Dauphine | PSL
UNIVERSITÉ PARIS

Mathematics for Data Science

Clément W. Royer

Lecture notes - M1 IDD - 2024/2025

- The last version of these notes can be found at:
  https://www.lamsade.dauphine.fr/~croyer/ensdocs/MDS/PolyMDS.pdf.
- Comments, typos, etc, can be sent to `clement.royer@lamsade.dauphine.fr`.
  Thanks to the students who provided feedback.
  These notes have greatly benefited from Alexandre Vérine, Julien Lesca and Gil Puig i Surroca.
- **Major updates:**
  - 2025.01.02: Corrected typos in Chapters 2 and 3.
  - 2024.11.08: Added strong convexity back.
  - 2024.10.30: Added strict convexity back.
  - 2024.10.28: Modification to Theorem 2.3.
  - 2024.10.25: Adjusted content of Chapter 1, fixed typos in Chapter 2.
  - 2024.10.21: Corrected typos in Chapter 1, added Chapter 2.
  - 2024.09.27: Revised contents of Chapter 1 as well as Example 0.2.
  - 2024.09.10: First version of the notes with Chapter 1.
- **Learning goals:**
  - Identify and use convex sets and convex functions.
  - Write convex formulations of optimization problems.
  - Use concentration inequalities to bound functions of random vectors and matrices.

Contents

Introduction

About data science

Data science tasks have grown to prominence in modern society. Numerous economical models are now based on the value of data, and the way this data is exploited. Handling massive amounts of data, as in biology, poses a number of mathematical and computational challenges. More globally, *data-driven* approaches are taking over *model-based approaches*, in that the former apply when the latter cannot be implemented.

Course summary

This course aims at describing the mathematical foundations of data science tasks. The underlying goal is for students to become comfortable with these models, not only for subsequent courses but also to leverage those tools in academic or industrial settings.

The first part of the course is centered around mathematical optimization. We focus on convex optimization problems, that remain the formulations of choice for many data science tasks. From a mathematical perspective, these problems possess a structure that allows for characterizing their solutions.

The second part of the course revolves around statistical aspects, that are prevalent in data science. We will present a series of results associated with estimation and regression problems, combining convex problems with statistics. We will also investigate concentration inequalities, and how those are used to provide statistical guarantees on certain problems.

Chapter 0

Notations and background

This chapter gathers all the notations and mathematical background that will be used throughout the course.

0.1 Notations

- Scalars (i.e. reals) are denoted by lowercase letters: $a, b, c, \alpha, \beta, \gamma$.
- Vectors are denoted by **bold** lowercase letters: $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$.
- Matrices are denoted by **bold** uppercase letters: $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.
- Sets are denoted by **bold** uppercase cursive letters : $\mathcal{A}, \mathcal{B}, \mathcal{C}$.
- A new operator or quantity is defined using $:=$.
- The following quantifiers are used throughout the notes: $\forall$ (for every), $\exists$ (it exists), $\exists!$ (it exists a unique), $\in$ (belongs to), $\subseteq$ (subset of), $\subset$ (proper subset).
- The Σ operator is used for sums. To lighten the notation, and in the absence of ambiguity, we may omit the first and last indices, or use one sum over multiple indices. As a result, the notations $\sum_{i=1}^{m} \sum_{j=1}^{n}$, $\sum_i \sum_j$ and $\sum_{i,j}$ may be used interchangeably.
- The Π operator is used for products. To lighten the notation, and in the absence of ambiguity, we may omit the first and last indices, or use one sum over multiple indices. As a result, the notations $\prod_{i=1}^{m} \prod_{j=1}^{n}$, $\prod_i \prod_j$ and $\prod_{i,j}$ may be used interchangeably.
- The notation $i = 1, \ldots, m$ indicates that the variable i takes all integer values between 1 and m.

0.1.1 Scalar and vector notations

- The set of natural numbers (nonnegative integers) is denoted by $\mathbb{N}$; the set of integers is denoted by $\mathbb{Z}$.
- The set of real numbers is denoted by $\mathbb{R}$. Our notations for the subset of nonnegative real numbers and the set of positive real numbers are $\mathbb{R}_+$ and $\mathbb{R}_{++}$, respectively. We also define the extended real line $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$.

- The notation $\mathbb{R}^n$ is used for the set of vectors with $n \in \mathbb{N}$ real components; although we do not explicitly indicate it in the rest of these notes, we always assume that $n \geq 1$.
- A vector $\boldsymbol{x} \in \mathbb{R}^n$ is thought as a column vector, with $x_i \in \mathbb{R}$ denoting its i-th coordinate in the canonical basis of $\mathbb{R}^n$. We thus write $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, or, in a compact form, $\boldsymbol{x} = [x_i]_{1 \leq i \leq n}$.
- Given a column vector $\boldsymbol{x} \in \mathbb{R}^n$, the corresponding row vector is denoted by $\boldsymbol{x}^{\mathrm{T}}$, so that $\boldsymbol{x}^{\mathrm{T}} = [x_1 \ \cdots \ x_n]$ and $[\boldsymbol{x}^{\mathrm{T}}]^{\mathrm{T}} = \boldsymbol{x}$.
- For any integer $n \geq 1$, the vectors $\mathbf{0}_n$ and $\mathbf{1}_n$ correspond to the vectors of $\mathbb{R}^n$ for which all elements are 0 or 1, respectively. For simplicity, we may write $\boldsymbol{x} \geq 0$ to indicate that all components of $\boldsymbol{x}$ are nonnegative.

0.1.2 Matrix notations

- We use $\mathbb{R}^{m \times n}$ to denote the set of real rectangular matrices with m rows and n columns, where m et n will always be assumed to be at least 1. If $m = n$, $\mathbb{R}^{n \times n}$ refers to the set of square matrices of size n.
- We identify a matrix in $\mathbb{R}^{m \times 1}$ with its corresponding column vector in $\mathbb{R}^m$.
- Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, A_{ij} or $[\boldsymbol{A}]_{ij}$ refers to the coefficient from the i-th row and the j-th column of $\boldsymbol{A}$. Provided this notation is not ambiguous, we use the notations $\boldsymbol{A}$, $[A_{ij}]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$ and $[A_{ij}]$ interchangeably.
- Depending on the context, we may use $\boldsymbol{a}_i^{\mathrm{T}}$ to denote the i-th row of $\boldsymbol{A}$ or $\boldsymbol{a}_j$ to denote the j-th column of $\boldsymbol{A}$, leading to $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{a}_m^{\mathrm{T}} \end{bmatrix}$ or $\boldsymbol{A} = [\boldsymbol{a}_1 \ \cdots \ \boldsymbol{a}_n]$, respectively.
- The diagonal of a square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is given by the coefficients A_{ii}. The trace of such a matrix is $\mathrm{trace}(\boldsymbol{A}) := \sum_{i=1}^{d} A_{ii}$.
- Given $\boldsymbol{A} = [\boldsymbol{A}_{ij}] \in \mathbb{R}^{m \times n}$, the *transpose of matrix* $\boldsymbol{A}$, denoted by $\boldsymbol{A}^{\mathrm{T}}$ (read "$\boldsymbol{A}$ transpose"), is defined as the matrix in $\mathbb{R}^{n \times m}$ (or "n-by-m matrix") such that

  $$\forall i = 1 \ldots m, \ \forall j = 1 \ldots n, \quad [\boldsymbol{A}^{\mathrm{T}}]_{ji} = A_{ij}.$$

  Note that this generalizes the notation used for row vectors.
- For every $n \geq 1$, $\mathbf{I}_n$ refers to the identity matrix in $\mathbb{R}^{n \times n}$ (with 1s on the diagonal and 0s elsewhere).

0.2 Linear algebra

This section provides useful linear algebra results for this course. Unlike general linear algebra classes, we focus on linear algebra in $\mathbb{R}^n$.

0.2.1 Vector linear algebra

We always consider vectors in the normed vector space $\mathbb{R}^n$, of dimension n. The following operations are defined in this space:

- For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the sum of $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted by $\boldsymbol{x} + \boldsymbol{y} = [x_i + y_i]_{1 \leq i \leq n}$;
- For any $\lambda \in \mathbb{R}$, we define $\lambda \boldsymbol{x} \overset{n}{=} \lambda \cdot \boldsymbol{x} = [\lambda x_i]_{1 \leq i \leq n}$. In this context, the real value λ is called a *scalar*.

Using these operations, we can build **linear combinations** of vectors in $\mathbb{R}^n$ that produce a vector in $\mathbb{R}^n$ of the form $\sum_{i=1}^p \lambda_i \boldsymbol{x}_i$, where $\boldsymbol{x}_i \in \mathbb{R}^n$ and $\lambda_i \in \mathbb{R}$ for any $i = 1, \ldots, p$.

The matrix space $\mathbb{R}^{m \times n}$ can also be endowed with a vector space structure of dimension mn:

- For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$, the sum of $\boldsymbol{A}$ and $\boldsymbol{B}$ is denoted by $\boldsymbol{A} + \boldsymbol{B} = [\boldsymbol{A}_{ij} + \boldsymbol{B}_{ij}]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$;
- For any scalar $\lambda \in \mathbb{R}$, we define $\lambda \boldsymbol{A} \overset{n}{=} \lambda \cdot \boldsymbol{A} = [\lambda \boldsymbol{A}_{ij}]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$.

Definition 0.1 *A set $\mathcal{S} \subseteq \mathbb{R}^n$ satisfying the conditions*

1. $\mathbf{0}_n \in \mathcal{S}$;
2. $\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S},\ \boldsymbol{x} + \boldsymbol{y} \in \mathcal{S}$;
3. $\forall \boldsymbol{x} \in \mathcal{S},\ \forall \lambda \in \mathbb{R},\ \lambda \boldsymbol{x} \in \mathcal{S}$.

is called a (linear) subspace *of $\mathbb{R}^n$.*

Definition 0.2 *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ be p vectors in $\mathbb{R}^n$. The span (or linear span) of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, denoted by $\mathrm{Span}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, is the linear subspace of $\mathbb{R}^n$ defined by*

$$\mathrm{Span}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) := \left\{ \boldsymbol{x} = \sum_{i=1}^p \alpha_i \boldsymbol{x}_i \middle| \alpha_i \in \mathbb{R}\ \forall i \right\}.$$

We now recall various properties of vector sets.

Definition 0.3

- *The vectors in a set $\{\boldsymbol{x}_i\}_{i=1}^k \subset \mathbb{R}^n$ are called linearly independent if for any scalars $\lambda_1, \ldots, \lambda_k$ satisfying $\sum_{i=1}^k \lambda_i \boldsymbol{x}_i = 0$, we have $\lambda_1 = \cdots = \lambda_k = 0$. In that case, $k \leq n$.*
- *If the above property does not hold, the vectors are called linearly dependent.*
- *A spanning set is a set of vectors $\{\boldsymbol{x}_i\} \subset \mathbb{R}^n$ such that their span is $\mathbb{R}^n$.*
- *A set of vectors $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^n$ is a basis if it is both linearly independent and a spanning set. In that case, any vector in $\mathbb{R}^n$ can be written as a uniquely defined linear combination of the $\boldsymbol{x}_i$s. Any basis in $\mathbb{R}^n$ has exactly n vectors.*

Since the size of a basis in $\mathbb{R}^n$ is n, we say that the dimension of the space is n. Consequently, any linear subspace of $\mathbb{R}^n$ has dimension at most n.

Example 0.1 *Any vector $\boldsymbol{x}$ in $\mathbb{R}^n$ can be written as $\boldsymbol{x} = \sum_{i=1}^n x_i \boldsymbol{e}_i$, where $\boldsymbol{e}_i = [0 \cdots 0\ 1\ 0 \cdots 0]^{\mathrm{T}}$ is the ith vector of the canonical basis (with a 1 in the ith coordinate).*

Norm and scalar product Using a Euclidean norm and its associated scalar product allows to compare vectors by measuring the distance between them. This ability is particularly useful to establish that a sequence of vector generated by an optimization method converges toward the solution of a given problem.

Definition 0.4 *The* **Euclidean norm** $\|\cdot\|$ *on* $\mathbb{R}^n$ *is defined by*

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \quad \|\boldsymbol{x}\| := \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Remark 0.1 *This is indeed a norm, since it fulfills the four axioms that define what a norm is:*

1. $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,\ \|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$;
2. $\|\boldsymbol{x}\| = 0 \ \Leftrightarrow \boldsymbol{x} = \mathbf{0}_{\mathbb{R}^n}$;
3. $\forall \boldsymbol{x},\ \|\boldsymbol{x}\| \geq 0$;
4. $\forall \boldsymbol{x} \in \mathbb{R}^n, \forall \lambda \in \mathbb{R}, \|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$.

A vector $\boldsymbol{x} \in \mathbb{R}^n$ is called a unit vector if $\|\boldsymbol{x}\| = 1$.

Definition 0.5 *For any vectors* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, *the* **scalar product** *derived from the Euclidean norm is a function of* $\boldsymbol{x}$ *and* $\boldsymbol{y}$, *denoted by* $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$, *defined as follows:*

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} := \sum_{i=1}^{n} x_i y_i.$$

Two vectors $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *are called* orthogonal *if* $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = 0$.

Note that $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$, hence the scalar product defines a "product" between a row vector and a column vector.

Proposition 0.1 *Let* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *be two vectors in* $\mathbb{R}^n$. *Then, the following properties hold*

i) $\|\boldsymbol{x}+\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} + \|\boldsymbol{y}\|^2$;

ii) $\|\boldsymbol{x}-\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} + \|\boldsymbol{y}\|^2$;

iii) $\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 = \frac{1}{4}\left(\|\boldsymbol{x}+\boldsymbol{y}\|^2 + \|\boldsymbol{x}-\boldsymbol{y}\|^2\right)$;

iv) **Cauchy-Schwarz inequality** *:*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \qquad \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|.$$

Remark 0.2 *The last inequality is a key result in both linear algebra and analysis.*

0.2.2 Matrix linear algebra

We can define the product of two matrices that have compatible dimensions. More precisely, for any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times p}$, the product matrix $\boldsymbol{AB}$ is defined as the matrix $\boldsymbol{C} \in \mathbb{R}^{m\times p}$ such that

$$\forall i = 1, \dots, m,\ \forall j = 1, \dots, p, \quad \boldsymbol{C}_{ij} = \sum_{k=1}^{n} \boldsymbol{A}_{ik}\boldsymbol{B}_{kj}.$$

Using this definition, the product of a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with a (column) vector $\boldsymbol{x} \in \mathbb{R}^n$ is the vector $\boldsymbol{y} \in \mathbb{R}^m$ given by

$$\forall i = 1, \dots, m,\ y_i = \sum_{j=1}^{n} \boldsymbol{A}_{ij}x_j.$$

Remark 0.3 *Note that the scalar product on $\mathbb{R}^n$ corresponds to the matrix product for matrices of sizes $1 \times n$ and $n \times 1$: the result of this operation is a 1×1 matrix, that is, a scalar.*

When one work with matrices, the following linear subspaces are of interest.

Definition 0.6 (Fundamental subspaces) *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.*

- *The* null space *of $\boldsymbol{A}$ is the linear subspace*

$$\mathrm{Null}(\boldsymbol{A}) \ := \ \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{Ax} = \boldsymbol{0}_m\}$$

- *The* range space *of $\boldsymbol{A}$ is the linear subspace*

$$\mathrm{Range}(\boldsymbol{A}) \ := \ \{\boldsymbol{y} \in \mathbb{R}^m \mid \exists \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} = \boldsymbol{Ax}\}$$

The dimension of this linear subspace is called the **rank** *of $\boldsymbol{A}$. We denote it by $\mathrm{rank}(\boldsymbol{A})$. One always has $\mathrm{rank}(\boldsymbol{A}) \leq \min\{m, n\}$.*

Theorem 0.1 (Rank-nullity theorem) *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then,*

$$\dim(\mathrm{Null}(\boldsymbol{A})) + \mathrm{rank}(\boldsymbol{A}) = n.$$

Definition 0.7 (Matrix norms) *Consider the space $\mathbb{R}^{m\times n}$. The operator norm $\|\cdot\|$ and the Frobenius norm $\|\cdot\|_F$ are defined by*

$$\forall \boldsymbol{A} \in \mathbb{R}^{m\times n}, \begin{cases} \|\boldsymbol{A}\| & := \ \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{x}\neq\boldsymbol{0}_n}} \frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|} = \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \|\boldsymbol{x}\|=1}} \|\boldsymbol{Ax}\| \\ \|\boldsymbol{A}\|_F & := \ \sqrt{\sum_{\substack{1\leq i\leq m \\ 1\leq j\leq n}} \boldsymbol{A}_{ij}^2}. \end{cases}$$

Definition 0.8 (Symmetric matrix) *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is called symmetric if $\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{A}$.*

The set of symmetric matrices in $\mathbb{R}^{n\times n}$ is denoted by $\mathcal{S}^n$.

Definition 0.9 (Invertible matrix) *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is called invertible if there exists $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{BA} = \boldsymbol{AB} = \boldsymbol{I}_n$ (where we recall that $\boldsymbol{I}_n$ denotes the identity matrix in $\mathbb{R}^{n\times n}$).*

When it exists, such a matrix $\boldsymbol{B}$ is unique. It is then called **the inverse of $\boldsymbol{A}$** *and denoted by $\boldsymbol{A}^{-1}$.*

Definition 0.10 (Positive (semi)definite matrix) *A square, symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is called positive semidefinite if*

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} \geq 0,$$

which we write $\boldsymbol{A} \succeq 0$.

Such a matrix is called positive definite when $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} > 0$ for any nonzero vector $\boldsymbol{x}$. We write this as $\boldsymbol{A} \succ 0$.

Example 0.2 *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be symmetric and* diagonally dominant with nonnegative diagonal entries, *i.e. $\boldsymbol{A}_{ii} \geq 0$ for any $i \in \{1, \dots, n\}$ and $\boldsymbol{A}_{ii} \geq \sum_{j\neq i} \boldsymbol{A}_{ij}$. Then $\boldsymbol{A}$ is positive semidefinite. Moreover, if the diagonal entries are positive and $\boldsymbol{A}_{ii} > \sum_{j\neq i} \boldsymbol{A}_{ij}$, the matrix is positive definite.*

This example shows in particular that the identity matrix is positive definite.

Definition 0.11 (Orthogonal matrix) *A square matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is called orthogonal if $\boldsymbol{P}^{\mathrm{T}} = \boldsymbol{P}^{-1}$.*

More generally, a matrix $\boldsymbol{Q} \in \mathbb{R}^{m\times n}$, where $m \leq n$, is called orthogonal if $\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}} = \boldsymbol{I}_m$ (the columns of $\boldsymbol{Q}$ are orthonormal in $\mathbb{R}^m$).

When $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is orthogonal, then so is its transpose $\boldsymbol{Q}^{\mathrm{T}}$ (this result only applies to square matrices). Orthogonal matrices have the following desirable property.

Lemma 0.1 *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{U} \in \mathbb{R}^{m\times m}$, $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ be two orthogonal matrices. Then,*

$$\|\boldsymbol{A}\| = \|\boldsymbol{U}\boldsymbol{A}\| = \|\boldsymbol{A}\boldsymbol{V}\| \quad \text{and} \quad \|\boldsymbol{A}\|_F = \|\boldsymbol{U}\boldsymbol{A}\|_F = \|\boldsymbol{A}\boldsymbol{V}\|_F,$$

i.e. multiplying by an orthogonal matrix preserves the norm.

As a corollary of the previous lemma, we observe that an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{m\times n}$ with $m \leq n$ must satisfy $\|\boldsymbol{Q}\| = 1$ and $\|\boldsymbol{Q}\|_F = \sqrt{m}$.

Definition 0.12 (Eigenvalue) *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. A scalar $\lambda \in \mathbb{R}$ is called an* **eigenvalue of $\boldsymbol{A}$** *if*

$$\exists \boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{v} \neq \boldsymbol{0}_n, \quad \boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}.$$

The vector $\boldsymbol{v}$ is called an **eigenvector associated with the eigenvalue λ**. *The set of eigenvalues of $\boldsymbol{A}$ is the spectrum of $\boldsymbol{A}$.*

The span of eigenvectors associated to the same eigenvalue is called the eigenspace. Its dimension corresponds to the multiplicity of the eigenvalue relatively to the matrix.

Proposition 0.2 *For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the following holds:*

- *$\boldsymbol{A}$ has n complex eigenvalues.*
- *If $\boldsymbol{A}$ is symmetric positive semidefinite (resp. definite), then its eigenvalues are real nonnegative (resp. real positive).*
- *The null space of $\boldsymbol{A}$ is spanned by the eigenvectors associated with the 0 eigenvalue.*

Theorem 0.2 (Eigenvalue decomposition theorem) *Any symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has an* **eigenvalue decomposition** *of the form*

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T,$$

where $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$ is a diagonal matrix that contains the n eigenvalues of $\boldsymbol{A}$ $\lambda_1, \ldots, \lambda_n$ on its diagonal.

The eigenvalue decomposition is not unique, but the set of eigenvalues that appears in the decomposition is uniquely defined.

Remark 0.4 *There are matrices that possess an eigenvalue decomposition of the form $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$, where $\boldsymbol{P}$ is invertible (but not necessarily orthogonal). Those matrices are called diagonalizable.*

Link with singular value decomposition Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. In general, $m \neq n$ and the notion of eigenvalue that we introduced above does not apply. However, we can always consider the eigenvalues of

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} \in \mathbb{R}^{n\times n} \qquad \text{and} \qquad \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} \in \mathbb{R}^{m\times m}.$$

These matrices are real and symmetric, hence they can be diagonalized. This property is what gives rise to the singular value decomposition (or *SVD*).

Definition 0.13 (SVD) *For any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, there exist a decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$ called* singular value decomposition, *or SVD, satisfying the following properties:*

i) The matrix $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ is orthogonal, i.e. its columns form an orthonormal basis of $\mathbb{R}^m$, hence $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}_m$.

ii) The matrix $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ is orthogonal, i.e. $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}_n$.

iii) The matrix $\boldsymbol{S} \in \mathbb{R}^{m\times n}$ has all entries equal to zero except for the first $r \leq \min\{m, n\}$ entries on its diagonal $\{\boldsymbol{S}_{ii} | 1 \leq i \leq \min\{m, n\}\}$, that are positive. Without loss of generality, we assume that these values appear in decreasing order, that is, $\boldsymbol{S}_{11} \geq \cdots \geq \boldsymbol{S}_{rr} > 0$.

The values $\boldsymbol{S}_{11}, \ldots, \boldsymbol{S}_{rr}$ are called the singular values *or $\boldsymbol{A}$, and the value r is called the* rank *of $\boldsymbol{A}$.*

Remark 0.5 *Some definitions of the singular value decomposition allow for zero singular values, others directly shrink the size of the decomposition by keeping the first r columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. The latter decomposition, called* truncated SVD, *gives*

$$\boldsymbol{X} = \boldsymbol{U}_r\boldsymbol{S}_r\boldsymbol{V}_r^{\mathrm{T}},$$

where $\boldsymbol{U}_r \in \mathbb{R}^{m\times r}$ consists of the first r columns of $\boldsymbol{U}$, $\boldsymbol{V}_r \in \mathbb{R}^{n\times r}$ consists of the first r columns of $\boldsymbol{V}$, and $\boldsymbol{S}_r \in \mathbb{R}^{r\times r}$ is a diagonal matrix with diagonal coefficients $\boldsymbol{S}_{11} \geq \cdots \geq \boldsymbol{S}_{rr} > 0$.

The singular value decomposition is widely used in image and signal processing, as it allows to compute *approximations* of the matrix $\boldsymbol{A}$ by using the information corresponding to the largest singular values.

0.3 Sets and basic topology

Rather than introducing topological notions in a general fashion, we will focus on topologies that are defined through norms. Norms are a fundamental component of mathematical thinking, as they allow to quantify the distance between two objects. The various concepts from topology that will be used throughout the course involve a norm in some capacity, along with the concepts form set theory below.

Definition 0.14 *1. The empty set in $\mathbb{R}^n$ will be denoted by $\emptyset$.*

2. *For any sets $\mathcal{A}$ and $\mathcal{B}$ in $\mathbb{R}^n$, we write $\mathcal{A} \subseteq \mathcal{B}$ to indicate that $\mathcal{A}$ is contained in $\mathcal{B}$. A strict inclusion (i.e. $\mathcal{A} \subseteq \mathcal{B}$ but the two sets are not equal), we write $\mathcal{A} \subset \mathcal{B}$. In both cases, we say that $\mathcal{A}$ is a subset of $\mathcal{B}$.*

3. *For any (sub)sets $\mathcal{A}$ and $\mathcal{B}$ in $\mathbb{R}^n$, we let $\mathcal{A} \cap \mathcal{B}$ denote the intersection of those two sets and $\mathcal{A} \cup \mathcal{B}$ denote their union. These definitions apply recursively to define intersections and unions of several sets.*

4. *For any $\mathcal{S} \subseteq \mathbb{R}^n$, the complement of $\mathcal{S}$ in $\mathbb{R}^n$, denoted by $\mathbb{R}^n \setminus \mathcal{S}$, is defined as*

$$\{ \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} \notin \mathcal{S} \}.$$

   Similarly, for any $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathbb{R}^n$, we define the complement of $\mathcal{A}$ in $\mathcal{B}$ as the set of all elements of $\mathcal{B}$ that do not belong to $\mathcal{A}$, and denote it by $\mathcal{B} \setminus \mathcal{A}$.

5. *The Minkowski sum of two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ in $\mathbb{R}^n$ is defined by*

$$\mathcal{S}_1 + \mathcal{S}_2 := \{ \boldsymbol{x} + \boldsymbol{y} \mid \boldsymbol{x} \in \mathcal{S}_1, \boldsymbol{y} \in \mathcal{S}_2 \}.$$

6. *When the set $\mathcal{S}_1$ consists in one vector $\boldsymbol{x} \in \mathbb{R}^n$, we write $\boldsymbol{x} + \mathcal{S}_2$ for the sum $\{\boldsymbol{x}\} + \mathcal{S}_2$.*

Using a norm (in our case, the Euclidean norm) allows for defining key topological concepts in $\mathbb{R}^n$.

Definition 0.15 (Closed and open balls) *Let $\boldsymbol{c}$ be a vector in $\mathbb{R}^n$.*

- *The* **closed ball** *centered at $\boldsymbol{c}$ with radius $r > 0$ is defined by*

$$\mathcal{B}_r(\boldsymbol{c}) = \{ \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{c}\| \leq r \}.$$

- *The* **open ball** *centered at $\boldsymbol{c}$ with radius $r > 0$ is defined by*

$$\mathcal{B}_r^O(\boldsymbol{c}) = \{ \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{c}\| < r \}.$$

Definition 0.16

- *An open set in $\mathbb{R}^n$ is a set that contains an open ball centered around each of its points.*
- *A closed set in $\mathbb{R}^n$ is a set such that its complement is open.*

Example 0.3

- *An open ball is an open set.*

- *A closed ball is a closed set.*
- *A linear subspace of $\mathbb{R}^n$ is closed.*

Proposition 0.3
- *An intersection of closed sets is a closed set.*
- *A union of open sets is an open set.*
- *A finite union of closed sets is a closed set.*

Definition 0.17 *Let $\mathcal{S} \subseteq \mathbb{R}^n$.*

- *The interior of $\mathcal{S}$, denoted by $\mathrm{int}(\mathcal{S})$, is the largest open set included in $\mathcal{S}$.*
- *The closure $\mathcal{S}$, denoted by $\mathrm{cl}(\mathcal{S})$, is the smallest closed set containing $\mathcal{S}$.*
- *The boundary of $\mathcal{S}$ is defined by $\partial(\mathcal{S}) = \mathrm{cl}(\mathcal{S}) \setminus \mathrm{int}(\mathcal{S})$.*
- *A set $\mathcal{S}$ is called bounded if it is contained in a ball (the ball can be open or closed).*
- *A set $\mathcal{S}$ is called compact if it is both closed and bounded.*

Sequences and connection to topology Given an arbitrary set $\mathcal{S}$, we let $\mathcal{S}^{\mathbb{N}}$ denote the set of *sequences* of elements of $\mathcal{S}$, i.e. families of elements of $\mathcal{S}$ indexed by nonnegative integers. Sequences are particularly relevant for convergence of certain algorithmic procedures, such as sampling in statistics. They also serve to characterize various topological concepts. We describe below the results in the case of vector sequences, but point out that similar results can be obtained in the context of matrix spaces.

Definition 0.18 (Subsequence) *Let $\mathcal{S} \subseteq \mathbb{R}^n$. A* subsequence *of a sequence $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}} \in \mathcal{S}^{\mathbb{N}}$ is a subset of the elements of the sequence indexed by $\{\phi(k)\}_{k\in\mathbb{N}}$, where $\phi(k) \geq k$ and $\phi(k+1) > \phi(k)$ for every $k \in \mathbb{N}$.*

Definition 0.19 (Convergence of a sequence in $\mathbb{R}^n$) *A sequence $\{\boldsymbol{x}_k\}_k \in (\mathbb{R}^n)^{\mathbb{N}}$ converges towards $\boldsymbol{x}^* \in \mathbb{R}^n$ if*

$$\forall \epsilon > 0,\ \exists K \in \mathbb{N},\ \forall k \geq K, \quad \|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \epsilon.$$

The vector $\boldsymbol{x}^*$ is called the limit *of the sequence.*

The limit of a converging subsequence of $\{\boldsymbol{x}_k\}_k$ is called a limit point, *or an* accumulation point.

Proposition 0.4 *Let $\mathcal{S} \subset \mathbb{R}^n$.*

- *The set $\mathcal{S}$ is closed if any converging sequence in $\mathcal{S}$ converges towards a point in $\mathcal{S}$.*
- *The set $\mathcal{S}$ is compact if it is closed and bounded (it follows that every sequence in $\mathcal{S}$ has a limit point).*
- *The closure of $\mathcal{S}$ is the set of limit points of sequences in $\mathcal{S}$.*

Supremum and maximum Given a set $\mathcal{A} \subseteq \mathbb{R}$, we say that a is an upper bound on $\mathcal{A}$ if $\forall x \in \mathcal{A},\ x \leq a$. The smallest upper bound on $\mathcal{A}$ is called the *supremum of* $\mathcal{A}$, and denoted by $\sup \mathcal{A}$ (by convention, $\sup \emptyset = -\infty$ and $\sup \mathcal{A} = \infty$ for any $\mathcal{A}$ without a finite upper bound). When $\sup \mathcal{A} \in \mathcal{A}$, the supremum is reached by an element in $\mathcal{A}$, in which case it is called a maximum of $\mathcal{A}$ and denoted by $\max \mathcal{A}$.

Similarly, we say that a is a *lower bound* on $\mathcal{A}$ if $\forall x \in \mathcal{A},\ x \geq a$. The largest lower bound is called the *infimum of* $\mathcal{A}$ and denoted by $\inf \mathcal{A}$ (by convention, $\inf \emptyset = \infty$ and $\inf \mathcal{A} = -\infty$ if $\mathcal{A}$ does not have a finite lower bound). When $\inf \mathcal{A} \in \mathcal{A}$, the infimum is reached by an element of $\mathcal{A}$, in which case it is called a minimum of $\mathcal{A}$ and denoted by $\min \mathcal{A}$.

0.4 Calculus

Definition 0.20 (Continuity) *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called* **continuous in $x \in \mathbb{R}^n$** *if*

$$\forall \epsilon > 0,\ \exists \delta > 0,\ \forall \boldsymbol{y} \in \mathbb{R}^n, \quad \|\boldsymbol{y} - \boldsymbol{x}\| < \delta \quad \Rightarrow \quad \|f(\boldsymbol{y}) - f(\boldsymbol{x})\| < \epsilon.$$

The function f is continuous on a set $\mathcal{A} \subseteq \mathbb{R}^n$ if it is continuous at every point of $\mathcal{A}$. When $\mathcal{A} = \mathbb{R}^n$, we simply say that f is continuous.

Remark 0.6 *In certain textbooks, the notion above is termed* uniform continuity*. For simplicity of exposure, we will use it as our definition of continuity.*

An alternate characterization of continuity based on sequences is given below. Sequences typically appear when considering iterative algorithms, hence the relevance of this notion here.

Definition 0.21 (Continuity (sequential definition)) *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $\boldsymbol{x} \in \mathbb{R}^n$ if*

$$\forall \{\boldsymbol{x}_n\} \in (\mathbb{R}^n)^{\mathbb{N}},\ \{\boldsymbol{x}_n\} \to \boldsymbol{x}, \quad \lim_{n \to \infty} f(\boldsymbol{x}_n) = f(\boldsymbol{x}).$$

Example 0.4 *A linear map $f : \mathbb{R}^n \to \mathbb{R}^m$, where $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ for any $\boldsymbol{x} \in \mathbb{R}^n$ with $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$, is a continuous function on $\mathbb{R}^n$.*

Definition 0.22 (Differentiability Jacobian matrix) *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called* **differentiable** *at a point $\boldsymbol{x} \in \mathbb{R}^n$ if there exists a matrix $\boldsymbol{J}_f(\boldsymbol{x}) \in \mathbb{R}^{m \times n}$ such that*

$$\lim_{\substack{\boldsymbol{z} \to \boldsymbol{x} \\ \boldsymbol{z} \neq \boldsymbol{x}}} \frac{\|f(\boldsymbol{z}) - f(\boldsymbol{x}) - \boldsymbol{J}_f(\boldsymbol{x})(\boldsymbol{z} - \boldsymbol{x})\|}{\|\boldsymbol{z} - \boldsymbol{x}\|} = 0.$$

- *$\boldsymbol{J}_f(\boldsymbol{x})$ is called the* **Jacobian** *of f at $\boldsymbol{x}$, and is uniquely defined.*
- *If $f(\cdot) = [f_1(\cdot), \ldots, f_m(\cdot)]^{\mathrm{T}}$, then*

$$\forall 1 \leq i \leq m,\ \forall 1 \leq j \leq n,\ [\boldsymbol{J}_f(\boldsymbol{x})]_{ij} = \frac{\partial f_i}{\partial x_j}(\boldsymbol{x}).$$

The following special cases are instrumental to optimization and basic analysis.

Corollary 0.1 • *When $m = 1$, we define the (column) vector $\nabla f(\boldsymbol{x}) \equiv \boldsymbol{J}_f(\boldsymbol{x})^{\mathrm{T}}$, called the* **gradient** *of f at $\boldsymbol{x}$. In this case, the gradient is the vector of partial derivatives of f:*

$$\forall i = 1, \ldots, n, \qquad \nabla f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_i}(\boldsymbol{x})\right]_{1 \leq i \leq n}.$$

- *When $n = m = 1$, both the Jacobian and the gradient are equivalent to a scalar $f'(x) \equiv \nabla f(x) \equiv \boldsymbol{J}_f(\boldsymbol{x})^{\mathrm{T}}$, called the derivative of f at $\boldsymbol{x}$.*

Remark 0.7 *When needed, we may consider f to be defined on a product space, in which case we will specify the variables on which we compute the gradient. For instance, given $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ with $f : (\boldsymbol{x}, \boldsymbol{y}) \mapsto f(\boldsymbol{x}, \boldsymbol{y})$, the notation $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})$ will denote the gradient of f with respect to the first n variables (it will thus be a vector in $\mathbb{R}^n$, computed at $(\boldsymbol{x}, \boldsymbol{y})$).*

In these notes, we assume familiarity with the common derivative formulas for functions from $\mathbb{R}$ to $\mathbb{R}$. More complex formulas are typically obtained thanks to the rule below.

Theorem 0.3 (Chain rule) *If $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ and $g : \mathbb{R}^m \mapsto \mathbb{R}^p$ are both differentiable, respectively on $\mathbb{R}^n$ and $\mathbb{R}^m$, then $h = g \circ f : \mathbb{R}^n \mapsto \mathbb{R}^p$ is differentiable on $\mathbb{R}^n$ and*

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{J}_h(\boldsymbol{x}) = \boldsymbol{J}_g(f(\boldsymbol{x}))\boldsymbol{J}_f(\boldsymbol{x}).$$

Remark 0.8 *Special cases of the chain rule:*

- $m = p = 1$ *:* $\nabla h(\boldsymbol{x}) = g'(f(\boldsymbol{x}))\nabla f(\boldsymbol{x})$*;*
- $n = m = p = 1$ *:* $h'(x) = g'(f(x))f'(x)$*.*

Theorem 0.4 (Mean-value theorem in dimension 1) *Let $f : [a, b] \to \mathbb{R}$. If f is continuous on $[a, b]$ and differentiable on (a, b), there exists $c \in (a, b)$ such that*

$$\frac{f(b) - f(a)}{b - a} = f'(c).$$

Definition 0.23 (Taylor expansion) *Let $f : [a, b] \mapsto \mathbb{R}$ be $\mathcal{C}^1$ on $[a, b]$, then*

$$\begin{aligned} f(b) &= f(a) + f'(c)(b - a) \qquad \text{where } c \in [a, b] \\ f(b) &= f(a) + \int_0^1 f'(a + t(b - a))(b - a)\, dt. \end{aligned}$$

Theorem 0.5 (Mean-value theorem in dimension d) *Let $f : \mathbb{R}^d \to \mathbb{R}$ $f \in \mathcal{C}^1(\mathbb{R}^d)$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\boldsymbol{x} \neq \boldsymbol{y}$, there exists $t \in (0, 1)$ such that*

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}).$$

Definition 0.24 (Function classes) *Let $f : \mathbb{R}^d \to \mathbb{R}$. We say that f is $\mathcal{C}^p(\mathbb{R}^d)$ (or simply $\mathcal{C}^p$) if it is differentiable p times with a continuous pth-order derivative (in which case all derivatives up to order p are continuous). The class of $\mathcal{C}^\infty$ functions is the intersection of all $\mathcal{C}^p$ with $p \in \mathbb{N}$.*

Theorem 0.6 (Taylor expansion of order 1) *Let $f \in \mathcal{C}^1(\mathbb{R}^d)$. For any vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of $\mathbb{R}^d$, we have*

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \int_0^1 \nabla f(\boldsymbol{x} + t\,(\boldsymbol{y} - \boldsymbol{x}))^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x})\, dt.$$

Chapter 1

Convexity

Convexity plays a major role in continuous optimization, as an indicator on the difficulty to solve a given problem. In computational mathematics, one often considers a gap between linear and nonlinear problems (for instance, a system of linear equations is usually easier to solve than a system of nonlinear equations). In optimization, this divide evolved during the second half of the 20th century, so that the key gap in problem solving is now between convex and nonconvex problems (this may still change in the future!).

1.1 Convex sets

Convexity is by essence a geometrical notion, that applies to sets rather than functions. In this section, we thus introduce convexity for subsets of $\mathbb{R}^n$.

1.1.1 Affine sets

As a warmup to convex sets, we first define a closely related notion from standard linear algebra.

Definition 1.1 (Affine set) *A set $\mathcal{X} \subseteq \mathbb{R}^n$ is called* **affine** *if*

$$\forall (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}^2, \ \forall \alpha \in \mathbb{R}, \quad \alpha\, \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2 \in \mathcal{X}. \tag{1.1.1}$$

Equivalently, the set $\mathcal{X}$ is affine if it can be written as $\mathcal{X} = \boldsymbol{x} + \mathcal{S}$, where $\mathcal{S}$ is a linear subspace of $\mathbb{R}^n$. This linear subspace is called the parallel subspace to $\mathcal{X}$, and the dimension *of $\mathcal{X}$ is defined as the dimension of its parallel subspace.*

Remark 1.1 *Any affine set contains every line passing by two of its points.*

Remark 1.2 *The second characterization of an affine set shows that such a set is implicitly defined by a subspace. In the literature, an affine set is sometimes called an affine subspace. However, we will adopt the* affine set *terminology, commonly used in optimization, and save the use of* subspace *to refer to linear subspaces.*

Because of their connections with subspaces, affine sets have the following desirable property.

Proposition 1.1 *Every affine set is closed.*

The property stated in Proposition 1.1 means that if a sequence of elements in an affine set converges, then it converges within the affine set.

Definition 1.2 (Affine combination) *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ be k vectors in $\mathbb{R}^n$. A vector $\boldsymbol{x} \in \mathbb{R}^n$ is an* affine combination *of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ if there exist k real values $\alpha_1, \ldots, \alpha_k$ satisfying*

$$\sum_{i=1}^{k} \alpha_i = 1 \quad \textit{and} \quad \boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i.$$

For any set of vectors, the previous definition allows to build an affine set as follows.

Definition 1.3 (Affine hull) *The* affine hull *of a set $\mathcal{X}$, denoted by $\mathrm{aff}(\mathcal{X})$, is the set of affine combinations of points in $\mathcal{X}$.*

The affine hull can be equivalently defined as the intersection of all affine sets containing $\mathcal{X}$, i.e. as the smallest affine set containing $\mathcal{X}$.

1.1.2 Convex sets and related properties

We now provide our first definition of convexity.

Definition 1.4 (Convex set) *A set $\mathcal{C} \subseteq \mathbb{R}^n$ is* **convex** *if*

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{C}^2, \ \forall \alpha \in [0,1], \quad \alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in \mathcal{C}. \tag{1.1.2}$$

A set is therefore convex if it contains any line segment connecting two of its points.

By convention, the empty set is considered convex. However, to avoid ambiguous definitions later on, we may restrain our study to non-empty convex sets.

Proposition 1.2 *i) The intersection of a family of convex sets is convex.*

ii) The sum of two convex sets is convex.

iii) For any $\lambda \in \mathbb{R}$ and any convex set $\mathcal{C} \subset \mathbb{R}^n$, the set $\lambda\mathcal{C}$ is convex.

Definition 1.5 (Convex combination) *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ be k vectors of $\mathbb{R}^n$. A vector $\boldsymbol{x} \in \mathbb{R}^n$ is a* convex combination *of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ if there exist k nonnegative real values $\alpha_1, \ldots, \alpha_k$ such that*

$$\sum_{i=1}^{k} \alpha_i = 1 \quad \textit{and} \quad \boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i.$$

Note that the concept of convex combination is a restriction of that of affine combination, since the coefficients are required to be nonnegative.

The counterpart to affine hull for convexity is given below.

Definition 1.6 (Convex hull) *The* convex hull *of a set $\mathcal{X} \subseteq \mathbb{R}^n$, denoted by $\mathrm{conv}(\mathcal{X})$, is the set of all convex combinations of points in $\mathcal{X}$.*

Convex hulls and affine hulls allow for defining related topological notions.

Definition 1.7 (Relative interior) *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a nonempty convex set. The relative interior of $\mathcal{C}$, denoted by $\mathrm{ri}(\mathcal{C})$, is the set of points $\boldsymbol{x} \in \mathcal{C}$ for which there exists $\epsilon > 0$ satisfying $\mathcal{B}_\epsilon^O(\boldsymbol{x}) \cap \mathrm{aff}(\mathcal{C}) \subseteq \mathcal{C}$.*

We say that $\mathcal{C}$ is relatively open when $\mathrm{ri}(\mathcal{C}) = \mathcal{C}$, and we will define the relative boundary of $\mathcal{C}$ as the set $\mathrm{cl}(\mathcal{C}) \setminus \mathrm{ri}(\mathcal{C})$.

Proposition 1.3 *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a nonempty convex set. The following properties hold:*

i) The set $\mathrm{cl}(\mathcal{C})$ is convex and nonempty.

ii) The set $\mathrm{int}(\mathcal{C})$ is convex.

iii) The set $\mathrm{ri}(\mathcal{C})$ is convex and nonempty.

1.1.3 Cones

Cones (and, in particular, convex cones) are mathematical objects that are instrumental in formulating optimality conditions for (convex) optimization problems, as seen in the next chapter.

Definition 1.8 (Cone) *A set $\mathcal{K} \subseteq \mathbb{R}^n$ is called a* **cone** *if*

$$\forall \boldsymbol{x} \in \mathcal{K}, \forall t > 0, \quad t\boldsymbol{x} \in \mathcal{K}.$$

Remark 1.3 *Several authors define a cone using any $t \geq 0$, implying that any (nonempty) cone of $\mathbb{R}^n$ must contain the zero vector in $\mathbb{R}^n$. Other authors follow Definition 1.8, and define pointed cones as cones that contain the zero vector. We follow the latter approach, and stress out that cones obtained via conic combinations will always be pointed.*

Classical examples of cones include the empty set $\emptyset$, the whole space $\mathbb{R}^n$, as well as any half-line of the form $\{t\boldsymbol{x} | t > 0\}$ for some $\boldsymbol{x} \in \mathbb{R}^n$. Any linear subspace of $\mathbb{R}^n$ is also a cone.

Per the definition, a cone does not necessarily contain the origin and need not be convex. The notion of convex cone must thus be defined in a separate fashion.

Definition 1.9 (Convex cone) *A set $\mathcal{K} \subseteq \mathbb{R}^n$ is called a* **convex cone** *if it is both a cone and a convex set.*

Example 1.1 *The set $\{\boldsymbol{x} \in \mathbb{R}^3 \mid x_1^2 + x_2^2 \leq x_3\}$ is a convex cone in $\mathbb{R}^3$.*

Any cone is associated with two other cones as follows.

Definition 1.10 *Let $\mathcal{K}$ be a cone in $\mathbb{R}^n$. The dual cone of $\mathcal{K}$, denoted by $\mathcal{K}^*$, is given by*

$$\mathcal{K}^* = \left\{ \boldsymbol{v} \in \mathbb{R}^n \middle| \boldsymbol{v}^{\mathrm{T}}\boldsymbol{x} \geq 0 \ \forall \boldsymbol{x} \in \mathcal{K} \right\}.$$

The set $\mathcal{K}^*$ is a closed convex cone.

Definition 1.11 *Let $\mathcal{K}$ be a cone in $\mathbb{R}^n$. The polar cone of $\mathcal{K}$, denoted by $\mathcal{K}^\circ$, is given by*

$$\mathcal{K}^\circ = \left\{ \boldsymbol{v} \in \mathbb{R}^n \middle| \boldsymbol{v}^{\mathrm{T}}\boldsymbol{x} \leq 0 \ \forall \boldsymbol{x} \in \mathcal{K} \right\}.$$

The set $\mathcal{K}^\circ$ is a closed convex cone.

An important result (due to the French mathematician Jean-Jacques Moreau) generalizes the orthogonal decomposition in linear subspaces using convex cones.

Theorem 1.1 (Moreau decomposition) *Let $\mathcal{K}$ be a closed convex cone in $\mathbb{R}^n$. For any $\boldsymbol{x} \in \mathbb{R}^n$, there exists $\boldsymbol{u} \in \mathcal{K}$ $\boldsymbol{v} \in \mathcal{K}^\circ$ such that*

$$\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v} \qquad \text{and} \qquad \boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} = 0.$$

To end this section, we provide below counterpart definitions to that of convex (resp. affine) combinations and convex (resp. affine) hull.

Definition 1.12 (Conic combination) *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ be k vectors in $\mathbb{R}^n$. A vector $\boldsymbol{x} \in \mathbb{R}^n$ is a conic combination of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ if there exist k* nonnegative *real values $\alpha_1, \ldots, \alpha_k$ such that*

$$\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i.$$

Conic combinations are more general that convex combinations, in that the coefficients of a conic combination are not required to sum at 1.

Definition 1.13 (Conic hull) *Let $\mathcal{X} \subset \mathbb{R}^n$ be a nonempty set. The cone spanned by $\mathcal{X}$, also known as the conic hull of $\mathcal{X}$ and denoted by $\mathrm{cone}(\mathcal{X})$, is the set of all conic combinations of points in $\mathcal{X}$.*

Remark 1.4 *Unlike arbitrary cones, a conic hull always contains the origin (hence it is pointed), and is a convex cone.*

1.1.4 Examples of convex sets

In this section, we provide several key examples of convex sets, that will be used in the rest of the lecture notes.

Proposition 1.4 *Let $\boldsymbol{x} \in \mathbb{R}^n$ and $r > 0$. The closed ball $\mathcal{B}_r(\boldsymbol{x})$ as defined in Definition 0.15 is a closed convex set.*

The notion of ball (and its convex nature) can be generalized to that of ellipsoid.

Definition 1.14 (Ellipsoid) *Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix and $\boldsymbol{c} \in \mathbb{R}^n$. The set*

$$\mathcal{E} = \{\boldsymbol{x} \in \mathbb{R}^n | (\boldsymbol{x} - \boldsymbol{c})^{\mathrm{T}} \boldsymbol{A}^{-1} (\boldsymbol{x} - \boldsymbol{c}) \leq 1\}$$

is called an ellipsoid.

Proposition 1.5 *Any ellipsoid in $\mathbb{R}^n$ is a compact convex set.*

Definition 1.15 *The second-order cone (also called Lorentz cone or ice-cream cone) associated with the Euclidean norm in $\mathbb{R}^n$ is the subset of $\mathbb{R}^{n+1}$ given by*

$$\{(\boldsymbol{x}, t) \mid \|\boldsymbol{x}\| \leq t\}.$$

The second-order cone is a convex cone in $\mathbb{R}^{n+1}$.

To end this list of examples, we provide an example of cone in $\mathcal{S}^n$, i.e. the set of symmetric matrices in $\mathbb{R}^{n\times n}$.

Proposition 1.6 *The set of symmetric positive semidefinite matrices*

$$\mathcal{S}^n_+ := \{\boldsymbol{X} \in \mathcal{S}^n | \boldsymbol{X} \succeq 0\},$$

and the set of symmetric positive definite matrices

$$\mathcal{S}^n_{++} := \{\boldsymbol{X} \in \mathcal{S}^n | \boldsymbol{X} \succ 0\}$$

are convex cones in $\mathbb{R}^{n\times n}$.

Hyperplanes and half-spaces We now relate convexity and linear (in)equalities, thanks to a few additional definitions.

Definition 1.16 (Hyperplane) *A set $\mathcal{H} \subseteq \mathbb{R}^n$ is a hyperplane if it can be written as*

$$\mathcal{H} = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = b\},$$

where $\boldsymbol{a} \in \mathbb{R}^n$ is a nonzero vector and $b \in \mathbb{R}$.

Theorem 1.2 (Separating hyperplane) *Let $\mathcal{X}$ and $\mathcal{Y}$ by two disjoint convex sets in $\mathbb{R}^n$. There exist $\boldsymbol{a} \neq \boldsymbol{0}$ and $b \in \mathbb{R}$ such that*

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \leq b \ \forall \boldsymbol{x} \in \mathcal{X} \quad \text{and} \quad \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \geq b \ \forall \boldsymbol{x} \in \mathcal{Y}.$$

The hyperplane $\{\boldsymbol{x} | \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = b\}$ is called a separating hyperplane *for $\mathcal{X}$ and $\mathcal{Y}$.*

Definition 1.17 (Vertically tangent hyperplane) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $\boldsymbol{x}_0 \in \partial\mathcal{X}$. A* vertically tangent hyperplane to $\mathcal{X}$ at $\boldsymbol{x}_0$ *is a set $\{\boldsymbol{x} | \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_0\}$, where $\boldsymbol{a} \in \mathbb{R}^n$ is a nonzero vector such that $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \leq \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_0$ for any $\boldsymbol{x} \in \mathcal{X}$.*

The next result will be used to formulate optimal conditions in the next chapter.

Theorem 1.3 *For any nonempty convex set $\mathcal{C} \subseteq \mathbb{R}^n$ and any $\boldsymbol{x} \in \partial\mathcal{C}$, there exists a vertically tangent hyperplane to $\mathcal{C}$ at $\boldsymbol{x}_0$.*

A hyperplane is an affine set (and even a linear subspace when $b = 0$) that "separates" $\mathbb{R}^n$ in two parts, each associated with a half-space in the following sense.

Definition 1.18 (Half-space) *A set $\mathcal{H} \subseteq \mathbb{R}^n$ is called a half-space if*

$$\mathcal{H} = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \leq b\},$$

where $\boldsymbol{a} \in \mathbb{R}^n$ is a nonzero vector and $b \in \mathbb{R}$.

Unlike hyperplanes, half-spaces are not affine sets. However, they satisfy another desirable property in certain cases.

Proposition 1.7 *A half-space of $\mathbb{R}^n$ of the form $\{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \leq 0\}$ is a convex cone.*

One can consider intersections of half-spaces, giving rise to the following definition.

Definition 1.19 (Polyhedral sets) *A set $\mathcal{P} \subseteq \mathbb{R}^n$ is called a* polyhedron *if*

$$\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} \leq b_i \text{ pour } i = 1, \ldots, m\},$$

where $m \geq 1$, $\{\boldsymbol{a}_i\}_{i=1}^m$ is a set of nonzero vectors in $\mathbb{R}^n$ and $\{b_i\}_{i=1}^m$ is a set of real values.
When $b_i = 0$ $\forall i = 1, \ldots, m$, the set $\mathcal{P}$ is a cone called a polyhedral cone.

The next result gives an example of polyhedral cone and justifies that linear subspaces have nice structure, in that they can be characterized by linear inequalities.

Proposition 1.8 *Every subspace is a polyhedral cone.*

Another example of polyhedral set is given below.

Definition 1.20 (Simplex) *Let $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_k$ be $k+1$ vectors in $\mathbb{R}^n$ such that the vectors $\boldsymbol{x}_1 - \boldsymbol{x}_0, \ldots, \boldsymbol{x}_k - \boldsymbol{x}_0$ are linearly independent (we say then that the $\boldsymbol{x}_i$s are affinely indepedent). The convex hull of $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_k\}$, given by*

$$\left\{ \alpha_0\boldsymbol{x}_0 + \cdots + \alpha_k\boldsymbol{x}_k \middle| \sum_{i=0}^{k} \alpha_i = 1,\ \alpha_i \geq 0 \forall i = 0, \ldots, k \right\},$$

is called a **simplex** *in dimension k.*

Simplices are used in particular to model discrete probability distributions.

1.2 Convex functions

1.2.1 Definitions and first properties

Definition 1.21 (Convex function) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. A function $f : \mathcal{X} \to \mathbb{R}$ is called a* **convex function** *if*

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X},\ \forall \alpha \in [0,1], \quad f(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \leq \alpha f(\boldsymbol{x}_1) + (1-\alpha)f(\boldsymbol{x}_2). \tag{1.2.1}$$

We say that f is strictly convex if the inequality (1.2.1) is strict for any $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ and $\alpha \in (0,1)$.

A function is convex whenever it lies "under" any line segment that connects two of its values (see Figure 1.1. When it lies "above" any line segment, as in Figure 1.2, we obtain the companion notion of concave function.

Definition 1.22 (Concave function) *Let $\mathcal{X} \subseteq \mathbb{R}^n$. A function $f : \mathcal{X} \to \mathbb{R}$ is called* **concave** *(resp. strictly concave) when its negative $-f$ is a convex (resp. strictly convex).*

The inequality (1.2.1) is a special case of a result of an inequality named after the Danish mathematican Johan Jensen.

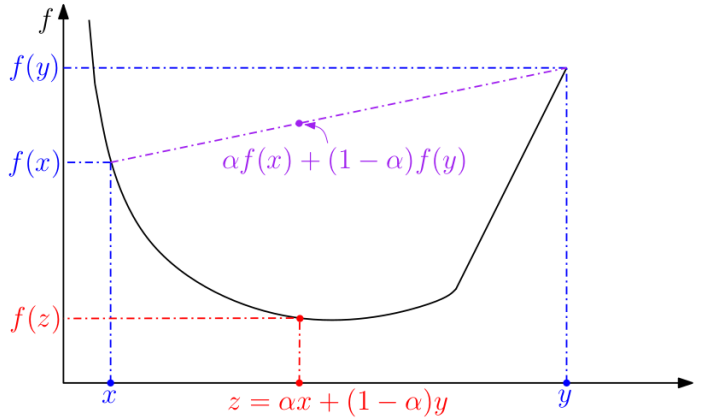

Figure 1.1: A convex function. *Source: D. P. Robinson [4].*

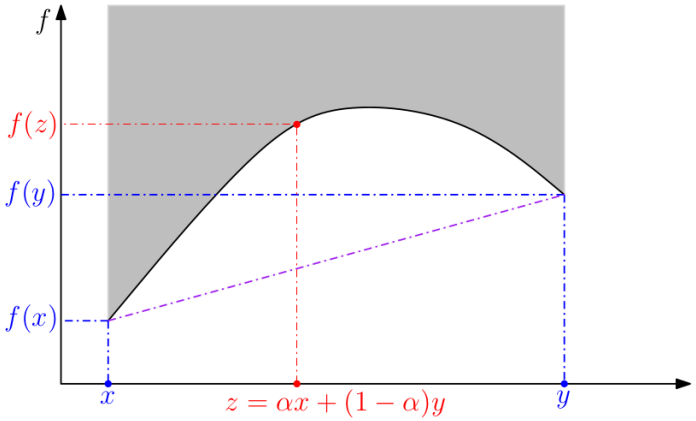

Figure 1.2: A concave function. *Source: D. P. Robinson [4].*

Theorem 1.4 (Jensen's inequality) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function on $\mathcal{X}$. Then, for any $x_1, \ldots, x_k$ in $\mathcal{X}$ and $\alpha_1, \ldots, \alpha_k$ in $\mathbb{R}_+$ such that $\sum_{i=1}^k \alpha_i = 1$, we have:*

$$f\left(\sum_{i=1}^k \alpha_i \boldsymbol{x}_i\right) \leq \sum_{i=1}^k \alpha_i f(\boldsymbol{x}_i).$$

Proposition 1.9 *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a convex set. A function $f : \mathcal{C} \to \mathbb{R}$ that is both convex and concave on $\mathcal{C}$ is called an affine function on $\mathcal{C}$. It then exists $\boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that $\forall \boldsymbol{x} \in \mathcal{C},\ f(\boldsymbol{x}) = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{x} + b$.*

Remark 1.5 *A function that is not convex need not be concave! For instance, $x \mapsto \sin(x)$ is neither convex nor concave on $\mathbb{R}$.*

Example 1.2 (Convex functions in one variable)

1. *The function $x \mapsto x^a$ is convex on $\mathbb{R}_{++}$ when $a \notin (0, 1)$ and concave when $a \in [0, 1]$.*

2. *The function $x \mapsto -\ln(x)$ (defined on $\mathbb{R}_{++}$) is convex on $\mathbb{R}_{++}$, and thus $x \mapsto \ln(x)$ is concave on $\mathbb{R}_{++}$.*
3. *The function $x \mapsto |x|$ is convex on $\mathbb{R}$.*
4. *The negative entropy function $x \mapsto x\ln(x)$ is convex on $\mathbb{R}_{++}$.*

Example 1.3 (Convex functions in dimension n)

1. *If f defines a norm on $\mathbb{R}^n$, then it is a convex function since we automatically have $f(\alpha\boldsymbol{x}) = \alpha f(\boldsymbol{x})$ and $f(\boldsymbol{x}+\boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ by properties of a norm.*
2. *A quadratic function of the form*
$$\boldsymbol{x} \mapsto \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}$$
where $\boldsymbol{b} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathcal{S}_+^n$ is postive semidefinite is a convex function.
3. *The function $\boldsymbol{x} \mapsto \ln(e^{x_1} + \cdots + e^{x_n})$ is convex.*

Remark 1.6 *Note that if a function from $\mathbb{R}^n$ to $\mathbb{R}$ is convex, then it is convex with respect to all of its variables, i.e. for fixed $i \in \{1,\ldots,n\}$ and fixed $x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n$ in $\mathbb{R}$, the function $x_i \mapsto f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)$ is convex. However, the converse is not true. For instance, the function $(x_1,x_2) \mapsto x_1^2 - 3x_1x_2 + x_2^2$ is convex in x_1 and x_2 but not jointly convex in both variables (the function violates (1.2.1) by taking $\boldsymbol{x} = [1\ 1]^{\mathrm{T}}$, $\boldsymbol{y} = [-1\ -1]^{\mathrm{T}}$ and $\alpha = 0.5$).*

Convexity is a property that is preserved through several operations, some of which are given below. For simplicity, we present the results assuming that those functions are defined on $\mathbb{R}^n$.

Proposition 1.10 1. *Let $f_1,\ldots,f_m$ be m convex functions from $\mathbb{R}^n$ to $\mathbb{R}$ and $\alpha_1,\ldots,\alpha_m$ are m nonnegative real values. Then the function $\sum_{i=1}^m \alpha_i f_i$ is convex (i.e. a conic combination of convex functions is convex).*
2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ defined by $g(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ with $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{b} \in \mathbb{R}^n$. If f is convex (resp. concave), then g is convex (resp. concave).*
3. *Let f_1 and f_2 be two convex functions from $\mathbb{R}^n$ to $\mathbb{R}$. Then the function $f : \boldsymbol{x} \to \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}$ is convex. More generally, if $f_1,\ldots,f_m$ are m convex functions from $\mathbb{R}^n$ to $\mathbb{R}$. The function $\boldsymbol{x} \to \max\{f_1(\boldsymbol{x}),\ldots,f_m(\boldsymbol{x})\}$ is also convex.*
4. *Let $\{f_i\}_{i\in\mathcal{I}}$ be a finite family of convex functions from $\mathbb{R}^n$ to $\mathbb{R}$, then*
$$f : \boldsymbol{x} \mapsto \sup_{i\in\mathcal{I}} f_i(\boldsymbol{x})$$
is a convex function. [1]
5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $g : \mathbb{R} \to \mathbb{R}$ be a nondecreasing, convex function. Then the function $h : \mathbb{R}^n \to \mathbb{R}$ defined by $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$ is convex.*
6. *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be m convex functions and $g : \mathbb{R}^m \to \mathbb{R}$ be a convex function nondecreasing with respect to each of its variables. Then, the function $h : \mathbb{R}^n \to \mathbb{R}$ defined by $h(\boldsymbol{x}) = g(f_1(\boldsymbol{x}),\ldots,f_m(\boldsymbol{x}))$ is convex.*

[1] This result generalizes to an arbitrary family of functions, thanks to the theory developed in Section 1.1.4.

1.2.2 Extended-value functions and convexity

Up to this point, we only considered real-valued functions, i.e. functions that always output a real number. Nevertheless, the theory of convex functions has long been adapted to accommodate for infinite values, and accounting for those values even allows for studying a broader class of problems.

Consider for instance a family of functions $\{f_i\}_{i\in\mathcal{I}}$ where $\mathcal{I}$ a given index set and $f_i : \mathbb{R}^n \to \mathbb{R}$ for any $i \in \mathcal{I}$. The supremum function $\boldsymbol{x} \mapsto \sup_{i\in\mathcal{I}} f_i(\boldsymbol{x})$ can take the value ∞[2]. Allowing infinite values enables us to analyze this supremum function.

We thus define as extended value functions as functions that can take any real value along with $-\infty$ and ∞. This new setup is however not compatible with Definition 1.21, since it introduces an ambiguity when considering two vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ such that $f(\boldsymbol{x}_1) = \infty$ and $f(\boldsymbol{x}_2) = -\infty$. For such functions, the concept of convexity must be redefined. To this end, we introduce several notions below.

Definition 1.23 *Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $f : \mathcal{X} \to \overline{\mathbb{R}}$, where we recall that $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Then, the effective domain of f, denoted by $\mathrm{dom}(f)$, is the subset of $\mathcal{X}$ of all points at which f does not take the value ∞, i.e.*

$$\mathrm{dom}(f) := \{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) < \infty\}.$$

This definition allows in particular to define the infimum and supremum of an extended value function. These concepts are instrumental in studying convex optimization problems.

Definition 1.24 (Infimum and supremum) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $f : \mathcal{X} \to \overline{\mathbb{R}}$. The* **infimum** *of f with respect to $\mathcal{X}$, denoted by $\inf_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ or $\inf f(\mathcal{X})$, is defined as*

$$\inf f(\mathcal{X}) := \begin{cases} \infty & \text{if } \mathcal{X} = \emptyset, \\ \inf_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) & \text{if } \mathcal{X} \neq \emptyset \text{ and } f(\mathcal{X}) \subset \mathbb{R}, \\ \infty & \text{if } \mathcal{X} \neq \emptyset \text{ and } \forall \boldsymbol{x} \in \mathcal{X},\ f(\boldsymbol{x}) = \infty, \\ -\infty & \text{if } \exists \boldsymbol{x} \in \mathcal{X},\ f(\boldsymbol{x}) = -\infty. \end{cases} \tag{1.2.2}$$

The **supremum** *of f with respect to $\mathcal{X}$, denoted by $\sup_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ or $\sup f(\mathcal{X})$, is defined as*

$$\sup f(\mathcal{X}) := \begin{cases} -\infty & \text{if } \mathcal{X} = \emptyset, \\ \sup_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) & \text{if } \mathcal{X} \neq \emptyset \text{ and } f(\mathcal{X}) \subset \mathbb{R}, \\ -\infty & \text{if } \mathcal{X} \neq \emptyset \text{ and } \forall \boldsymbol{x} \in \mathcal{X},\ f(\boldsymbol{x}) = -\infty, \\ \infty & \text{if } \exists \boldsymbol{x} \in \mathcal{X},\ f(\boldsymbol{x}) = \infty. \end{cases} \tag{1.2.3}$$

Another key concept for studying extended value functions is that of epigraph.

Definition 1.25 (Epigraph) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $f : \mathcal{X} \to \overline{\mathbb{R}}$. The* **epigraph** *of f $\mathrm{epi}(f)$ is the set of all vectors in $\mathbb{R}^{n+1}$ that lie "above" the graph of f:*

$$\mathrm{epi}(f) := \left\{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} \;\middle|\; \boldsymbol{x} \in \mathcal{X} \text{ and } y \geq f(\boldsymbol{x})\right\}.$$

Using epigraphs, one can generalize convexity (and concavity) to extended value functions.

Definition 1.26 (Convex extended value function) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. A function $f : \mathcal{X} \to \overline{\mathbb{R}}$ is called convex if its epigraph $\mathrm{epi}(f)$ is a convex set.*

[2] Unless there is a need to emphasize this, we will always write ∞ for $+\infty$.

Definition 1.27 (Concave extended value function) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. A function $f : \mathcal{X} \to \overline{\mathbb{R}}$ is called* **concave** *if its negative $-f$ is a convex function on $\mathcal{X}$.*

Note that the notion of epigraph also allows for extending other results to extended value functions, such as that of Proposition 1.10. It is also helpful in defining the notion of a closed extended value function.

Definition 1.28 *A function $f : \mathcal{X} \subseteq \mathbb{R}^n \to \overline{\mathbb{R}}$ is called* closed *if is epigraph is a closed set (in the sense of Definition 0.16).*

One can show that both closedness and convexity are preserved through conic combinations, maximum, supremum, and composition by an affine function.

Indicator functions Thanks to Definition 1.26, we can determine whether a function is convex by looking at its epigraph. Conversely, one can determine whether a set is convex by checking convexity of an appropriate extended-value function. To this end, we introduce the notion of indicator function, that has numerous applications in constrained optimization.

Definition 1.29 (Indicator function) *Let $\mathcal{X} \subseteq \mathbb{R}^n$. The* **indicator function** *of $\mathcal{X}$, denoted by $\delta_{\mathcal{X}} : \mathbb{R}^n \to (-\infty, \infty]$, is given by*

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \quad \delta_{\mathcal{X}}(\boldsymbol{x}) := \begin{cases} 0 & \text{if } \boldsymbol{x} \in \mathcal{X} \\ \infty & \text{otherwise.} \end{cases}$$

As claimed above, we have a direct relationship between the convexity of a set and that of its indicator function.

Theorem 1.5 *A set $\mathcal{X} \subseteq \mathbb{R}^n$ is convex if and only if its characteristic function is convex on $\mathcal{X}$.*

To end this section, we investigate continuity properties of convex, extended value functions, by first restricting ourselves to a subclass of extended value functions.

Definition 1.30 (Proper function) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $f : \mathcal{X} \to \overline{\mathbb{R}}$. The function f is called* proper *(for the infimum) if $f(\boldsymbol{x}) > -\infty$ for all $\boldsymbol{x} \in \mathcal{X}$ and there exists $\bar{\boldsymbol{x}} \in \mathcal{X}$ such that $f(\bar{\boldsymbol{x}}) < \infty$.*

The notion of proper function correspond to saying that $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ with $\mathrm{dom}(f) \neq \emptyset$. This class of functions conveniently excludes functions that only take infinite values.

Remark 1.7 *Definition 1.21 applies with no ambiguity to proper convex functions.*

The following result illustrates the interest of considering proper convex functions.

Theorem 1.6 *i) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then f is continuous.*

ii) Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a proper convex function. Then the restriction of f to its domain $\mathrm{dom}(f)$ is continuous on its relative interior.

1.2.3 Caracterizing convexity through derivatives

In this section, we consider convex differentiable functions, and show that convexity can be characterized through first- and second-order derivatives. Such results have numerous consequences on the behavior of optimization algorithms applied to convex optimization problems.

Theorem 1.7 *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on an open set containing $\mathcal{X}$. Then, f is a convex function on $\mathcal{X}$ if and only if*

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^2, \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}). \tag{1.2.4}$$

The function f is strictly convex on $\mathcal{X}$ if and only if the inequality (1.2.4) is strict when $\boldsymbol{x} \neq \boldsymbol{y}$.

Proof. We only prove the result in the case of convex functions, as the proof readily adapts to the strictly convex case.

Suppose first that f satisfies (1.2.4). Our goal is to prove that f is convex on $\mathcal{X}$. Let thus $\boldsymbol{x}$ and $\boldsymbol{y}$ be two points in $\mathcal{X}$, and let $\alpha \in [0, 1]$. Defining $\boldsymbol{z} = \alpha\boldsymbol{x} + (1 - \alpha)\boldsymbol{y}$, we have $\boldsymbol{z} \in \mathcal{X}$ by convexity of $\mathcal{X}$. Applying (1.2.4) to the pairs $(\boldsymbol{z}, \boldsymbol{x})$ and $(\boldsymbol{z}, \boldsymbol{y})$ gives

$$\begin{aligned} f(\boldsymbol{x}) &\geq f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{z}) \\ f(\boldsymbol{y}) &\geq f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{z}). \end{aligned}$$

We multiply the first inequality by α, the second one by $(1-\alpha)$ and sum the two resulting inequalities. We thus obtain

$$\alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^{\mathrm{T}}(\alpha\boldsymbol{x} + (1 - \alpha)\boldsymbol{y} - \boldsymbol{z}) = f(\boldsymbol{z}) = f(\alpha\boldsymbol{x} + (1 - \alpha)\boldsymbol{y}),$$

where the equalities come from the definition of $\boldsymbol{z}$. We have thus derived the very definition of convexity (1.2.1), from which we conclude that f is convex.

Conversely, suppose that the function f is convex on $\mathcal{X}$, and let $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^2$. If $\boldsymbol{x} = \boldsymbol{y}$, then (1.2.4) trivially holds. In the rest of the proof, we will therefore assume that $\boldsymbol{x} \neq \boldsymbol{y}$. Let $g : (0, 1] \to \mathbb{R}$ be defined as

$$g(\alpha) = \frac{f(\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x})}{\alpha}.$$

Since f is differentiable at $\boldsymbol{x}$, we have

$$\lim_{\substack{\boldsymbol{x}+\alpha(\boldsymbol{y}-\boldsymbol{x}) \to \boldsymbol{x} \\ \boldsymbol{x}+\alpha(\boldsymbol{y}-\boldsymbol{x}) \neq \boldsymbol{x}}} \frac{|f(\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^{\mathrm{T}}(\alpha(\boldsymbol{y} - \boldsymbol{x}))|}{\|\alpha(\boldsymbol{y} - \boldsymbol{x})\|} = \lim_{\substack{\boldsymbol{x}+\alpha(\boldsymbol{y}-\boldsymbol{x}) \to \boldsymbol{x} \\ \boldsymbol{x}+\alpha(\boldsymbol{y}-\boldsymbol{x}) \neq \boldsymbol{x}}} \frac{|g(\alpha) - \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x})|}{\|\boldsymbol{y} - \boldsymbol{x}\|} = 0.$$

Consequently, using $\boldsymbol{y} \neq \boldsymbol{x}$ implies that

$$\lim_{\alpha \downarrow 0} g(\alpha) = \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}).$$

Let us show now that g is nondecreasing. For any α_1 and α_2 such that $0 < \alpha_1 < \alpha_2 < 1$, it holds that

$$\bar{\alpha} = \frac{\alpha_1}{\alpha_2} \in (0, 1) \quad \text{and} \quad \boldsymbol{z} = \boldsymbol{x} + \alpha_2(\boldsymbol{y} - \boldsymbol{x}) \in \mathcal{X}.$$

Applying Definition (1.2.1) at the point $\boldsymbol{x} + \bar{\alpha}(\boldsymbol{z} - \boldsymbol{x}) \in \mathcal{X}$, we get

$$\begin{array}{rcl} f(\boldsymbol{x} + \bar{\alpha}(\boldsymbol{z} - \boldsymbol{x})) & \leq & \bar{\alpha} f(\boldsymbol{z}) + (1 - \bar{\alpha}) f(\boldsymbol{x}) \\ \dfrac{f(\boldsymbol{x} + \bar{\alpha}(\boldsymbol{z} - \boldsymbol{x})) - f(\boldsymbol{x})}{\bar{\alpha}} & \leq & f(\boldsymbol{z}) - f(\boldsymbol{x}) \\ \dfrac{f(\boldsymbol{x} + \alpha_1(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x})}{\alpha_1} & \leq & \dfrac{f(\boldsymbol{x} + \alpha_2(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x})}{\alpha_2}, \end{array}$$

hence $g(\alpha_1) \leq g(\alpha_2)$. As a result, we have shown that

$$\lim_{\alpha \downarrow 0} g(\alpha) = \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \leq g(1) = f(\boldsymbol{y}) - f(\boldsymbol{x}),$$

which corresponds to (1.2.4). □

Remark 1.8 *The property (1.2.4) is useful to show that a quadratic function based on a positive semidefinite matrix is convex. Indeed, letting $f : \boldsymbol{x} \mapsto \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}$ denote such a quadratic, we have $\nabla f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ and thus, for any $(\boldsymbol{x}, \boldsymbol{y}) \in (\mathbb{R}^n)^2$, we obtain*

$$\begin{array}{rcl} f(\boldsymbol{y}) & = & \dfrac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} \\ & = & \dfrac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} + \dfrac{1}{2}(\boldsymbol{y} + \boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) + \boldsymbol{b}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \\ & = & f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \dfrac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) \\ & \geq & f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \end{array}$$

using that $\boldsymbol{A} \succeq \boldsymbol{0}$.

Although the next property is less important for the purpose of these notes, it provides a nice characterization of convexity for twice differentiable functions.

Theorem 1.8 *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable on an open set containing $\mathcal{X}$. Then, the following properties hold:*

1. *If $\nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{0}$ for any $\boldsymbol{x} \in \mathcal{X}$, then f is convex on $\mathcal{X}$;*
2. *If $\mathcal{X}$ is open and f is convex on $\mathcal{X}$, then $\nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{0}$ for any $\boldsymbol{x} \in \mathcal{X}$.*

Remark 1.9 *Recall the function $f : \boldsymbol{x} \mapsto \ln(\sum_{i=1}^n e^{x_i})$ from Example 1.3. This function is twice continuously differentiable on $\mathbb{R}^n$, and for any $\boldsymbol{x} \in \mathbb{R}^n$, one has*

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{\left(\sum_{j=1}^n e^{x_j}\right)^2} \begin{bmatrix} \sum_{i \neq 1} e^{x_1 + x_i} & -e^{x_1 + x_2} & \cdots & -e^{x_1 + x_n} \\ -e^{x_1 + x_2} & \sum_{i \neq 2} e^{x_2 + x_i} & \ddots & -e^{x_2 + x_n} \\ \vdots & \ddots & \ddots & \vdots \\ -e^{x_1 + x_{n-1}} & \cdots & \sum_{i \neq n-1} e^{x_{n-1} + x_i} & -e^{x_1 + x_n} \\ -e^{x_1 + x_n} & \cdots & -e^{x_{n-1} + x_n} & \sum_{i \neq n} e^{x_n + x_i} \end{bmatrix},$$

and this matrix is positive semidefinite as diagonally dominant with non-negative diagonal entries (for more on this notion, see Example 0.2).

Corollary 1.1 *The results of Theorems 1.7 and 1.8 also apply to an extended value function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ when its domain $\mathrm{dom}(f)$ is a convex set, assuming appropriate differentiability properties on this domain.*

1.2.4 Strongly convex functions

For sake of generality, we introduce this concept (a special case of convexity) in the framework of extended value functions.

Definition 1.31 (Strongly convex functions) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathcal{X} \to \overline{\mathbb{R}}$ be a proper convex function. Let $\mu > 0$. The function f is called* **μ-strongly convex** *(or strongly convex with parameter μ) if*

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^2,\ \forall \alpha \in [0,1], \quad f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) - \mu \frac{\alpha(1-\alpha)}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Remark 1.10 *Any strongly convex function is strictly convex and convex, however the converse is false. For instance, the function $x \mapsto x^4$ is convex on $\mathbb{R}$ but not strongly convex. Similarly, the function $x \mapsto \exp(-x)$ is strictly convex but not strongly convex.*

As in the case of convex functions, one can characterize convexity using the derivatives of f when they exist. We present those results in the context of real-valued functions.

Theorem 1.9 *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on an open set containing $\mathcal{X}$. Then, f is μ-strongly convex on $\mathcal{X}$ if and only if*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}^2, \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{1.2.5}$$

Theorem 1.10 *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable on an open set containing $\mathcal{X}$. Then, f is μ-strongly convex with $\mu > 0$ if and only if*

$$\forall \boldsymbol{x} \in \mathcal{X}, \quad \nabla^2 f(\boldsymbol{x}) \succeq \mu \boldsymbol{I}_n.$$

Chapter 2

Convex optimization

In Chapter 1, we have explored convex sets and convex functions in detail. We will now leverage these concepts to define convex optimization problems, which are commonly used to model data science tasks. We begin by defining the key components of an optimization problem in Section 2.1, together with several results for proving that a given problem has a solution. Of particular interest of us is the case of convex optimization problems, that arises naturally through the notion of duality. We explore that notion in detail in Section 2.2.

2.1 Definitions and examples

2.1.1 Optimization problem

An optimization problem is a mathematical model of taking the best decision out of a set of alternatives. For this course, we will consider optimization problems under the following canonical form:

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) \\ \text{s. t.} & g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \dots, m \\ & h_i(\boldsymbol{x}) = 0, \quad i = 1, \dots, \ell, \end{array} \tag{2.1.1}$$

The mathematical object (2.1.1) represents a **minimization** problem, symbolized by the word *minimize*. The goal is to determine the lowest possible value of an **objective function** $f : \mathbb{R}^n \to \overline{\mathbb{R}}$. To this end, we act on the inputs of the function, gathered in a vector $\boldsymbol{x} \in \mathbb{R}^n$ of **decision variables**[1] In order to be acceptable, a vector of decision variables must satisfy a set of **constraints** (we say that it is *subject to* the constraints), that we describe through m inequality constraints (of the form $g_i(\boldsymbol{x}) \leq 0$ with $g_i : \mathbb{R}^n \to \overline{\mathbb{R}}$) and ℓ equality constraints (of the form $h_i(\boldsymbol{x}) = 0$ with $h_i : \mathbb{R}^n \to \overline{\mathbb{R}}$). We allow $m = \ell = 0$, in which case there are no constraints on the problem. We then simply write $\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$, and we say that this is an **unconstrained problem**. Otherwise, we say that it is a **constrained problem**.

Remark 2.1 *By introducing the vector-valued functions $\boldsymbol{g} : \mathbb{R}^n \to (\overline{\mathbb{R}})^m$ and $\boldsymbol{h} : \mathbb{R}^n \to (\overline{\mathbb{R}})^\ell$ such*

[1] In this course, we will only consider real-valued decision variables, which is the most common setting in a variety of fields, including data science (where it corresponds to a parametric approach).

that $\boldsymbol{g}(\boldsymbol{x}) = [g_i(\boldsymbol{x})]$ *and* $\boldsymbol{h}(\boldsymbol{x}) = [h_i(\boldsymbol{x})]$*, the problem* (2.1.1) *can be rewritten:*

$$\begin{cases} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0} \\ & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \end{cases} \tag{2.1.2}$$

The *domain* of the optimization problem (2.1.1) is defined by

$$\mathcal{X} := \operatorname{dom}(f) \cap \bigcap_{i=1}^{m} \operatorname{dom}(g_i) \cap \bigcap_{i=1}^{\ell} \operatorname{dom}(h_i).$$

Discarding points where any function value (objective or constraint function) is equal to $+\infty$ makes sense given that such points would not satisfy the constraints nor lead to the smallest objective value in general.

The **feasible set** of the optimization problem (2.1.1) is defined by

$$\mathcal{F} := \left\{ \boldsymbol{x} \in \bigcap_{i=1}^{m} \operatorname{dom}(g_i) \cap \bigcap_{i=1}^{\ell} \operatorname{dom}(h_i) \;\middle|\; \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0},\ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} \right\}. \tag{2.1.3}$$

Remark 2.2 *When the objective function is constant, solving problem* (2.1.1) *amounts to finding points in the feasible set. We then call this problem a* feasibility problem*. Finding a feasible point (i.e. a point in the feasible set) can be as difficult as minimizing a function over that set, as we will see later.*

We define the **optimal value** of problem (2.1.1) as

$$f^* := \inf \{ f(\boldsymbol{x}) \mid \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}_m,\ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}_\ell \}. \tag{2.1.4}$$

Note that $f^* \in \overline{\mathbb{R}}$ .If the feasible set is empty, we set $f^* = \infty$.

A vector $\boldsymbol{x}^* \in \mathbb{R}^n$ is called a **solution**[2] of problem (2.1.1) if

$$\boldsymbol{x}^* \in \mathcal{F} \quad \text{and} \quad f(\boldsymbol{x}^*) = f^*,$$

where $\mathcal{F}$ and f^* are defined in (2.1.3) and (2.1.4), respectively. The set of such optimal points is denoted by

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \{ f(\boldsymbol{x}) \mid \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0},\ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} \}.$$

Note that this set can be empty even if the feasible set is not empty.

2.1.2 Reformulations

In the previous section, we presented problem (2.1.1) under a canonical form, where inequality constraints are written as $g_i(\boldsymbol{x}) \leq 0$ and equality constraints have a 0 right-hand side. It is always possible to write a problem under this form, as illustrated by the following example.

[2] Other terms include optimal point, optimum or minimum.

Example 2.1 (Bound constraints) *Consider the following bound-constrained problem:*

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) \\ \text{subject to} & l_i \leq x_i \leq u_i \quad i = 1, \ldots, n, \end{array}$$

where $l_i \leq u_i$ are (finite) bounds on the variable x_i. The problem can be rewritten as

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \end{array}$$

where $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^{2n}$ is defined as

$$\boldsymbol{g}(\boldsymbol{x}) = \begin{bmatrix} l_1 - x_1 \\ \vdots \\ l_n - x_n \\ x_1 - u_1 \\ \vdots \\ x_n - u_n \end{bmatrix}.$$

This new problem has the same feasible set, optimal value and optimal solutions than the original one.

There exist infinitely many ways to formulate the same optimization problem, typically by changing the functions describing the feasible set. In addition, an optimization problem can be connected to other problems, even though they may have different feasible sets, optimal values, optimal solutions or even different decision variables. To highlight this connection, we will say that two optimization problems are equivalent when the solution of one is immediately obtained from that of the other, and vice-versa.

With that definition, one can see that problem (2.1.1) is equivalent to

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & \alpha_0\, f(\boldsymbol{x}) \\ \text{subject to} & \alpha_i\, g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, m \\ & \beta_i\, h_i(\boldsymbol{x}) = 0 \quad i = 1, \ldots, \ell, \end{array}$$

for any $\alpha_i \in \mathbb{R}_{++}\ \forall i = 0, \ldots, m$ and $\beta_i \neq 0\ \forall i = 1, \ldots, \ell$. Indeed, the feasible sets, problem domains and sets of optimal solutions coincide for both problems, although they do not have the same objective and optimal value.

Remark 2.3 *The previous observation can be extended further. The optimization problem* (2.1.1) *is equivalent to any problem of the form*

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & \phi_0(f(\boldsymbol{x})) \\ \text{subject to} & \phi_i(g_i(\boldsymbol{x})) \leq 0, \quad i = 1, \ldots, m \\ & \phi_{m+i}(h_i(\boldsymbol{x})) = 0 \quad i = 1, \ldots, \ell, \end{array}$$

where $\phi_0 : \mathbb{R} \to \mathbb{R}$ is monotone increasing, $\phi_i : \mathbb{R} \to \mathbb{R}$ satisfies $\phi_i(t) \leq 0 \Leftrightarrow t \leq 0$ for any $i = 1, \ldots, m$ and $\phi_i : \mathbb{R} \to \mathbb{R}$ satisfies $\phi_i(t) = 0 \Leftrightarrow t = 0$ for any $i = m+1, \ldots, m+\ell$.

In general, we say that an optimization problem admits a reformulation as an equivalent problem. We review popular reformulation techniques below.

Reformulations based on constraints One classical way to reformulate an optimization problem is by modifying the description of its feasible set. A common transformation is based on the observation that $g_i(\boldsymbol{x}) \leq 0$ if and only if it exists $s_i \geq 0$ such that $g_i(\boldsymbol{x}) + s_i = 0$. Using this observation, we can reformulate the problem (2.1.1) as

$$\begin{array}{lll} \text{minimize}_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{s}\in\mathbb{R}^m}} & f(\boldsymbol{x}) & \\ \text{subject to} & s_i \geq 0, & i = 1, \dots, m \\ & g_i(\boldsymbol{x}) + s_i = 0, & i = 1, \dots, m \\ & h_i(\boldsymbol{x}) = 0 & i = 1, \dots, \ell, \end{array} \tag{2.1.5}$$

The problem now features m additional variables s_i, that are called slack variables. Their use leads to a problem in which constraints are either nonnegativity (bound) constraints or equalities. This problem is equivalent to (2.1.1): if $(\boldsymbol{x}^*, \boldsymbol{s}^*)$ is a solution of (2.1.5), then $\boldsymbol{x}^*$ solves (2.1.5).

Similarly, one can transform an equality constraint $h_i(\boldsymbol{x}) = 0$ into two constraints $h_i(\boldsymbol{x}) \leq 0$ and $-h_i(\boldsymbol{x}) \leq 0$, creating a reformulation of problem (2.1.1) with $m + 2\ell$ inequality constraints.

Epigraph reformulation The **epigraph formulation** of problem (2.1.1) is given by

$$\begin{cases} \begin{array}{lll} \text{minimize}_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ t\in\mathbb{R}}} & t & \\ \text{subject to} & f(\boldsymbol{x}) - t \leq 0 & \\ & g_i(\boldsymbol{x}) \leq 0, & i = 1, \dots, m \\ & h_i(\boldsymbol{x}) = 0 & i = 1, \dots, \ell. \end{array} \end{cases} \tag{2.1.6}$$

The feasible set of (2.1.6) involves the epigraph of the function f introduced in Section 1.2. This reformulation is equivalent to the original one, and can be viewed as bringing the objective function into the constraints.

Indicator function Problem (2.1.1) is equivalent to the unconstrained optimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\boldsymbol{x}) + \delta_{\mathcal{F}}(\boldsymbol{x}),$$

where $\delta_{\mathcal{F}}(\cdot)$ is the indicator function of the feasible set (recall Definition 1.29). This reformulation can be viewed as bringing the constraints into the objective. Extended value functions can often be formulated as the sum of a real-valued function plus the indicator function of a given set in $\mathbb{R}^n$.

Maximum and optimization problems The canonical form (2.1.9) is that of a *minimization* problem, in that we seek the smallest possible objective value. It is also possible to define a maximization problem, written as

$$\begin{array}{lll} \text{maximize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) & \\ \text{subject to} & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}_m, & i = 1, \dots, m \\ & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}_\ell, & i = 1, \dots, \ell, \end{array} \tag{2.1.7}$$

as a reformulation of

$$\begin{array}{lll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & -f(\boldsymbol{x}) & \\ \text{subject to} & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}_m, & i = 1, \dots, m \\ & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}_\ell, & i = 1, \dots, \ell, \end{array} \tag{2.1.8}$$

where $-f$ denotes the opposite function to f. [3] Those problems have the same domain and feasible set. We then define the optimal value of (2.1.7) as

$$g^* = -\inf_{\boldsymbol{x}\in\mathbb{R}^n} \{-f(\boldsymbol{x}) | \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\},$$

which we will write

$$g^* = \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x}) | \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\}.$$

Similarly, we define the set of optimal solutions of problem (2.1.7) as

$$\operatorname*{argmax}_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x}) | \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\} := \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \{-f(\boldsymbol{x}) | \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\}.$$

2.1.3 Convex optimization problems

A convex optimization problem essentially consists in a convex feasible set on which the objective function is convex. Mathematically, this means that there exists a reformulation of the problem as

$$\begin{array}{lll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) & \\ \text{subject to} & g_i(\boldsymbol{x}) \leq 0, & i = 1, \dots, m \\ & \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} - b_i = 0, & i = 1, \dots, \ell, \end{array} \tag{2.1.9}$$

where $f, g_1, \dots, g_m$ are convex (extended-value functions) and $(\boldsymbol{a}_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$ for any $i = 1, \dots, \ell$. Under these assumptions, the problem domain and the feasible set are convex sets.

The formulation (2.1.9) is called a **standard form** of a convex optimization problem[4]. Note that convex problems are not necessarily written in standard form. For instance, the problem

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^2} & x_1^2 + x_2^2 \\ \text{subject to} & \frac{x_1}{1+x_2^2} \leq 0 \\ & (x_1 + x_2)^2 = 0, \end{array}$$

is not written in standard form. However, by looking at the feasible set, one observes that it is equal $\{\boldsymbol{x} | x_1 \leq 0, x_1 + x_2 = 0\}$. As a result, the problem admits the reformulation

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^2} & x_1^2 + x_2^2 \\ \text{subject to} & x_1 \leq 0 \\ & x_1 + x_2 = 0, \end{array}$$

which is a convex optimization problem in standard form.

We will see numerous examples of convex problems in the rest of the course. To end this section, we describe the most classical (and most studied) category of such problems.

Example 2.2 (Linear programming) *A linear program, or linear optimization problem, can be put in the form*

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0}, \end{array} \tag{2.1.10}$$

[3] One can of course define maximization problems independently of minimization problems. However, we adopt this unified view for consistency.

[4] Standard forms can actually be formulated using upper inequalities $g_i(\boldsymbol{x}) \geq 0$ and putting b_is on the right-hand side of the inequalities.

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{c} \in \mathbb{R}^n$. This formulation is called the standard form of a linear program. Note that it matches the standard form of convex programs in the sense of (2.1.9) up to a sign flip of the inequalities.

Example 2.3 (Quadratic programming) *A quadratic program, or quadratic optimization problem, can be put in the form*

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{x} \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0}, \end{array} \tag{2.1.11}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{c} \in \mathbb{R}^n$ and $\boldsymbol{H} \succeq \boldsymbol{0}$. This formulation is called the standard form of a quadratic program.

Example 2.4 (Semidefinite programming) *A semidefinite program (or SDP) is an optimization problem that can be put under the form*

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{X}\in\mathcal{S}^n} & \operatorname{trace}(\boldsymbol{C}\boldsymbol{X}) \\ \text{subject to} & \operatorname{trace}(\boldsymbol{A}_i\boldsymbol{X}) = b_i, \quad i = 1,\dots,m \\ & \boldsymbol{X} \succeq \boldsymbol{0}, \end{array} \tag{2.1.12}$$

where $\boldsymbol{C} \in \mathcal{S}^n$, $\boldsymbol{A}_i \in \mathcal{S}^n$ and $b_i \in \mathbb{R}$ for every $i \in \{1,\dots,m\}$.

Remark 2.4 *By convention, a maximization problem of the form*

$$\begin{array}{lll} \text{maximize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) & \\ \text{subject to} & g_i(\boldsymbol{x}) \leq 0, & i = 1,\dots,m \\ & \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} - b_i = 0, & i = 1,\dots,\ell \end{array} \tag{2.1.13}$$

where f is a concave function and $g_1,\dots,g_m$ are convex functions is viewed as a convex optimization problem, in the sense that it can be reformulated as

$$\begin{array}{lll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & -f(\boldsymbol{x}) & \\ \text{subject to} & g_i(\boldsymbol{x}) \leq 0, & i = 1,\dots,m \\ & \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} - b_i = 0, & i = 1,\dots,\ell, \end{array} \tag{2.1.14}$$

which is a convex optimization problem in standard form.

2.1.4 Existence of solutions

Optimization theory is concerned with proving that a given problem possesses a solution. In continuous optimization, one of the most classical results is the following theorem, attributed to the German mathematician Karl Weierstrass (Weierstraß in German).

Theorem 2.1 (Weierstrass) *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be continuous and $\mathcal{C} \subset \mathbb{R}^n$ be a nonempty compact set. Then, the problem*

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{C} \end{array}$$

has at least one solution.

The previous theorem is restricted to compact feasible domains, which is restrictive as it excludes (in particular) unconstrained problems. To establish existence results on unbounded domains, we leverage the following concept.

Definition 2.1 (Coercive function) *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a unbounded, nonempty set. A function $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is called* coercive *if*

$$\lim_{\substack{\boldsymbol{x}\in\mathcal{X} \\ \|\boldsymbol{x}\|\to\infty}} f(\boldsymbol{x}) = \infty.$$

A classical example of a coercive function is $\boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|^2$ on any unbounded subset of $\mathbb{R}^n$.
Coercive objective functions are amenable to minimization, as shown by the following result.

Theorem 2.2 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a coercive, continuous function and let $\mathcal{F} \subseteq \mathbb{R}^n$ be a nonempty closed set. Then, the problem*

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{F} \end{array}$$

has at least one solution.

Finally, in the context of a convex problem, even stronger results can be provided.

Theorem 2.3 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex on a nonempty closed convex set $\mathcal{C} \subseteq \mathbb{R}^n$.*

(i) The solution set

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{C}\}$$

is either empty or convex.

(ii) If f is strictly convex on $\mathcal{C}$, the problem has at most one solution.

(iii) If f is strongly convex on $\mathcal{C}$, the problem has a unique solution.

Finally, if we further assume differentiability in addition to convexity, we can provide checkable conditions for a point to be a solution of the problem, that involve the derivative of f and are called *optimality conditions*.

Theorem 2.4 (First-order optimality conditions) *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a nonempty, convex set, and $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable on an open set containing $\mathcal{C}$. A point $\boldsymbol{x}^* \in \mathbb{R}^n$ is a solution of the problem*

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{C}, \end{array} \tag{2.1.15}$$

if and only if

$$\boldsymbol{x}^* \in \mathcal{C} \quad \textit{and} \quad \nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{z} - \boldsymbol{x}^*) \geq 0 \quad \forall \boldsymbol{z} \in \mathcal{C}. \tag{2.1.16}$$

In practice, condition (2.1.16) is not necessarily straightforward to check, since it depends on the structure of $\mathcal{C}$. In certain cases, however, the condition can be simplified.

Corollary 2.1 (Special first-order conditions) *Let the assumptions of Theorem 2.4 hold.*

i) If $\mathcal{C}$ is a linear subspace of $\mathbb{R}^n$, then condition (2.1.16) is equivalent to $\nabla f(\boldsymbol{x}^*) \in \mathcal{C}^\perp$, i.e.

$$\boldsymbol{x}^* \in \mathcal{C} \quad \text{and} \quad \forall \boldsymbol{y} \in \mathcal{C}, \quad \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{y} = 0.$$

ii) If $\mathcal{C}$ is an affine set, then condition (2.1.16) is equivalent to

$$\boldsymbol{x}^* \in \mathcal{C} \quad \text{and} \quad \forall \boldsymbol{y} \in \mathcal{C}, \quad \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{x}^*) = 0.$$

iii) If $\mathcal{C}$ is a convex cone containing $\mathbf{0}$, then condition (2.1.16) is equivalent to

$$\boldsymbol{x}^* \in \mathcal{C}, \quad \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{x}^* = 0 \quad \text{and} \quad \forall \boldsymbol{y} \in \mathcal{C}, \ \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{y} \geq 0.$$

In the next section, we will leverage the description of the feasible set based on inequalities and equalities to derive more existence results and characterizations of solutions.

2.2 Duality

Duality is a fundamental concept in constrained optimization. For any given problem (called "primal problem"), one defines another problem (called "dual problem") that provides information about solving the original problem. In particular, it allows to state special optimality conditions called the KKT conditions.

For the rest of this chapter, we will focus on a generic optimization problem of the form

$$\begin{cases} \text{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} & f(\boldsymbol{x}) \\ \text{subject to} & g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \dots, m \\ & h_i(\boldsymbol{x}) = 0, \quad i = 1, \dots, \ell, \end{cases} \tag{2.2.1}$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$, $\boldsymbol{g} = [g_i] : \mathbb{R}^n \to (-\infty, \infty]^m$ and $\boldsymbol{h} = [h_i] : \mathbb{R}^n \to (-\infty, \infty]^\ell$. We will suppose that the problem domain $\mathcal{X} := \operatorname{dom}(f) \cap \bigcap_{i=1}^m \operatorname{dom}(g_i) \cap \bigcap_{i=1}^\ell \operatorname{dom}(h_i)$ is not empty, and we let p^* denote the optimal value problem, i.e.

$$p^* = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ f(\boldsymbol{x}) \mid \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}, \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0} \right\}.$$

Note that we do not assume convexity of any of the functions defining problem (2.2.1).

2.2.1 Lagrangian function and dual problem

In this course, we work with one of several notions of duality, namely *Lagrangian duality*. This theory is based on the **Lagrangian function**, named after the French-Italian mathematician Joseph-Louis Lagrange.

Definition 2.2 (Lagrangian function) *The* **Lagrangian** *function (or Lagrangian in short) associated with problem (2.2.1) is the function*
$\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^\ell \to \overline{\mathbb{R}}$ defined by

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := f(\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{h}(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x}) + \sum_{i=1}^\ell \mu_i h_i(\boldsymbol{x}). \tag{2.2.2}$$

In that setting, the components of the vector $\boldsymbol{x}$ are called the primal variables, while that of the vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are called the dual variables, or the Lagrange multipliers associated with inequality and equality constraints, respectively.

Note that the Lagrangian is an extended value function, with domain $\mathcal{X} \times \mathbb{R}^m \times \mathbb{R}^\ell$ (recall that $\mathcal{X}$ is the domain of problem (2.2.1)).

For any $\boldsymbol{x} \in \mathcal{X}$ that is feasible for the primal problem (2.2.1), it is straightforward to see that

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq f(\boldsymbol{x})$$

for any $\boldsymbol{\lambda} \in (\mathbb{R}_+)^m$ and any $\boldsymbol{\mu} \in \mathbb{R}^\ell$. This property shows that the Lagrangian function is an underapproximation of the objective function under certain conditions. This is formalized in the following definition.

Definition 2.3 *Consider problem (2.2.1) and define the set*

$$\mathcal{Y} := \left\{ (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^m \times \mathbb{R}^\ell \;\middle|\; \lambda_i \geq 0 \; \forall i = 1, \ldots, m \right\}.$$

The **primal function** *of the problem the function $p : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by*

$$p(\boldsymbol{x}) := \begin{cases} \sup_{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) & \text{if } \boldsymbol{x} \in \mathcal{X} \\ \infty & \text{if } \boldsymbol{x} \notin \mathcal{X}. \end{cases} \tag{2.2.3}$$

Similarly, the **dual function** *of the problem is the function $d : \mathbb{R}^m \times \mathbb{R}^\ell \to \overline{\mathbb{R}}$ defined by*

$$d(\boldsymbol{\lambda}, \boldsymbol{\mu}) := \begin{cases} \inf_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) & \text{if } (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathcal{Y} \\ -\infty & \text{if } (\boldsymbol{\lambda}, \boldsymbol{\mu}) \notin \mathcal{Y}. \end{cases} \tag{2.2.4}$$

Both the primal and the dual functions are extended value functions. As their name suggest, they represent two dual views of the optimization problem of interest. We first describe the connection between the primal function and the primal problem (2.2.1).

Proposition 2.1 *The primal problem $\operatorname{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} p(\boldsymbol{x})$ is equivalent to the problem (2.2.1) when the latter is feasible, i.e. when the feasible set is not empty.*

The result of Proposition 2.1 justifies that problem (2.2.1) is often called the primal problem.

In general, the dual function has a more intricate expression than the primal function, and its connection with the original problem is less immediate. To build towards such a connection, we begin by an important property of the dual function.

Proposition 2.2 *The dual function (2.2.4) is concave.*

Recall that a function is concave if its negative is convex. As a result, we can define a convex optimization problem involving the dual function.

Definition 2.4 *The* **dual problem** *(or simply dual) of problem (2.2.1) is*

$$\underset{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^m \times \mathbb{R}^\ell}{\text{maximize}} \; d(\boldsymbol{\lambda}, \boldsymbol{\mu}) \quad \text{subject to} \quad \boldsymbol{\lambda} \geq \mathbf{0}. \tag{2.2.5}$$

Since the constraint on $\boldsymbol{\lambda}$ can be rewritten as $-\boldsymbol{\lambda} \leq \mathbf{0}$, which is a convex constraint, and $-d$ is a convex function per Proposition 2.2, the problem (2.2.5) is a convex optimization problem. Its domain is $\mathrm{dom}(-d)$. A point $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathrm{dom}(-d)$ is called dual feasible if $\boldsymbol{\lambda} \geq \mathbf{0}$. Finally, the optimal value of problem (2.2.5) is given by

$$d^* := \sup_{\boldsymbol{\lambda},\boldsymbol{\mu}} \{d(\boldsymbol{\lambda}, \boldsymbol{\mu}) \mid \boldsymbol{\lambda} \geq \mathbf{0}\} = -\inf_{\boldsymbol{\lambda},\boldsymbol{\mu}} \{-d(\boldsymbol{\lambda}, \boldsymbol{\mu}) \mid \boldsymbol{\lambda} \geq \mathbf{0}\} \tag{2.2.6}$$

Example 2.5 (Dual linear program) *Consider the linear program*

$$\begin{cases} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} & x_i \geq 0, \quad i = 1, \ldots, n, \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \end{cases} \tag{2.2.7}$$

where $\boldsymbol{A} \in \mathbb{R}^{\ell\times n}$ and $\boldsymbol{b} \in \mathbb{R}^{\ell}$. The dual function of this problem is the function $d : \mathbb{R}^n \times \mathbb{R}^{\ell} \to \mathbb{R}$ by

$$d(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{cases} -\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\mu} & \text{if } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu} - \boldsymbol{\lambda} + \boldsymbol{c} = \mathbf{0} \text{ and } \boldsymbol{\lambda} \geq \mathbf{0} \\ -\infty & \text{otherwise.} \end{cases}$$

As a result, the dual problem can be reformulated as

$$\begin{cases} \text{maximize}_{(\boldsymbol{\lambda},\boldsymbol{\mu})\in\mathbb{R}^n\times\mathbb{R}^{\ell}} & -\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\mu} \\ \text{subject to} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu} - \boldsymbol{\lambda} + \boldsymbol{c} = \mathbf{0}, \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \end{cases}$$

which is itself equivalent to

$$\begin{cases} \text{minimize}_{(\boldsymbol{\lambda},\boldsymbol{\mu})\in\mathbb{R}^n\times\mathbb{R}^{\ell}} & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\mu} \\ \text{subject to} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu} - \boldsymbol{\lambda} + \boldsymbol{c} = \mathbf{0}, \\ & \boldsymbol{\lambda} \geq \mathbf{0}. \end{cases}$$

We then see that the dual problem to a linear program has the same structure than the original problem, in that it is also a linear program.

2.2.2 Weak duality and strong duality

The interest of deriving dual formulations lies in their link with the primal problem. The first key property of dual problems, called weak duality, is described below.

Theorem 2.5 (Weak duality) *Let p and d be the primal and dual functions associated with problem (2.2.1). Then,*

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \ \forall (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^m \times \mathbb{R}^{\ell}, \qquad d(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq p(\boldsymbol{x}). \tag{2.2.8}$$

In particular,

$$\forall \boldsymbol{x} \in \mathcal{X}, \ \forall (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathcal{Y}, \qquad d(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq f(\boldsymbol{x}) \quad \Leftrightarrow \quad d^* \leq p^*. \tag{2.2.9}$$

In general, the above weak duality result is the best one can obtain, in the sense that $d^* < p^*$ in general. This naturally leads to the following concept.

Definition 2.5 (Duality gap) *The duality gap of problem (2.2.1) is given by the quantity $p^* - d^*$.*

We stay that **strong duality** holds for problem (2.2.5) when the duality gap is empty, i.e. when $p^* = d^*$.

Remark 2.5 *With strong duality, one can "swap" the infimum and supremum in the expressions of the primal and dual optimal value. It makes then sense to talk about a primal-dual solution* $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ *such that*

$$\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \inf_{\boldsymbol{x}\in\mathbb{R}^n} \sup_{\substack{(\boldsymbol{\lambda},\boldsymbol{\mu})\in\mathbb{R}^m\times\mathbb{R}^\ell \\ \boldsymbol{\lambda}\geq\mathbf{0}}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \sup_{\substack{(\boldsymbol{\lambda},\boldsymbol{\mu})\in\mathbb{R}^m\times\mathbb{R}^\ell \\ \boldsymbol{\lambda}\geq\mathbf{0}}} \inf_{\boldsymbol{x}\in\mathbb{R}^n} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

As a result, $\boldsymbol{x}^* \in \mathcal{X}$ *and* $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \in \mathcal{Y}$ *are solutions of the primal and dual problems, respectively, if*

$$\forall (\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}), \quad \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \leq \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$$

for any triplet $(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$*. A primal-dual solution is in fact a saddle point of the Lagrangian (minimum with respect to* $\boldsymbol{x}$ *and maximum with respect to* $\boldsymbol{\lambda}$ *and* $\boldsymbol{\mu}$*).*

As mentioned above, strong duality does not hold for all problems. However, for convex optimization problems, strong duality typically holds under certain conditions called constraint qualifications that we illustrate in the next section.

2.2.3 Karush-Kuhn-Tucker conditions

In this section, we show how duality theory can be combined with regularity properties of the problem to derive optimality conditions. Those form an alternative to the conditions derived in Section 2.1.4.

We again consider problem (2.2.1), and recall that the Lagrangian for this problem is given by

$$\mathcal{L} : (\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \mapsto f(\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{h}(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{i=1}^{\ell} \mu_i h_i(\boldsymbol{x}).$$

We now make the following assumption on the objective and constraint functions.

Assumption 2.1 *The functions* f*,* $\{g_i\}_{i=1}^m$*,*$\{h_i\}_{i=1}^\ell$ *appearing in problem* (2.2.1) *are differentiable on an open set containing the problem domain.*

Thanks to Assumption 2.1, one can compute the gradient of $\mathcal{L}$ with respect to $\boldsymbol{x}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$[5]. For the primal variables, we obtain

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}) + \sum_{i=1}^{\ell} \mu_i \nabla h_i(\boldsymbol{x}),$$

which is a linear combination of the objective gradient and the gradients for the constraint functions. In addition, given that the Lagrangian function is linear in the dual variables $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ (which is directly connected to the concave nature of the dual function), we also have

$$\begin{cases} \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{g}(\boldsymbol{x}) \\ \nabla_{\boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{h}(\boldsymbol{x}). \end{cases}$$

[5]Note that these gradients correspond to partial gradients, in the sense of Remark 0.7.

As explained in the previous section, a primal-dual solution maximizes the Lagrangian function in the dual variables and minimizes the Lagrangian function with respect to the primal variables, hence all gradients must be 0. To form optimality conditions, one must distinguish two situations for the inequality constraints: if $\boldsymbol{x}^*$ is a solution, then for any $i = 1, \dots, m$, either $g_i(\boldsymbol{x}^*) = 0$ or $g_i(\boldsymbol{x}^*) < 0$. When the latter property holds, this implies that the constraint does not matter for characterizing the solution in the optimality conditions, and therefore the corresponding dual variable must be 0. This crucial observation is at the heart of KKT conditions [6], that are stated in Theorem 2.6.

Theorem 2.6 *Consider the problem* (2.2.1) *and its dual* (2.2.5) *under Assumption 2.1. Suppose that strong duality holds. Then, for any primal-dual solution* $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$*, we have*

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0} \tag{2.2.10a}$$
$$\boldsymbol{g}(\boldsymbol{x}^*) \leq \mathbf{0} \tag{2.2.10b}$$
$$\boldsymbol{h}(\boldsymbol{x}^*) = \mathbf{0} \tag{2.2.10c}$$
$$\boldsymbol{\lambda}^* \geq \mathbf{0} \tag{2.2.10d}$$
$$\lambda_i^* g_i(\boldsymbol{x}^*) = 0 \quad \forall i = 1, \dots, m. \tag{2.2.10e}$$

The system of equations (2.2.10) *is called the (first-order)* **KKT conditions**. *A solution of these equations is called a (first-order)* **KKT point**.

Note that a vector $\boldsymbol{x}^*$ of primal variables is sometimes called a KKT point, which then means that there exist dual variables $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ such that $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ solves (2.2.10)

Remark 2.6
- *The first KKT condition* (2.2.10a) *means that the gradient of the objective and that of the constraints are linearly dependent when the dual variables are non all zeros.*
- *The last KKT condition* (2.2.10e) *is called the* **complementarity condition**, *and showcases the difficulty of dealing with inequality constraints. Note that when* $g_i(\boldsymbol{x}^*) < 0$ *for some* $i = 1, \dots, m$, *this condition implies that* $\lambda_i^* = 0$.

The KKT conditions are *necessary* optimality conditions, that are not sufficient to characterize solutions in general. However, in the convex setting, these conditions become necessary *and sufficient*.

Theorem 2.7 *Consider problem* (2.2.1) *and its dual* (2.2.5) *under Assumption 2.1. Suppose that the functions* f *and* g_i *are convex, while the functions* h_i *are linear, and suppose further that strong duality holds. Then, a triplet* $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^\ell$ *is a primal-dual solution of the problem if and only if it satisfies the KKT conditions* (2.2.10), *in which case we have* $p(\boldsymbol{x}^*) = p^* = d^* = d(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$.

Theorem 2.7 thus shows that the KKT conditions are both necessary and sufficient for optimality in the convex setting.

[6] KKT stands for Karush-Kuhn-Tucker. Harold Kuhn (USA) and Albert Tucker (Canada) published these conditions in 1951, but it was found out years later that William Karush (USA) had already obtained these conditions in his 1939 Master thesis!

Constraint qualification and KKT conditions The results of both Theorem 2.6 and 2.7 are valid under strong duality. There exists various assumptions, termed **constraint qualifications**, under which strong duality is guaranteed (in which case we say that constraint qualification holds). We provide several examples below.

The most classical case of constraint qualifications is the case of linear constraints.

Theorem 2.8 *Suppose that all constraint functions in problem (2.2.1) are linear. Then, constraint qualification and strong duality hold.*

In particular, strong duality holds for linear programming.

When the constraint functions are differentiable, other conditions can be provided, that tie to the first KKT condition (2.2.10a). We state the two most classical ones below.

Definition 2.6 (LICQ) *Consider problem (2.2.1) under Assumption 2.1. Let $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ be a feasible point for the problem, and consider the set $\mathcal{A}(\bar{\boldsymbol{x}}) = \{i \in \{1, \dots, m\} \mid g_i(\bar{\boldsymbol{x}}) = 0\}$. We say that Linear Independence Constraint Qualification (LICQ) holds at $\bar{\boldsymbol{x}}$ if the vectors*

$$\{\nabla g_i(\bar{\boldsymbol{x}})\}_{i \in \mathcal{A}(\bar{\boldsymbol{x}})} \bigcup \{\nabla h_i(\bar{\boldsymbol{x}})\}_{i=1,\dots,\ell}$$

are linearly independent.

Definition 2.7 (MFCQ) *Consider problem (2.2.1) under Assumption 2.1. Let $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ be a feasible point for the problem, and consider the set $\mathcal{A}(\bar{\boldsymbol{x}}) = \{i \in \{1, \dots, m\} \mid g_i(\bar{\boldsymbol{x}}) = 0\}$. We say that Mangasarian-Fromovitz Constraint Qualification (MFCQ) holds at $\bar{\boldsymbol{x}}$ if*

i) The vectors $\{\nabla h_i(\bar{\boldsymbol{x}})\}_{i=1,\dots,\ell}$ are linearly independent, and

ii) There exists $\boldsymbol{d} \in \mathbb{R}^n$ such that

$$\begin{cases} \nabla g_i(\bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{d} < 0 & \forall i \in \mathcal{A}(\bar{\boldsymbol{x}}), \\ \nabla h_i(\bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{d} = 0 & \forall i = 1, \dots, \ell. \end{cases}$$

The Mangasarian-Fromovitz condition[7] implies the Linear Independence Constraint Qualification, but the converse is not true, as shown by the following example.

Example 2.6 *Consider the problem*

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x} \in \mathbb{R}^2} & \|\boldsymbol{x}\|^2 \\ \text{subject to} & x_1 \leq 0 \\ & x_1 \leq 0 \\ & x_2 = 0. \end{array}$$

The feasible set does not satisfy LICQ at $\boldsymbol{x}^* = \boldsymbol{0}$, however it satisfies MFCQ at this point (consider $\boldsymbol{d} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$).

As a final comment for this chapter, we note that it is theoretically possible to compute the primal-dual solution of an optimization problem by solving its KKT equations. However, the nature of these equations is such that the solution cannot be found in closed form in general. In practice, we thus resort to iterative algorithms to compute approximate solutions.

[7] Developed at Stanford by Olvi Mangasarian (born in Iraq) and Stan Fromovitz (born in Poland).

Chapter 3

Statistics and concentration inequalities

Probability theory and statistics are of primary importance in the context of data science. On one hand, it is generally useful to think of the data at hand as originating from a certain distribution, in order to infer future behavior of other samples emanating from the same distribution. On the other hand, the use of randomized algorithms in data science has become standard, due in part to the cheaper cost of these alternatives.

This chapter is concerned with deriving useful inequalities on random quantities, called **concentration inequalities**. Those are particularly useful to analyze random data and algorithms, and will be presented for random variables, vectors as well as matrices.

3.1 Basics of probability theory

The concept of probability originates from measure theory. All results in probability and statistics implicitly rely on probability spaces, i.e. triplets $(\Omega, \mathcal{A}, \mathbb{P})$, where

- Ω is a set of possible values, or outcomes;
- $\mathcal{A}$ is a family of subsets of Ω called set of events, that satisfy certain properties that make it a σ-algebra;
- $\mathbb{P} : \mathcal{A} \to [0, 1]$ is a probability measure, that satisfies in particular $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.

Given this definition, a random variable is a mapping from a probability space to another space that induces a new probability measure on the latter. The term *random variable* is often used for scalar quantities, thus we will make a distinction between several random quantities

- random variables y defined on a probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ by
$$\forall B \in \mathcal{B}(\mathbb{R}), \quad \mathbb{P}(y \in B) = \mathbb{P}(B);$$
- random vectors $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \cdots \\ y_d \end{bmatrix}$ of size d, defined on the probability space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P})$;
- random matrices $\boldsymbol{Y} = [Y_{ij}]$ of size $n \times d$, defined on the probability space $(\mathbb{R}^{n \times d}, \mathcal{B}(\mathbb{R}^d), \mathbb{P})$.

In both cases, the set of events will be the Borel σ-algebra $\mathcal{B}(\cdot)$.

3.1.1 Random variables

Although a generic study of random variables can be performed by considering them as taking a continuum of values, we begin by providing the more elementary definition of discrete random variables.

Definition 3.1 (Discrete random variable) *A discrete random variable y is defined by*

- *A discrete set of possible values $\mathcal{Y} = \{y_i\} \subset \mathbb{R}$;*
- *An associated set of probabilities $p = \{p_i\}$ such that $p_i \geq 0$, $\sum_i p_i = 1$ and*

$$\forall \mathcal{S} \subset \mathcal{Y}, \quad \mathbb{P}(y \in \mathcal{S}) = \sum_{y_i \in \mathcal{S}} p_i.$$

Definition 3.2 (Continuous random variable) *A continuous random variable y is defined by*

- *A continuous set of possible values $\mathcal{Y} \subset \mathbb{R}$;*
- *An associated probability density $p : \mathcal{Y} \to \mathbb{R}^+$ such that $\int_{\mathbb{R}} p(y)\, dy = 1$ and*

$$\forall \mathcal{S} \subset \mathcal{Y}, \quad \mathbb{P}(y \in \mathcal{S}) = \int_{y \in \mathcal{S}} p(y)\, dy.$$

For both continuous and discrete random variables, we will say that y follows a distribution characterized by $(p, \mathcal{Y})$, or simply p when the set of possible values is implicit from the definition of p.

Example 3.1 *A Gaussian/normally distributed random variable y of law $\mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma > 0$ is defined by the density*

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y-\mu)^2) \quad \forall y \in \mathbb{R}.$$

3.1.2 Moments

To understand the behavior of random variables, one can look at the moments of their distribution (provided they are well defined). The canonical example of such a quantity is the mean (also called the expected value) of a random variable.

Definition 3.3 (Expected value/Mean) *Let y be a random variable with a distribution $(p, \mathcal{Y})$, which we indicate as $y \sim p$. The* **expected value of** y *is defined by*

$$\mathbb{E}[y] = \mathbb{E}_y[y] = \begin{cases} \sum_{y_i \in \mathcal{Y}} y_i\, p(y = y_i) & \textit{(discrete case)} \\ \\ \int_{\mathcal{Y}} y\, p(y)\, dy & \textit{(continuous case).} \end{cases}$$

The expected value has several desirable properties that facilitate its use, especially the following.

Proposition 3.1 *The expected value is a linear operator: that is, for every random variable y and every $\alpha, \beta \in \mathbb{R}$, one has:*

$$\mathbb{E}[\alpha\, y + \beta] = \alpha\, \mathbb{E}[z] + \beta;$$

The expected value has several nice properties. In particular, for any convex function f, we have $\mathbb{E}[f(y)] \geq f(\mathbb{E}[y])$.

Definition 3.4 (Variance and standard deviation) *Let y be a random variable.*

- *The variance of y is defined by*

$$\mathrm{Var}[y] = \mathbb{E}\left[y^2\right] - \mathbb{E}[y]^2.$$

- *The standard deviation of y is the square root of the variance.*

Lemma 3.1

- *If y is a discrete random variable, then $\mathrm{Var}[y] = \sum_i p_i y_i^2 - [\sum_i p_i y_i]^2$;*
- *If y has zero mean, i.e. $\mathbb{E}[y] = 0$, then $\mathrm{Var}[y] = \mathbb{E}\left[y^2\right]$.*

3.2 From random variables to random vectors and matrices

Statistics are not only based on multiple instances of random variables, that can be structured under the form of vectors and/or matrices. In this section, we provide the tools necessary to understand such quantities.

3.2.1 Pair of random variables

When two random variables possess the same distribution on the same probability space, we say that those variables are **identically distributed**. In a general setting, one can study the distribution of the pair formed by two random variables.

Definition 3.5 (Joint distribution (discrete case)) *Let y and z be two discrete random variables taking values in $\mathcal{Y} = \{y_i\}$ and $\mathcal{Z} = \{z_j\}$, respectively. The distribution of the pair of random variables (y, z) is defined by*

- *The set of possible values $\mathcal{Y} \times \mathcal{Z} = \{(y_i, z_j)\}$;*
- *The discrete probability density $p = \{p_{i,j}\}$, where*

$$p_{i,j} = \mathbb{P}(y = y_i, z = z_j).$$

Definition 3.6 (Joint distribution (continuous case)) *Let y and z be two continuous random variables taking values in $\mathcal{Y}$ and $\mathcal{Z}$. The distribution of the pair of random variables (y, z) is defined by*

- *The set of possible values $\mathcal{Y} \times \mathcal{Z}$;*

- *The continuous probability density* $p : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^+$ *such that*

$$\int_y \int_z p(y,z)\, dy\, dz \;=\; 1.$$

In the above definitions, we started from two random variables to obtain the joint distribution of the pair formed by these variables. It is also possible to go the other way around, by defining marginal laws.

Definition 3.7 (Marginal laws (discrete case)) *Let* y *and* z *be two discrete random variables taking values in* $\mathcal{Y} = \{y_i\}$ *and* $\mathcal{Z} = \{z_j\}$*, respectively. Let* $\{p_{i,j}\}$ *be the joint distribution of* (y, z)*.*

- *The marginal law of* y *is given by* $\{p_{i\bullet}\}_i$*, where*

$$p_{i\bullet} := \mathbb{P}(y = y_i) = \sum_{j | z_j \in \mathcal{W}} \mathbb{P}(y = y_i,\ z = z_j) = \sum_j p_{i,j}.$$

- *Similarly, the marginal law of* w *is given by* $\{p_{\bullet j}\}_j$*, where*

$$p_{\bullet j} := \mathbb{P}(z = z_j) = \sum_{i | y_i \in \mathcal{Y}} \mathbb{P}(y = y_i,\ z = z_j) = \sum_i p_{i,j}.$$

Definition 3.8 (Marginal laws (continuous case)) *Let* y *and* z *be two continuous random variables taking values in* $\mathcal{Y}$ *and* $\mathcal{Z}$*, respectively. Let* $p : (y, z) \mapsto p(y, z)$ *be the joint density of* (y, z)*.*

- *The marginal law of* y*, denoted by* p_y *or* $p(y, \bullet)$*, is the function* $p_y : \mathcal{Y} \to \mathbb{R}^+$ *given by*

$$\forall y \in \mathcal{Y}, \quad p_y(y) = \int_{\mathcal{Z}} p(y,z)\, dz.$$

- *The marginal law of* z*, denoted by* p_z *or* $p(\bullet, z)$*, is the function* $p_z : \mathcal{Z} \to \mathbb{R}^+$ *given by*

$$\forall z \in \mathcal{Z}, \quad p_z(z) = \int_{\mathcal{Y}} p(y,z)\, dy.$$

Definition 3.9 (Covariance and correlation) *Let* y *and* z *be two random variables. The covariance of* y *and* z *is defined by*

$$\mathrm{Cov}[y,z] = \mathbb{E}_{y,z}\left[(y - \mathbb{E}[y])\,(z - \mathbb{E}[z])\right].$$

The correlation of y *and* z *is*

$$\mathrm{Corr}[y,z] = \frac{\mathrm{Cov}[y,z]}{\sqrt{\mathrm{Var}_y[y]}\sqrt{\mathrm{Var}_z[z]}}.$$

Independent random variables Independence is widely used in statistics, where it is often combined with the notion of identically distributed variables: we then say that the random variables are **i.i.d.**, which stands for "independent, identically distributed".

Definition 3.10 (Independent variables) *Let y and z be two random variables with distributions $(p_y, \mathcal{Y})$ and $(p_z, \mathcal{Z})$, respectively. The variables y and z are called* **independent** *if the pair (y, z) satisfies*

$$\forall \mathcal{S} \times \mathcal{T} \subset \mathcal{Y} \times \mathcal{Z}, \quad \mathbb{P}(y \in \mathcal{S}, z \in \mathcal{T}) \;=\; \mathbb{P}(y \in \mathcal{S})\,\mathbb{P}(z \in \mathcal{T}).$$

Independence allows for an easy characterization of the joint distribution, as illustrated by the following result.

Proposition 3.2 *Let y and z be two independent random variables. Then, their joint distribution is obtained as the product of the marginal distributions. We thus have*

$$\begin{cases} p_{ij} = p_{i\bullet} \times p_{\bullet j} & \text{(discrete case)} \\ p(y,z) = p_y(y) \times p_z(z) & \text{(continuous case).} \end{cases}$$

Proposition 3.3 *Let y and z be two independent random variables. Then, these values are decorrelated, i. e. $\mathrm{Cov}[y,z] = \mathrm{Corr}[y,z] = 0$.*

3.2.2 Random vectors

Most of the previous results on random variables can be extended to the case of **random vectors**, i.e. multidimensional random quantities. We provide below the basic concepts.

Definition 3.11 (Law of a random vector) *Let $\boldsymbol{y} = [y_i]_i$ be a random vector in $\mathbb{R}^n$: the law (or the distribution) of $\boldsymbol{y}$ is given by the joint distribution of its components. In particular, we define the following moments of this distribution:*

- *the expected value of $\boldsymbol{y}$ is the vector of the expected values of each component:*

$$\mathbb{E}[\boldsymbol{y}] = \{\mathbb{E}[y_i]\}_i \in \mathbb{R}^n;$$

  where the expected value is taken with respect to $\boldsymbol{y}$;

- *the* **covariance** *matrix of $\boldsymbol{y}$, denoted by $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ is the matrix of the covariances between each component*

$$\forall 1 \le i,j \le n, \quad [\boldsymbol{\Sigma}_{\boldsymbol{y}}]_{i,j} := \mathbb{E}[(y_i - \mathbb{E}[y_i])(y_j - \mathbb{E}[y_j])].$$

Note that the covariance matrix can be written as

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \mathbb{E}\left[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^{\mathrm{T}}\right] \in \mathbb{R}^{n \times n}.$$

Lemma 3.2 *If the components of a random vector are independent, then its covariance matrix is diagonal.*

Example 3.2 (Gaussian vectors) *A vector $\boldsymbol{y} \in \mathbb{R}^n$ is a Gaussian vector $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \in \mathcal{S}^n_{++}$ if its density is given by*

$$f(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^n (\det \boldsymbol{\Sigma})}} \exp\left\{-\tfrac{1}{2}\boldsymbol{y}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}\right\}.$$

When the components of $\boldsymbol{y}$ are Gaussian i.i.d. variables $\mathcal{N}(0, \sigma^2)$ with $\sigma > 0$, we further have

$$f(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left\{-\tfrac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2\right\}.$$

About random matrices In these notes, we will mostly not require specific distributions of random matrices, but will rather consider random matrices produced from random vectors or random variables. With that perspective, we note that the probability distribution of a matrix can be understood either as the joint distribution of its entries or that of its columns/rows.

3.3 Scalar concentration inequalities

Concentration inequalities are used to show that random variables concentrate around a certain interval of values.

3.3.1 Markov's inequality

Markov's inequality[1] is the most classical concentration inequality for random variable with finite expected value.

Theorem 3.1 (Markov's inequality) *Let $y \in \mathbb{R}$ be a random variable with $\mathbb{E}[|y|] < \infty$. Then, for any $t > 0$,*

$$\mathbb{P}(|y| \geq t) \leq \frac{\mathbb{E}[|y|]}{t}. \tag{3.3.1}$$

In the literature, one can find several results termed Markov's inequality. This includes the result below.

Theorem 3.2 (Markov's inequality (alternate form)) *Let $y \in \mathbb{R}_+$ be a random variable with $\mathbb{E}[y] < \infty$. Then, for any $t > 0$,*

$$\mathbb{P}(y \geq t) \leq \frac{\mathbb{E}[y]}{t}. \tag{3.3.2}$$

Although simple proofs of this result exist, we provide below a proof based on convex optimization, that shows that (3.3.1) provides the best bound based on the expected value.

Proof. We seek the tightest possible inequality of the form

$$\mathbb{P}(y \geq t) \leq x_1 + x_2 \mathbb{E}[y],$$

where $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$. This can be modeled as the following optimization problem:

$$\begin{cases} \text{minimize}_{\boldsymbol{x} \in \mathbb{R}^2} & x_1 + x_2 \mathbb{E}[z] \\ \text{s.t.} & f(z) = x_1 + x_2 z \geq 1 \quad \forall z \geq t \\ & f(z) = x_1 + x_2 z \geq 0 \quad \forall z \in [0, t). \end{cases} \tag{3.3.3}$$

Although the feasible set of this problem has infinitely many constraints, it admits the following simple description:

$$\left\{ \boldsymbol{x} \in \mathbb{R}^2 \;\middle|\; x_1 + t x_2 \geq 1,\ x_1 \geq 0, x_2 \geq 0 \right\}.$$

It can then be shown that problem (3.3.3) is convex, and that it has a unique solution given by $x_1^* = 0$ and $x_2^* = \frac{1}{t}$. Thus, the optimal inequality is

$$\mathbb{P}(y \geq t) \leq x_1^* + x_2^* \mathbb{E}[y] = \frac{\mathbb{E}[y]}{t}.$$

[1] Also called *Markov inequality*.

□

A related inequality, called Chebyshev's inequality, provides a bound based on variance.

Theorem 3.3 (Chebyshev's inequality) *Let y be a real random variable with $\mathrm{Var}[y] < \infty$. Then, for any $t > 0$,*

$$\mathbb{P}(|y - \mathbb{E}[y]| \geq t) \leq \frac{\mathrm{Var}[y]}{t^2}. \tag{3.3.4}$$

Proof. Since

$$\mathbb{P}(|y - \mathbb{E}[y]| \geq t) = \mathbb{P}\left(|y - \mathbb{E}[y]|^2 \geq t^2\right),$$

it suffices to apply Markov's inequality to the random variable $|y - \mathbb{E}[y]|^2$. This gives

$$\mathbb{P}\left(|y - \mathbb{E}[y]|^2 \geq t^2\right) \leq \frac{\mathbb{E}\left[|y - \mathbb{E}[y]|^2\right]}{t^2} = \frac{\mathrm{Var}[y]}{t^2},$$

where the last equality uses the definition of $\mathrm{Var}[y]$. □

Corollary 3.1 *Let y be a random variable such that $\mathbb{E}[y] < \infty$ and $\mathrm{Var}[y] = \sigma^2 \in \mathbb{R}_{++}$. Then, for any $t > 0$,*

$$\mathbb{P}(|y - \mu| \geq t\sigma) \leq \frac{1}{t^2}. \tag{3.3.5}$$

Example 3.3 *Let $y_1, \ldots, y_m$ be m random variables i.i.d. following a distribution with mean μ and variance $\sigma^2 \in \mathbb{R}_{++}$. Then, for any $t > 0$, we have*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} y_i - \mu\right| \geq t\right) \leq \frac{\sigma^2}{nt^2}.$$

3.3.2 Hoeffding's inequality

Hoeffding's inequality applies to binary-valued variables.

Definition 3.12 (Bernoulli variables)

- *A random variable y follows a* Bernoulli distribution *of parameter $p \in [0,1]$ if $\mathcal{Y} = \{0,1\}$ and $\mathbb{P}(y = 0) = 1 - p$, $\mathbb{P}(y = 1) = p$.*
- *A random variable y follows a* symmetric Bernoulli *or* Rademacher *distribution if $\mathcal{Y} = \{-1, 1\}$ and $\mathbb{P}(y = 1) = \mathbb{P}(y = -1) = \frac{1}{2}$.*

Theorem 3.4 (Hoeffding's inequality) *Let $y_1, \ldots, y_N$ be i.i.d. Rademacher variables. Then, for any $t \geq 0$ and any $\boldsymbol{a} \in \mathbb{R}^N$,*

$$\mathbb{P}\left(\sum_{i=1}^{N} a_i y_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\|\boldsymbol{a}\|^2}\right). \tag{3.3.6}$$

Since the right-hand side of (3.3.6) goes to zero exponentially fast as $t \to \infty$, the result of Theorem 3.4 guarantees that $\mathbb{P}\left(\sum_{i=1}^{N} a_i y_i \geq t\right) \ll \mathbb{P}\left(\sum_{i=1}^{N} a_i y_i < t\right)$ for sufficiently large t.

Remark 3.1 *The proof of Hoeffding's inequality is based on Markov's inequality applied to*

$$\mathbb{P}\left(\exp\left(\lambda\sum_{i=1}^{N}a_iy_i\right)\geq\exp\left(\lambda t\right)\right).$$

Similarly to the proof of Markov's inequality, it also involves a convex optimization problem over λ.

Similarly to Markov's inequality, a number of variants of Theorem 3.4 are also referred to as Hoeffding's inequality. We provide below two of these variants, the first one being a direct corollary of Theorem 3.4.

Corollary 3.2 *Let* $y_1,\dots,y_N$ *be i.i.d. Rademacher variables. Then, for any* $t\geq 0$,

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}y_i\geq t\right)\leq\exp\left(-\frac{N\,t^2}{2}\right). \tag{3.3.7}$$

The right-hand side of (3.3.7) goes to zero when $t\to\infty$, but also when $N\to\infty$. In the limit, this non-asymptotic result shows that $\frac{1}{N}\sum_{i=1}^{N}y_i\to 0$, i. e. the empirical mean converges to the mean of the y_is.

The second variant on Hoeffding's inequality shows that it applies to other distributions than Bernoulli distributions.

Theorem 3.5 *Let* $y_1,\dots,y_N$ *be i.i.d. variables such that* $m\leq y_i\leq M$ *for* $i=1,\dots,N$. *Then, for any* $t\geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{N}\left(y_i-\mathbb{E}\left[y_i\right]\right)\geq t\right)\leq\exp\left(-\frac{2t^2}{N\left(M-m\right)^2}\right). \tag{3.3.8}$$

3.3.3 Sub-gaussian random variables

So far we have obtained inequalities for certain distributions.

Definition 3.13 *A random variable* y *is called a* **sub-gaussian** *random variable if there exists a constant* $K>0$ *such that*

- $\mathbb{E}\left[\exp\left(\frac{y^2}{K^2}\right)\right]\leq 2$;
- *It exists* $c>0$ *such that for any* $t\geq 0$,

$$\mathbb{P}\left(|y|\geq t\right)\leq 2\exp\left(-c\frac{t^2}{K^2}\right).$$

In that case, we say that y *follows a sub-gaussian distribution.*

The smallest positive constant K *such that the two properties hold is denoted by* $\|y\|_{\Psi_2}$, *and called the sub-gaussian norm of* y.[2]

[2] This is actually a norm on the space of sub-gaussian distributions.

Note that the two items in the definition are actually equivalent.

The family of sub-gaussian distributions include classical examples of probabilty distributions, some of which are given below.

Example 3.4

- *Any Gaussian variable* $y \sim \mathcal{N}(0, \sigma^2)$ *is sub-gaussian, and* $\|y\|_{\Psi_2} \leq C\sigma$*, where* C *is a bound on the value of* $\|\cdot\|_{\Psi_2}$ *for a standard Gaussian variable.*
- *A Rademacher variable* y *is sub-gaussian with* $\|y\|_{\Psi_2} = \frac{1}{\sqrt{\ln(2)}}$.
- *Any bounded random variable* y *is sub-gaussian with* $\|y\|_{\Psi_2} \leq \frac{1}{\sqrt{\ln(2)}}\|y\|_\infty$*, where* $\|y\|_\infty = \max\{|y| \,|\, y \in \mathcal{Y}\}$.

Lemma 3.3 *If* y *is a sub-gaussian random variable and* $a \in \mathbb{R}$*, then* $y + a$ *is also a sub-gaussian random variable.*

The properties of sub-gaussian random variables allow for deriving yet other variants of Hoeffding's inequality.

Theorem 3.6 *Let* $y_1, \ldots, y_N$ *be independent, sub-gaussian random variables with zero mean (*$\mathbb{E}[y_1] = \cdots = \mathbb{E}[y_N] = 0$*). Then, for any* $t \geq 0$ *and any* $\boldsymbol{a} \in \mathbb{R}^N$*,*

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i y_i\right| \geq t\right) \leq 2\exp\left(-\frac{c\,t^2}{\|\boldsymbol{a}\|^2 \sum_{i=1}^N \|y_i\|_{\Psi_2}^2}\right), \tag{3.3.9}$$

where $c > 0$ *is a universal positive constant, that does not depend on* n *nor* t*.*

A more general version of this inequality applies to sub-gaussian variables with nonzero mean, by applying Theorem 3.6 to the variables $\{y_i - \mathbb{E}[y_i]\}_{i=1}^N$.

3.3.4 Sub-exponential random variables

Sub-gaussian random variables cover many classical distributions, but not all. In particular, the square of a sub-gaussian random variable is not sub-gaussian, yet we expect the square variable to concentrate if the original variable does. This leads to the following concept.

Definition 3.14 *A random variable* y *is called a* **sub-exponential** *random variable if there exists a constant* $K > 0$ *such that*

- $\mathbb{E}\left[\exp\left(\frac{|y|}{K}\right)\right] \leq 2$;
- *It exists* $c > 0$ *such that for any* $t \geq 0$*,*

$$\mathbb{P}(|y| \geq t) \leq 2\exp\left(-c\frac{t}{K}\right).$$

In that case, we say that y follows a sub-exponential distribution.

The smallest positive constant K such that the two properties hold is denoted by $\|y\|_{\Psi_1}$, and called the sub-exponential norm of y.[3]

As expected, the square of a sub-gaussian random variable is sub-exponential. The connection between the two families is even stronger, as shown by the following result.

Proposition 3.4 *A random variable y is sub-gaussian if and only if the random variable y^2 is sub-exponential.*

We now give a concentration inequality for sub-exponential variables.

Theorem 3.7 (Bernstein's inequality) *Let $y_1, \dots, y_N$ be independent, sub-exponential random variables with zero mean ($\mathbb{E}[y_1] = \cdots = \mathbb{E}[y_N] = 0$). Then, for any $t \geq 0$ and any $\boldsymbol{a} \in \mathbb{R}^N$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i y_i\right| \geq t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2}{\|\boldsymbol{a}\|^2 \max_{1\leq i\leq N}\|y_i\|_{\Psi_1}^2}, \frac{t}{\|\boldsymbol{a}\|_\infty \max_{1\leq i\leq N}\|y_i\|_{\Psi_1}}\right\}\right), \tag{3.3.10}$$

where $c > 0$ is a universal positive constant, that does not depend on n nor t.

Similarly to Hoeffding's inequality, there exist several variants of Bernstein's inequality.

Corollary 3.3 *Let $y_1, \dots, y_N$ be independent, sub-exponential random variables with zero mean ($\mathbb{E}[y_1] = \cdots = \mathbb{E}[y_N] = 0$). Then, for any $t \geq 0$,*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N y_i\right| \geq t\right) \leq 2\exp\left(-cN\min\left\{\frac{t^2}{\max_{1\leq i\leq N}\|y_i\|_{\Psi_1}^2}, \frac{t}{\max_{1\leq i\leq N}\|y_i\|_{\Psi_1}}\right\}\right), \tag{3.3.11}$$

[3] This is actually a norm on the space of sub-exponential distributions.

Bibliography

[1] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, United Kingdom, 2004.

[2] M. Delfour. *Introduction à l'optimisation et au calcul semi-différentiel*. Mathématiques appliquées pour le Master/SMAI. Dunod, 2012.

[3] A. C. Gilbert M. W. Mahoney, J. C. Duchi, editor. *The mathematics of data*. Number 25 in IAS/Park City Mathematics Series. AMS, IAS/Park City Mathematics Institute, and Society for Industrial and Applied Mathematics, Princeton, 2018.

[4] D. P. Robinson. Convex analysis. Department of Applied Mathematics and Statistics, The Johns Hopkins University, 2017.

[5] R. Vershynin. *High-dimensional probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Appendix A

English VS French: Mathematical terminology

We provide here a short dictionary for the key mathematical terms used in these notes (commonly used in academia as well as industry). Note that we adopted the American wording and spelling rather than the British ones.

Accumulation point/Limit point	*Valeur d'adhérence*
Affine/Linear (depend on authors)	*Affine*
Chain rule	*Dérivée d'une composition*
Classifier	*Classificateur*
Closure	*Adhérence*
Coercive	*Croissante à l'infini/*0*-coercive*
Convex conjugate (function)	*Fonction conjuguée*
Convex/affine hull	*Enveloppe affine/convexe*
Feasible	*Réalisable (pour un problème)/Admissible (pour un point)*
Infimum/supremum	*Borne inférieure/supérieure*
Inner product/Dot product	*Produit scalaire*
Kernel space/Null space	*Noyau*
Likelihood	*Vraisemblance*
Linear subspace	*Sous-espace vectoriel*
(L-)Lipschitz continuous function	*Fonction (*L*-)lipschitzienne*
Map	*Application*
Mean squared error	*Erreur quadratique moyenne*
Nonnegative/Non-negative	*Positif ou nul*
Positive	*Strictement positif*
Proper function	*Fonction propre*
Quasiconvex/Quasiconcave	*Quasi-convexe/quasi-concave*
Quasilinear	*Quasi-linéaire*
Range space	*Image*
Ridge regression	*Régularisation écrêtée (*ℓ_2*)*
Self-concordant	*Autoconcordant(e)*
Sequence (subsequence)	*Suite (extraite)*
Slack variables	*Variable d'écart*
Sublevel/Superlevel set	*Section inférieure/supérieure*.

Remark A.1 *There exists another notion of* fonction propre *corresponding to the English eigenfunction, which we will not use in this course.*

Remark A.2 *The terminology* Section inférieure/supérieure *follows Delfour [2]. Other authors sometimes use* sous/sur-ensembles de niveau*, i.e. a direct translation from the English terms.*

Remark A.3 *The term* quasi-linéaire *might appear as* quasi-affine *in the literature.s*

Notations The $\log$ notation is sometimes used in American (and computer science more globally) literature to denote the base e logarithm. In the French literature, the former is almost exclusively used for base 10 logarithm.