A refresher on Matrix Analysis and Optimization

Hamza Ennaji
Univ. Grenoble Alpes, CNRS, Grenoble INP*, LJK
hamza.ennaji@univ-grenoble-alpes.fr

September 21, 2026

Abstract

These notes accompany the refresher course on Matrix Analysis and Optimization. The first chapter reviews the essentials of linear algebra and matrix analysis (vector spaces, linear maps, norms and inner products, linear systems, eigenvalues and the SVD). The second chapter introduces the basics of convex optimization (convex sets and functions, optimality conditions) and presents the two fundamental minimization algorithms: gradient descent and Newton's method.

Contents

These are short lecture notes for the refresher course on Matrix Analysis and Optimization. Their aim is to recall the essential notions and to give pointers to prepare your main courses; they are not a complete mathematics course and contain no proofs. Only basic knowledge of mathematics and programming is assumed.

These notes are mainly based on the following references (full details at the end of the document):

- [1] for linear algebra and matrix analysis (Chapter 1), from a machine learning perspective;
- [2, 3] and [4] for convex optimization (Chapter 2);
- [3] for algorithms and numerical aspects;
- [5, 6] and [7] for those who want to go deeper into convex geometry, convex analysis and variational analysis.

French-speaking students who want the essentials of convex analysis and optimization, with proofs and exercises, without diving into these textbooks, may also consult my lecture notes (in French) for the Ensimag course *Bases d'analyse convexe et optimisation* [8], of which Chapter 2 is a condensed version.

These notes are regularly updated. Please do not hesitate to report any typo, error or suggestion for improvement.

Chapitre 1.

Matrix Analysis

1.1 Matrices: first definitions

1.1.1 What is a matrix?

Let $m, n \in \mathbb{N}^*$. A real (m, n) matrix is an array of $m \times n$ real numbers arranged in m rows and n columns:

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}.$$

We write $A = (a_{ij})_{1 \le i \le m,\, 1 \le j \le n}$ in a compact way, and $\mathcal{M}_{m,n}(\mathbb{R})$ for the set of all such matrices. When $m = n$ we say that A is **square** and write $\mathcal{M}_n(\mathbb{R}) := \mathcal{M}_{n,n}(\mathbb{R})$. A vector $x \in \mathbb{R}^n$ is identified with a column matrix in $\mathcal{M}_{n,1}(\mathbb{R})$.

1.1.2 Matrix operations

- **Addition & scaling:** for $A, B \in \mathcal{M}_{m,n}(\mathbb{R})$ and $\lambda \in \mathbb{R}$, $(A+B)_{ij} = a_{ij} + b_{ij}$ and $(\lambda A)_{ij} = \lambda a_{ij}$.
- **Product:** if $A \in \mathcal{M}_{m,p}(\mathbb{R})$ and $B \in \mathcal{M}_{p,n}(\mathbb{R})$, then $C = AB \in \mathcal{M}_{m,n}(\mathbb{R})$ with

  $$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj} \qquad \text{(row } i \text{ of } A \text{ “times” column } j \text{ of } B\text{)}.$$

  The product is only defined when the number of columns of A equals the number of rows of B. Moreover $\mathbf{I}_m A = A \mathbf{I}_p = A$ for $A \in \mathcal{M}_{m,p}(\mathbb{R})$, where

  $$\mathbf{I}_m = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \in \mathcal{M}_m(\mathbb{R})$$

  is the **identity matrix** of size m.
- **Associativity:** if $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $B \in \mathcal{M}_{n,p}(\mathbb{R})$ and $C \in \mathcal{M}_{p,q}(\mathbb{R})$, then $(AB)C = A(BC)$.
- **Distributivity:** if $A, B \in \mathcal{M}_{m,n}(\mathbb{R})$ and $C, D \in \mathcal{M}_{n,p}(\mathbb{R})$, then $(A + B)C = AC + BC$ and $A(C + D) = AC + AD$.
- **Transpose:** $A^\top \in \mathcal{M}_{n,m}(\mathbb{R})$ is obtained by interchanging the rows and columns of A, i.e., $(A^\top)_{ij} = a_{ji}$.
- **Trace:** for $A \in \mathcal{M}_n(\mathbb{R})$, $\mathrm{Tr}(A) = \sum_{i=1}^{n} a_{ii}$ (sum of the diagonal entries). It is linear and satisfies $\mathrm{Tr}(A^\top) = \mathrm{Tr}(A)$ and $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $B \in \mathcal{M}_{n,m}(\mathbb{R})$.
- **Inverse:** a square matrix $A \in \mathcal{M}_n(\mathbb{R})$ is **invertible** (regular, non-singular) if there exists $B \in \mathcal{M}_n(\mathbb{R})$ such that $AB = BA = \mathbf{I}_n$. Such a B is unique; we denote it A^{-1}.

Example 1 (Matrix product). Let

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 3 & -1 \end{pmatrix} \in \mathcal{M}_{3,2}(\mathbb{R}), \qquad B = \begin{pmatrix} 1 & 0 & 2 \\ -1 & 1 & 0 \end{pmatrix} \in \mathcal{M}_{2,3}(\mathbb{R}).$$

Then

$$AB = \begin{pmatrix} -1 & 2 & 2 \\ -1 & 1 & 0 \\ 4 & -1 & 6 \end{pmatrix} \in \mathcal{M}_3(\mathbb{R}), \qquad BA = \begin{pmatrix} 7 & 0 \\ -1 & -1 \end{pmatrix} \in \mathcal{M}_2(\mathbb{R}).$$

For instance $(AB)_{11} = 1 \cdot 1 + 2 \cdot (-1) = -1$. Note that AB and BA do not even have the same size. One can check that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA) = 6$.

Remark 1 (Pitfalls of the matrix product). Matrix multiplication behaves differently from the multiplication of real numbers. With

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

we have

- $AB = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \neq BA = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$: the product is **not commutative**;
- $A^2 = AA = 0$ although $A \neq 0$: a product can vanish without any of the factors being zero. In particular $AB = AC$ does not imply $B = C$ in general.

Proposition 1.
1. For $A, B \in \mathcal{M}_{m,n}(\mathbb{R})$ and $C \in \mathcal{M}_{n,p}(\mathbb{R})$: $(AC)^\top = C^\top A^\top$, $(A+B)^\top = A^\top + B^\top$ and $(A^\top)^\top = A$.
2. If $A, C \in \mathcal{M}_n(\mathbb{R})$ are invertible, then AC and $A^\top$ are invertible and

$$(AC)^{-1} = C^{-1}A^{-1}, \qquad (A^\top)^{-1} = (A^{-1})^\top, \qquad (A^{-1})^{-1} = A.$$

Remark 2. In general $(A+B)^{-1} \neq A^{-1} + B^{-1}$. Even worse, $A+B$ need not be invertible when A and B are: take $A = \mathbf{I}_n$ and $B = -\mathbf{I}_n$. And for $A = B = \mathbf{I}_n$, $(A+B)^{-1} = \frac{1}{2}\mathbf{I}_n$ whereas $A^{-1} + B^{-1} = 2\mathbf{I}_n$.

1.1.3 Determinant

The determinant is a number $\det(A)$ associated with every square matrix $A \in \mathcal{M}_n(\mathbb{R})$. We only recall how to compute it and its main properties.

- $n = 2$: $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$.
- $n = 3$ (expansion along the first row):

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

More generally, a determinant can be expanded along any row or column (cofactor expansion).

Proposition 2. Let $A, B \in \mathcal{M}_n(\mathbb{R})$ and $\lambda \in \mathbb{R}$.

1. $\det(AB) = \det(A)\det(B)$, $\det(A^\top) = \det(A)$, $\det(\lambda A) = \lambda^n \det(A)$, $\det(\mathbf{I}_n) = 1$.
2. If A is triangular (see below), $\det(A) = \prod_{i=1}^n a_{ii}$.
3. A is invertible if and only if $\det(A) \neq 0$, and then $\det(A^{-1}) = 1/\det(A)$.

Remark 3. In general $\det(A + B) \neq \det(A) + \det(B)$.

Example 2 (Inverse of a 2×2 matrix)**.** If $ad - bc \neq 0$,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

For instance, $A = \begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix}$ has $\det(A) = 1$ and $A^{-1} = \begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix}$; one checks that $AA^{-1} = \mathbf{I}_2$.

1.1.4 Special square matrices

Let $A \in \mathcal{M}_n(\mathbb{R})$.

- **Diagonal:** $a_{ij} = 0$ for $i \neq j$. We write $A = \mathrm{diag}(a_{11}, \ldots, a_{nn})$.
- **Upper triangular:** $a_{ij} = 0$ for $i > j$.
- **Lower triangular:** $a_{ij} = 0$ for $i < j$.
- **Symmetric matrices:** $\mathcal{S}^n := \{A \in \mathcal{M}_n(\mathbb{R}) \mid A = A^\top\}$.
- **Positive semidefinite matrices:** $\mathcal{S}^n_+ := \{A \in \mathcal{S}^n \mid x^\top Ax \geq 0,\ \forall x \in \mathbb{R}^n\}$.
- **Positive definite matrices:** $\mathcal{S}^n_{++} := \{A \in \mathcal{S}^n \mid x^\top Ax > 0,\ \forall x \in \mathbb{R}^n \setminus \{0\}\}$.
- **Orthogonal matrices:** $\mathcal{O}(n) := \{A \in \mathcal{M}_n(\mathbb{R}) \mid AA^\top = A^\top A = \mathbf{I}_n\}$, i.e., A is invertible and $A^{-1} = A^\top$.

We will write $A \succeq 0$ for $A \in \mathcal{S}^n_+$ and $A \succ 0$ for $A \in \mathcal{S}^n_{++}$.

Example 3.

- $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} \in \mathcal{S}^2$, whereas $\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$ is upper triangular but not symmetric.
- $\mathbf{I}_n \in \mathcal{S}^n_{++}$ since $x^\top \mathbf{I}_n x = \sum_i x_i^2 > 0$ for $x \neq 0$. More generally $\mathrm{diag}(d_1, \ldots, d_n) \in \mathcal{S}^n_{++}$ iff $d_i > 0$ for all i.
- $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \in \mathcal{S}^2_+$ since $x^\top Ax = (x_1 + x_2)^2 \geq 0$, but $A \notin \mathcal{S}^2_{++}$ because $x^\top Ax = 0$ for $x = (1, -1)^\top$.
- The rotation of angle θ, $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, is orthogonal. So is any permutation matrix, e.g. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

1.2 Vector spaces

1.2.1 Groups

Definition 1. Let G be a set endowed with an operation $\star : G \times G \to G$. We say that $(G, \star)$ is a **group** if

1. (closure) $\forall x, y \in G$, $x \star y \in G$;
2. (associativity) $\forall x, y, z \in G$, $x \star (y \star z) = (x \star y) \star z$;
3. (neutral element) there exists $e \in G$ such that $x \star e = e \star x = x$ for all $x \in G$;
4. (inverse) $\forall x \in G$, $\exists y \in G$ such that $x \star y = y \star x = e$. We denote y by x^{-1}, the inverse of x w.r.t. $\star$.

If in addition $x \star y = y \star x$ for all $x, y \in G$, we say that $(G, \star)$ is an **Abelian** (commutative) group.

Example 4.
- $(\mathbb{Z}, +)$, $(\mathbb{R}^*, \cdot)$, $(\mathbb{R}^n, +)$ and $(\mathcal{M}_{m,n}(\mathbb{R}), +)$ are Abelian groups.
- $(\mathbb{N}, +)$ is not a group: 1 has no inverse ($1 + y = 0$ has no solution in $\mathbb{N}$). $(\mathbb{Z}, \cdot)$ is not a group: 2 has no inverse in $\mathbb{Z}$.
- The set $GL_n(\mathbb{R})$ of invertible matrices of $\mathcal{M}_n(\mathbb{R})$, with the matrix product, is a group with neutral element $\mathbf{I}_n$; it is *not* Abelian for $n \geq 2$. The orthogonal matrices $\mathcal{O}(n)$ also form a group for the product.

1.2.2 Vector spaces

Definition 2. A **real vector space** $(E, +, \cdot)$ is a set E, whose elements are called vectors, equipped with an inner operation $+ : E \times E \to E$ (addition of vectors) and an outer operation $\cdot : \mathbb{R} \times E \to E$ (multiplication by scalars), such that:

- $(E, +)$ is an Abelian group;
- for all $\lambda, \mu \in \mathbb{R}$ and $x, y \in E$: $\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y$, $(\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot x$ and $(\lambda\mu) \cdot x = \lambda \cdot (\mu \cdot x)$;
- for all $x \in E$, $1 \cdot x = x$.

Remark 4. The neutral element of the group $(E, +)$ is called the zero vector and denoted 0_E. It depends on the space: $0_{\mathbb{R}^n} = (0, \ldots, 0)^\top$, the zero of $\mathcal{M}_{m,n}(\mathbb{R})$ is the matrix with all entries equal to 0, and the zero of $\mathbb{R}_n[X]$ is the zero polynomial.

Example 5.
- $\mathbb{R}^n$, $\mathcal{M}_{m,n}(\mathbb{R})$ and $\mathbb{R}_n[X]$ (polynomials of degree at most n) are real vector spaces.
- The set $\mathcal{C}([a, b])$ of continuous functions $f : [a, b] \to \mathbb{R}$, with $(f + g)(t) = f(t) + g(t)$ and $(\lambda f)(t) = \lambda f(t)$, is a real vector space.

1.2.3 Vector subspaces

Definition 3. Let $(E, +, \cdot)$ be a real vector space and $F \subset E$ with $F \neq \emptyset$. We say that F is a **vector subspace** of E if $(F, +, \cdot)$ is a vector space with the restricted operations $+_{|F\times F}$ and $\cdot_{|\mathbb{R}\times F}$.

In practice, to show that $F \subset E$ is a vector subspace of E, it suffices to check that

1. $0_E \in F$,
2. $\forall \lambda \in \mathbb{R},\ \forall x, y \in F$: $\lambda x + y \in F$.

Example 6.
- $F = \{x \in \mathbb{R}^3 \mid x_1 + x_2 + x_3 = 0\}$ is a subspace of $\mathbb{R}^3$ (a plane through the origin).
- $G = \{x \in \mathbb{R}^2 \mid x_1 + x_2 = 1\}$ is *not* a subspace of $\mathbb{R}^2$, since $0 \notin G$.
- $\mathcal{S}^n$ is a subspace of $\mathcal{M}_n(\mathbb{R})$. On the other hand, $\mathcal{S}^n_+$ is *not* a subspace: $\mathbf{I}_n \in \mathcal{S}^n_+$ but $-\mathbf{I}_n \notin \mathcal{S}^n_+$. (It is a *convex cone*, a notion which will be important in the optimization chapter.)
- For $A \in \mathcal{M}_{m,n}(\mathbb{R})$, the set $\{x \in \mathbb{R}^n \mid Ax = 0\}$ is a subspace of $\mathbb{R}^n$.

1.2.4 Linear (in)dependence, bases, dimension

Definition 4. A **linear combination** of a family of vectors $\{v_1, \dots, v_k\}$ of E is a vector of the form $v = \sum_{i=1}^k \lambda_i v_i$, with $\lambda_i \in \mathbb{R}$.

Definition 5. A family of vectors $\{v_1, \dots, v_k\}$ of E is said to be **linearly independent** if

$$\sum_{i=1}^k \lambda_i v_i = 0_E,\ \lambda_i \in \mathbb{R} \quad \Longrightarrow \quad \lambda_1 = \cdots = \lambda_k = 0.$$

Otherwise, i.e., if there exists $(\lambda_1, \dots, \lambda_k) \neq (0, \dots, 0)$ such that $\sum_{i=1}^k \lambda_i v_i = 0_E$, we say that $\{v_1, \dots, v_k\}$ is **linearly dependent**.

Remark 5. The question of linear (in)dependence can be reformulated as: is it possible to write 0_E as a nontrivial linear combination of the v_i's? Equivalently, a family is dependent iff one of its vectors is a linear combination of the others.

Example 7.
- In $\mathbb{R}^3$, $v_1 = (1, 0, 1)^\top$, $v_2 = (0, 1, 1)^\top$, $v_3 = (1, 1, 2)^\top$ are linearly dependent since $v_1 + v_2 - v_3 = 0$.
- $v_1 = (1, 0, 1)^\top$ and $v_2 = (0, 1, 1)^\top$ are independent: $\lambda_1 v_1 + \lambda_2 v_2 = (\lambda_1, \lambda_2, \lambda_1 + \lambda_2)^\top = 0$ forces $\lambda_1 = \lambda_2 = 0$.
- Any family containing 0_E is dependent.

Remark 6 (In practice, in $\mathbb{R}^m$)**.** Put the vectors $v_1, \dots, v_k \in \mathbb{R}^m$ as the columns of a matrix $V = [v_1, \dots, v_k] \in \mathcal{M}_{m,k}(\mathbb{R})$. Since $V\lambda = \sum_i \lambda_i v_i$, the family is independent iff $V\lambda = 0$ has the unique solution $\lambda = 0$. When $k = m$, this holds iff $\det(V) \neq 0$.

Definition 6. Let $S = \{v_1, \dots, v_k\} \subset E$. The set of all linear combinations of elements of S is a subspace of E called the **span** of S, denoted $\text{span}(S)$. We say that S is a **generating set** (or that S spans E) if $\text{span}(S) = E$, i.e., if for any $x \in E$ there exist $\lambda_1, \dots, \lambda_k \in \mathbb{R}$ such that $x = \sum_{i=1}^{k} \lambda_i v_i$.

Example 8. In $\mathbb{R}^3$, $\text{span}\{e_1, e_2\} = \{x \in \mathbb{R}^3 \mid x_3 = 0\}$ (the horizontal plane), and $\text{span}\{e_1, e_2, e_3\} = \mathbb{R}^3$, where $e_1 = (1,0,0)^\top$, $e_2 = (0,1,0)^\top$, $e_3 = (0,0,1)^\top$.

Definition 7. A family $\{v_1, \dots, v_n\}$ is a **basis** of E if it is linearly independent and spans E. In this case E is said to be finite-dimensional; all the bases of E have the same number of elements, called the **dimension** of E:

$$\dim(E) = \text{card}(\{v_1, \dots, v_n\}) = n.$$

By convention $\dim(\{0_E\}) = 0$.

Definition 8. A generating set S of E is **minimal** if no proper subset $T \subsetneq S$ spans E. An independent family S is **maximal** if adding any vector of E to it makes it dependent.

Proposition 3. Let E be a vector space of dimension n.

1. A basis is exactly a minimal generating set, or equivalently a maximal linearly independent family.
2. Any independent family has at most n elements; any generating set has at least n elements.
3. A family of exactly n vectors is a basis $\Leftrightarrow$ it is independent $\Leftrightarrow$ it spans E.
4. If F is a subspace of E, then $\dim(F) \leq \dim(E)$, with equality iff $F = E$.

Example 9 (Canonical bases).

- $\{e_1, \dots, e_n\}$, with $e_i = (0, \dots, 0, 1, 0, \dots, 0)^\top$ (1 in position i), is the **canonical basis** of $\mathbb{R}^n$: $\dim(\mathbb{R}^n) = n$.
- The matrices E_{ij} (with 1 in position (i,j) and 0 elsewhere) form a basis of $\mathcal{M}_{m,n}(\mathbb{R})$: $\dim(\mathcal{M}_{m,n}(\mathbb{R})) = mn$.
- $\{1, X, \dots, X^n\}$ is a basis of $\mathbb{R}_n[X]$: $\dim(\mathbb{R}_n[X]) = n+1$.
- $\dim(\mathcal{S}^n) = \frac{n(n+1)}{2}$ (a symmetric matrix is determined by its entries on and above the diagonal).
- $\mathcal{C}([a,b])$ is infinite-dimensional: $\{1, t, t^2, \dots, t^k\}$ is independent for every k.

Example 10. $\{(1,1)^\top, (1,-1)^\top\}$ is a basis of $\mathbb{R}^2$: it has $2 = \dim(\mathbb{R}^2)$ vectors and they are independent since $\det \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = -2 \neq 0$.

1.2.5 Rank of a matrix

Definition 9. The number of linearly independent columns of a matrix $A \in \mathcal{M}_{m,n}(\mathbb{R})$ equals the number of linearly independent rows. This number is called the **rank** of A and is denoted by $\text{rk}(A)$.

Proposition 4. Let $A \in \mathcal{M}_{m,n}(\mathbb{R})$.

1. $\mathrm{rk}(A) = \mathrm{rk}(A^\top)$ (column rank = row rank), and $\mathrm{rk}(A) \leq \min(m, n)$.
2. The columns of A span a subspace $U \subset \mathbb{R}^m$ with $\dim(U) = \mathrm{rk}(A)$. This subspace is called the *image* or *range* of A (see Section 1.3). A basis of U is obtained by Gaussian elimination on A: take the columns of A corresponding to the pivots.
3. The rows of A span a subspace $W \subset \mathbb{R}^n$ with $\dim(W) = \mathrm{rk}(A)$. A basis of W is obtained by Gaussian elimination on $A^\top$.
4. The solutions of $Ax = 0$ form a subspace of $\mathbb{R}^n$ of dimension $n - \mathrm{rk}(A)$, called the *kernel* or *null space* of A.
5. For $A \in \mathcal{M}_n(\mathbb{R})$: A is invertible $\Leftrightarrow \mathrm{rk}(A) = n$.
6. For $b \in \mathbb{R}^m$, the linear system $Ax = b$ has a solution iff $\mathrm{rk}(A) = \mathrm{rk}(A|b)$, where $(A|b) \in \mathcal{M}_{m,n+1}(\mathbb{R})$ denotes the augmented matrix.
7. $\mathrm{rk}(AB) \leq \min(\mathrm{rk}(A), \mathrm{rk}(B))$ for $B \in \mathcal{M}_{n,p}(\mathbb{R})$, and $\mathrm{rk}(A^\top A) = \mathrm{rk}(A)$.

Definition 10. A matrix $A \in \mathcal{M}_{m,n}(\mathbb{R})$ has **full rank** if $\mathrm{rk}(A) = \min(m, n)$, the largest possible rank for a matrix of this size. Otherwise it is **rank deficient**. When $\mathrm{rk}(A) = n$ (resp. m) we say that A has full column (resp. row) rank.

Example 11 (Rank).

- $A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ has two linearly independent rows/columns (the third column is the sum of the first two), so $\mathrm{rk}(A) = 2$.
- $\mathrm{rk}(\mathbf{I}_3) = 3$, since the columns e_1, e_2, e_3 are independent (they span $\mathbb{R}^3$).
- $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 0 & 1 \end{pmatrix}$: the second row is twice the first, and rows 1 and 3 are not proportional, so $\mathrm{rk}(A) = 2$ and $\det(A) = 0$.
- For nonzero $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$, the matrix $uv^\top \in \mathcal{M}_{m,n}(\mathbb{R})$ has rank 1: all its columns are multiples of u.

1.3 Linear maps

Linear maps are the maps which preserve the vector space structure.

1.3.1 Definitions and examples

Definition 11. Let V, W be real vector spaces. A map $\Phi : V \to W$ is **linear** (or a *homomorphism* of vector spaces) if

$$\forall x, y \in V,\ \forall \lambda, \mu \in \mathbb{R} : \qquad \Phi(\lambda x + \mu y) = \lambda \Phi(x) + \mu \Phi(y).$$

The set of linear maps from V to W is denoted $\mathcal{L}(V, W)$; it is itself a vector space.

Remark 7. A linear map always satisfies $\Phi(0_V) = 0_W$. This gives a quick way to detect non-linear maps.

Example 12.
- For $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $\Phi : \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto Ax$ is linear. We will see that every linear map between finite-dimensional spaces is of this form once bases are fixed.
- The rotation $x \mapsto R_\theta x$ and the orthogonal projection $(x_1, x_2, x_3) \mapsto (x_1, x_2, 0)$ are linear.
- The derivative $D : \mathbb{R}_n[X] \to \mathbb{R}_n[X]$, $p \mapsto p'$, is linear; so are $\mathrm{Tr} : \mathcal{M}_n(\mathbb{R}) \to \mathbb{R}$ and $A \mapsto A^\top$.
- $x \mapsto x + b$ with $b \neq 0$ (a translation) is *not* linear since it does not map 0 to 0; it is *affine*. The maps $x \mapsto x^2$ on $\mathbb{R}$ and $x \mapsto \|x\|$ are not linear either.

Definition 12 (Injective, surjective, bijective)**.** Consider a mapping $\Phi : \mathcal{V} \to \mathcal{W}$, where $\mathcal{V}, \mathcal{W}$ are arbitrary sets. Then Φ is called
- **injective** (one-to-one) if $\forall x, y \in \mathcal{V}$: $\Phi(x) = \Phi(y) \Rightarrow x = y$;
- **surjective** (onto) if $\Phi(\mathcal{V}) = \mathcal{W}$, i.e., every $w \in \mathcal{W}$ is the image of some $x \in \mathcal{V}$;
- **bijective** if it is injective and surjective. In this case Φ has an inverse map $\Phi^{-1} : \mathcal{W} \to \mathcal{V}$.

We then introduce the following special cases of linear mappings between vector spaces V and W:
- **Isomorphism:** $\Phi : V \to W$ linear and bijective;
- **Endomorphism:** $\Phi : V \to V$ linear;
- **Automorphism:** $\Phi : V \to V$ linear and bijective;
- $\mathrm{id}_V : V \to V$, $x \mapsto x$, is the **identity mapping** (identity automorphism) of V.

Example 13 (Homomorphism)**.** The mapping $\Phi : \mathbb{R}^2 \to \mathbb{C}$, $\Phi(x) = x_1 + ix_2$, is a homomorphism (here $\mathbb{C}$ is seen as a real vector space):

$$\Phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) = (x_1 + y_1) + i(x_2 + y_2) = \Phi\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \Phi\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \qquad \Phi\left(\lambda\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \lambda\Phi\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

It is moreover bijective (its inverse is $z \mapsto (\mathrm{Re}\, z, \mathrm{Im}\, z)$), hence an isomorphism. This justifies why complex numbers can be represented as pairs in $\mathbb{R}^2$.

Proposition 5. Finite-dimensional vector spaces V and W are isomorphic if and only if $\dim(V) = \dim(W)$.

Remark 8. Consequently every real vector space of dimension n "is" $\mathbb{R}^n$: once a basis is chosen, a vector is identified with its coordinates (Section 1.3.3). For instance $\mathcal{M}_{m,n}(\mathbb{R})$ can be identified with $\mathbb{R}^{mn}$ (this is what `reshape`/`flatten` do in `NumPy`) and $\mathbb{R}_n[X]$ with $\mathbb{R}^{n+1}$.

1.3.2 Kernel and image

Definition 13. Let $\Phi \in \mathcal{L}(V, W)$. The **kernel** and the **image** of Φ are

$$\mathrm{Ker}(\Phi) = \{x \in V \mid \Phi(x) = 0_W\} \subset V, \qquad \mathrm{Im}(\Phi) = \Phi(V) = \{\Phi(x) \mid x \in V\} \subset W.$$

They are vector subspaces of V and W respectively. For a matrix A, $\mathrm{Ker}(A)$ and $\mathrm{Im}(A)$ are those of $x \mapsto Ax$.

Proposition 6. Let $\Phi \in \mathcal{L}(V, W)$ with V finite-dimensional.

1. Φ is injective $\Leftrightarrow \mathrm{Ker}(\Phi) = \{0_V\}$; Φ is surjective $\Leftrightarrow \mathrm{Im}(\Phi) = W$.
2. **Rank–nullity theorem:** $\dim(V) = \dim(\mathrm{Ker}(\Phi)) + \dim(\mathrm{Im}(\Phi))$. For $A \in \mathcal{M}_{m,n}(\mathbb{R})$: $n = \dim(\mathrm{Ker}(A)) + \mathrm{rk}(A)$.
3. If $\dim(V) = \dim(W)$ (e.g. for an endomorphism): Φ injective $\Leftrightarrow$ Φ surjective $\Leftrightarrow$ Φ bijective.

Example 14. Let $\Phi : \mathbb{R}^3 \to \mathbb{R}^2$, $\Phi(x) = (x_1 + x_2,\ x_2 + x_3)^\top$. Then $\Phi(x) = 0$ iff $x = t(1, -1, 1)^\top$, $t \in \mathbb{R}$, so $\mathrm{Ker}(\Phi) = \mathrm{span}\{(1, -1, 1)^\top\}$ has dimension 1. By rank–nullity $\dim(\mathrm{Im}(\Phi)) = 3 - 1 = 2$, so $\mathrm{Im}(\Phi) = \mathbb{R}^2$: Φ is surjective but not injective.

1.3.3 Matrix representation of linear maps

Definition 14 (Coordinates). Let V be an n-dimensional vector space with an ordered basis $B = (b_1, \ldots, b_n)$. Every $x \in V$ has a *unique* representation

$$x = \sum_{i=1}^{n} \alpha_i b_i.$$

The scalars $\alpha_1, \ldots, \alpha_n$ are called the **coordinates** of x w.r.t. B, and $\hat{x} = (\alpha_1, \ldots, \alpha_n)^\top \in \mathbb{R}^n$ is the **coordinate vector** of x w.r.t. B.

Example 15. In $\mathbb{R}^2$, the vector $x = (3, 1)^\top$ has coordinates $(3, 1)^\top$ in the canonical basis, but $(2, 1)^\top$ in the basis $B = ((1, 1)^\top, (1, -1)^\top)$, since $x = 2(1, 1)^\top + 1\,(1, -1)^\top$. The same vector has different coordinates in different bases.

Definition 15 (Transformation matrix). Consider vector spaces V, W with ordered bases $B = (b_1, \ldots, b_n)$ and $C = (c_1, \ldots, c_m)$, and a linear map $\Phi : V \to W$. For $j \in \{1, \ldots, n\}$, let

$$\Phi(b_j) = \alpha_{1j} c_1 + \cdots + \alpha_{mj} c_m = \sum_{i=1}^{m} \alpha_{ij} c_i$$

be the unique representation of $\Phi(b_j)$ w.r.t. C. The $m \times n$ matrix A_Φ with entries $A_\Phi(i, j) = \alpha_{ij}$ is called the **transformation matrix** of Φ (w.r.t. the bases B of V and C of W). In words: *the j-th column of A_Φ contains the coordinates of $\Phi(b_j)$ in the basis C.*

Proposition 7.

1. If $\hat{x}$ are the coordinates of $x \in V$ w.r.t. B and $\hat{y}$ those of $y = \Phi(x)$ w.r.t. C, then $\hat{y} = A_\Phi \hat{x}$.
2. For $\Phi(x) = Ax$ with $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and the canonical bases of $\mathbb{R}^n$ and $\mathbb{R}^m$, $A_\Phi = A$.
3. The composition of linear maps corresponds to the product of matrices: $A_{\Psi \circ \Phi} = A_\Psi A_\Phi$ (with compatible bases). Φ is an isomorphism iff A_Φ is square and invertible, and then $A_{\Phi^{-1}} = A_\Phi^{-1}$.

Example 16. Let $\Phi : V \to W$ with V having basis $B = (b_1, b_2, b_3)$ and W having basis $C = (c_1, \dots, c_4)$, and

$$\Phi(b_1) = c_1 - c_2 + 3c_3 - c_4, \qquad \Phi(b_2) = 2c_1 + c_2 + 7c_3 + 2c_4, \qquad \Phi(b_3) = 3c_2 + c_3 + 4c_4.$$

The transformation matrix A_Φ w.r.t. B and C satisfies $\Phi(b_k) = \sum_{i=1}^{4} \alpha_{ik} c_i$ for $k = 1, 2, 3$ and is given by

$$A_\Phi = [\alpha_1, \alpha_2, \alpha_3] = \begin{pmatrix} 1 & 2 & 0 \\ -1 & 1 & 3 \\ 3 & 7 & 1 \\ -1 & 2 & 4 \end{pmatrix},$$

where the α_j, $j = 1, 2, 3$, are the coordinate vectors of $\Phi(b_j)$ w.r.t. C.

Example 17 (The derivative as a matrix)**.** Let $D : \mathbb{R}_2[X] \to \mathbb{R}_2[X]$, $p \mapsto p'$, with the basis $B = (1, X, X^2)$ on both sides. Since $D(1) = 0$, $D(X) = 1$ and $D(X^2) = 2X$,

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

For $p = 1 + 3X + 5X^2$, $\hat{p} = (1, 3, 5)^\top$ and $A_D \hat{p} = (3, 10, 0)^\top$, i.e., $p' = 3 + 10X$. Here $\mathrm{rk}(A_D) = 2$ and $\mathrm{Ker}(D)$ is the set of constant polynomials (dimension 1), in agreement with rank–nullity: $3 = 1 + 2$.

Remark 9 (Change of basis)**.** Let B and B' be two bases of $\mathbb{R}^n$ and let S be the invertible matrix whose columns are the coordinates of the vectors of B' in B. Then the coordinates are related by $\hat{x}_B = S\,\hat{x}_{B'}$. If an endomorphism has matrix A in B, its matrix in B' is $A' = S^{-1}AS$. Such matrices are called **similar**; they share the same rank, determinant, trace and eigenvalues (Section 1.6). Choosing a "good" basis to make A' as simple as possible (e.g. diagonal) is the idea behind diagonalization.

1.4 Norms, inner products and geometry

In this section E is a real vector space.

1.4.1 Norms

When we think of geometric vectors, i.e., directed segments starting at the origin, the length of a vector is intuitively the distance from its "end" to the origin. Norms formalize this notion of length.

Definition 16. A **norm** on E is a function $\|\cdot\| : E \to \mathbb{R}_+$ such that, for all $x, y \in E$ and $\lambda \in \mathbb{R}$:

(i) Positive definiteness: $\|x\| \geq 0$ and $\|x\| = 0 \Leftrightarrow x = 0_E$;

(ii) Absolute homogeneity: $\|\lambda x\| = |\lambda|\,\|x\|$;

(iii) Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$.

Geometrically, the triangle inequality says that in any triangle the length of one side is at most the sum of the lengths of the other two. A useful consequence is the reverse triangle inequality $\big|\,\|x\| - \|y\|\,\big| \leq \|x - y\|$.

Common norms on $\mathbb{R}^n$.

- The Manhattan norm (1-norm, ℓ_1 norm): $\|x\|_1 = \sum_{i=1}^n |x_i|$.
- The Euclidean norm (2-norm, ℓ_2 norm): $\|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{1/2} = \sqrt{x^\top x}$. It computes the Euclidean distance of x from the origin.
- For $p \geq 1$, the p-norm: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$.
- The ∞-norm (max norm): $\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$ (it is the limit of $\|x\|_p$ as $p \to \infty$).

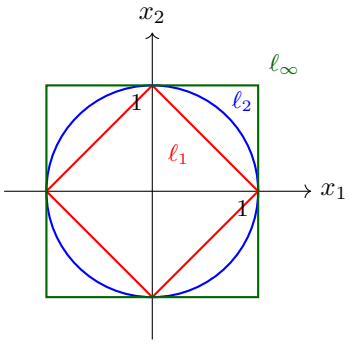


Figure 1.1: Unit spheres $\{x \in \mathbb{R}^2 \mid \|x\| = 1\}$ for the ℓ_1 (red), ℓ_2 (blue) and ℓ_∞ (green) norms.

Example 18. For $x = (3, -4)^\top \in \mathbb{R}^2$: $\|x\|_1 = 3 + 4 = 7$, $\|x\|_2 = \sqrt{9 + 16} = 5$ and $\|x\|_\infty = 4$. Different norms give different "lengths" for the same vector.

Remark 10. For $0 < p < 1$, the formula $(\sum_i |x_i|^p)^{1/p}$ does *not* define a norm: with $p = \frac{1}{2}$, $x = (1, 0)^\top$ and $y = (0, 1)^\top$, we get $\|x + y\| = (1 + 1)^2 = 4 > \|x\| + \|y\| = 2$, so the triangle inequality fails.

Proposition 8 (Equivalence of norms)**.** In a finite-dimensional space all norms are *equivalent*: for any two norms $\|\cdot\|_a, \|\cdot\|_b$ there exist constants $c, C > 0$ such that $c\|x\|_a \leq \|x\|_b \leq C\|x\|_a$ for all x. On $\mathbb{R}^n$:

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2 \leq n\,\|x\|_\infty .$$

In particular, the notion of convergence of a sequence (x^k) in $\mathbb{R}^n$ does not depend on the chosen norm (but the constants, and thus numerical behaviour, may depend on n).

Norms on $\mathcal{M}_{m,n}(\mathbb{R})$.

- **Frobenius norm:** $\|A\|_F = \sqrt{\mathrm{Tr}(A^\top A)} = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$ (the Euclidean norm of A seen as a vector of $\mathbb{R}^{mn}$).
- **Induced (operator) norm:** given norms $\|\cdot\|_a$ on $\mathbb{R}^n$ and $\|\cdot\|_b$ on $\mathbb{R}^m$, $\|A\|_{a,b} = \max_{\|x\|_a \leq 1} \|Ax\|_b$. It measures by how much A can "stretch" a vector, and satisfies $\|Ax\|_b \leq \|A\|_{a,b}\,\|x\|_a$. When $a = b = p$ we simply write $\|A\|_p$. Classical formulas:

$$\|A\|_1 = \max_j \sum_{i=1}^m |a_{ij}| \text{ (max column sum)}, \qquad \|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}| \text{ (max row sum)}.$$

- **Spectral norm (2-norm):** $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)} = \sigma_{\max}(A)$, where $\lambda_{\max}$ is the largest eigenvalue and $\sigma_{\max}$ the largest singular value (Section 1.6 and Section 1.7).

Induced norms and the Frobenius norm are **submultiplicative**: $\|AB\| \leq \|A\| \, \|B\|$. Moreover $\|A\|_2 \leq \|A\|_F \leq \sqrt{\mathrm{rk}(A)} \, \|A\|_2$.

Example 19. Let $A = \begin{pmatrix} 1 & -2 \\ 3 & 4 \end{pmatrix}$. Then $\|A\|_1 = \max(1{+}3,\, 2{+}4) = 6$, $\|A\|_\infty = \max(1{+}2,\, 3{+}4) = 7$, $\|A\|_F = \sqrt{1+4+9+16} = \sqrt{30} \approx 5.48$. For the spectral norm, $A^\top A = \begin{pmatrix} 10 & 10 \\ 10 & 20 \end{pmatrix}$ has eigenvalues $15 \pm 5\sqrt{5}$, so $\|A\|_2 = \sqrt{15+5\sqrt{5}} \approx 5.12 \leq \|A\|_F$.

1.4.2 Inner products

Inner products allow us to introduce intuitive geometric concepts such as the length of a vector and the angle or distance between two vectors. A major purpose of inner products is to determine whether vectors are orthogonal to each other.

Definition 17 (Standard dot product)**.** The **dot product** (scalar product) on $\mathbb{R}^n$ is

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

The dot product is a particular case of a more general notion. A **bilinear map** $\mathrm{b} : E \times E \to \mathbb{R}$ is a map with two arguments which is linear in each argument: for all $x, y, z \in E$ and $\lambda, \psi \in \mathbb{R}$,

$$\mathrm{b}(\lambda x + \psi y, z) = \lambda \mathrm{b}(x, z) + \psi \mathrm{b}(y, z), \qquad \mathrm{b}(x, \lambda y + \psi z) = \lambda \mathrm{b}(x, y) + \psi \mathrm{b}(x, z).$$

It is **symmetric** if $\mathrm{b}(x, y) = \mathrm{b}(y, x)$ for all x, y, and **positive definite** if $\mathrm{b}(x, x) > 0$ for all $x \neq 0_E$ (note that $\mathrm{b}(0_E, 0_E) = 0$ always holds by bilinearity).

Definition 18. An **inner product** on E is a symmetric, positive definite bilinear map $\langle \cdot, \cdot \rangle : E \times E \to \mathbb{R}$. Equivalently, for all $x, y, z \in E$ and $\alpha, \beta \in \mathbb{R}$:

(i) Symmetry: $\langle x, y \rangle = \langle y, x \rangle$;

(ii) Linearity with respect to the first variable: $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ (linearity in the second variable then follows from symmetry);

(iii) Positive definiteness: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Leftrightarrow x = 0_E$.

The pair $(E, \langle \cdot, \cdot \rangle)$ is called an **inner product space**.

Definition 19. A *finite-dimensional* real inner product space is called a **Euclidean space**. For instance, $\mathbb{R}^n$ with the dot product is the (standard) Euclidean space.

Example 20 (Inner products)**.**

- The dot product on $\mathbb{R}^n$.
- The **Frobenius inner product** on $\mathcal{M}_{m,n}(\mathbb{R})$: $\langle A, B \rangle_F = \mathrm{Tr}(A^\top B) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}$.
- On $\mathcal{C}([a, b])$: $\langle f, g \rangle = \int_a^b f(t) g(t) \, \mathrm{d}t$ (an inner product on an infinite-dimensional space).

1.4.3 Symmetric positive definite matrices and inner products

The previous example is a particular case of the following construction. Recall that $A \in \mathcal{S}_{++}^n$ means

$$A = A^\top \quad \text{and} \quad \forall x \in \mathbb{R}^n \setminus \{0\} : \; x^\top A x > 0,$$

and $A \in \mathcal{S}_+^n$ if only $\geq$ holds.

Example 21 (Symmetric positive definite matrices). Consider the symmetric matrices

$$A_1 = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad A_3 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

- $A_1 \in \mathcal{S}_{++}^2$: $x^\top A_1 x = 2x_1^2 - 2x_1x_2 + 2x_2^2 = x_1^2 + x_2^2 + (x_1 - x_2)^2 > 0$ for all $x \neq 0$.
- $A_2 \in \mathcal{S}_{+}^2$ but $A_2 \notin \mathcal{S}_{++}^2$: $x^\top A_2 x = (x_1 - x_2)^2 \geq 0$, and it vanishes for $x = (1,1)^\top \neq 0$.
- $A_3 \notin \mathcal{S}_{+}^2$: $x^\top A_3 x = x_1^2 + 4x_1x_2 + x_2^2$ equals $-2 < 0$ for $x = (1,-1)^\top$, although all the entries of A_3 are positive.

The last case shows that positive entries do not imply positive definiteness (and conversely, A_1 has negative entries).

Definition 20 (Q-inner product and Q-norm). Given $Q \in \mathcal{S}_{++}^n$, the formula

$$\langle x, y \rangle_Q = x^\top Q y$$

defines an inner product on $\mathbb{R}^n$. Its associated **Q-norm** (or energy norm) is $\|x\|_Q = \sqrt{x^\top Q x}$. For $Q = \mathbf{I}_n$ we recover the dot product and the Euclidean norm ($\langle x, x \rangle_{\mathbf{I}_2} = x_1^2 + x_2^2$).

Proposition 9. Conversely, every inner product on $\mathbb{R}^n$ is of the form $\langle x, y \rangle = x^\top Q y$ with $Q \in \mathcal{S}_{++}^n$, namely $Q_{ij} = \langle e_i, e_j \rangle$. More generally, if V is a Euclidean space with basis B, then $\langle x, y \rangle = \hat{x}^\top Q \hat{y}$, where $\hat{x}, \hat{y}$ are the coordinates of x, y w.r.t. B and $Q_{ij} = \langle b_i, b_j \rangle$.

How to check that a symmetric matrix is positive (semi)definite?

Proposition 10. Let $A \in \mathcal{S}^n$.

1. (Eigenvalues) $A \in \mathcal{S}_{++}^n \Leftrightarrow$ all eigenvalues of A are > 0; $A \in \mathcal{S}_{+}^n \Leftrightarrow$ all eigenvalues of A are ≥ 0 (see Section 1.6).
2. (Sylvester's criterion) $A \in \mathcal{S}_{++}^n \Leftrightarrow$ all leading principal minors are positive: $\det(A_k) > 0$ for $k = 1, \ldots, n$, where $A_k = (a_{ij})_{1 \leq i,j \leq k}$ is the upper-left $k \times k$ block.
3. For $n = 2$: $\begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathcal{S}_{++}^2 \Leftrightarrow a > 0$ and $ac - b^2 > 0$.
4. If $A \in \mathcal{S}_{++}^n$ then $a_{ii} > 0$ for all i, $\det(A) > 0$ and A is invertible with $A^{-1} \in \mathcal{S}_{++}^n$.
5. For any $B \in \mathcal{M}_{m,n}(\mathbb{R})$, $B^\top B \in \mathcal{S}_{+}^n$ since $x^\top B^\top B x = \|Bx\|_2^2 \geq 0$; moreover $B^\top B \in \mathcal{S}_{++}^n$ iff $\mathrm{rk}(B) = n$.

Example 22. For the matrices of the previous example: A_1 has leading minors $2 > 0$ and $4 - 1 = 3 > 0$, so $A_1 \succ 0$ (its eigenvalues are 1 and 3); A_2 has eigenvalues 0 and 2, so $A_2 \succeq 0$ but $A_2 \not\succ 0$; A_3 has $\det(A_3) = 1 - 4 = -3 < 0$ (eigenvalues 3 and -1), so A_3 is indefinite. The matrix $\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$ of the inner product given in the Inner products example has minors $1 > 0$ and $2 - 1 = 1 > 0$.

Remark 11. Sylvester's criterion does *not* extend naively to semidefiniteness: $A = \mathrm{diag}(0, -1)$ has leading minors 0 and 0 (both ≥ 0) but $A \notin \mathcal{S}_{+}^2$. Use the eigenvalue criterion instead (or check all principal minors).

Remark 12. Positive (semi)definite matrices are ubiquitous in optimization: the Hessian matrix of a twice differentiable convex function is positive semidefinite, and a critical point where the Hessian is positive definite is a strict local minimum (see Chapter 2).

1.4.4 Lengths and distances

Any inner product induces a norm in a natural way:

$$\|x\| := \sqrt{\langle x, x\rangle},$$

so that we can compute lengths of vectors using the inner product. However, not every norm is induced by an inner product: the Manhattan norm $\|\cdot\|_1$ and the max norm $\|\cdot\|_\infty$ are examples of norms without a corresponding inner product.

Proposition 11 (Cauchy–Schwarz inequality)**.** In an inner product space, for all $x, y \in E$,

$$|\langle x, y\rangle| \leq \|x\| \, \|y\|,$$

with equality if and only if x and y are linearly dependent (collinear).

Proposition 12 (Parallelogram law)**.** A norm is induced by an inner product if and only if $\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$ for all x, y. In that case the inner product is recovered by $\langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$.

Example 23. For $x = (1,0)^\top$, $y = (0,1)^\top$: $\|x+y\|_1^2 + \|x-y\|_1^2 = 4+4 = 8 \neq 2\|x\|_1^2 + 2\|y\|_1^2 = 4$. Hence $\|\cdot\|_1$ does not come from any inner product.

Example 24 (Lengths of vectors using inner products)**.** Let $x = (1,1)^\top \in \mathbb{R}^2$. With the dot product, $\|x\| = \sqrt{x^\top x} = \sqrt{1^2+1^2} = \sqrt{2}$. Now choose the inner product

$$\langle x, y\rangle := x^\top \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix} y = x_1 y_1 - \tfrac{1}{2}(x_1 y_2 + x_2 y_1) + x_2 y_2.$$

It returns smaller values than the dot product when x_1 and x_2 have the same sign ($x_1 x_2 > 0$), and greater values otherwise. With this inner product $\langle x, x\rangle = 1 - 1 + 1 = 1$, so $\|x\| = 1$: x is "shorter" than with the dot product.

Definition 21 (Distance and metric)**.** In an inner product space (or more generally a normed space) E,

$$d(x, y) := \|x - y\| = \sqrt{\langle x - y, x - y\rangle}$$

is called the **distance** between x and y. If we use the dot product, it is the *Euclidean distance*. The mapping $d : E \times E \to \mathbb{R}$, $(x, y) \mapsto d(x, y)$, is called a **metric**.

A metric satisfies, for all $x, y, z \in E$:

1. positive definiteness: $d(x, y) \geq 0$ and $d(x, y) = 0 \Leftrightarrow x = y$;
2. symmetry: $d(x, y) = d(y, x)$;
3. triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

Remark 13. The lists of properties of inner products and metrics look similar, but $\langle x, y\rangle$ and $d(x, y)$ behave in opposite directions: very similar x and y give a large inner product (relative to their norms) and a small distance.

1.4.5 Angles and orthogonality

Inner products also capture the geometry of a vector space by defining the angle between two vectors. Let $x \neq 0$, $y \neq 0$. By Cauchy–Schwarz,

$$-1 \leq \frac{\langle x, y\rangle}{\|x\|\ \|y\|} \leq 1,$$

so there exists a unique $\omega \in [0, \pi]$ such that

$$\cos\omega = \frac{\langle x, y\rangle}{\|x\|\ \|y\|}.$$

The number ω is the **angle** between x and y. It tells us how similar the orientations of x and y are; in $\mathbb{R}^2$ and $\mathbb{R}^3$ with the dot product it coincides with the usual angle. For example, the angle between x and $y = 4x$ is 0. (In data science, $\cos\omega$ is known as the *cosine similarity*.)

Example 25 (Angle between vectors). Let $x = (1, 1)^\top$ and $y = (1, 2)^\top$, with the dot product. Then

$$\cos\omega = \frac{x^\top y}{\sqrt{x^\top x\ y^\top y}} = \frac{3}{\sqrt{2 \cdot 5}} = \frac{3}{\sqrt{10}},$$

so $\omega = \arccos(3/\sqrt{10}) \approx 0.32$ rad, about 18°.

Definition 22 (Orthogonality). Two vectors x and y are **orthogonal** if and only if $\langle x, y\rangle = 0$, and we write $x \perp y$. If additionally $\|x\| = \|y\| = 1$, i.e., the vectors are unit vectors, then x and y are **orthonormal**.

An implication of this definition is that 0_E is orthogonal to every vector. Orthogonality generalizes perpendicularity to inner products that are not the dot product.

Proposition 13 (Pythagoras). If $x \perp y$ then $\|x + y\|^2 = \|x\|^2 + \|y\|^2$.

Example 26 (Orthogonality depends on the inner product). Consider $x = (1, 1)^\top$, $y = (-1, 1)^\top \in \mathbb{R}^2$. With the dot product, $x^\top y = 0$: the angle is 90° and $x \perp y$. Now choose the inner product

$$\langle x, y\rangle = x^\top \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} y.$$

Then $\langle x, y\rangle = -2 + 1 = -1$ and $\|x\|^2 = \|y\|^2 = 3$, so

$$\cos\omega = \frac{\langle x, y\rangle}{\|x\|\ \|y\|} = -\frac{1}{3} \Rightarrow \omega \approx 1.91 \text{ rad} \approx 109.5°,$$

and x and y are not orthogonal. Vectors orthogonal w.r.t. one inner product need not be orthogonal w.r.t. another one.

Orthonormal bases. A basis $(u_1, \ldots, u_n)$ of a Euclidean space is **orthonormal** if $\langle u_i, u_j\rangle = 0$ for $i \neq j$ and $\|u_i\| = 1$. Coordinates are then easy to compute:

$$x = \sum_{i=1}^{n} \langle x, u_i\rangle\ u_i \qquad \text{and} \qquad \|x\|^2 = \sum_{i=1}^{n} \langle x, u_i\rangle^2 .$$

Any basis can be turned into an orthonormal one (Gram–Schmidt process).

Example 27. The canonical basis of $\mathbb{R}^n$ is orthonormal for the dot product. So is $u_1 = \frac{1}{\sqrt{2}}(1,1)^\top$, $u_2 = \frac{1}{\sqrt{2}}(1,-1)^\top$ in $\mathbb{R}^2$: for $x = (3,1)^\top$, $\langle x, u_1\rangle = 2\sqrt{2}$ and $\langle x, u_2\rangle = \sqrt{2}$, and indeed $x = 2\sqrt{2}\,u_1 + \sqrt{2}\,u_2$.

Orthogonal matrices. $U \in \mathcal{M}_n(\mathbb{R})$ is orthogonal ($U \in \mathcal{O}(n)$) iff its columns form an orthonormal basis of $\mathbb{R}^n$ (for the dot product). Orthogonal matrices preserve the geometry:

$$\langle Ux, Uy\rangle = \langle x, y\rangle\,, \qquad \|Ux\|_2 = \|x\|_2\,, \qquad \det(U) = \pm 1.$$

Rotations ($\det = 1$) and reflections ($\det = -1$) are the typical examples, e.g. R_θ and $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

Orthogonal complement and orthogonal projection. Let F be a subspace of a Euclidean space E. Its **orthogonal complement** is $F^\perp = \{x \in E \mid \langle x, y\rangle = 0\ \forall y \in F\}$; it is a subspace with $\dim(F) + \dim(F^\perp) = \dim(E)$. Every $x \in E$ decomposes uniquely as $x = p + r$ with $p \in F$ and $r \in F^\perp$; $p = P_F(x)$ is the **orthogonal projection** of x onto F, and it is the point of F closest to x: $\|x - P_F(x)\| = \min_{y\in F} \|x - y\|$. In $\mathbb{R}^n$ with the dot product:

- onto the line spanned by $a \neq 0$: $P(x) = \dfrac{a^\top x}{a^\top a}\, a$;
- if the columns of $U \in \mathcal{M}_{n,k}(\mathbb{R})$ form an orthonormal basis of F: $P_F(x) = UU^\top x$;
- if $F = \mathrm{Im}(A)$ with $A \in \mathcal{M}_{n,k}(\mathbb{R})$ of full column rank: $P_F(x) = A(A^\top A)^{-1}A^\top x$.

For any matrix A, $\mathrm{Ker}(A) = \mathrm{Im}(A^\top)^\perp$.

Example 28. Projecting $x = (2,0)^\top$ onto the line spanned by $a = (1,1)^\top$ gives $P(x) = \frac{2}{2}(1,1)^\top = (1,1)^\top$. The residual $r = x - P(x) = (1,-1)^\top$ is indeed orthogonal to a.

1.5 Linear systems

We consider the problem of finding $x \in \mathbb{R}^n$ such that $Ax = b$, where $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and $b \in \mathbb{R}^m$.

- **Existence:** a solution exists if and only if $b \in \mathrm{Im}(A) = \{Ax \mid x \in \mathbb{R}^n\}$, i.e., iff $\mathrm{rk}(A) = \mathrm{rk}(A|b)$.
- **Structure of the solution set:** if x_0 is one solution, the set of all solutions is $x_0 + \mathrm{Ker}(A) = \{x_0 + z \mid Az = 0\}$. Hence, when a solution exists, it is unique iff $\mathrm{Ker}(A) = \{0\}$, i.e., iff $\mathrm{rk}(A) = n$.
- **Square invertible case:** if $m = n$, the following are equivalent: A is invertible; $\det(A) \neq 0$; $\mathrm{rk}(A) = n$; $\mathrm{Ker}(A) = \{0\}$; $\mathrm{Im}(A) = \mathbb{R}^n$; 0 is not an eigenvalue of A. Then, for every b, the unique solution is $x = A^{-1}b$.

Example 29. Consider $x_1 + 2x_2 = 3$, $2x_1 + 4x_2 = c$. Here $\mathrm{rk}(A) = 1$.

- If $c = 6$, $\mathrm{rk}(A|b) = 1$: infinitely many solutions $x = (3,0)^\top + t(-2,1)^\top$, $t \in \mathbb{R}$ (a particular solution plus $\mathrm{Ker}(A)$).
- If $c = 7$, $\mathrm{rk}(A|b) = 2$: no solution.

Remark 14. In practice one (almost) never computes A^{-1} to solve $Ax = b$: it is more expensive and less accurate than a factorization (see below). In `NumPy`, use `np.linalg.solve(A, b)` rather than `np.linalg.inv(A) @ b`.

Least squares. If no exact solution exists (typically for overdetermined systems, $m > n$), we look for x minimizing the residual $\|Ax - b\|_2^2$. The minimizers are exactly the solutions of the **normal equations**

$$A^\top Ax = A^\top b.$$

If A has full column rank, $A^\top A \in \mathcal{S}_{++}^n$ and the solution is unique: $x^\star = (A^\top A)^{-1} A^\top b$. Geometrically, $Ax^\star$ is the orthogonal projection of b onto $\mathrm{Im}(A)$, and the residual $Ax^\star - b$ is orthogonal to $\mathrm{Im}(A)$.

Example 30 (Fitting a line). We look for a line $y = \beta_0 + \beta_1 t$ through the points $(t, y) = (0, 1), (1, 2), (2, 2)$. This leads to the overdetermined system

$$\underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}}_{A} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}}_{b}, \qquad A^\top A = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}, \quad A^\top b = \begin{pmatrix} 5 \\ 6 \end{pmatrix}.$$

Solving the normal equations gives $\beta_1 = \frac{1}{2}$ and $\beta_0 = \frac{7}{6}$. The residual $A\beta - b = (\frac{1}{6}, -\frac{1}{3}, \frac{1}{6})^\top$ is orthogonal to both columns of A.

Conditioning. For A invertible, the **condition number** $\kappa(A) = \|A\| \left\|A^{-1}\right\| \geq 1$ measures the sensitivity of the solution of $Ax = b$ to perturbations of the data: roughly, relative errors on b can be amplified by a factor $\kappa(A)$. For the 2-norm, $\kappa_2(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$, and for $A \in \mathcal{S}_{++}^n$, $\kappa_2(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$. A matrix with a large condition number is called *ill-conditioned*. Note that $\kappa_2(A^\top A) = \kappa_2(A)^2$, which is why the normal equations are avoided for ill-conditioned least squares problems (QR or SVD are used instead). The condition number will also govern the speed of gradient methods in Chapter 2.

Direct factorizations.

- **LU decomposition:** $PA = LU$, where P is a permutation matrix, L is lower triangular with 1's on the diagonal and U is upper triangular (this is Gaussian elimination written in matrix form). To solve $Ax = b$, solve the two triangular systems $Ly = Pb$ (forward substitution) and $Ux = y$ (backward substitution).
- **Cholesky decomposition:** if $A \in \mathcal{S}_{++}^n$, there exists a unique lower triangular matrix L with positive diagonal entries such that $A = LL^\top$. It costs about half of an LU decomposition and is the method of choice for symmetric positive definite systems. (It also provides a practical test: the factorization succeeds iff $A \succ 0$.)
- **QR decomposition:** any $A \in \mathcal{M}_{m,n}(\mathbb{R})$ with $m \geq n$ can be written $A = QR$, where $Q \in \mathcal{M}_{m,n}(\mathbb{R})$ has orthonormal columns and $R \in \mathcal{M}_n(\mathbb{R})$ is upper triangular. If A has full column rank, the least squares solution is obtained from $Rx = Q^\top b$.

Example 31.

- (LU) $\begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$ (here $P = \mathbf{I}_2$): the multiplier $2 = 4/2$ used to eliminate the entry $(2, 1)$ is stored in L.
- (Cholesky) $\begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & \sqrt{2} \end{pmatrix}$.

Iterative methods. Iterative methods are preferred for very large, sparse systems. Many of them are based on a splitting of A:

$$A = P - (P - A),$$

where P is an easily invertible matrix (a *preconditioner*). The system $Ax = b$ is equivalent to $Px = (P - A)x + b$, which suggests the iteration

$$Px^{k+1} = (P - A)x^k + b \Leftrightarrow x^{k+1} = Bx^k + g,$$

with iteration matrix $B = P^{-1}(P - A) = \mathbf{I}_n - P^{-1}A$ and $g = P^{-1}b$. The sequence (x^k) converges to the solution for every initial guess x^0 if and only if the spectral radius satisfies $\rho(B) < 1$ (Section 1.6); the smaller $\rho(B)$, the faster the convergence. Write $A = D + L + U$ where D is the diagonal part and L, U the strictly lower and upper parts of A (not to be confused with the LU factors). Classical examples:

- **Jacobi:** $P = D$, i.e., $x_i^{k+1} = \frac{1}{a_{ii}}\big(b_i - \sum_{j \neq i} a_{ij} x_j^k\big)$;
- **Gauss–Seidel:** $P = D + L$ (the updated components are used as soon as they are available).

Both methods converge if A is strictly diagonally dominant ($|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ for all i); Gauss–Seidel also converges if $A \in \mathcal{S}_{++}^n$.

Example 32. For $A = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$ (strictly diagonally dominant), the Jacobi iteration matrix is $B = -D^{-1}(L+U) = \begin{pmatrix} 0 & -\frac{1}{4} \\ -\frac{1}{3} & 0 \end{pmatrix}$, whose eigenvalues are $\pm\frac{1}{\sqrt{12}}$; hence $\rho(B) \approx 0.29 < 1$ and the method converges (the error is roughly divided by 3.5 at each iteration).

1.6 Eigenvalues and eigenvectors

Definition 23. Let $A \in \mathcal{M}_n(\mathbb{R})$. A scalar $\lambda \in \mathbb{C}$ is an **eigenvalue** of A and a *nonzero* vector $x \in \mathbb{C}^n$ is an associated **eigenvector** if $Ax = \lambda x$.

- **Spectrum:** $\sigma(A)$ is the set of all eigenvalues of A.
- **Characteristic polynomial:** $\chi_A(\lambda) = \det(A - \lambda \mathbf{I}_n)$, a polynomial of degree n. The roots of χ_A are exactly the eigenvalues of A; hence A has n eigenvalues in $\mathbb{C}$, counted with multiplicity.
- **Eigenspace:** for $\lambda \in \sigma(A)$, $\mathrm{Ker}(A - \lambda \mathbf{I}_n)$ is the eigenspace associated with λ (all eigenvectors together with 0).
- **Spectral radius:** $\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$.

Geometrically, an eigenvector is a direction which is only stretched (by the factor λ) and not rotated by A.

Proposition 14. Let $A \in \mathcal{M}_n(\mathbb{R})$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ (in $\mathbb{C}$, with multiplicity).

1. $\mathrm{Tr}(A) = \sum_{i=1}^n \lambda_i$ and $\det(A) = \prod_{i=1}^n \lambda_i$. In particular A is invertible iff $0 \notin \sigma(A)$.
2. The eigenvalues of a triangular matrix are its diagonal entries.
3. If $Ax = \lambda x$, then $A^k x = \lambda^k x$, $(A + c\mathbf{I}_n)x = (\lambda + c)x$ and, if A is invertible, $A^{-1}x = \lambda^{-1}x$.
4. A and $A^\top$ have the same eigenvalues; similar matrices have the same eigenvalues.
5. $\rho(A) \leq \|A\|$ for any induced matrix norm.

Example 33. Let $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Then $\chi_A(\lambda) = (2-\lambda)^2 - 1 = (\lambda - 1)(\lambda - 3)$, so $\sigma(A) = \{1, 3\}$ and $\rho(A) = 3$. Solving $(A - \lambda \mathbf{I}_2)x = 0$ gives the eigenvectors $(1, -1)^\top$ for $\lambda = 1$ and $(1, 1)^\top$ for $\lambda = 3$. Check: $\mathrm{Tr}(A) = 4 = 1 + 3$ and $\det(A) = 3 = 1 \cdot 3$.

Example 34 (Real matrices may have complex eigenvalues). The rotation by 90°, $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, has $\chi_A(\lambda) = \lambda^2 + 1$, hence $\sigma(A) = \{i, -i\}$: no real direction is left invariant by a rotation.

Diagonalization. $A \in \mathcal{M}_n(\mathbb{R})$ is **diagonalizable** if there exist an invertible P and a diagonal D such that $A = PDP^{-1}$; the columns of P are then eigenvectors of A and the diagonal of D contains the eigenvalues. This holds iff A has n linearly independent eigenvectors, which is the case for instance when its n eigenvalues are distinct. Not every matrix is diagonalizable: $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has the single eigenvalue 1 but its eigenspace $\mathrm{span}\{e_1\}$ has dimension 1 only. When $A = PDP^{-1}$, powers are easy to compute: $A^k = PD^kP^{-1}$.

Spectral theorem for symmetric matrices.

Proposition 15 (Spectral theorem). If $A \in \mathcal{S}^n$, then all its eigenvalues are real and A is diagonalizable in an orthonormal basis: there exist $U \in \mathcal{O}(n)$, whose columns $u_1, \dots, u_n$ are orthonormal eigenvectors, and $D = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ containing the (real) eigenvalues, such that

$$A = UDU^\top = \sum_{i=1}^{n} \lambda_i \, u_i u_i^\top .$$

Example 35. For $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, take $u_1 = \frac{1}{\sqrt{2}}(1, -1)^\top$, $u_2 = \frac{1}{\sqrt{2}}(1, 1)^\top$:

$$A = \underbrace{\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}}_{U} \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} \underbrace{\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}}_{U^\top} = 1 \cdot u_1 u_1^\top + 3 \cdot u_2 u_2^\top .$$

Proposition 16 (Consequences for $A \in \mathcal{S}^n$). Let $\lambda_{\min}$ and $\lambda_{\max}$ be the smallest and largest eigenvalues of $A \in \mathcal{S}^n$.

1. (Rayleigh quotient) For all $x \in \mathbb{R}^n$: $\lambda_{\min} \|x\|_2^2 \le x^\top A x \le \lambda_{\max} \|x\|_2^2$, with equality for the corresponding eigenvectors. Hence $\lambda_{\max} = \max_{\|x\|_2 = 1} x^\top A x$ and $\lambda_{\min} = \min_{\|x\|_2 = 1} x^\top A x$.
2. $A \in \mathcal{S}^n_{++} \Leftrightarrow \lambda_{\min} > 0$, and $A \in \mathcal{S}^n_{+} \Leftrightarrow \lambda_{\min} \ge 0$ (Proposition 10).
3. $\|A\|_2 = \rho(A) = \max(|\lambda_{\min}|, |\lambda_{\max}|)$.

Remark 15 (Geometric interpretation). For $A \in \mathcal{S}^n_{++}$, the level set $\{x \in \mathbb{R}^n \mid x^\top A x = 1\}$ is an ellipse (ellipsoid) whose axes are directed by the eigenvectors u_i, with semi-axis lengths $1/\sqrt{\lambda_i}$. The ratio $\lambda_{\max}/\lambda_{\min} = \kappa_2(A)$ measures how "elongated" it is; this picture is central to understanding gradient descent on quadratic functions in Chapter 2.

1.7 Singular Value Decomposition (SVD)

Eigen-decompositions only exist for (some) square matrices. The SVD works for *any* matrix.

Proposition 17 (SVD). Let $A \in \mathcal{M}_{m,n}(\mathbb{R})$ with rank r. There exist $U \in \mathcal{O}(m)$, $V \in \mathcal{O}(n)$ and a "diagonal" matrix $\Sigma \in \mathcal{M}_{m,n}(\mathbb{R})$ (i.e., $\Sigma_{ij} = 0$ for $i \neq j$) such that

$$A = U\Sigma V^\top, \qquad \Sigma_{ii} = \sigma_i,$$

where $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0 = \sigma_{r+1} = \dots$ are the **singular values** of A. The nonzero

singular values are the square roots of the nonzero eigenvalues of $A^\top A$ (or equivalently of $AA^\top$); the columns v_i of V (right singular vectors) are eigenvectors of $A^\top A$ and the columns u_i of U (left singular vectors) are eigenvectors of $AA^\top$, with $Av_i = \sigma_i u_i$.

Proposition 18 (Properties).
1. $\mathrm{rk}(A) = r$ is the number of nonzero singular values, and $A = \sum_{i=1}^{r} \sigma_i\, u_i v_i^\top$ (sum of r rank-one matrices).
2. $\|A\|_2 = \sigma_1$ and $\|A\|_F = \left(\sum_{i=1}^{r} \sigma_i^2\right)^{1/2}$.
3. $(u_1, \dots, u_r)$ is an orthonormal basis of $\mathrm{Im}(A)$ and $(v_{r+1}, \dots, v_n)$ an orthonormal basis of $\mathrm{Ker}(A)$.
4. If $A \in \mathcal{M}_n(\mathbb{R})$ is invertible, $\kappa_2(A) = \sigma_1/\sigma_n$. If $A \in \mathcal{S}_+^n$, its singular values are its eigenvalues (and its SVD is its spectral decomposition).
5. (Eckart–Young) For $k < r$, the truncated SVD $A_k = \sum_{i=1}^{k} \sigma_i\, u_i v_i^\top$ is the best approximation of A by a matrix of rank at most k, both in $\|\cdot\|_2$ and in $\|\cdot\|_F$, with $\|A - A_k\|_2 = \sigma_{k+1}$.
6. The **pseudo-inverse** $A^+ = V\Sigma^+ U^\top$, where $\Sigma^+ \in \mathcal{M}_{n,m}(\mathbb{R})$ has diagonal entries $1/\sigma_i$ ($i \le r$) and 0 elsewhere, gives the minimum-norm least squares solution $x = A^+ b$ of $Ax = b$. If A has full column rank, $A^+ = (A^\top A)^{-1} A^\top$.

Example 36. Let $A = \begin{pmatrix} 3 & 0 \\ 4 & 0 \end{pmatrix}$. Then $A^\top A = \begin{pmatrix} 25 & 0 \\ 0 & 0 \end{pmatrix}$, so $\sigma_1 = 5$, $\sigma_2 = 0$ and $\mathrm{rk}(A) = 1$. With $v_1 = (1,0)^\top$, $u_1 = Av_1/\sigma_1 = \frac{1}{5}(3,4)^\top$, and completing with $v_2 = (0,1)^\top$, $u_2 = \frac{1}{5}(-4,3)^\top$:

$$A = \underbrace{\frac{1}{5}\begin{pmatrix} 3 & -4 \\ 4 & 3 \end{pmatrix}}_{U} \begin{pmatrix} 5 & 0 \\ 0 & 0 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{V^\top} = 5\, u_1 v_1^\top.$$

Here $\|A\|_2 = \|A\|_F = 5$.

Remark 16 (Geometric meaning of the SVD). Reading $Ax = U\big(\Sigma(V^\top x)\big)$ from right to left, every linear map is the composition of three simple operations: a *rotation* (or reflection) $V^\top$, a *stretching* Σ of the coordinate axes by the factors σ_i, and another rotation U. In particular, A maps the unit sphere onto an ellipsoid whose semi-axes are the vectors $\sigma_i u_i$ (Fig. 1.2); $\|A\|_2 = \sigma_1$ is the length of the longest semi-axis.

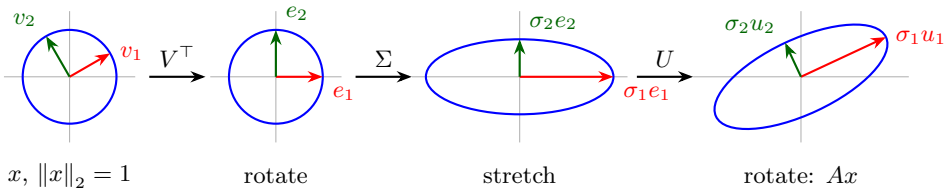


Figure 1.2: The SVD $A = U\Sigma V^\top$ of a 2×2 matrix: the unit circle is rotated by $V^\top$, stretched by Σ and rotated by U. Its image is an ellipse with semi-axes $\sigma_1 u_1$ and $\sigma_2 u_2$, and $Av_i = \sigma_i u_i$.

Chapitre 2.

Optimization

An optimization problem consists in finding the "best" value of a decision variable x according to a criterion f:

$$\min_{x \in C} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the **objective function** (cost, loss, energy…) and $C \subset \mathbb{R}^n$ is the **feasible set** (the constraints). When $C = \mathbb{R}^n$ the problem is **unconstrained**. Maximizing is not a different problem: $\max_C f = -\min_C(-f)$.

Example 37 (Optimization is everywhere).

- *Least squares / linear regression:* $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2$ (Chapter 1).
- *Machine learning:* training a model with parameters θ on data $(a_i, y_i)_{i=1}^N$ amounts to minimizing an empirical risk $\min_\theta \frac{1}{N} \sum_{i=1}^N \ell(h_\theta(a_i), y_i)$, possibly plus a regularization term such as $\lambda \|\theta\|_2^2$.
- *Portfolio selection:* $\min_{x \in \Delta_n} x^\top \Sigma x - \gamma \mu^\top x$, where $\Sigma \in \mathcal{S}_+^n$ is a covariance matrix, μ the expected returns and Δ_n the unit simplex.

Remark 17 (The guiding picture). Throughout this chapter, think of the graph of f as a landscape and of minimizing f as a hiker looking for the lowest point of a valley. The gradient tells the hiker which way is uphill, the Hessian describes the shape of the ground (bowl, ridge, saddle), convexity guarantees that there is only one valley, and algorithms are strategies to walk down, often in the fog, i.e., using only local information.

Definition 24 (Minimizers). Let $f : C \to \mathbb{R}$ and $x^* \in C$.

- x^* is a **global minimizer** of f on C if $f(x^*) \leq f(x)$ for all $x \in C$; it is **strict** if $f(x^*) < f(x)$ for all $x \in C \setminus \{x^*\}$.
- x^* is a **local minimizer** if there exists $\varepsilon > 0$ such that $f(x^*) \leq f(x)$ for all $x \in C$ with $\|x - x^*\| \leq \varepsilon$.

The optimal value is $f^* = \inf_{x \in C} f(x)$ and the set of minimizers is denoted $\operatorname{argmin}_C f$. Local and global maximizers are defined similarly.

Remark 18. The infimum always exists in $[-\infty, +\infty)$, but it need not be attained: $\inf_{\mathbb{R}} e^x = 0$ while $e^x > 0$ for every x, so $\operatorname{argmin}_{\mathbb{R}} e^x = \emptyset$. For $f(x) = x$ on $\mathbb{R}$, $f^* = -\infty$.

2.1 Differential calculus toolbox

2.1.1 Gradient, Jacobian, Hessian

Let $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}$.

- The **partial derivative** of f with respect to x_k at a is $\frac{\partial f}{\partial x_k}(a) = \lim_{t \to 0} \frac{f(a + te_k) - f(a)}{t}$, i.e., the ordinary derivative of f when all variables but x_k are frozen.

- If f is differentiable, its **gradient** is the column vector $\nabla f(a) = \left(\frac{\partial f}{\partial x_1}(a), \dots, \frac{\partial f}{\partial x_n}(a)\right)^\top \in \mathbb{R}^n$, and the **directional derivative** of f at a in the direction d is

$$f'(a;d) = \lim_{t \to 0^+} \frac{f(a+td) - f(a)}{t} = \langle \nabla f(a), d \rangle .$$

- For a vector-valued map $F = (F_1, \dots, F_p)^\top : U \to \mathbb{R}^p$, the **Jacobian matrix** $J_F(a) \in \mathcal{M}_{p,n}(\mathbb{R})$ has entries $\big(J_F(a)\big)_{ij} = \frac{\partial F_i}{\partial x_j}(a)$. For $p = 1$, $J_f(a) = \nabla f(a)^\top$.

- If f is twice differentiable, its **Hessian matrix** is $\nabla^2 f(a) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(a)\right)_{i,j=1}^{n} \in \mathcal{M}_n(\mathbb{R})$. If f is C^2 (Schwarz's theorem), $\nabla^2 f(a) \in \mathcal{S}^n$ is symmetric. Note that $\nabla^2 f = J_{\nabla f}$.

f is of class C^1 (resp. C^2) on U if all its partial derivatives of order 1 (resp. up to order 2) exist and are continuous on U.

Proposition 19 (Taylor expansions). Let $f \in C^2(U)$, $x \in U$ and h small enough. Then

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|),$$
$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^\top \nabla^2 f(x)\, h + o(\|h\|^2).$$

Proposition 20 (Chain rule). If $F : \mathbb{R}^n \to \mathbb{R}^m$ and $G : \mathbb{R}^m \to \mathbb{R}^p$ are C^1, then $J_{G \circ F}(x) = J_G(F(x))\, J_F(x)$. In particular, for $g : \mathbb{R}^m \to \mathbb{R}$ and $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^m$: $\nabla\big[g(Ax+b)\big] = A^\top \nabla g(Ax+b)$ and $\nabla^2\big[g(Ax+b)\big] = A^\top \nabla^2 g(Ax+b) A$.

Remark 19 (Intuition). The first-order expansion says that, zooming in, the graph of f looks like its tangent hyperplane; the gradient is the slope. The second-order expansion adds the curvature: near x, the landscape looks like a quadratic "bowl", "dome" or "saddle", whose shape is given by the eigenvalues and eigenvectors of $\nabla^2 f(x)$ (see the spectral theorem in Chapter 1).

Proposition 21 (Geometric meaning of the gradient). Let $\nabla f(x) \neq 0$.

1. Among all unit directions d, the slope $f'(x;d) = \langle \nabla f(x), d \rangle$ is maximal for $d = \nabla f(x) / \|\nabla f(x)\|$ and minimal for $d = -\nabla f(x) / \|\nabla f(x)\|$ (by Cauchy–Schwarz). The gradient points in the direction of *steepest ascent*, and $-\nabla f(x)$ in the direction of *steepest descent*.
2. $\nabla f(x)$ is orthogonal to the level set $\{y \mid f(y) = f(x)\}$ at x.

2.1.2 Useful gradients

The following formulas are used constantly; they can all be obtained from the first-order expansion by identifying the term linear in h. Here $A \in \mathcal{M}_n(\mathbb{R})$, $Q \in \mathcal{S}^n$, $B \in \mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}^m$.

$f(x)$	$\nabla f(x)$	$\nabla^2 f(x)$
$b^\top x$	b	0
$x^\top A x$	$(A + A^\top)x$	$A + A^\top$
$\frac{1}{2} x^\top Q x + b^\top x + \alpha$	$Qx + b$	Q
$\|x\|_2^2$	$2x$	$2\mathbf{I}_n$
$\|Bx - c\|_2^2$	$2B^\top(Bx - c)$	$2B^\top B$
$\|x\|_2 \quad (x \neq 0)$	$x / \|x\|_2$	$\frac{1}{\|x\|_2}\left(\mathbf{I}_n - \frac{xx^\top}{\|x\|_2^2}\right)$

Example 38. For $f(x) = \|Bx - c\|_2^2$: $f(x+h) = \|Bx - c\|^2 + 2\langle Bx - c, Bh\rangle + \|Bh\|^2 = f(x) + \langle 2B^\top(Bx - c), h\rangle + o(\|h\|)$, hence $\nabla f(x) = 2B^\top(Bx - c)$. Setting it to zero gives the normal equations $B^\top Bx = B^\top c$ of Chapter 1.

Example 39. Let $f(x, y) = x^2 + 3xy + y^3$. Then

$$\nabla f(x, y) = \begin{pmatrix} 2x + 3y \\ 3x + 3y^2 \end{pmatrix}, \qquad \nabla^2 f(x, y) = \begin{pmatrix} 2 & 3 \\ 3 & 6y \end{pmatrix}.$$

2.2 Existence of minimizers

Before looking for a minimizer, one should make sure that it exists. We use the following vocabulary: a set $S \subset \mathbb{R}^n$ is *closed* if it contains the limits of its convergent sequences, *bounded* if $S \subset B(0, R)$ for some $R > 0$, and *compact* if it is closed and bounded. For instance $[0, 1]$ and the unit sphere are compact, $(0, 1]$ is not closed and $\mathbb{R}^n_+$ is not bounded.

Proposition 22 (Weierstrass). If $f : S \to \mathbb{R}$ is continuous and $S \subset \mathbb{R}^n$ is nonempty and compact, then f attains its minimum and its maximum on S.

When S is not bounded (e.g. $S = \mathbb{R}^n$), we need f to "go up at infinity".

Definition 25. A function $f : \mathbb{R}^n \to \mathbb{R}$ is **coercive** if $f(x) \to +\infty$ as $\|x\| \to +\infty$.

Proposition 23. If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and coercive and $S \subset \mathbb{R}^n$ is nonempty and closed, then f has a global minimizer on S.

Example 40.

- $f(x) = \|x\|^2$ and more generally $f(x) = \frac{1}{2}x^\top Qx + b^\top x$ with $Q \in \mathcal{S}^n_{++}$ are coercive (since $x^\top Qx \geq \lambda_{\min}(Q)\|x\|^2$).
- e^x on $\mathbb{R}$ and x_1^2 on $\mathbb{R}^2$ are not coercive: the first has no minimizer, the second has infinitely many.
- The least squares objective $\|Ax - b\|^2$ always has a minimizer, although it is coercive only if $\mathrm{rk}(A) = n$.

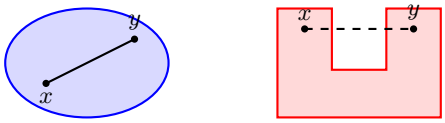
2.3 Convex sets

Definition 26. A set $C \subset \mathbb{R}^n$ is **convex** if for all $x, y \in C$ and $\lambda \in [0, 1]$, $\lambda x + (1 - \lambda)y \in C$, i.e., the segment $[x, y]$ is contained in C.

Remark 20 (Intuition). A set is convex if every point can "see" every other point: a straight line between two points of C never leaves C. A disk is convex; a crescent, a star or a ring are not.

Example 41 (Convex sets).

- Vector subspaces, affine sets $\{x \mid Ax = b\}$ (e.g. lines and planes), hyperplanes $\{x \mid a^\top x = \beta\}$ and half-spaces $\{x \mid a^\top x \leq \beta\}$.
- Balls $\{x \mid \|x - a\| \leq r\}$ for *any* norm (see the unit balls of Chapter 1).
- Polyhedra $\{x \mid Ax \leq b\}$ (componentwise inequalities), e.g. boxes $[a_1, b_1] \times \cdots \times [a_n, b_n]$, the orthant $\mathbb{R}^n_+$ and the simplex Δ_n.
- $\mathcal{S}^n_+$ and $\mathcal{S}^n_{++}$ are convex subsets of $\mathcal{S}^n$.



convex not convex

Figure 2.1: A convex set contains every segment joining two of its points.

- Non-examples: the sphere $\{x \mid \|x\| = 1\}$, the set $\{x \in \mathbb{R}^2 \mid x_1 x_2 \geq 1\}$, $\mathbb{Z}^n$.

Proposition 24 (Operations preserving convexity). 1. Any intersection (even infinite) of convex sets is convex. For instance a polyhedron is an intersection of half-spaces.

2. If C_1, C_2 are convex and $\lambda \in \mathbb{R}$, then $C_1 + C_2 = \{x + y \mid x \in C_1, y \in C_2\}$, λC_1 and $C_1 \times C_2$ are convex.

3. Images and preimages of convex sets by affine maps $x \mapsto Ax + b$ are convex.

Unions of convex sets are in general *not* convex.

Definition 27 (Convex combinations and convex hull). A **convex combination** of $x_1, \ldots, x_m$ is a point $\sum_{i=1}^m \lambda_i x_i$ with $\lambda \in \Delta_m$ ($\lambda_i \geq 0$, $\sum_i \lambda_i = 1$): a weighted average. The **convex hull** $\mathrm{conv}(S)$ of a set S is the set of all convex combinations of points of S; it is the smallest convex set containing S.

Remark 21 (Intuition). In the plane, the convex hull of a finite set of points is obtained by stretching a rubber band around them and letting it go. For example $\mathrm{conv}\{\pm e_1, \pm e_2\}$ is the ℓ_1 unit ball (a diamond), and the convex hull of the 2^n points $\{\pm 1\}^n$ is the ℓ_∞ unit ball (a cube).

Definition 28. A set K is a **cone** if $\lambda x \in K$ for all $x \in K$ and $\lambda \geq 0$. Examples of convex cones: $\mathbb{R}^n_+$, $\mathcal{S}^n_+$, and the second-order ("ice-cream") cone $\{(x, t) \in \mathbb{R}^n \times \mathbb{R} \mid \|x\|_2 \leq t\}$.

2.4 Convex functions

2.4.1 Definition and examples

In this section $C \subset \mathbb{R}^n$ is a nonempty convex set.

Definition 29. A function $f : C \to \mathbb{R}$ is **convex** if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \qquad \forall x, y \in C,\ \forall \lambda \in [0,1].$$

It is **strictly convex** if the inequality is strict for $x \neq y$ and $\lambda \in (0,1)$, and **concave** if $-f$ is convex.

Remark 22 (Intuition). Geometrically, the chord joining $(x, f(x))$ and $(y, f(y))$ lies above the graph: a convex function is shaped like a bowl, with no bumps and no "second valley". Equivalently, its *epigraph* $\mathrm{epi}(f) = \{(x, t) \mid f(x) \leq t\}$ (the region above the graph) is a convex set: f is convex iff $\mathrm{epi}(f)$ is convex.

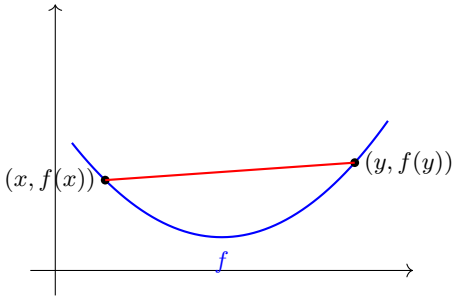


Figure 2.2: The chord lies above the graph of a convex function.

Example 42 (Convex functions).

- Affine functions $a^\top x + b$ (both convex and concave) and all norms $\|x\|$.
- On $\mathbb{R}$: x^2, $|x|^p$ $(p \geq 1)$, $e^{\alpha x}$ $(\alpha \in \mathbb{R})$; on $(0, +\infty)$: $-\log x$, $x \log x$, $1/x$.
- The quadratic function $f(x) = \frac{1}{2} x^\top Q x + b^\top x + c$ with $Q \in \mathcal{S}^n$ is convex iff $Q \in \mathcal{S}^n_+$, and strictly convex iff $Q \in \mathcal{S}^n_{++}$.
- $x \mapsto \max_i x_i$, the log-sum-exp $x \mapsto \log \sum_i e^{x_i}$ (a "smooth max" used in machine learning) and $X \mapsto -\log\det X$ on $\mathcal{S}^n_{++}$.
- Non-examples: x^3 on $\mathbb{R}$, $\sqrt{|x|}$, $\sin x$, $x_1 x_2$ on $\mathbb{R}^2$.

Proposition 25 (Jensen's inequality). If f is convex, $x_1, \ldots, x_m \in C$ and $\lambda \in \Delta_m$, then $f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i)$: "$f$ of an average is at most the average of f". In probabilistic terms, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.

Proposition 26 (Operations preserving convexity).

1. Nonnegative weighted sums: if $f_1, \ldots, f_m$ are convex and $\alpha_i \geq 0$, then $\sum_i \alpha_i f_i$ is convex.
2. Pointwise supremum: if $(f_i)_{i \in I}$ are convex, then $x \mapsto \sup_{i \in I} f_i(x)$ is convex.
3. Composition with an affine map: if f is convex, then $x \mapsto f(Ax + b)$ is convex.
4. If f is convex and $g : \mathbb{R} \to \mathbb{R}$ is convex and nondecreasing, then $g \circ f$ is convex (e.g. $e^{f(x)}$).

Example 43. The regularized least squares (ridge) objective $\|Ax - b\|_2^2 + \lambda \|x\|_2^2$ and the LASSO objective $\|Ax - b\|_2^2 + \lambda \|x\|_1$ $(\lambda \geq 0)$ are convex, as sums of convex functions composed with affine maps.

Proposition 27. If f is convex, its sublevel sets $\{x \in C \mid f(x) \leq \alpha\}$ are convex for every $\alpha \in \mathbb{R}$. (The converse is false: $\sqrt{|x|}$ has convex sublevel sets but is not convex.)

2.4.2 Characterizations for differentiable functions

Proposition 28 (First-order characterization). Let f be C^1 on an open convex set C. Then f is convex iff

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \qquad \forall x, y \in C,$$

i.e., the graph of f lies above all its tangent hyperplanes. Equivalently, ∇f is *monotone*: $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$ for all $x, y \in C$.

Remark 23. This "gradient inequality" is the key to convex optimization: if $\nabla f(x^*) = 0$, it gives $f(y) \geq f(x^*)$ for all y. For a convex function, *local information (the tangent plane) yields a global lower bound.*

Proposition 29 (Second-order characterization)**.** Let f be C^2 on an open convex set C. Then

- f is convex $\Leftrightarrow \nabla^2 f(x) \succeq 0$ for all $x \in C$;
- if $\nabla^2 f(x) \succ 0$ for all $x \in C$, then f is strictly convex (the converse fails: x^4 is strictly convex but $f''(0) = 0$).

Example 44.

- $f(x) = \|Ax - b\|_2^2$ has $\nabla^2 f(x) = 2A^\top A \succeq 0$: it is convex, and strictly convex iff $\mathrm{rk}(A) = n$ (Chapter 1).
- $f(x, y) = x^2 + 3xy + y^3$ (see above) has $\nabla^2 f = \begin{pmatrix} 2 & 3 \\ 3 & 6y \end{pmatrix}$ with determinant $12y - 9$, which is negative for $y < 3/4$: f is not convex on $\mathbb{R}^2$.
- $f(x, y) = \frac{x^2}{y}$ on $\mathbb{R} \times (0, +\infty)$ has $\nabla^2 f = \frac{2}{y^3} \begin{pmatrix} y^2 & -xy \\ -xy & x^2 \end{pmatrix} = \frac{2}{y^3} \begin{pmatrix} y \\ -x \end{pmatrix} \begin{pmatrix} y \\ -x \end{pmatrix}^\top \succeq 0$: it is convex.

2.4.3 Strong convexity and smoothness

Two quantitative notions govern the behaviour of optimization algorithms.

Definition 30. Let f be C^1 on $\mathbb{R}^n$ and $0 < \mu \leq L$.

- f is **μ-strongly convex** if $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex, or equivalently $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$ for all x, y.
- f is **L-smooth** if ∇f is L-Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all x, y. We write $f \in C_L^{1,1}$.

If f is C^2: f is μ-strongly convex iff $\nabla^2 f(x) \succeq \mu \mathbf{I}_n$, and f is L-smooth iff $\left\|\nabla^2 f(x)\right\|_2 \leq L$, for all x. For a convex C^2 function, the two conditions read

$$\mu\, \mathbf{I}_n \preceq \nabla^2 f(x) \preceq L\, \mathbf{I}_n \quad \forall x, \quad \text{i.e.} \quad \mu \leq \lambda_{\min}(\nabla^2 f(x)) \leq \lambda_{\max}(\nabla^2 f(x)) \leq L.$$

Remark 24 (Intuition)**.** Strong convexity says that the bowl is curved *at least* by μ in every direction (no flat bottom), and smoothness says that it is curved *at most* by L (no sharp turns). The ratio $\kappa = L/\mu \geq 1$ is the **condition number** of f: $\kappa \approx 1$ means a round bowl, $\kappa \gg 1$ a long narrow valley. A strongly convex function is strictly convex and coercive, hence has a unique minimizer.

Proposition 30 (Descent lemma)**.** If f is L-smooth, then for all $x, y \in \mathbb{R}^n$:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

In words: an L-smooth function lies below a quadratic upper model built at any point x (and, if convex, above its tangent hyperplane at x). This sandwich is the basis of the analysis of gradient descent.

Example 45 (The quadratic case). For $f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$ with $Q \in \mathcal{S}^n_{++}$, $\nabla^2 f = Q$, so f is μ-strongly convex and L-smooth with

$$\mu = \lambda_{\min}(Q), \qquad L = \lambda_{\max}(Q), \qquad \kappa = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} = \kappa_2(Q).$$

Its level sets are ellipses (ellipsoids) with axes along the eigenvectors of Q (Chapter 1): the condition number of the function is exactly the condition number of the matrix.

2.5 Optimality conditions

2.5.1 Unconstrained problems

Proposition 31 (First-order necessary condition, Fermat's rule). Let f be differentiable on an open set U. If $x^* \in U$ is a local minimizer (or maximizer) of f, then $\nabla f(x^*) = 0$.

A point with $\nabla f(x^*) = 0$ is called a **critical** (or stationary) point. The converse is false: $f(x) = x^3$ has $f'(0) = 0$, but 0 is neither a minimizer nor a maximizer.

Remark 25 (Intuition). At the bottom of a valley (or at a summit, or at a mountain pass) the ground is flat: the tangent plane is horizontal. Being flat is necessary to be a minimum, not sufficient.

Proposition 32 (Second-order conditions). Let f be C^2 on an open set U and let $x^* \in U$ be a critical point.

1. (Necessary) If x^* is a local minimizer, then $\nabla^2 f(x^*) \succeq 0$; if it is a local maximizer, then $\nabla^2 f(x^*) \preceq 0$.
2. (Sufficient) If $\nabla^2 f(x^*) \succ 0$, then x^* is a strict local minimizer; if $\nabla^2 f(x^*) \prec 0$, it is a strict local maximizer.
3. If $\nabla^2 f(x^*)$ has both a positive and a negative eigenvalue, then x^* is a **saddle point**: neither a local minimizer nor a local maximizer.

If $\nabla^2 f(x^*)$ is only semidefinite (some zero eigenvalues), no conclusion can be drawn from the Hessian alone.

In practice the Hessian at a critical point is classified with the criteria of Chapter 1 (eigenvalues, or for $n = 2$ the sign of the determinant and of the trace).

Example 46. Let $f(x, y) = x^3 + y^3 - 3x - 3y$. Then $\nabla f = (3x^2 - 3,\ 3y^2 - 3)^\top$ vanishes at the four points $(\pm 1, \pm 1)$, and $\nabla^2 f(x, y) = \mathrm{diag}(6x, 6y)$.

- At $(1, 1)$: $\mathrm{diag}(6, 6) \succ 0$, strict local minimizer (but not global: $f(x, 0) \to -\infty$ as $x \to -\infty$).
- At $(-1, -1)$: $\mathrm{diag}(-6, -6) \prec 0$, strict local maximizer.
- At $(1, -1)$ and $(-1, 1)$: eigenvalues of opposite signs, saddle points.

Similarly $f(x, y) = x^2 - y^2$ has a saddle at the origin: it is a minimum along the x-axis and a maximum along the y-axis (the shape of a horse saddle or a mountain pass).

2.5.2 The convex case: local is global

Convexity removes all the difficulties above.

Proposition 33. Let C be convex and $f : C \to \mathbb{R}$ convex.

1. Every local minimizer of f on C is a global minimizer.
2. The set $\operatorname{argmin}_C f$ is convex (possibly empty). If f is strictly convex, it contains at most one point.
3. If f is differentiable on $C = \mathbb{R}^n$: x^* is a global minimizer $\Leftrightarrow \nabla f(x^*) = 0$.

Remark 26 (Intuition). A convex landscape has a single valley (possibly with a flat bottom): as soon as the hiker finds a flat spot, she knows she has reached the lowest point. In nonconvex problems (e.g. training neural networks) one can get stuck in local minima or slowed down near saddle points.

Example 47 (Quadratic functions). Let $f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$ with $Q \in \mathcal{S}^n$.

- If $Q \succ 0$: f is strongly convex and the unique global minimizer solves $\nabla f(x) = Qx + b = 0$, i.e., $x^* = -Q^{-1}b$. Minimizing a strongly convex quadratic is *the same as* solving an SPD linear system.
- If $Q \succeq 0$ is singular: minimizers exist iff $b \in \operatorname{Im}(Q)$, and they form the affine set $\{x \mid Qx = -b\}$.
- If Q has a negative eigenvalue: f is unbounded below ($f^* = -\infty$).

For least squares, $f(x) = \|Ax - b\|^2$ is convex, so its minimizers are exactly the solutions of the normal equations $A^\top Ax = A^\top b$.

2.5.3 A glimpse at constrained problems

When $C \neq \mathbb{R}^n$, the minimizer may lie on the boundary of C, where ∇f need not vanish.

Proposition 34. Let C be closed and convex and f be C^1 and convex. Then $x^* \in C$ minimizes f on C iff

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \qquad \forall x \in C,$$

i.e., no feasible direction is a descent direction. For $C = \mathbb{R}^n$ this reduces to $\nabla f(x^*) = 0$.

Example 48. Minimize $f(x, y) = x + y$ on the unit disk. Since $\nabla f = (1, 1)^\top$ never vanishes, the minimizer is on the boundary: by Cauchy–Schwarz $x + y \geq -\sqrt{2}\,\|(x, y)\| \geq -\sqrt{2}$, with equality at $x^* = -\frac{1}{\sqrt{2}}(1, 1)^\top$. There, $-\nabla f(x^*)$ points outwards, orthogonally to the circle: going downhill would mean leaving C. The general tool for constrained problems (Lagrange multipliers, KKT conditions) formalizes this picture and is beyond the scope of this refresher.

Definition 31 (Orthogonal projection onto a convex set). If C is nonempty, closed and convex, every $x \in \mathbb{R}^n$ has a unique closest point in C:

$$P_C(x) = \operatorname{argmin}_{y \in C} \|y - x\|_2 .$$

It is characterized by $\langle x - P_C(x), y - P_C(x) \rangle \leq 0$ for all $y \in C$ (the angle at $P_C(x)$ is obtuse),

and P_C is 1-Lipschitz.

Example 49.
- Box $C = [a_1, b_1] \times \cdots \times [a_n, b_n]$: $P_C(x)_i = \min(\max(x_i, a_i), b_i)$ ("clipping", `np.clip`).
- Ball $C = \{\|y\|_2 \leq r\}$: $P_C(x) = x$ if $\|x\|_2 \leq r$ and $P_C(x) = r\,x/\|x\|_2$ otherwise.
- Subspace $C = \mathrm{Im}(A)$: P_C is the linear orthogonal projection of Chapter 1.

2.6 Descent algorithms

Except for quadratic functions, the equation $\nabla f(x) = 0$ cannot be solved in closed form. Iterative algorithms build a sequence $(x_k)_{k\geq 0}$ with $f(x_k) \to f^*$ and, hopefully, $x_k \to x^*$.

2.6.1 General descent scheme

Definition 32. A nonzero vector d is a **descent direction** of f at x if $f'(x; d) = \langle \nabla f(x), d\rangle < 0$. Then $f(x + td) < f(x)$ for all $t > 0$ small enough.

Geometrically, d is a descent direction iff it makes an obtuse angle with $\nabla f(x)$.

Descent method

- Initialization: choose $x_0 \in \mathbb{R}^n$ and a tolerance $\varepsilon > 0$.
- For $k = 0, 1, 2, \ldots$
  - choose a descent direction d_k;
  - choose a step size $t_k > 0$ such that $f(x_k + t_k d_k) < f(x_k)$;
  - set $x_{k+1} = x_k + t_k d_k$;
  - stop if $\|\nabla f(x_{k+1})\| \leq \varepsilon$.

Remark 27 (Intuition). This is the hiker in the fog: she only sees the ground at her feet (the gradient, possibly the curvature). At each step she chooses a direction that goes down (d_k) and a stride length (t_k). Different choices of d_k give different methods: gradient descent, Newton, quasi-Newton, conjugate gradient…

Choosing the step size (line search).

- **Constant step:** $t_k = \bar{t}$. Simple, but too small a step gives slow progress and too large a step can overshoot the valley and make f increase (or diverge).
- **Exact line search:** $t_k \in \mathrm{argmin}_{t\geq 0} f(x_k + t d_k)$, a one-dimensional problem. Rarely computable in closed form, except for quadratics: for $f(x) = \frac{1}{2}x^\top Q x + b^\top x$ and $g_k = \nabla f(x_k)$, the exact step along $d_k = -g_k$ is $t_k = \frac{\|g_k\|^2}{g_k^\top Q g_k}$.
- **Backtracking (Armijo):** fix $s > 0$ and $\alpha, \beta \in (0, 1)$ (typically $\alpha = 10^{-4}$, $\beta = \frac{1}{2}$). Start with $t = s$ and, while
$$f(x_k + t d_k) > f(x_k) + \alpha\, t\, \langle \nabla f(x_k), d_k\rangle\,,$$
replace t by βt. This loop always terminates for a descent direction and guarantees a *sufficient* decrease of f, not just any decrease. It is the default choice in practice.

2.6.2 Gradient descent

Gradient descent uses the steepest descent direction $d_k = -\nabla f(x_k)$, which is a descent direction as long as $\nabla f(x_k) \neq 0$ since $f'(x_k; -\nabla f(x_k)) = -\|\nabla f(x_k)\|^2 < 0$.

Gradient descent

- Initialization: choose $x_0 \in \mathbb{R}^n$ and a tolerance $\varepsilon > 0$.
- For $k = 0, 1, 2, \ldots$
  - choose t_k (constant, exact or backtracking line search);
  - set $x_{k+1} = x_k - t_k \nabla f(x_k)$;
  - stop if $\|\nabla f(x_{k+1})\| \leq \varepsilon$.

Remark 28 (Why the step $1/L$?). If f is L-smooth, the descent lemma at $x = x_k$ gives the quadratic upper model $f(y) \leq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|^2$. Minimizing this model in y gives exactly $y = x_k - \frac{1}{L}\nabla f(x_k)$, and then

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 .$$

Gradient descent with step $1/L$ is thus "minimize a safe quadratic over-estimate of f, and move to its minimizer". More generally, any constant step $t \in (0, 2/L)$ ensures that $f(x_k)$ decreases.

Proposition 35 (Convergence of gradient descent). Let f be L-smooth and bounded below, and let (x_k) be generated by gradient descent with $t_k = 1/L$ (similar results hold for backtracking).

1. (Nonconvex case) $f(x_k)$ is nonincreasing, $\nabla f(x_k) \to 0$, and $\min_{0 \leq j \leq k} \|\nabla f(x_j)\| \leq \sqrt{\dfrac{2L(f(x_0) - f^*)}{k+1}}$. Only convergence to a critical point is guaranteed.
2. (Convex case) If f is convex and has a minimizer x^*, then $f(x_k) - f^* \leq \dfrac{L\|x_0 - x^*\|^2}{2k}$, and $x_k \to$ some minimizer.
3. (Strongly convex case) If f is μ-strongly convex with minimizer x^*, then

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{1}{\kappa}\right)^k \|x_0 - x^*\|^2 ,$$

$$f(x_k) - f^* \leq \left(1 - \frac{1}{\kappa}\right)^k \big(f(x_0) - f^*\big), \qquad \kappa = \frac{L}{\mu}.$$

Remark 29 (Reading convergence rates). In the convex case, reaching $f(x_k) - f^* \leq \varepsilon$ requires $\mathcal{O}(1/\varepsilon)$ iterations (*sublinear* rate). In the strongly convex case the error is multiplied by at most $(1 - 1/\kappa)$ at each iteration (*linear* rate, a straight line on a log-scale plot), and about $\kappa \log(1/\varepsilon)$ iterations suffice. The condition number is therefore the key quantity: gradient descent is fast on round bowls and slow in narrow valleys.

Example 50 (Gradient descent in a narrow valley). Let $f(x_1, x_2) = \frac{1}{2}(x_1^2 + \gamma x_2^2)$ with $\gamma \geq 1$, so that $\nabla f(x) = (x_1, \gamma x_2)^\top$, $\mu = 1$, $L = \gamma$ and $\kappa = \gamma$. With a constant step t, the iteration decouples:

$$x_{1,k+1} = (1 - t)\, x_{1,k}, \qquad x_{2,k+1} = (1 - t\gamma)\, x_{2,k}.$$

- It converges for every x_0 iff $|1 - t| < 1$ and $|1 - t\gamma| < 1$, i.e., $0 < t < 2/\gamma = 2/L$. For $t > 2/L$ the x_2-component blows up: the hiker overshoots from one side of the valley to

the other, higher and higher.

- With $t = 1/L = 1/\gamma$, x_2 is exact after one step but $x_{1,k} = (1 - 1/\gamma)^k x_{1,0}$. For $\gamma = 100$, reducing the error by a factor 100 requires about 460 iterations, since $0.99^{458} \approx 0.01$.

With exact line search, starting from $x_0 = (\gamma, 1)$, one can show that $x_k = \left(\frac{\gamma-1}{\gamma+1}\right)^k (\gamma, (-1)^k)$: successive steps are orthogonal, and the iterates zigzag across the valley (Fig. 2.3).

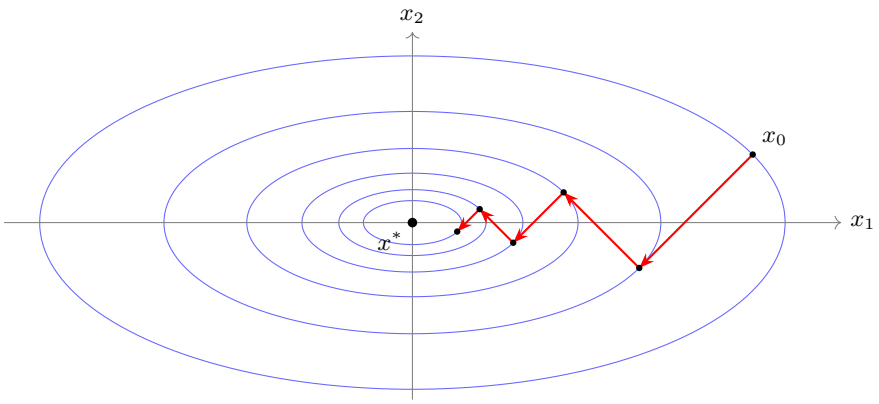


Figure 2.3: Gradient descent with exact line search on $f(x) = \frac{1}{2}(x_1^2 + 5x_2^2)$: the iterates zigzag across the elongated level sets.

Remark 30 (Going further). Many popular methods are refinements of gradient descent. *Momentum* and *Nesterov's accelerated gradient* add inertia (a ball rolling downhill rather than a cautious hiker) and improve the dependence on the condition number from κ to $\sqrt{\kappa}$. *Preconditioning* changes variables to make the valley rounder.

Remark 31 (Stochastic gradient descent and machine learning). In machine learning the objective is an average over the training data, $f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$, where f_i is the loss on the i-th sample and N can be in the millions or billions. A single exact gradient $\nabla f(x) = \frac{1}{N}\sum_i \nabla f_i(x)$ then requires a full pass over the dataset. *Stochastic gradient descent* (SGD) replaces it by the gradient on a small random *mini-batch* $B_k \subset \{1, \dots, N\}$ of size $b \ll N$:

$$x_{k+1} = x_k - t_k\, g_k, \qquad g_k = \frac{1}{b}\sum_{i \in B_k} \nabla f_i(x_k).$$

The estimate g_k is noisy but unbiased ($\mathbb{E}[g_k] = \nabla f(x_k)$), and its cost depends on b only, not on N: this is what makes training on massive datasets possible. The price to pay is the noise, which forces the use of small or decreasing step sizes (the "learning rate") and yields slower convergence per iteration than exact gradient descent, a trade-off largely in favour of SGD when N is huge. Variants such as Adam are the default optimizers of deep learning libraries.

2.6.3 Newton's method

Gradient descent uses a first-order (linear) model of f. Newton's method uses the second-order model: at x_k, replace f by its Taylor quadratic approximation

$$m_k(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}(x - x_k)^\top \nabla^2 f(x_k)(x - x_k),$$

and, if $\nabla^2 f(x_k) \succ 0$, jump to the minimizer of m_k, obtained by solving $\nabla m_k(x) = 0$:

$$x_{k+1} = x_k + d_k, \qquad \text{where} \quad \nabla^2 f(x_k)\, d_k = -\nabla f(x_k).$$

Remark 32 (Intuition). Instead of taking a step in the steepest direction, the hiker fits a bowl to the ground under her feet (slope *and* curvature) and jumps directly to the bottom of this bowl. Equivalently, Newton's method for minimization is Newton's root-finding method $x_{k+1} = x_k - J_F(x_k)^{-1}F(x_k)$ applied to the equation $F(x) = \nabla f(x) = 0$; in 1D, $x_{k+1} = x_k - f'(x_k)/f''(x_k)$.

Newton's method

- Initialization: choose $x_0 \in \mathbb{R}^n$ and a tolerance $\varepsilon > 0$.
- For $k = 0, 1, 2, \ldots$
  - compute the Newton direction d_k by solving the linear system $\nabla^2 f(x_k)\, d_k = -\nabla f(x_k)$ (e.g. with a Cholesky factorization, Chapter 1);
  - set $x_{k+1} = x_k + d_k$;
  - stop if $\|\nabla f(x_{k+1})\| \le \varepsilon$.

Remark 33. If $\nabla^2 f(x_k) \succ 0$, the Newton direction is a descent direction, since $\langle \nabla f(x_k), d_k \rangle = -\nabla f(x_k)^\top \nabla^2 f(x_k)^{-1} \nabla f(x_k) < 0$. It is the gradient direction measured in the Q-norm of Chapter 1 with $Q = \nabla^2 f(x_k)$: Newton's method automatically "rounds the valley", which is why it is insensitive to conditioning.

Proposition 36 (Local quadratic convergence). Let $f \in C^2(\mathbb{R}^n)$ with $\nabla^2 f(x) \succeq m\mathbf{I}_n$ for all x ($m > 0$) and $\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \le L_H \|x - y\|$ for all x, y, and let x^* be its unique minimizer. Then Newton's iterates satisfy

$$\|x_{k+1} - x^*\| \le \frac{L_H}{2m} \|x_k - x^*\|^2 .$$

In particular, if $\|x_0 - x^*\| \le m/L_H$, then $\|x_k - x^*\| \le \frac{2m}{L_H} \left(\frac{1}{2}\right)^{2^k}$.

Remark 34. Quadratic convergence means that the number of correct digits roughly *doubles* at each iteration: once close to the solution, a handful of iterations gives machine precision. For a strongly convex quadratic $f(x) = \frac{1}{2}x^\top Q x + b^\top x$, the model m_k is f itself and Newton's method finds $x^* = -Q^{-1}b$ in *one* step, from any x_0.

Example 51 (Fast, but only locally). Minimize $f(x) = x - \log x$ on $(0, +\infty)$; its minimizer is $x^* = 1$. Since $f'(x) = 1 - \frac{1}{x}$ and $f''(x) = \frac{1}{x^2}$, the Newton iteration reads

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} = 2x_k - x_k^2, \qquad \text{so that} \qquad 1 - x_{k+1} = (1 - x_k)^2.$$

- From $x_0 = 0.5$, the errors $1 - x_k$ are $0.5,\ 0.25,\ 0.0625,\ 3.9 \cdot 10^{-3},\ 1.5 \cdot 10^{-5},\ 2.3 \cdot 10^{-10}$: the error is squared at each step.
- From $x_0 = 3$, we get $x_1 = 6 - 9 = -3$, which is outside the domain of f: far from the solution, the quadratic model is a poor guide and the jump can be disastrous.

Remark 35 (Newton in practice). The pure Newton method can fail when the starting point is far from x^*, or when $\nabla^2 f(x_k)$ is not positive definite (then d_k may not even be a descent direction, e.g. near a saddle point). The standard safeguard is the *damped Newton method*, $x_{k+1} = x_k + t_k d_k$ with t_k chosen by backtracking: far from the solution it behaves like a descent method, and close to it the full step $t_k = 1$ is accepted and quadratic convergence is recovered. Each iteration requires forming and solving an $n \times n$ linear system, which costs $\mathcal{O}(n^3)$ operations for a dense Hessian. *Quasi-Newton* methods (e.g. BFGS, L-BFGS) replace $\nabla^2 f(x_k)$ by an approximation B_k built only from successive gradients, updated at each iteration so as to

satisfy the *secant equation* $B_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k)$ (a finite-difference version of $\nabla^2 f\,\Delta x \approx \Delta \nabla f$). They offer a compromise between the two methods (see the table below); L-BFGS, which stores only a few vector pairs, is the default choice for smooth problems of moderate to large size (`scipy.optimize.minimize`).

2.6.4 Gradient descent vs. Newton vs. quasi-Newton

	Gradient descent	**Quasi-Newton (BFGS)**	**Newton**
Model of f	linear (+ step size)	quadratic, approximate Hessian B_k	quadratic (Taylor order 2)
Information used	∇f	∇f only (curvature learned from gradients)	∇f and $\nabla^2 f$
Cost per iteration	$\mathcal{O}(n)$ (+ one gradient)	$\mathcal{O}(n^2)$; $\mathcal{O}(mn)$ for L-BFGS with memory m	$\mathcal{O}(n^3)$ (solve a linear system)
Convergence	linear, rate $1 - 1/\kappa$ (strongly convex)	superlinear (locally)	quadratic (locally)
Sensitivity to conditioning	strong (zigzag in narrow valleys)	weak (learns the curvature)	none (affine invariant)
Robustness far from x^*	good with a suitable step	needs a line search	needs damping / line search
Typical use	very large n, stochastic setting (ML, imaging)	moderate to large n, smooth problems	moderate n, high accuracy

2.6.5 Handling simple constraints: projected gradient

For $\min_{x \in C} f(x)$ with C closed and convex, a feasible version of gradient descent is obtained by projecting back onto C after each step:

$$x_{k+1} = P_C\big(x_k - t_k \nabla f(x_k)\big).$$

The minimizers are exactly the fixed points $x^* = P_C(x^* - t\nabla f(x^*))$ $(t > 0)$, and for convex L-smooth f with $t_k = 1/L$ one recovers the rate $f(x_k) - f^* \leq \frac{L\|x_0 - x^*\|^2}{2k}$ of the unconstrained case. The method is only practical when P_C is cheap, e.g. for boxes, balls or the nonnegative orthant, where it amounts to a clipping or a rescaling.

References

[1] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020. (p. 1)

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. (p. 1)

[3] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. MOS-SIAM Series on Optimization. SIAM, 2014. (p. 1)

[4] Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2018. (p. 1)

[5] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001. (p. 1)

[6] Valeriu Soltan. *Lectures on Convex Sets*. World Scientific, 2022. (p. 1)

[7] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer, 1998. (p. 1)

[8] Hamza Ennaji. *Bases d'analyse convexe et optimisation*. Lecture notes (in French), Ensimag, Grenoble INP. 2025. (p. 1)