consider functions f defined on a normed vector space E of finite dimension.

Theorem 7.1: Extreme Values

Let V be a compact (closed and bounded) subset of E. If f is a continuous function on V, then f is bounded and reaches its bounds.

This theorem guarantees the existence of a minimum for a continuous function on a compact set. This result can be extended to unbounded sets using the notion of coercivity.

Definition 7.1: Coercivity

A function $f : E \to \mathbb{R}$ is coercive if

$$\lim_{\|x\|\to+\infty} f(x) = +\infty.$$

Theorem 7.2: Extreme Values (Extension)

Let V be a non-empty closed subset of E, and let f be a continuous and coercive function on E. Then f is bounded below and reaches its infimum on V. In other words, there exists $x^* \in V$ such that

$$f(x^*) \le f(x), \quad \forall x \in V.$$

Proof: Let $x_0 \in V$. Since f is coercive, there exists a constant $c > 0$ such that $f(x) > f(x_0)$ for all $x \in V$ with $\|x\| > c$. Consider the set

$$B = \{x \in V : \|x\| \le c\}.$$

This set is the intersection of the closed ball centered at 0 with radius c and the set V, which is closed by assumption. Thus, B is a closed and bounded (hence compact) set containing x_0. Since f is continuous, the *Extreme Value Theorem* guarantees the existence of a point $x^* \in B$ minimizing f on B. Let $f^* = \min(f(x_0), f(x^*))$. From the above, $f(x) \ge f^*$ for all $x \in V$, and the minimum is reached at x_0 or x^*. Hence, f is bounded below and reaches its infimum. □

Remark.

Minimizing sequences

Consider a sequence $(x_n)_{n\in\mathbb{N}}$ such that $f(x_n) \underset{n\to\infty}{\longrightarrow} I$. Such a sequence is called **minimizing sequence**. Thanks to the boundness of V, or the coercivity of f, the sequence is bounded. Thus, Bolzanno-Weierstrass theorem induces the existence of a sub-sequence $x_{\phi(n)}$ that converges towards $x \in V$, as V is closed. Moreover, as f is continuous, $f(x) = \lim_{n\to\infty} f(x_{\phi(n)}) = I$, so that $x_{\phi(n)}$ converges towards x in which the minimum of f on V is reached.

7.2.2 Characterization of Local Minima

Theorem 7.3

Let f be a function defined on an open set $U \subset E$ and differentiable at a point $x^* \in U$. If f has a local minimum at x^*, then

$$df_{x^*}(h) = 0, \quad \forall h \in E.$$

This equation is called the **Euler equation**. Points x^* satisfying this equation are referred to as **critical points** of f. The theorem asserts that points achieving a local minimum of f must satisfy the Euler condition. Thus, the local minima of f can be sought amongs its critical points. However, the theorem does not provide a sufficient condition for a critical point to achieve a local minimum. Therefore, each critical point must be verified to determine whether it reaches a local minimum or not.

If the gradient is associated with the differential, the Euler condition can also be written as:

$$\langle \nabla f(x^*), h\rangle = 0, \quad \forall h \in E,$$

or equivalently,

$$\nabla f(x^*) = 0.$$

Proof: Let $h \in E$. Define the function φ such that $\varphi(t) = f(x^* + th)$. Since f is differentiable at x^*, φ is differentiable at $t = 0$, and

$$\varphi'(0) = df_{x^*}(h).$$

Moreover, for sufficiently small t, $x^* + th \in U$ and

$$\varphi(t) - \varphi(0) = f(x^* + th) - f(x^*) \geq 0,$$

since f has a local minimum at x^*. Thus,

$$\frac{\varphi(t) - \varphi(0)}{t} \begin{cases} \geq 0 & \text{if } t > 0, \\ \leq 0 & \text{if } t < 0. \end{cases}$$

Taking the limits, we obtain:

$$\varphi'(0) = \lim_{t \to 0} \frac{\varphi(t) - \varphi(0)}{t} = 0.$$

Therefore, $df_{x^*}(h) = 0$. □

Theorem 7.4

Let f be twice differentiable on an open set $U \subset E$. If f has a local minimum at $x^* \in U$, then

$$d^2 f_{x^*}(h,h) \geq 0, \quad \forall h \in E.$$

This condition is not sufficient, as demonstrated by the counterexample $f(x) = x^3$ at $x^* = 0$.

Proof: Let h be any point in E. For t sufficiently small, we have $x^* + th \in U$ and $f(x^* + th) - f(x^*) \geq 0$. Moreover, using the Taylor theorem of order 2 at x^*, we have

$$f(x^* + th) - f(x^*) = df_{x^*}(th) + \frac{1}{2} d^2 f_{x^*}(th, th) + o(|th|^2).$$

Now, from the first-order characterization of local minima, $df_{x^*} \equiv 0$, giving

$$f(x^* + th) - f(x^*) = \frac{t^2}{2} d^2 f_{x^*}(h,h) + o(t^2).$$

Thus, for t sufficiently small,

$$\frac{t^2}{2}\left(d^2 f_{x^*}(h,h) + \epsilon(t)\right) \geq 0,$$

for some function ϵ tending to 0 as t approaches 0.

Assume, for contradiction, that $d^2 f_{x^*}(h,h) < 0$. For t small enough,

$$\frac{t^2}{2}(df_{x^*}(h,h) + \epsilon(t)) < 0,$$

which contradicts the above. Consequently, $d^2 f_{x^*}(h,h) \geq 0$. □

Theorem 7.5

Let f be a differentiable function on an open set U of E such that the Euler condition is satisfied at a point x^*: $df_{x^*}(h) = 0, \forall h \in E$.
Condition 1. If f is twice differentiable at x^* and there exists $\alpha > 0$ such that

$$d^2 f_{x^*}(h,h) \geq \alpha |h|^2, \forall h \in E,$$

then f has a strict local minimum at x^*.
Condition 2. If f is twice differentiable on U and there exists a ball $B \subset U$ containing x^* such that

$$d^2 f_x(h,h) \geq 0, \forall x \in B, \forall h \in E,$$

then f has a strict local minimum at x^*.

Proof: Condition 1. Applying the Taylor theorem of order 2 to f at x^* and using the Euler condition, we have

$$f(x^*+h)-f(x^*)=\frac{1}{2}d^2f_{x^*}(h,h)+|h|^2\epsilon(h).$$

Using the assumption on the second-order differential, it follows that

$$f(x^*+h)-f(x^*)\geq |h|^2\left(\frac{\alpha}{2}+\epsilon(h)\right).$$

For h sufficiently small, $\frac{\alpha}{2}+\epsilon(h)>0$. Therefore, $f(x^*+h)-f(x^*)>0$ for h in the neighborhood of 0.

Condition 2. According to the second-order Taylor expansion for f at x^* and the Euler condition, there exists $c\in[x^*,x^*+h]$,

$$f(x^*+h)-f(x^*)=\frac{1}{2}d^2f_c(h,h).$$

Thus, using the assumption on the second-order differential, $f(x^*+h)-f(x^*)\geq 0$. □

7.3 Minimization of Convex Functions

In this section, we focus on the minimization of convex functions.

7.3.1 Definitions and Properties

Definition 7.2: Convex Set

A set C in a vector space E is convex if:

$$\forall x,y\in C, \forall t\in[0,1],\quad tx+(1-t)y\in C,$$

which is equivalent to saying that:

$$\forall x,y\in C,\ [x,y]\subset C.$$

For example, in $\mathbb{R}^d$, balls are convex.

Proposition 7.6

The intersection of any collection of convex sets is a convex set.

Definition 7.3: Convex function

Let C be a convex set in E, and let $f:C\longrightarrow\mathbb{R}$ be a function. We say that f is convex on C if:

$$\forall x,y\in C, \forall t\in[0,1],\quad f(tx+(1-t)y)\leq tf(x)+(1-t)f(y).$$

The function f is strictly convex if:

$$\forall x,y\in C, \forall t\in(0,1),\quad f(tx+(1-t)y)< tf(x)+(1-t)f(y).$$

This definition captures the geometric intuition that between any two points x and y, the graph of a convex function on $[x,y]$ lies below the line segment $[f(x),f(y)]$.

Example 7.5: Linear map

A linear map on E is convex on E. However, it is not strictly convex.

Example 7.6: Norm

A norm on E is convex on E. This follows from the triangle inequality.

Numerous characterizations of convex functions exist. Here, we examine a few for differentiable functions.

Theorem 7.7: Characterization of Differentiable Convex Functions (Order 1)

Let U be an open subset of E, f a differentiable function on U, and C a convex subset of U.

1. f is convex on C if and only if:

$$f(y) \geq f(x) + df_x(y-x), \quad \forall x, y \in C.$$

2. f is strictly convex on C if and only if:

$$f(y) > f(x) + df_x(y-x), \quad \forall x, y \in C, x \neq y.$$

Proof: We limit ourselves to proving the first characterization.

Assume f is convex. For any $t \in (0,1)$, we have:

$$f(x + t(y-x)) \leq (1-t)f(x) + tf(y),$$

which implies:

$$\frac{f(x+t(y-x)) - f(x)}{t} \leq f(y) - f(x).$$

By letting $t \longrightarrow 0^+$, we get:

$$\lim_{t \to 0^+} \frac{f(x+t(y-x)) - f(x)}{t} \leq f(y) - f(x).$$

Since f is differentiable, this yields:

$$df_x(y-x) \leq f(y) - f(x).$$

Conversely, suppose $f(y) \geq f(x) + df_x(y-x)$, $\forall x, y \in C$. Fix $t \in [0,1]$ and deduce that:

$$f(y) \geq f(tx + (1-t)y) + df_{tx+(1-t)y}(-t(x-y)),$$

and:

$$f(x) \geq f(tx + (1-t)y) + df_{tx+(1-t)y}((1-t)(x-y)).$$

Adding these inequalities gives:

$$(1-t)f(y) + tf(x) \geq f(tx + (1-t)y),$$

proving the convexity of f. □

Theorem 7.8: Characterization of Differentiable Convex Functions (Second-Order)

Let U be an open subset of a vector space E, f a twice-differentiable function on U, and C a convex subset of U.

- **Characterization 1:** f is convex on C if and only if:

$$d^2f_x(y-x, y-x) \geq 0, \quad \forall x, y \in C.$$

- **Characterization 2:** f is strictly convex on C if and only if:

$$d^2f_x(y-x, y-x) > 0, \quad \forall x, y \in C, x \neq y.$$

Proof: We restrict ourselves to proving Characterization 1.

Assume f is convex on C. Define:

$$G(y) = f(y) - df_x(y).$$

The function G is twice differentiable, and:

$$dG_y(h) = df_y(h) - df_x(h), \quad d^2G_y(h,h) = d^2f_y(h,h).$$

In particular, we note that $dG_x \equiv 0$. Applying the second-order Taylor expansion to G at x, we get:

$$G(x+h) - G(x) = d^2f_x(h,h) + ||h||^2\epsilon(h).$$

Now, $G(x+h) - G(x) = f(x+h) - f(x) + df_x(h) \geq 0$, since f is convex. Thus:

$$d^2f_x(h,h) + ||h||^2\epsilon(h) \geq 0.$$

For $h = t(y-x)$, this gives:

$$t^2\left(d^2 f_x(y-x, y-x) + \|y-x\|^2 \epsilon(t)\right) \geq 0,$$

implying:

$$d^2 f_x(y-x, y-x) \geq 0.$$

Conversely, assume $d^2 f_x(y-x, y-x) \geq 0 \, \forall x, y \in C$. Applying the second-order Taylor expansion to f at x, there exists $c \in]x, y[$ such that:

$$f(y) - f(x) = df_x(y-x) + \frac{1}{2} d^2 f_c(y-x, y-x).$$

Since $c \in]x, y[$, we have $y - x = \alpha(y-c)$ for some $\alpha \in]0,1[$. Thus:

$$f(y) - f(x) - df_x(y-x) = \frac{\alpha^2}{2} d^2 f_c(y-c, y-c).$$

Using the assumption, we obtain:

$$f(y) - f(x) - df_x(y-x) \geq 0,$$

for all $x, y \in C$. According to the first-order characterization of convex functions, this implies that f is convex on C. □

7.3.2 Existence and Uniqueness Conditions

Lemma 7.9: Uniqueness

Let E be a normed vector space and f a strictly convex function from E to $\mathbb{R}$. There exists at most one point x^* in E that achieves a global minimum of f on E.

Proof: Let x^* be a minimizer of f and $y \in E \setminus \{x^*\}$ arbitrary. Set $z = y - x^*$ and $t \in (0,1]$. Due to the strict convexity of f, we have:

$$f(x^* + tz) = f((1-t)x^* + ty) < (1-t)f(x^*) + tf(y).$$

Thus,

$$f(x^* + tz) - f(x^*) < t(f(y) - f(x^*)),$$

or equivalently,

$$f(y) - f(x^*) > \frac{f(x^* + tz) - f(x^*)}{t}.$$

Now, since $f(x^* + tz) - f(x^*) \geq 0$ (as x^* is a minimizer of f), it follows that $f(y) > f(x^*)$. In other words, y is not a minimizer of f. □

By combining this results with the extreme value theorem, we obtain a fundamental result of existence and uniqueness for convex functions.

Theorem 7.10: Existence and Uniqueness

Let E be a finite-dimensional Hilbert space and f a continuous, strictly convex, and coercive function from E to $\mathbb{R}$. There exists a unique element $x^* \in E$ that achieves a global minimum of f on E.

Moreover, when it exists, the minimizer of a differentiable convex function can be characterized as follows:

Theorem 7.11

Let f be a function defined on E with values in $\mathbb{R}$, convex on E, and differentiable at a point x^* in E. The function f achieves a global minimum at x^* if and only if

$$df_{x^*}(h) = 0, \quad \forall h \in E.$$

Proof: The implication has already been proven (it corresponds to Euler's condition).

Let us prove the converse. Due to the first-order characterization of convex functions,

$$f(y) \geq f(x^*) + df_{x^*}(y - x^*).$$

Now, by hypothesis, $df_{x^*}(h) = 0, \forall h$. Hence,

$$f(y) \geq f(x^*).$$

□

7.4 Exercises

Exercise 7.1: Quadratic Functional

Let A be a symmetric positive-definite matrix of size $p \times p$ and $b \in \mathbb{R}^p$. For all $\theta \in \mathbb{R}^p$, define the quadratic functional:

$$J(\theta) = \frac{1}{2}\langle A\theta, \theta\rangle - \langle b, \theta\rangle.$$

We are interested in minimizing J over $\mathbb{R}^p$.

1. Show that for all $\theta \in \mathbb{R}^p$,
$$\langle A\theta, \theta\rangle \geq \alpha|\theta|^2,$$
and then that
$$J(\theta) \geq \frac{\alpha}{2}((|\theta| - \beta)^2 - \beta^2),$$
where $\alpha > 0$ and β are constants to be determined.
2. Deduce that J is coercive.
3. Show that J admits a global minimum on $\mathbb{R}^p$, which is reached at a point $\theta \in \mathbb{R}^p$.
4. Compute the differential of J and verify that J is strictly convex.
5. Deduce that the global minimum of J is reached at a unique point $\theta \in \mathbb{R}^p$. Characterize this point.
6. Is the existence of a global minimum guaranteed if A is positive semidefinite (not positive definite)? Consider the case where b is nonzero and satisfies $Ab = 0$.

Exercise 7.2: Linear Model

Consider the linear model:

$$Y = X\theta + \epsilon,$$

where ϵ is a Gaussian vector of size n, centered, with variance matrix $\sigma^2 I$, θ represents parameters in $\mathbb{R}^p$, and X is a matrix of size $n \times p$ with rank p. Using the exercise on quadratic functionals, show that the maximum likelihood estimator for this model exists and is unique.

Exercise 7.3: Logistic Regression

Consider a sample of n independent random variables Y_i following a Bernoulli distribution with parameter p_i. Assume that for each i, the parameter p_i depends on explanatory variables $x_i \in \mathbb{R}^p$ and unknown parameters $\theta \in \mathbb{R}^p$ through the relation $p_i = \sigma(x_i^T\theta)$, where $\sigma(t) = \frac{1}{1+e^{-t}}$. Assume that the family formed by the variables x_i has rank p.

1. Show that if it exists, the maximum likelihood estimator of θ for this statistical model is unique.
2. Characterize this estimator.
3. Assume there exists θ such that for all i, $y_i = 1$, $x_i^T\theta > 0$, and for all j, $y_j = 0$, $x_j^T\theta < 0$. Show that in this case, the maximum likelihood estimator does not exist.

Exercise 7.4: Multinomial Regression / Softmax

Let $K \in \mathbb{N}$ and Y be a random variable taking values in the set of labels (classes) $\{1, \cdots, K\}$. Assume that the probability distribution of Y depends on a regression vector $x \in \mathbb{R}^p$ and a set of $\theta = (\theta^{(k)})_{k=1}^K$, where $\theta^{(k)} \in \mathbb{R}^{p+1}$, through the softmax model:

$$\forall k \in \{1, \cdots, K\}, \quad \mathbb{P}(Y = k) = \frac{\exp(\langle\theta^{(k)}, \tilde{x}\rangle)}{\sum_{m=1}^K \exp(\langle\theta^{(m)}, \tilde{x}\rangle)},$$

with $\tilde{x} = \begin{pmatrix} 1 \\ x \end{pmatrix}$.

1. Verify that this model generalizes logistic regression.
2. Consider a sample of independent random variables Y_i satisfying the relation above for associated variables x_i. Specify the likelihood of the sample.
3. Show that finding the maximum likelihood estimator for the sample is equivalent to finding parameters θ that minimize the cross-entropy criterion:

   $$\mathcal{S}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} z_i^{(k)} \log\left(p_i^{(k)}(\theta)\right),$$

   where

   $$z_i^{(k)} = \mathbb{I}_{y_i=k},$$

   and

   $$p_i^{(k)}(\theta) = \frac{\exp(\langle \theta^{(k)}, \tilde{x}_i \rangle)}{\sum_{m=1}^{K} \exp(\langle \theta^{(m)}, \tilde{x}_i \rangle)}.$$

4. Show that there cannot be uniqueness of the solution to this problem.
5. Show that a solution to this problem necessarily satisfies the equation:

   $$\frac{1}{n}\sum_{i=1}^{n} z_i^{(k)}\left(p_i^{(k)}(\theta) - 1\right)\tilde{x}_i = 0.$$

Chapter 8 Algorithms for unconstrained optimisation

8.1 Descent Methods

Given a function f defined on a Hilbert space E with values in $\mathbb{R}$, the goal is to numerically approximate a minimizer of the function:

$$x^* \in \arg\min_{x \in E} f(x).$$

To achieve this, descent methods are employed, which involve constructing a sequence $(x_k)_k$ of points in E using an iterative process of the form:

$$x_{k+1} = x_k + \rho_k d_k,$$

where $\rho_k > 0$ is a step size and $d_k \in E$ is a descent direction, ensuring that:

$$f(x_{k+1}) < f(x_k).$$

A fundamental question in designing descent methods is finding appropriate descent directions. For convex and differentiable functions, this is addressed by the following characterization:

Proposition 8.1

Let $x \in E$ and f be a convex function on E, differentiable at x. A point $d \in E$ is a descent direction at x if and only if:

$$\langle \nabla f(x), d \rangle < 0.$$

Proof: By the first-order characterization of convex functions:

$$f(x) \geq f(x + \rho d) - \langle \nabla f(x), \rho d \rangle.$$

Thus, $f(x + \rho d) > f(x)$ holds if and only if $\langle \nabla f(x), d \rangle < 0$. □

We can verify that the direction

$$d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$$

satisfies the characterization. In other words, the direction opposite to the gradient of f at x is a descent direction for f. Consequently, many descent methods for convex and differentiable functions are constructed based on the gradient of f. These are referred to as gradient descent methods. In the following, we focus on a form of gradient descent expressed as:

$$x_{k+1} = x_k - \rho_k \nabla f(x_k).$$

This is known as gradient descent with variable step. A particular case arises when the step size is fixed across iterations ($\forall k, \rho_k = \rho$), referred to as fixed-step gradient descent.

Another specific case involves determining an optimal step at each iteration. The step size is optimal when it produces the greatest decrease in a fixed direction. Specifically, for a direction d, a step size ρ^* is optimal at $x \in E$ if the function $\varphi(\rho) = f(x) - f(x+\rho d)$ reaches a minimum at ρ^*. If f is convex and differentiable, a necessary and sufficient condition for ρ^* to be optimal is that it satisfies the Euler condition $\varphi'(\rho) = 0$, i.e.,

$$\langle \nabla f(x+\rho d), d\rangle = 0.$$

In the case of gradient descent, this condition translates at each iteration into:

$$\langle \nabla f(x_k - \rho \nabla f(x_k)), \nabla f(x_k)\rangle = 0.$$

Using a descent method to minimize a function raises several questions:

- Does the method converge to a point x^* in $\arg\min_x f(x)$?
- If so, how fast is the convergence ?

Answering these questions has practical significance. It ensures the user that the method will work correctly and helps establish suitable conditions. It may also provide an estimate of the method precision.

8.2 Convergence of Gradient Descent Methods

The study of the convergence of gradient descent methods requires strong assumptions that go beyond those ensuring the existence and uniqueness of the solution to the optimization problem. A key assumption in this context is ellipticity.

Definition 8.1: Elliptic function

Let f be a function defined on a Hilbert space E with values in $\mathbb{R}$, and let $\alpha > 0$. The function f is said to be α-elliptic if it is continuously differentiable on E and satisfies:

$$\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \alpha \|x-y\|^2, \quad \forall x, y \in E.$$

As the following results demonstrate, ellipticity is sufficient to guarantee the existence and uniqueness of a minimization problem.

Theorem 8.2: Properties of elliptic functions

If a function f is α-elliptic on E, then

$$f(y) - f(x) \geq \langle \nabla f(x), y-x\rangle + \frac{\alpha}{2}\|y-x\|^2 \tag{8.1}$$

holds for any x, y.

Corollary 8.3: Ellipticity, convexity and ellipticity

An α-elliptic function f on E is strictly convex and coercive on E.

Corollary 8.4

If f is an elliptic function on E, then it admits a global minimum, which is reached at a unique point.

Proof: Since f is assumed to be C^1, the Taylor expansion with an integral remainder gives:

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y-x)), y-x\rangle dt.$$

This also leads to:

$$f(y) - f(x) = \langle \nabla f(x), y-x\rangle + \int_0^1 \langle \nabla f(x+t(y-x)) - \nabla f(x), y-x\rangle dt.$$

Since f is elliptic, we deduce:

$$f(y) - f(x) \geq \langle \nabla f(x), y-x\rangle + \int_0^1 \alpha t \|y-x\|^2 dt \geq \langle \nabla f(x), y-x\rangle + \frac{\alpha}{2}\|y-x\|^2.$$

Thus, $f(y) - f(x) > \langle \nabla f(x), y - x \rangle$ whenever $x \neq y$, which is equivalent to f being strictly convex. Furthermore, by taking $x = 0$, we see that:

$$f(y) \geq f(0) - \|\nabla f(0)\|\|y\| + \frac{\alpha}{2}\|y\|^2.$$

As a result, $\lim_{\|y\| \to +\infty} f(y) = +\infty$, meaning that f is coercive. □

Theorem 8.5: Convergence of the gradient descent with optimal steps

Let $E = \mathbb{R}^n$ and f be an elliptic function on E. The sequence $(x_k)_k$, defined iteratively by:

$$x_{k+1} = x_k - \rho_k \nabla f(x_k), \quad \text{with} \quad \rho_k = \arg\min_{\rho > 0} f(x_k - \rho \nabla f(x_k)),$$

converges to:

$$x^* = \arg\min_{x \in E} f(x).$$

Proof: Without loss of generality, assume $\nabla f(x_k) \neq 0$ for all k. If this is not the case, the algorithm converges in a finite number of iterations.

Since f is elliptic, it is coercive and strictly convex. Thus, the function $\varphi(\rho) = f(x_k - \rho \nabla f(x_k))$ is also coercive and strictly convex. Therefore, it admits a global minimum at a unique point ρ_k such that:

$$\varphi'(\rho) = 0 \Leftrightarrow \langle \nabla f(x_k - \rho \nabla f(x_k)), \nabla f(x_k) \rangle = 0.$$

From $x_{k+1} = x_k - \rho_k \nabla f(x_k)$, it follows that:

$$\langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle = 0. \tag{8.2}$$

This implies successive descent directions are orthogonal. Additionally, using ellipticity, we obtain:

$$f(x_k) - f(x_{k+1}) \geq \frac{\alpha}{2}\|x_k - x_{k+1}\|^2. \tag{8.3}$$

Hence, $f(x_k) \geq f(x_{k+1})$, meaning $(f(x_k))_k$ is decreasing. Since it is bounded below, it converges. Thus:

$$\lim_{k \to +\infty} f(x_k) - f(x_{k+1}) = 0.$$

From inequality (8.3), it follows that:

$$\lim_{k \to +\infty} \|x_k - x_{k+1}\|^2 = 0.$$

This implies $(x_k)_k$ is bounded. Using the continuity of ∇f on compact sets:

$$\lim_{k \to +\infty} \|\nabla f(x_{k+1}) - \nabla f(x_k)\| = 0.$$

Moreover, combining (8.2) and the Cauchy-Schwarz inequality yields:

$$\|\nabla f(x_k)\|^2 = \langle \nabla f(x_k), \nabla f(x_k) - \nabla f(x_{k+1}) \rangle \leq \|\nabla f(x_k)\|\|\nabla f(x_k) - \nabla f(x_{k+1})\|.$$

Consequently:

$$\|\nabla f(x_k)\| \leq \|\nabla f(x_k) - \nabla f(x_{k+1})\|.$$

Thus:

$$\lim_{k \to +\infty} \|\nabla f(x_k)\| = 0.$$

Finally, the sequence $(x_k)_k$ converges to x^*, where $\nabla f(x^*) = 0$, completing the proof. □

Theorem 8.6: Convergence of the gradient descent with variable steps

Let E be a Hilbert space and $f : E \to \mathbb{R}$ a differentiable function. Assume that there exist $\alpha > 0$ and $M > 0$ such that the following conditions are satisfied:

$$\begin{aligned} &\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \alpha \|y - x\|^2, && \forall x, y \in E, \\ &\|\nabla f(y) - \nabla f(x)\| \leq M\|y - x\|, && \forall x, y \in E. \end{aligned}$$

If there exist constants $a > 0$ and $b > 0$ such that, for all k,

$$0 < a \leq \rho_k \leq b < \frac{2\alpha}{M^2},$$

then the sequence (x_k) defined by

$$x_{k+1} = x_k - \rho_k \nabla f(x_k),$$

converges to

$$x^* = \arg\min_{x \in E} f(x).$$

Moreover, there exists $\beta \in (0,1)$ such that

$$\|x_k - x^*\| \leq \beta^k \|x_0 - x^*\|.$$

Proof: Since $x_{k+1} = x_k - \rho_k \nabla f(x_k)$ and $\nabla f(x^*) = 0$, we have

$$x_{k+1} - x^* = x_k - x^* - \rho_k(\nabla f(x_k) - \nabla f(x^*)).$$

Expanding the squared norm yields:

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\rho_k \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^*\rangle + \rho_k^2 \|\nabla f(x_k) - \nabla f(x^*)\|^2.$$

Using the assumptions, it follows that:

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 (1 - 2\alpha\rho_k + M^2\rho_k^2).$$

Let $g(\rho) = 1 - 2\alpha\rho + M^2\rho^2$. Analyzing the variations of g shows that, over any interval $[a,b] \subset (0, \frac{2\alpha}{M^2})$,

$$g(\rho) \leq \max(g(a), g(b)) < 1.$$

Letting $\beta = \max(g(a), g(b))$, we deduce:

$$\|x_{k+1} - x^*\|^2 \leq \beta \|x_k - x^*\|^2,$$

for $\rho_k \in [a,b]$. By induction, it follows that:

$$\|x_{k+1} - x^*\|^2 \leq \beta^{k+1} \|x_0 - x^*\|^2.$$

Hence,

$$\lim_{k \to \infty} x_k = x^*.$$

□

8.3 Exercises

8.3.1 Quadratic Functional

We consider the minimization of the quadratic functional:

$$J(\theta) = \frac{1}{2}\langle A\theta, \theta\rangle - \langle b, \theta\rangle,$$

where A is a symmetric positive definite matrix of size $p \times p$, and $b \in \mathbb{R}^p$. For this purpose, we study a gradient descent algorithm of the form:

$$\theta_{k+1} = \theta_k - \rho_k \nabla J(\theta_k). \tag{8.4}$$

Gradient Descent with Variable Step Size

1. Show that the functional J is α-elliptic for a specific value of α that needs to be determined.

2. Prove that, for any $u, v \in \mathbb{R}^p$,

$$\|\nabla J(v) - \nabla J(u)\|^2 \leq M\|v - u\|^2,$$

   where M is a constant to be determined.

3. Deduce a condition on ρ_k such that the sequence defined by Equation (8.4) converges to the minimizer of J.

Gradient Descent with Optimal Step Size In the sequence defined by Equation (8.4), let us define:

$$\rho_k = \arg\min_{\rho\in\mathbb{R}} J(\theta_k - \rho\nabla J(\theta_k)). \tag{8.5}$$

1. Assume that $\nabla J(\theta_k) \neq 0$ and define:

$$g_k(\rho) = J(\theta_k - \rho\nabla J(\theta_k)).$$

2. Show that g_k is strictly convex and coercive.

3. Deduce that the descent step size ρ_k in Equation (8.5) is well-defined (i.e., exists and is unique) and is given by:

$$\rho_k = \frac{\|\nabla J(\theta_k)\|^2}{\langle A\nabla J(\theta_k), \nabla J(\theta_k)\rangle}.$$