Dauphine | PSL
UNIVERSITÉ PARIS

Optimization for Machine Learning

Clément W. Royer

Lecture notes - M2 MIAGE ID Apprentissage - 2023/2024

- The last version of these notes can be found at:
  https://www.lamsade.dauphine.fr/~croyer/ensdocs/OML/PolyOML.pdf.
- Comments, typos, etc, can be sent to `clement.royer@lamsade.dauphine.fr`.
  Thanks to the students who sent feedback. Thanks also to Sébastien Kerleau and Florian Le Bronnec for their feedback.
- **Version history:**
  - 2023.11.21: Fixed typo in Section 3.2.3.
  - 2023.11.02: Revised version with significant changes in Chapter 2 (corresponding to the second lecture and tutorial).
  - 2023.08.28: First version of these notes.
- **Learning goals:**
  - Understand the specifics of optimization problems, and the interest of certain formulations over others.
  - Given an optimization problem, select an algorithm well suited for solving the problem.
  - Analyze the theoretical and practical properties of a particular algorithm.
  - Identify challenges posed by optimization in a data science context, and ways to address these challenges.

Note: This course was previously taught in M2 MIAGE ID Apprentissage under the name "Optimization for Data and Decision Sciences".

Contents

Chapter 1

Introduction to optimization

1.1 About optimization

Optimization is a field of study that is concerned with making the best possible decision out of a set of alternatives. *Mathematical* optimization provides a framework to model optimization problems using mathematical language. We provide below a broad definition of optimization problems in the mathematical sense, and connect it to the concept of solving such a problem.

1.1.1 The optimization process

Most optimization problems are readily formulated in plain language, and it can be difficult to extract the core ideas behind the optimization task. On the contrary, a mathematical optimization problem is a well-identified object, that corresponds to the following definition.

Definition 1.1.1 *A mathematical optimization problem consists of three key components:*

- *An* **objective function***, that quantifies the quality of a decision. If a better decision yields a smaller objective value, the problem is called a* minimization *problem. On the other hand, if a better decision yields a larger objective value, the problem is called a* maximization *problem.*
- *A number of* **decision variables***: those are the knobs that can be turned to change the decision and, as a result, the value of the objective function. The goal of optimization is to determine what are the best (optimal) values of the decision variables relatively to the objective function.*
- *A set of* **constraints***, that specify requirements on the decision variables that must be satisfied for the decision to be valid.*

An optimization procedure consists in minimizing or maximizing an objective function with respect to the decision variables subject to constraints on those variables.

Remark 1.1.1 *In practice, many optimization problems are subject to hidden or implicit constraints, that may not appear in the optimization formulation yet can have a significant effect on an algorithm's performance. The absence of these constraints in the formulation may be due to misspecification (not including a positivity constraint for a variable corresponding to a mass, for instance), but this phenomenon often has a more subtle cause. In simulation-based optimization, where the objective function corresponds to running expensive computational codes, an error may be triggered by certain*

values of the decision variables that may not be known a priori. *Such constraints must then be addressed by an algorithm in real time.*

Any concrete problem goes through a **modeling phase** that produces a mathematical optimization problem. In order to solve this problem, we then use **optimization algorithms**, often designed to be implemented and run on a computer. Such methods typically output a candidate solution (or, possibly, that no solution exists), which may or may not form an appropriate answer to the problem. If needed, the optimization process can go over another modeling phase followed by another solve. In practice, it is typical that the optimization process includes a **discussion phase** with experts from the application domain.

Remark 1.1.2 *Numerical optimization typically obeys the following principles:*

- *There is no universal algorithm. Every method can be efficient on a certain class of problems and perform poorly on another class. Studying the structure of a problem of interest helps in choosing the best algorithm for solving this problem.*

- *There may be a big gap between theory and practice. Finite-precision arithmetic can introduce round-off errors that result in worse practical performance than what the theory predicts. Conversely, theory (such as complexity bounds introduced below) can be overly pessimistic, and much better results can be observed in practice.*

- *Theory informs practice, and vice-versa. For most optimization problems, one can use mathematical formulas to assess whether a solution has been found: these expressions are at the heart of most of the methods we will describe in this course. More broadly, theoretical guarantees can help guiding the optimization process beyond pure heuristics and guesses. Meanwhile, some algorithms have demonstrated success without being endowed with theoretical properties: such phenomena are common, and motivate researchers to explain this behavior using mathematical tools.*

1.1.2 Modern optimization

Numerical optimization started during the 1940s. Major theoretical advances were achieved during the 1980s, along with significant algorithmic developments, that leverage the limited computing power of this era. The next two decades saw the rise of computing power, leading to numerous successes of optimization techniques. This trend continues to this day, albeit with a key change of paradigm.

Indeed, as data becomes more prevalent, optimization problems now involve massive amounts of data in the calculation of their objective. In addition to the challenges posed by the use of such datasets, possibly in a distributed fashion, other difficulties arise due to the fact that the solutions of these problems must generalize to yet unseen data.

In this course, we will go on a *tour d'horizon* of optimization problems and algorithms, with a focus on methods that have been successful in classical industrial settings as well as new paradigms such as machine learning. Our presentation will cover theoretical, algorithmic and practical aspects.

1.2 The optimization problem

We now introduce the mathematical foundations behind optimization. As a field of study, optimization is defined as the process of making the best decision out of a set of alternatives.

Mathematically, we write an optimization problem using three components:

- An **objective function**, i.e. a criterion that measures how good a given decision is, that we want to minimize or maximize depending on the context;
- **Decision variables**, that represent the knobs we can turn to change the decision;
- **Constraints**, i.e. conditions that the decision variables must satisfy in order for the decision to be acceptable.

The general form of the optimization problems considered in these notes will be the following

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w}) \quad \text{subject to} \quad \boldsymbol{w} \in \mathcal{F}. \tag{1.2.1}$$

In problem (1.2.1), f is the objective function (that we want to minimize), $\boldsymbol{w}$ is the vector of decision variables defined over a definition set $\mathbb{R}^d$, and $\mathcal{F}$ is a set encompassing all the constraints on the decision variables. This set is called the feasible set, and is often described using mathematical expressions.

There are multiple ways of formulating an optimization problem given a description of this problem in plain language: constraints can be expressed in the definition set or in the objective, maximization can be preferred over minimization, redundant constraints can be added to the problem, etc. Some formulations will be well suited for theoretical analysis, while others will be efficiently solvable by modern software.

Definition 1.2.1 *Two mathematical optimization problems are called* equivalent *if the solution of one is readily obtained from the solution of the other, and vice-versa.*

For instance, the minimization problem (1.2.1) has an equivalent reformulation as a maximization problem when $f : \mathbb{R}^d \to \mathbb{R}$ (i.e. f outputs numerical values):

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{maximize}} -f(\boldsymbol{w}) \quad \text{subject to} \quad \boldsymbol{w} \in \mathcal{F}.$$

1.2.1 Mathematical background

Optimization draws from several fields of mathematics, mostly pertaining to linear algebra, topology and differential calculus. We briefly review the key definitions below.

We will always consider $\mathbb{R}^d$ and $\mathbb{R}^{n\times d}$ as endowed with their canonical normed vector space structure; in particular, this means that we will be able to add two vectors (or two matrices), and to multiply a vector (or a matrix) by a scalar value. It also implies that we can measure the distance between two elements of these spaces using norms, the most classical of which is defined below.

Definition 1.2.2 (Euclidean norm on $\mathbb{R}^d$) *The Euclidean norm (or ℓ_2 norm) of a vector $\boldsymbol{w} \in \mathbb{R}^d$ is given by:*

$$\|\boldsymbol{w}\| := \sqrt{\sum_{i=1}^{d} w_i^2}.$$

Definition 1.2.3 (Scalar product on $\mathbb{R}^d$) *The scalar product is defined for every $\boldsymbol{w}, \boldsymbol{z} \in \mathbb{R}^d$ by:*

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z} := \sum_{i=1}^{d} w_i\, z_i.$$

One thus has $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z} = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{w}$ and $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = \|\boldsymbol{w}\|^2$.

There are natural counterparts to the Euclidean norm and its associated scalar product in a matrix space.

Definition 1.2.4 (Frobenius norm) *For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, the Frobenius norm is defined by*

$$\|\boldsymbol{A}\|_F := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} A_{ij}^2} = \sqrt{\mathrm{trace}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}.$$

The associated scalar product is the mapping

$$(\boldsymbol{A}, \boldsymbol{B}) \in \left(\mathbb{R}^{n\times d}\right)^2 \longmapsto \mathrm{trace}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B}).$$

In this course, we will mainly rely on Euclidean norms to measure distance between vectors, and algorithmic behavior. However, other norms will be used.

Definition 1.2.5 (ℓ_1 norm) *The ℓ_1 norm of a vector $\boldsymbol{w} \in \mathbb{R}^d$ is defined by:*

$$\|\boldsymbol{w}\|_1 := \sum_{i=1}^{d} |w_i|.$$

Both the ℓ_1 and ℓ_2 norms are special cases of the ℓ_p norm, defined for any $p \in [1, \infty)$ by

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \quad \|\boldsymbol{w}\|_p := \left[\sum_{i=1}^{d} |w_i|^p\right]^{1/p}.$$

Definition 1.2.6 (ℓ_∞ norm) *The ℓ_∞ norm of a vector $\boldsymbol{w} \in \mathbb{R}^d$ is defined by:*

$$\|\boldsymbol{w}\|_\infty := \max_{1\le i\le d} |w_i|.$$

Remark 1.2.1 *Interestingly, all norms can be defined as optimal values of optimization problems. This is clear from the definition of the ℓ_∞ norm. Below are two optimization problems that yield this optimal value:*

$$\underset{i\in\mathbb{R}}{\text{maximize}}\, |w_i| \quad \textit{s.t.} \quad i \in \{1, \dots, d\}.$$

$$\underset{v\in\mathbb{R}}{\text{maximize}}\, |v| \quad \textit{s.t.} \quad v \in \{w_1, \dots, w_d\}.$$

Notice that these two problems do not have the same decision variables or constraints, however they both possess a combinatorial structure. [1]

[1] By convention, in these notes, we will always consider optimization problems with real, continuous decision variables. As illustrated by the ℓ_∞ examples, it is indeed possible to model a discrete decision-making problem using continuous formulations. Our focus in these notes in on continuous optimization problems.

Definition 1.2.7 (Matrix inversion) *A matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is* invertible *if it exists $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ such that $\boldsymbol{BA} = \boldsymbol{AB} = \mathbf{I}_d$, where $\mathbf{I}_d$ is the identity matrix of $\mathbb{R}^{d\times d}$.*

In this case, $\boldsymbol{B}$ is the unique matrix with this property: $\boldsymbol{B}$ is called the inverse matrix of $\boldsymbol{A}$*, and is denoted by $\boldsymbol{A}^{-1}$.*

Definition 1.2.8 (Positive (semi-)definiteness) *A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is* positive semidefinite *if*

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

It is called positive definite *when $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} > 0$ for every nonzero vector $\boldsymbol{x}$.*

Definition 1.2.9 (Eigenvalues and eigenvectors) *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. A real λ is called an* eigenvalue of $\boldsymbol{A}$ *if*

$$\exists \boldsymbol{v} \in \mathbb{R}^d, \|\boldsymbol{v}\| \neq 0, \qquad \boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}.$$

The vector $\boldsymbol{v}$ is then called an eigenvector of $\boldsymbol{A}$ *associated to the eigenvalue λ.*

Theorem 1.2.1 *Any symmetric matrix in $\mathbb{R}^{d\times d}$ possesses d real eigenvalues.*

Notation 1.2.1 *Given two symmetric matrices $(\boldsymbol{A}, \boldsymbol{B}) \in \mathbb{R}^{d\times d}$, we introduce the following notations:*

- $\lambda_{\min}(\boldsymbol{A})/\lambda_{\max}(\boldsymbol{A})$*: smallest/largest eigenvalue of $\boldsymbol{A}$;*
- $\boldsymbol{A} \succeq \boldsymbol{B} \;\Leftrightarrow\; \lambda_{\min}(\boldsymbol{A}) \geq \lambda_{\max}(\boldsymbol{B})$*;*
- $\boldsymbol{A} \succ \boldsymbol{B} \Leftrightarrow \lambda_{\min}(\boldsymbol{A}) > \lambda_{\max}(\boldsymbol{B})$.

Following these notations, a matrix $\boldsymbol{A}$ is positive semi-definite (resp. positive definite) if and only if $\boldsymbol{A} \succeq 0$ (resp. $\boldsymbol{A} \succ 0$).

Differential calculus We will mostly consider minimization problems involving a smooth objective function: the term "smooth" can be loosely defined in the optimization or learning literature, but generally means that the function is as regular as needed for the desired algorithms and analysis to be applicable. In these notes, we will consider that a smooth function is at least continuously differentiable.

Definition 1.2.10 (Continuous function) *A function $f : \mathbb{R}^d \to \mathbb{R}^m$ is continuous at $\boldsymbol{w} \in \mathbb{R}^d$ if for every $\epsilon > 0$, it exists $\delta > 0$ such that*

$$\forall \boldsymbol{v} \in \mathbb{R}^d, \; \|\boldsymbol{v} - \boldsymbol{w}\| \leq \delta \implies \|f(\boldsymbol{v}) - f(\boldsymbol{w})\| \leq \epsilon.$$

Definition 1.2.11 (Lipschitz continuous function) *A function $f : \mathbb{R}^d \to \mathbb{R}^m$ is L-Lipschitz continuous over $\mathbb{R}^d$ if*

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \left(\mathbb{R}^d\right)^2, \quad \|f(\boldsymbol{u}) - f(\boldsymbol{v})\| \leq L \|\boldsymbol{u} - \boldsymbol{v}\|,$$

where $L > 0$ is called a Lipschitz constant.

Note that every Lipschitz continuous function is continuous.

Derivatives are ubiquitous in continuous optimization, as they allow to characterize the local behavior of a function. We assume that the reader is familiar with the concept of derivative of a function from $\mathbb{R} \to \mathbb{R}$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called *differentiable* at $\boldsymbol{w} \in \mathbb{R}^d$ if all its partial derivatives at $\boldsymbol{w}$ exist.

Definition 1.2.12 (Classes of functions)
- *A function $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable if its first-order derivative exists and is continuous. The set of continously differentiable functions is denoted by $\mathcal{C}^1(\mathbb{R}^d)$.*
- *A function $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable if $f \in \mathcal{C}^1(\mathbb{R}^d)$, its second-order derivative of f exists and is continuous. The set of twice continously differentiable functions is denoted by $\mathcal{C}^2(\mathbb{R}^d)$.*

Definition 1.2.13 (First-order derivative) *Let $f \in \mathcal{C}^1(\mathbb{R}^d)$ be a continuously differentiable function. For any $\boldsymbol{w} \in \mathbb{R}^d$, the* **gradient of f at $\boldsymbol{w}$** *is given by*

$$\nabla f(\boldsymbol{w}) := \left[\frac{\partial f}{\partial w_i}(\boldsymbol{w})\right]_{1 \leq i \leq d} \in \mathbb{R}^d.$$

Definition 1.2.14 (Second-order derivative) *Let $f \in \mathcal{C}^2(\mathbb{R}^d)$ be a twice continuously differentiable function. For any $\boldsymbol{w} \in \mathbb{R}^d$, the* **Hessian of f at $\boldsymbol{w}$** *is given by*

$$\nabla^2 f(\boldsymbol{w}) := \left[\frac{\partial^2 f}{\partial w_i \partial w_j}(\boldsymbol{w})\right]_{1 \leq i,j \leq d} \in \mathbb{R}^{d \times d}.$$

The Hessian matrix is symmetric.

Finally, we define an important class of problems involving a Lipschitz continuity assumption.

Definition 1.2.15 (Smooth functions with Lipschitz derivatives)
- *Given $L > 0$, the set $\mathcal{C}_L^{1,1}(\mathbb{R}^d)$ represents the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that belong to $\mathcal{C}^1(\mathbb{R}^d)$ such that ∇f is L-Lipschitz continuous.*
- *Given $L > 0$, the set $\mathcal{C}_L^{2,2}(\mathbb{R}^d)$ represents the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that belong to $\mathcal{C}^2(\mathbb{R}^d)$ such that $\nabla^2 f$ is L-Lipschitz continuous.*

An important property of such functions is that one can derive upper approximations on their values, as shown by the following theorem.

Theorem 1.2.2 (First-order Taylor expansion) *Let $f \in C_L^{1,1}(\mathbb{R}^d)$ with $L > 0$. For any vectors $\boldsymbol{w}, \boldsymbol{z} \in \mathbb{R}^d$, one has:*

$$f(\boldsymbol{z}) \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{z} - \boldsymbol{w}) + \frac{L}{2}\|\boldsymbol{z} - \boldsymbol{w}\|^2. \tag{1.2.2}$$

1.2.2 First definitions

Having defined an optimization problem as a mathematical object, we are now able to consider solving this problem. For simplicity, we will focus on minimization problems of the form:

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w}) \quad \text{s.t.} \boldsymbol{w} \in \mathcal{F}, \tag{1.2.3}$$

with $f : \mathbb{R}^d \to \mathbb{R}$ and $\mathcal{F} \subseteq \mathbb{R}^d$. Similar definitions hold for maximization problems.

Definition 1.2.16 (Feasible point) *A point $\boldsymbol{w} \in \mathbb{R}^d$ is called a feasible point of the optimization problem (1.2.3) if $\boldsymbol{w} \in \mathcal{F}$.*

If $\boldsymbol{w} \notin \mathcal{F}$, we say that $\boldsymbol{w}$ is infeasible. If $\mathcal{F}$ is empty, the problem (1.2.3) is said to be infeasible. For some optimization problems, the objective function may not be defined at infeasible points.

A solution of problem (1.2.3) must be a feasible point by definition. It should also lead to the best value in terms of cost function, hence the following definition.

Definition 1.2.17 (Global minimum) *A point $\boldsymbol{w}^* \in \mathbb{R}^d$ is called a solution or a global minimum of problem (1.2.3) if*

1. *$\boldsymbol{w}^*$ is feasible, i.e. $\boldsymbol{w}^* \in \mathcal{F}$;*
2. *$f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})$ for any feasible point $\boldsymbol{w} \in \mathcal{F}$.*

The set of global minima of problem (1.2.3) will be denoted by

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{argmin}} \{f(\boldsymbol{w}) \mid \boldsymbol{w} \in \mathcal{F}\} \subset \mathbb{R}^d, \tag{1.2.4}$$

and the minimal value of the problem will be defined as

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \{f(\boldsymbol{w}) \mid \boldsymbol{w} \in \mathcal{F}\} \in \mathbb{R} \cup \{-\infty, +\infty\}. \tag{1.2.5}$$

By convention, if f is unbounded below on $\mathcal{F} \neq \emptyset$, the minimal value is set to $-\infty$, while it is set to $+\infty$ if the problem is infeasible.

Definition 1.2.18 (Local minimum) *A point $\boldsymbol{w}^* \in \mathbb{R}^d$ is called a local minimum of problem (1.2.3) if*

1. *$\boldsymbol{w}^*$ is feasible, i.e. $\boldsymbol{w}^* \in \mathcal{F}$;*
2. *There exists $\epsilon > 0$ such that $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})$ for any feasible point $\boldsymbol{w} \in \mathcal{F}$ close to $\boldsymbol{w}^*$ in the sense that $\|\boldsymbol{w} - \boldsymbol{w}^*\| \leq \epsilon$.*

The notion of local minimum is weaker than that of global minimum. In practice, however, it is often more reasonable to seek local minima, as those can be characterized by mathematical conditions.

Optimality conditions In general, finding global or even local minima is a hard problem. For this reason, researchers in optimization have developed optimality conditions. These are mathematical expressions that can be checked at a given point (unlike the conditions above) and help assessing whether a given point is a local minimum or not.

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w}), \tag{1.2.6}$$

Theorem 1.2.3 (First-order necessary condition) *Suppose that the objective function f in problem (1.2.6) belongs to $\mathcal{C}^1(\mathbb{R}^d)$. Then,*

$$[\,\boldsymbol{w}^* \text{ is a local minimum of } f\,] \implies \|\nabla f(\boldsymbol{w}^*)\| = 0. \tag{1.2.7}$$

Note that this condition is only necessary: there may exist points with zero gradient that are not local minima. Indeed, the set of points with zero gradient, called *first-order stationary points*, also includes local maxima and saddle points[2].

Provided we strengthen our smoothness requirements on f, we can establish stronger optimality conditions for problem (1.2.6).

Theorem 1.2.4 (Second-order necessary condition) *Suppose that the objective function f in problem (1.2.6) belongs to $\mathcal{C}^2(\mathbb{R}^d)$. Then,*

$$[\,\boldsymbol{w}^* \text{ is a local minimum of } f\,] \implies \left[\|\nabla f(\boldsymbol{w}^*)\| = 0 \quad \text{and} \quad \nabla^2 f(\boldsymbol{w}^*) \succeq \mathbf{0}\right]. \tag{1.2.8}$$

From Theorem 1.2.3, first-order stationary points that violate the condition $\nabla^2 f(\boldsymbol{w}^*) \succeq \mathbf{0}$ cannot be local minima. This new condition is thus more precise than the first-order condition, although it remains a necessary. For an arbitrary function, there may exist points with zero gradient and positive semidefinite Hessian (termed **second-order stationary points**) that are not local minima.

On the other hand, and unlike first-order conditions, it is possible to derive a *sufficient* version of the second-order optimality conditions, that can be used to certify local optimality.

Theorem 1.2.5 (Second-order sufficient condition) *Suppose that the objective function f in problem (1.2.6) belongs to $\mathcal{C}^2(\mathbb{R}^d)$. Then,*

$$\left[\|\nabla f(\boldsymbol{w}^*)\| = 0 \quad \text{and} \quad \nabla^2 f(\boldsymbol{w}^*) \succ \mathbf{0}\right] \implies [\,\boldsymbol{w}^* \text{ is a local minimum of } f\,] \tag{1.2.9}$$

By exploiting the second-order derivative, it is thus possible to certify whether a point is a local minima (note that there could be local minima such that $\nabla^2 f(\boldsymbol{w}^*) \succeq \mathbf{0}$). With further assumptions on the structure of the problem, these optimality conditions can be more informative about minima. This is the case when the objective function is convex: we detail this property in the next section.

1.2.3 Convexity

Convexity is at its core a geometric notion: before defining what a convex function is, we describe the corresponding property for a set.

[2] A vector is a saddle point of a function if it is a local minimum with respect to certain directions and a local maximum with respect to other directions of the space.

Definition 1.2.19 (Convex set) *A set $\mathcal{C} \in \mathbb{R}^d$ is called* **convex** *if*

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{C}^2, \ \forall t \in [0,1], \qquad t\boldsymbol{u} + (1-t)\boldsymbol{v} \in \mathcal{C}.$$

Example 1.2.1 (Examples of convex sets) *The following sets are convex:*

- $\mathbb{R}^d$;
- *Every line segment of the form* $\{t\boldsymbol{w} | t \in \mathbb{R}\}$ *for some* $\boldsymbol{w} \in \mathbb{R}^d$;
- *Every (Euclidean) ball of the form* $\left\{ \boldsymbol{w} \in \mathbb{R}^d \,\middle|\, \|\boldsymbol{w}\|_2^2 = \sum_{i=1}^d [\boldsymbol{w}]_i^2 \leq 1 \right\}$.

We now provide the basic definition of a convex function.

Definition 1.2.20 (Convex function) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is* **convex** *if*

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2, \ \forall t \in [0,1], \qquad f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \leq t\, f(\boldsymbol{u}) + (1-t)\, f(\boldsymbol{v}). \tag{1.2.10}$$

Example 1.2.2 *The following functions are convex :*

- *Linear functions of the form* $\boldsymbol{w} \mapsto \boldsymbol{a}^{\mathrm{T}}\boldsymbol{w} + b$, *with* $\boldsymbol{a} \in \mathbb{R}^d$ *and* $b \in \mathbb{R}$;
- *Squared Euclidean norm:* $\boldsymbol{w} \mapsto \|\boldsymbol{w}\|_2^2 = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$.

If we consider differentiable functions, it is possible to characterize convexity using the derivatives of the function.

Theorem 1.2.6 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an element of $\mathcal{C}^1(\mathbb{R}^d)$. Then, the function f is convex if and only if*

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \quad f(v) \geq f(u) + \nabla f(u)^{\mathrm{T}}(v-u). \tag{1.2.11}$$

Theorem 1.2.7 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an element of $\mathcal{C}^2(\mathbb{R}^d)$. Then, the function f is convex if and only if*

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \quad \nabla^2 f(\boldsymbol{w}) \succeq \mathbf{0}. \tag{1.2.12}$$

Convex functions are particularly suitable for minimization problems as they satisfy the following property.

Theorem 1.2.8 *If f is a convex function, then every local minimum of f is a global minimum.*

If the function is differentiable, the optimality conditions as well as the characterization of convexity lead us to the following result.

Corollary 1.2.1 *If f is continuously differentiable, every point $\boldsymbol{w}^*$ such that $\|\nabla f(\boldsymbol{w}^*)\| = 0$ is a global minimum of f.*

Strong convexity The results above can be further improved by assuming that a convex function is strongly convex, as defined below.

Definition 1.2.21 (Strongly convex function) *A function $f : \mathbb{R}^d \to \mathbb{R}$ in $\mathcal{C}^1$ is μ-strongly convex (or* strongly convex of modulus $\mu > 0$*) if for all $(\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2$ and $t \in [0, 1]$,*

$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \;\leq\; t\, f(\boldsymbol{u}) + (1-t) f(\boldsymbol{v}) - \frac{\mu}{2} t(1-t)\|\boldsymbol{v} - \boldsymbol{u}\|^2.$$

Theorem 1.2.9 *Any strongly convex function has a unique global minimizer.*

As for convex functions, there exist characterizations of strong convexity that involve derivatives.

Theorem 1.2.10 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an element of $\mathcal{C}^1(\mathbb{R}^d)$. Then, the function f is μ-strongly convex if and only if*

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \quad f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{v} - \boldsymbol{u}) + \frac{\mu}{2}\|\boldsymbol{v} - \boldsymbol{u}\|^2. \tag{1.2.13}$$

Theorem 1.2.11 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an element of $\mathcal{C}^2(\mathbb{R}^d)$. Then, the function f is μ-strongly convex if and only if*

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \quad \nabla^2 f(\boldsymbol{w}) \succeq \mu \mathbf{I}. \tag{1.2.14}$$

1.3 Optimization algorithms

The field of optimization can be broadly divided into three categories:

- Mathematical optimization is concerned with the theoretical study of complex optimization formulations, and the proof of well-posedness of such problems (for instance, prove that their exist solutions);
- Computational optimization deals with the development of software that can solve a family of optimization problems, through careful implementation of efficient methods;
- **Algorithmic** optimization lies in-between the previous two categories, and aims at proposing new algorithms that address a particular issue, with theoretical guarantees and/or validation of their practical interest.

These notes cover material from the third category of optimization activities. The design of optimization algorithms (also called methods, or schemes) is a particularly subtle process, as an algorithm must exploit the theoretical properties of the problem while being amenable to implementation on a computer.

1.3.1 The algorithmic process

Most numerical optimization algorithms do not attempt to find a solution of a problem in a direct way, and rather proceed in an *iterative* fashion. Given a current point, that represents the current approximation to the solution, an optimization procedure attempts to move towards a (potentially) better point: to this end, the method generally requires a certain amount of calculation.

Suppose we apply such a process to the problem $\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$, resulting in a sequence of iterates $\{\boldsymbol{w}_k\}_k$. Ideally, these iterates obey one of the scenarios below:

1. The iterates produced get increasingly close to a solution, i. e.

$$\|\boldsymbol{w}_k - \boldsymbol{w}^*\| \to 0 \quad \text{when } k \to \infty.$$

   Although $\boldsymbol{w}^*$ is generally not known in practice, such results can be guaranteed by the theory, for instance on strongly convex problems.

2. The function values associated with the iterates get increasingly close to the optimum, i. e.

$$f(\boldsymbol{w}_k) \to f^* \quad \text{when } k \to \infty,$$

   As for the case above, f^* may not be known, but it can still be possible to prove convergence for certain algorithms and function classes (typically strongly convex, smooth functions).

3. The first-order optimality condition gets close to being satisfied, that is, $f \in \mathcal{C}^1(\mathbb{R}^d)$ and

$$\|\nabla f(\boldsymbol{w}_k)\| \to 0 \quad \text{when } k \to \infty.$$

Out of the three conditions, the last one is the easiest to track as the algorithm unfolds: it is, however, only a necessary condition, and does not guarantee convergence to a local minimum for generic, nonconvex functions. On the other hand, the first two conditions can only be measured approximately (by looking at the behavior of the iterates and enforcing decrease in the function values), but lead to stronger guarantees.

1.3.2 Convergence and convergence rates

The typical theoretical results that optimizers aim at proving for algorithms are asymptotic, as shown above: they only provide a guarantee in the limit. In practice, one may want to obtain more precise guarantees, that relate to a certain accuracy target that the practitioner would like to achieve. This led to the development of **global convergence rates**.

Example 1.3.1 (Global convergence rate for the gradient norm) *Given an algorithm applied to* $\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$ *that produces a sequence of iterates* $\{\boldsymbol{w}_k\}$*, we say that the method is* $\mathcal{O}(1/k)$ *for the gradient norm, or* $\|\nabla f(\boldsymbol{w}_k)\| = \mathcal{O}\left(\frac{1}{k}\right)$ *if*

$$\exists C > 0, \|\nabla f(\boldsymbol{w}_k)\| \leq \frac{C}{k} \ \forall k.$$

Such rates allow to quantify how much effort (in terms of iterations) is needed to reach a certain target accuracy $\epsilon > 0$. This leads to the companion notion of **worst-case complexity bound**.

Example 1.3.2 (Worst-case complexity for the gradient norm) *Given an algorithm applied to* $\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$ *that produces a sequence of iterates* $\{\boldsymbol{w}_k\}$*, we say that the method has a worst-case complexity of* $\mathcal{O}\left(\epsilon^{-1}\right)$ *for the gradient norm if*

$$\exists C > 0, \ \|\nabla f(\boldsymbol{w}_k)\| \leq \epsilon \text{ when } k \geq \frac{C}{\epsilon}.$$

Such results are quite common in theoretical computer science or statistics, which partly explain their popularity in machine learning. In optimization, they have been developed for a number of years in the context of convex optimization but have only gained momentum in general optimization over the last decade.

1.3.3 Popular optimization packages

Although a thorough numerical study is out of the scope of this course, we briefly mention popular choices for implementing optimization methods, either for industrial use or as research prototypes.

The most popular programming languages for optimization are C/C++/Fortran for high performance implementations, with Python and Julia raising increasing interest. The use of MATLAB/Octave is also widespread throughout the optimization community for prototyping (i.e. rapid and simple validation of an implementation), along with Python and Julia,

In addition to programming languages, optimizers have developed **modeling** languages that help bringing the code and the mathematical formulation of a problem closer. The broad-spectrum languages GAMS/AMPL/CVX are reknown examples; other languages, that are more domain-oriented, include MATPOWER and PyTorch.

Finally, there are many commercial solvers available, with CPLEX and Gurobi being arguably some of the most efficient for certain classes of problems. Some of those solvers are implemented in standard tools such as Microsoft Excel. Open-source codes are also quite popular, again fueled by the massive production of such implementations in the learning community. As far as optimization is concerned, the COIN-OR platform provides a good interface to all of these methods.

Chapter 2

Data fitting using standard optimization problems

In this chapter, we tackle the problem of fitting a linear model to some given data through the lens of two classical optimization formulations: linear least squares and linear programming. Those classes of optimization problems have been widely studied and are considered to be solvable for a large number of variables, yet modern data science problems challenge this conventional wisdom, partly because of the importance of the data defining the problem.

Motivation: linear models We consider a dataset of n elements (samples, individuals, etc). Each element possesses d features, represented by numerical values and gathered in a vector in $\mathbb{R}^d$. Letting $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be these vectors, we consider the data matrix:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{n \times d}. \tag{2.0.1}$$

In a supervised learning setup, each vector $\boldsymbol{x}_i$ is mapped to a label $y_i \in \mathbb{R}$, resulting in a vector $\boldsymbol{y} \in \mathbb{R}^n$. Our goal is then to find a relationship between the characteristics in $\boldsymbol{X}$ and the labels $\boldsymbol{y}$. The most simple relationship that can be thought of is linear: we thus seek a mapping $h : \mathbb{R}^d \to \mathbb{R}$ of the form $h(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}$ such that

$$h(\boldsymbol{x}_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{y}_i \qquad \forall i = 1, \ldots, n \;\Leftrightarrow\; \boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}.$$

This is a linear system of equations, but there is no a priori guarantee that this system always possesses a solution, and the problem cannot be reduced to that of solving a linear system. In this chapter, we consider the (often more realistic) objective of finding $h/\boldsymbol{w}$ that minimizes some function $\phi(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$, where ϕ will typically be a norm function. The subsequent sections focus on two variants of this problem, and motivate the interest of linear least squares and linear programming, respectively. We will be able to formulate and provide a solution for this problem using the tools from linear algebra presented in the next section.

2.1 Regression via linear least squares

Linear least squares are a particular class of (quadratic) optimization programs that are particularly useful for data fitting. They bear a close connection with linear systems and, as such, are able to leverage the power of linear algebra. In addition, they are related to statistical tasks such as maximum likelihood estimation, which is why linear least-squares problems are sometimes called (ℓ_2) linear regression problems.

2.1.1 Linear algebra tools

Square matrices are characterized by their eigenvalues: those are particularly useful to express solutions of linear systems (by inversion, for instance).

Definition 2.1.1 (Eigenvalue) *Let $\boldsymbol{X} \in \mathbb{R}^{d\times d}$. A value $\lambda \in \mathbb{C}$ is called an* **eigenvalue of $\boldsymbol{X}$** *if*

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

Such a vector $\boldsymbol{v}$ is called an **eigenvector associated with the eigenvalue** λ.

Eigenvectors and eigenvalues provide key information on the behavior of a square matrix, and on the way it acts on vectors in $\mathbb{R}^d$. In a very important case recalled below, we can express $\boldsymbol{X}$ as a diagonal operation on eigenvectors.

Theorem 2.1.1 (Spectral decomposition) *Let $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Then, there exists a matrix decomposition of $\boldsymbol{X}$ called* **spectral decomposition** *of the form*

$$\boldsymbol{X} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1},$$

where $\boldsymbol{P} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix with columns $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ form an orthonormal basis *of eigenvectors, and $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the d eigenvalues of $\boldsymbol{X}$ denoted by $\lambda_1, \ldots, \lambda_n$ on the diagonal.*

Consider now an arbitrary, rectangular matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$: the notion of eigenvector no longer makes sense here, since the dimensions of $\boldsymbol{X}\boldsymbol{v}$ and $\boldsymbol{v}$ (for any $\boldsymbol{v} \in \mathbb{R}^d$) may not agree. However, both matrices $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ are real symmetric matrices, and therefore the spectral theorem applies. Combining those two decompositions (plus additional manipulations) leads to a more general decomposition of $\boldsymbol{X}$.

Theorem 2.1.2 (Singular value decomposition) *Any matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has a* **singular value decomposition** *(or SVD) of the form*

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

where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ is orthogonal ($\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = I_n$), $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ is orthogonal ($\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = I_d$), and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times d}$ satisfies $\boldsymbol{\Sigma}_{ij} = 0$ if $i \neq j$ and $\boldsymbol{\Sigma}_{ii} \geq 0$.

The set of values $\{[\boldsymbol{\Sigma}]_{ii}\}_{i=1,\ldots,\min\{n,d\}}$ is called the set of singular values of $\boldsymbol{X}$. The maximum index $r \leq \min\{n, d\}$ such that $[\boldsymbol{\Sigma}]_{ii} > 0$ is called the rank of $\boldsymbol{X}$.

The SVD is instrumental to image and signal compression, and has also proven quite useful in matrix optimization problems. In the context of least squares (and linear systems of equations), it allows to define an operator that "inverts" the matrix $\boldsymbol{X}$.

Theorem 2.1.3 (Pseudo-inverse) *Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ be a singular value decomposition of the matrix $\boldsymbol{X}$, with $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ being of the form*

$$\left[\begin{array}{ccc|c} \sigma_1 & 0\cdots & 0 & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & \cdots 0 & \sigma_r & 0 \\ \hline 0 & \cdots & \cdots & 0 \end{array}\right]$$

with $\sigma_1 \geq \cdots \geq \sigma_r > 0$ (thus $\operatorname{rank}(\boldsymbol{X}) = r$).

The **pseudo-inverse** *of $\boldsymbol{X}$ is given by*

$$\boldsymbol{X}^{\dagger} = \boldsymbol{V}\boldsymbol{\Sigma}^{\dagger}\boldsymbol{U}^{\mathrm{T}}, \tag{2.1.1}$$

where $\boldsymbol{\Sigma}^{\dagger} \in \mathbb{R}^{n\times m}$ is the pseudo-inverse $\boldsymbol{\Sigma}$ defined explicitly by

$$\boldsymbol{\Sigma}^{\dagger} = \left[\begin{array}{ccc|c} \frac{1}{\sigma_1} & 0\cdots & 0 & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & \cdots 0 & \frac{1}{\sigma_r} & 0 \\ \hline 0 & \cdots & \cdots & 0 \end{array}\right].$$

This operator can be used to compute solutions of linear least-squares problems, as highlighted in the next section.

2.1.2 Linear least-squares optimization

As explained in the introduction of this chapter, we are given data under the form $(\boldsymbol{X}, \boldsymbol{y})$ where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$, and our goal consists in finding a vector $\boldsymbol{w} \in \mathbb{R}^d$ such that $\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \approx \boldsymbol{0}$. We would like the vector $\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}$ to be as close as possible to zero: an optimization formulation is particularly well suited for such a problem, and this gives rise to the following definition.

Definition 2.1.2 (Linear least squares) *A* **linear least-squares** *optimization problem is of the form*

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i)^2, \tag{2.1.2}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$ represent the problem data.

Note that the factor $\frac{1}{2}$ is introduced for normalization purposes, and does not affect the solution set of the problem[1].

We now express a solution of the problem using the pseudo-inverse formula that we introduced in the previous section.

Theorem 2.1.4 (Solution of linear least squares) *Given problem (2.1.2), we define the vector $\boldsymbol{w}^* = \boldsymbol{X}^{\dagger}\boldsymbol{y}$. The following properties hold:*

(i) The vector $\boldsymbol{w}^*$ is always a solution of the optimization problem (2.1.2).

[1] For any optimization problem $\operatorname{minimize}_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$ and any value $a > 0$, the set of solutions of $\operatorname{minimize}_{\boldsymbol{w}\in\mathbb{R}^d} a\, f(\boldsymbol{w})$ is identical to that of the original problem.

(ii) Among the solutions of the optimization problem, it is the solution of minimal norm, i.e.

$$\forall \hat{\boldsymbol{w}} \in \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2, \quad \|\boldsymbol{w}^*\| \leq \|\hat{\boldsymbol{w}}\|.$$

(iii) If $\operatorname{rank}(\boldsymbol{X}) = d \leq n$*, the vector* $\boldsymbol{w}^*$ *is the unique solution to the problem* (2.1.2)*, and it is also a solution of the linear system* $\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}$*.*

Theorem 2.1.4 implies that solving a linear least-squares problem can be done by computing an SVD of the data matrix $\boldsymbol{X}$. As we will see later in these notes, this may be deemed too expensive in large dimensions, thereby motivating the need for other optimization strategies.

2.1.3 Link with linear regression

Linear regression is a classical paradigm in data analysis, that aims at building a linear model from a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. It is typically assumed that the data comes from an underlying linear trend corrupted with noise: there exists $\boldsymbol{w}^* \in \mathbb{R}^d$ such that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a vector with i.i.d. (independent, identically distributed) entries following a Gaussian distribution (of mean zero and identity covariance matrix). Figure 2.1 illustrates the behavior of such data samples.

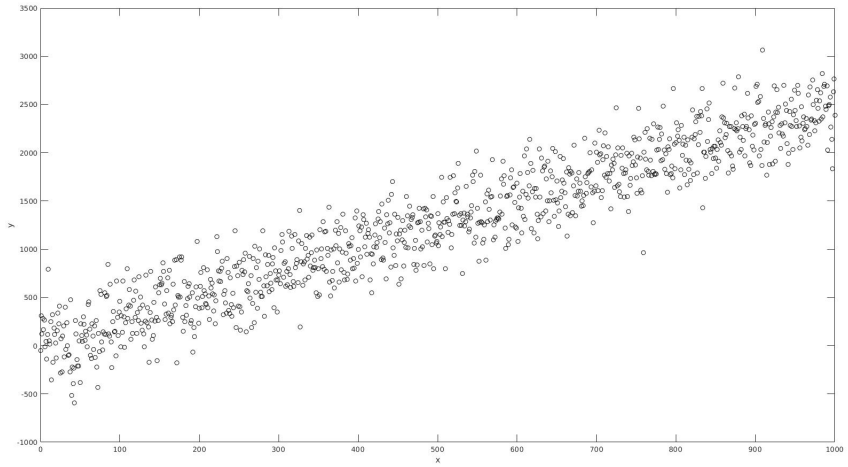

Figure 2.1: Data generated from a linear model corrupted with Gaussian noise.

We seek the most likely value of $\boldsymbol{w}^*$ given the samples, which is obtained by solving the following optimization problem:

$$\operatorname*{maximize}_{\boldsymbol{w} \in \mathbb{R}^d} L(y_1, \ldots, y_n; \boldsymbol{w}) := \left[\frac{1}{\sqrt{2\pi}}\right]^m \exp\left(-\frac{1}{2}\sum_{i=1}^m (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i)^2\right). \tag{2.1.3}$$

It can be shown (by taking the logarithm and the opposite of the objective function) that problems (2.1.2) and (2.1.3) are equivalent in the sense of Definition 1.2.1, and thus the solution of the latter (which involves a highly nonlinear objective) can be found by solving the former (with a much nicer objective function).

In particular, if we assume that $\mathrm{rank}(\boldsymbol{X}) = d \ll n$, problem (2.1.3) possesses a unique solution, called the **maximum likelihood estimator** and given by $\boldsymbol{X}^\dagger \boldsymbol{y} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$. This estimator has desirable statistical properties (for instance, it is an unbiased estimator of $\boldsymbol{w}^*$.

2.2 Linear programming

Linear programming is the most prevalent class of optimization problems, and has numerous applications such as economics and energy systems. It also offers a simple mathematical model of an optimization problem, for which it is often easy to draw interpretation from. Finally, large scale Linear Programs (LPs) are routinely solved by modern software with millions of decision variables.

2.2.1 Linear optimization problem

Definition 2.2.1 (Linear program) *A linear optimization problem, or* **linear program**, *is an optimization problem where the objective function is linear and the constraint set can be represented using a set of linear equalities and inequalities.*

A linear program is expressed in standard form when it is written as

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{w} \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b},\ \boldsymbol{w} \geq 0, \tag{2.2.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times d},\ \boldsymbol{b} \in \mathbb{R}^n,\ \boldsymbol{c} \in \mathbb{R}^d$.

By convention, an LP is in standard form regardless of whether the objective is maximized or maximized, and the nonnegative bounds on $\boldsymbol{w}$ can be applied to some of its components only.

Note that any linear optimization problem has an equivalent reformulation (in the sense of Definition 1.2.1) in standard form. For instance, the problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}}\ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}$$

is equivalent to

$$\underset{\substack{\boldsymbol{x}\in\mathbb{R}^d\\ \boldsymbol{s}\in\mathbb{R}^n}}{\text{minimize}}\ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} - \boldsymbol{s} = \boldsymbol{b},\ \boldsymbol{s} \geq \boldsymbol{0}.$$

Indeed, for any solution $\boldsymbol{x}^*$ of the former leads to a solution $(\boldsymbol{x}^*, \boldsymbol{s}^* = \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$ of the latter. Conversely, if $(\boldsymbol{x}^*, \boldsymbol{s}^*)$ is a solution of the latter problem, then $\boldsymbol{x}^*$ is also a solution of the former problem (in particular, $\boldsymbol{x}^*$ is feasible because $\boldsymbol{s}^* \geq \boldsymbol{0}$).

Remark 2.2.1 *Not all optimization problems can be reformulated as linear programs, but it can be quite useful to do so when possible. Indeed, there exist very efficient numerical solvers for linear programming that can solve problems with a very large number of both variables and constraints.*

Remark 2.2.2 *Linear programs are typically solved using a great deal of linear algebra. Tools such as that described in Section 2.1.1 can be employed to solve linear systems arising in the context of linear programming algorithms.*

2.2.2 Robust linear regression and linear program

The formulation of Section 2.1.2 usually leads to solutions that are affected by outliers in the data. An alternative consists in using **robust linear regression** formulations by replacing the ℓ_2 norm with other norms. We present below the results obtained using the ℓ_1 norm.

Definition 2.2.2 *Given $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$, the ℓ_1 (linear) regression problem is defined as*

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_1 = \sum_{i=1}^{n} |\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i|. \tag{2.2.2}$$

Problem (2.2.2) known to produce solutions (and thus, linear models) that are less sensitive to outliers in the data compared to that obtained with ℓ_2 regression. However, the objective function involves an absolute value, which is not a linear (nor a quadratic) function. Although we will see ways to tackle such functions in Chapter 5, we discuss here a reformulation of this problem as a linear program. This reformulation is based on the fact that any real number t can be written as $t = t^+ - t^-$ with $t^+ = \max\{t, 0\} \geq 0$ and $t^- = \max\{-t, 0\} \geq 0$. With that notation, $|t| = t^+ + t^-$, i.e. we can write the absolute value as a linear function of t^+ and t^-. This observation is at the heart of the reformulation

$$\begin{array}{lll} \underset{\substack{\boldsymbol{w}\in\mathbb{R}^d \\ \boldsymbol{t}^+\in\mathbb{R}^n \\ \boldsymbol{t}^-\in\mathbb{R}^n}}{\text{minimize}} & \sum_{i=1}^n (t_i^+ + t_i^-) & \\ \text{subject to} & \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i = t_i^+ - t_i^- \quad \forall i = 1, \ldots, n \\ & \boldsymbol{t}^+ \geq 0 \\ & \boldsymbol{t}^- \geq 0, \end{array} \tag{2.2.3}$$

which is a linear program. Solving this linear program readily gives the solution of the original problem (2.2.2) through the vector $\boldsymbol{w}$, while the optimal value is recovered (with no additional access to the data!) by $\sum_{i=1}^n (t_i^+ + t_i^-)$.

Chapter 3

Unconstrained optimization

In this chapter, we study more general nonlinear optimization problems of the form

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}). \tag{3.0.1}$$

We make the following assumption on the objective function.

Assumption 3.0.1 *The objective function f in (3.0.1) is $\mathcal{C}_L^{1,1}(\mathbb{R}^d)$ for $L > 0$, and bounded below by $f_{low} \in \mathbb{R}$ (i.e. $f(\boldsymbol{w}) \geq f_{low}\ \forall \boldsymbol{w} \in \mathbb{R}^d$).*

We will design and analyze algorithms that exploit gradient information to move towards better points. Our theoretical results will consist in complexity bounds and convergence rates.

3.1 Gradient descent

The gradient descent algorithm is arguably the most classical technique in unconstrained, smooth optimization. It is based on the following principle, derived from the first-order optimality condition (1.2.7).

For any vector $\boldsymbol{w} \in \mathbb{R}^d$, two cases can occur:

1. Either $\nabla f(\boldsymbol{w}) = 0$ and $\boldsymbol{w}$ is possibly a local minimum (when f is convex, we know that $\boldsymbol{w}$ is necessarily a global minimum);
2. Or $\nabla f(\boldsymbol{w}) \neq 0$, and we can show that f must decrease *locally* in the direction of $-\nabla f(\boldsymbol{w})$.

The second property is formalized below, and is at the core of the gradient descent framework.

3.1.1 Algorithm

The gradient descent algorithm is an iterative process wherein every iteration has the following form:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha \nabla f(\boldsymbol{w}), \tag{3.1.1}$$

where $\alpha > 0$ is a parameter called stepsize or steplength. When $\nabla f(\boldsymbol{w}) = 0$, note that the formula (3.1.1) does not change the value of $\boldsymbol{w}$: this is consistent with the fact that the gradient

cannot be used to determine a better point in that case. On the contrary, when $\nabla f(\boldsymbol{w}) \neq 0$, there will exist values for α leading to a lower function value for the point $\boldsymbol{w} - \alpha \nabla f(\boldsymbol{w})$.

Repeated applications of the update (3.1.1) lead to Algorithm 1. This method is called **gradient descent**, or (less frequently) *steepest descent*.[1]

Algorithm 1: Gradient descent for minimizing the function f.

Initialization: Choose $\boldsymbol{w}_0 \in \mathbb{R}^d$.
For $k = 0, 1, ...$
1. Compute the gradient $\nabla f(\boldsymbol{w}_k)$.
2. Define a stepsize $\alpha_k > 0$.
3. Set $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)$.

EndFor

Algorithm 1 actually describes a framework rather than a specific method. There exist numerous variants upon the gradient descent paradigm: we review the main characteristics of these methods below.

Stopping criterion In practice, an algorithm is often subject to budget requirements (in terms of arithmetic opterations, running time, iteration number, etc). These limitations are typically enforced using a stopping criterion. In the context of Algorithm 1, it is typical to terminate the method after $k_{\max}$ iterations, in which case $\boldsymbol{w}_{k_{\max}}$ is returned as an approximate solution to the problem.

In addition, stopping criteria can be used to check whether the method converged to a solution of the problem (or an approximation thereof). Such criteria are often based on optimality conditions. For gradient descent, the most classical criterion is based on the gradient norm: one stops the algorithm as soon as

$$\|\nabla f(\boldsymbol{w}_k)\| < \epsilon, \tag{3.1.2}$$

where $\epsilon > 0$ represents a given, desired accuracy level (the condition becomes more expensive to satisfy as ϵ gets smaller).

Finally, it is always possible (and often recommended) to add "safety checks", that stop the method whenever no visible progress is measured. For instance, if the difference between two successive iterates ($\|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\|$) is of the order of machine precision, the algorithm is likely stalling and no further improvement is expected: in that case, it is often better to stop the method.

Choosing the initial point The performance of a given algorithm can be significantly improved using a suitable starting point. However, finding such a point can be a difficult task without domain expertise, but if such expertise exists, it should be leveraged to obtain a good initial point that the method tries to improve upon. Alternatively, one could use several starting points drawn at random and run a few iterations of gradient descent so as to determine a good starting point.

[1] The direction of $-\nabla f(\boldsymbol{w}_k)$, i.e. the unit vector $\frac{-\nabla f(\boldsymbol{w}_k)}{\|\nabla f(\boldsymbol{w}_k)\|}$ (or the zero vector if the gradient is zero) is called the steepest descent direction.

3.1.2 Choosing the stepsize

In this section, we describe the main techniques for selecting the step size sequence in gradient descent. We provide generic principles, and emphasize that any information about a particular problem can be extremely valuable in designing a better step size.

Constant step size One of the most common approaches consists in fixing the step size to a single value for all iterations, i.e. setting $\alpha_k = \alpha > 0$ for all k. Depending on the computational budget, one could run gradient descent with several values of α and select the best value for future use: this practice of *tuning* the step size is commonly adopted in data science problems.

Provided f satisfies Assumption 3.0.1, there exists an interval of values that lead to convergence of gradient descent. In particular, the choice

$$\alpha_k = \alpha = \tfrac{1}{L}, \tag{3.1.3}$$

where L is the Lipschitz constant for the gradient, is well suited for that problem. Note however that this choice requires the knowledge of the Lipschitz constant: this information is not always available in practice.

Decreasing step size Another classical technique for selecting the stepsize consists in defining a decreasing sequence $\{\alpha_k\}$ such that $\alpha_k \to 0$ *prior to running the method*. This choice can also lead to converging method, but it risks producing steps that are unnecessarily small in norm. In fact, a good decreasing strategy should drive α_k to 0 quickly enough for convergence, but slowly enough that the norm of the steps do not approach 0 too rapidly.

Adaptive choice using a line search Line-search techniques are widely used in continuous optimization and scientific computing (though less popular in data science, for reasons that we will detail in Chapter 4 of these notes). At a given iteration of index k, we seek a stepsize α_k that leads to a decrease in the objective function along a suitably chosen direction (in the case of Algorithm 1, this would be the direction of $-\nabla f(\boldsymbol{w}_k)$). An exact line search results in the best possible decrease, but may be costly to perform. In practice, *inexact* approaches are preferred: Algorithm 2 details the most popular of such techniques, called backtracking.

Algorithm 2: Backtracking line search in direction d.

Inputs: $\boldsymbol{w} \in \mathbb{R}^d$, $\boldsymbol{d} \in \mathbb{R}^d$, $\alpha_0 \in \mathbb{R}^d$.
Initialization: Choose $\alpha = \alpha_0$.
While $f(\boldsymbol{w} + \alpha \boldsymbol{d}) > f(\boldsymbol{w})$
  $\alpha \to \frac{\alpha}{2}$.
End
Output: α.

The line-search procedure of Algorithm 2 can be used at Step 2 of Algorithm 1 using $\boldsymbol{w} = \boldsymbol{w}_k$, $\boldsymbol{d} = -\nabla f(\boldsymbol{w}_k)$ and (for instance) $\alpha_0 = 1$ as inputs. Many variants of this simple idea have been proposed in the literature: those are generally designed to guarantee that a better point (in terms of the objective function) is found. Still, these techniques require at least one additional function evaluation per stepsize value (possibly more), leading to an overall more expensive method.

3.1.3 Theoretical analysis for gradient descent

In this section, we present several convergence rates for gradient descent, in the case of a smooth objective function. We will see that the nonconvex, convex and strongly convex cases exhibit different behavior.

Proposition 3.1.1 *Consider the k-th iteration of Algorithm 1 applied to $f \in \mathcal{C}_L^{1,1}(\mathbb{R}^d)$, and suppose that $\nabla f(\boldsymbol{w}_k) \neq 0$. Then, if $0 < \alpha_k < \frac{2}{L}$, we have*

$$f(\boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)) < f(\boldsymbol{w}_k).$$

In particular, choosing $\alpha_k = \frac{1}{L}$ leads to

$$f(\boldsymbol{w}_k - \frac{1}{L}\nabla f(\boldsymbol{w}_k)) < f(\boldsymbol{w}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{w}_k)\|^2. \tag{3.1.4}$$

Proof. We use the inequality (1.2.2) with the vectors $(\boldsymbol{w}_k, \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k))$:

$$\begin{aligned} f(\boldsymbol{w}_k - \alpha_l \nabla f(\boldsymbol{w}_k)) &\leq f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^{\mathrm{T}} [-\alpha_k \nabla f(\boldsymbol{w}_k)] + \frac{L}{2}\| - \alpha_k \nabla f(\boldsymbol{w}_k)\|^2 \\ &= f(\boldsymbol{w}_k) - \alpha_k \nabla f(\boldsymbol{w}_k)^{\mathrm{T}} \nabla f(\boldsymbol{w}_k) + \frac{L}{2}\alpha_k^2 \|\nabla f(\boldsymbol{w}_k)\|^2 \\ &= f(\boldsymbol{w}_k) + \left(-\alpha_k + \frac{L}{2}\alpha_k^2\right) \|\nabla f(\boldsymbol{w}_k)\|^2. \end{aligned}$$

If $-\alpha_k + \frac{L}{2}\alpha_k^2 < 0$, the second term on the right-hand side will be negative, thus we will have $f(\boldsymbol{w}_k - \alpha_l \nabla f(\boldsymbol{w}_k)) < f(\boldsymbol{w}_k)$. Since $-\alpha_k + \frac{L}{2}\alpha_k^2 < 0 \Leftrightarrow \alpha_k < \frac{2}{L}$ and $\alpha_k > 0$ by definition, this proves the first part of the result.

To obtain (3.1.4), one simply needs to use $\alpha_k = \frac{1}{L}$ in the series of equations above. □

The result of Proposition 3.1.1 will be instrumental to obtain complexity guarantees on Algorithm 1 in three different settings: nonconvex, convex, and strongly convex.

Nonconvex case In the nonconvex case, we aim at bounding the number of iterations required to drive the gradient norm below some threshold $\epsilon > 0$: this means that we should be able to show that the gradient norm actually goes below this threshold, which is a guarantee of convergence.

Theorem 3.1.1 (Complexity of gradient descent for nonconvex functions) *Let f be a nonconvex function satisfying Assumption 3.0.1. Suppose that Algorithm 1 is applied with $\alpha_k = \frac{1}{L}$. Then, for any $K \geq 1$, we have*

$$\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\| \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right). \tag{3.1.5}$$

Proof. Let K be an iteration index such that for every $k = 0, \ldots, K-1$, we have $\|\nabla f(\boldsymbol{w}_k)\| > \epsilon$. From Proposition 3.1.1, we have that

$$\forall k = 0, \ldots, K-1, \quad f(\boldsymbol{w}_{k+1}) \leq f(\boldsymbol{w}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{w}_k)\|^2 \leq f(\boldsymbol{w}_k) - \frac{1}{2L}\left(\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\|\right)^2.$$

By summing across all such iterations, we obtain :

$$\sum_{k=0}^{K-1} f(\boldsymbol{w}_{k+1}) \leq \sum_{k=0}^{K-1} f(\boldsymbol{w}_k) - \frac{K}{2L} \left(\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\| \right)^2.$$

Removing identical terms on both sides yields

$$f(\boldsymbol{w}_K) \leq f(\boldsymbol{w}_0) - \frac{K}{2L} \left(\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\| \right)^2.$$

Using $f(\boldsymbol{w}_K) \geq f_{low}$ (which holds by Assumption 3.0.1) and re-arranging the terms leads to

$$\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\| \leq \left[\frac{2L(f(\boldsymbol{w}_0) - f_{low})}{K} \right]^{1/2} = \mathcal{O}\left(\frac{1}{\sqrt{K}} \right).$$

□

Equivalently, we say that the worst-case complexity of gradient descent is $\mathcal{O}\left(\epsilon^{-2}\right)$, because for any $\epsilon > 0$, a reasoning similar to the proof of Theorem 3.1.1 guarantees that $\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\| \leq \epsilon$ after at most

$$\left\lceil 2L(f(\boldsymbol{w}_0) - f_{low})\epsilon^{-2} \right\rceil = \mathcal{O}(\epsilon^{-2})$$

iterations.

Convex/Strongly convex case In addition to Assumption 3.0.1, if we further assume that the objective is convex or strongly convex, we can show that stronger guarantees than that of the nonconvex case can be obtained at a lower cost. This improvement illustrates the interest of convex functions in optimization.

In this paragraph, we let $f^* = \min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$ denote the minimal value of f (note that $f^* \geq f_{low}$) and we assume that there exists $\boldsymbol{w}^* \in \mathbb{R}^d$ such that $f(\boldsymbol{w}^*) = f^*$ (i.e. the set of minima is not empty). Given an accuracy threshold $\epsilon > 0$, we are interested in bounding the number of iterations necessary to reach an iterate $\boldsymbol{w}_k$ such that $f(\boldsymbol{w}_k) - f^* \leq \epsilon$.

Theorem 3.1.2 *Convergence of gradient descent for convex functions Let f be a convex function satisfying Assumption 3.0.1. Suppose that Algorithm 1 is applied with $\alpha_k = \frac{1}{L}$. Then, for any $K \geq 1$, the iterate $\boldsymbol{w}_K$ satisfies*

$$f(\boldsymbol{w}_k) - f^* \leq \mathcal{O}\left(\frac{1}{K} \right). \tag{3.1.6}$$

Proof. Let K be an index such that for every $k = 0, \ldots, K-1$, $f(\boldsymbol{w}_k) - f^* > \epsilon$.
For any $k = 0, \ldots, K-1$, the characterization of convexity (1.2.11) at $\boldsymbol{w}_k$ and $\boldsymbol{w}^*$ gives

$$f(\boldsymbol{w}^*) \geq f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{w}^* - \boldsymbol{w}_k).$$

Combining this property with (3.1.4), we obtain:

$$\begin{aligned} f(\boldsymbol{w}_{k+1}) &\leq f(\boldsymbol{w}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{w}_k)\|^2 \\ &\leq f(\boldsymbol{w}^*) + \nabla f(\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{w}_k - \boldsymbol{w}^*) - \frac{1}{2L}\|\nabla f(\boldsymbol{w}_k)\|^2. \end{aligned}$$

To proceed onto the next step, one notices that

$$\nabla f(\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{w}_k - \boldsymbol{w}^*) - \frac{1}{2L}\|\nabla f(\boldsymbol{w}_k)\|^2 = \frac{L}{2}\left(\|\boldsymbol{w}_k - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_k - \boldsymbol{w}^* - \frac{1}{L}\nabla f(\boldsymbol{w}_k)\|^2\right).$$

Thus, recalling that $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{1}{L}\nabla f(\boldsymbol{w}_k)$, we arrive at

$$\begin{aligned} f(\boldsymbol{w}_{k+1}) &\leq f(\boldsymbol{w}^*) + \frac{L}{2}\left(\|\boldsymbol{w}_k - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_k - \boldsymbol{w}^* - \frac{1}{L}\nabla f(\boldsymbol{w}_k)\|^2\right) \\ &= f(\boldsymbol{w}^*) + \frac{L}{2}\left(\|\boldsymbol{w}_k - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_{k+1} - \boldsymbol{w}^*\|^2\right). \end{aligned}$$

Hence,

$$f(\boldsymbol{w}_{k+1}) - f(\boldsymbol{w}^*) \leq \frac{L}{2}\left(\|\boldsymbol{w}_k - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_{k+1} - \boldsymbol{w}^*\|^2\right). \tag{3.1.7}$$

By summing (3.1.7) on all indices k between 0 and $K-1$, we obtain

$$\sum_{k=0}^{K-1} f(\boldsymbol{w}_{k+1}) - f(\boldsymbol{w}^*) \leq \frac{L}{2}\left(\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_K - \boldsymbol{w}^*\|^2\right) \leq \frac{L}{2}\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2.$$

Finally, using $f(\boldsymbol{w}_0) \geq f(\boldsymbol{w}_1) \geq \ldots \geq f(\boldsymbol{w}_K)$ (a consequence of Proposition 3.1.1, we obtain that

$$\sum_{k=0}^{K-1} f(\boldsymbol{w}_{k+1}) - f(\boldsymbol{w}^*) \geq K\left(f(\boldsymbol{w}_K) - f^*\right).$$

Injecting this formula into the previous equation finally yields the desired outcome:

$$f(\boldsymbol{w}_k) - f(\boldsymbol{w}^*) \leq \frac{L\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{2}\frac{1}{K}.$$

□

Equivalently, we say that the worst-case complexity of gradient descent is $\mathcal{O}\left(\epsilon^{-1}\right)$, which means here that there exist a positive constant C (that depends on $\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|$ and L) such that

$$f(\boldsymbol{w}_K) - f_{low} \leq \epsilon.$$

after at most $C\epsilon^{-1} = \mathcal{O}(\epsilon^{-1})$ iterations.

We now turn to the strongly convex case.

Theorem 3.1.3 *Convergence of gradient descent for strongly convex functions Let f be a μ-strongly convex function satisfying Assumption 3.0.1, with $\mu \in (0, L]$. Suppose that Algorithm 1 is applied with $\alpha_k = \frac{1}{L}$. Then, for any $K \in \mathbb{N}$, we have*

$$f(\boldsymbol{w}_k) - f^* \leq \mathcal{O}\left((1 - \tfrac{\mu}{L})^k\right). \tag{3.1.8}$$

We say that the convergence rate of gradient descent is $\mathcal{O}\left((1 - \frac{\mu}{L})^k\right)$.

Proof. We exploit the strong convexity property (1.2.13). For any $(\boldsymbol{x}, \boldsymbol{y}) \in (\mathbb{R}^n)^2$, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

Minimizing both sides with respect to $\boldsymbol{y}$ lead to $\boldsymbol{y} = \boldsymbol{w}^*$ on the left-hand side, and $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{\mu}\nabla f(\boldsymbol{x})$ on the right-hand side[2]. As a result, we obtain

$$\begin{aligned} f^* &\geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}\left[-\frac{1}{\mu}\nabla f(\boldsymbol{x})\right] + \frac{\mu}{2}\|-\frac{1}{\mu}\nabla f(\boldsymbol{x})\|^2 \\ f^* &\geq f(\boldsymbol{x}) - \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|^2. \end{aligned}$$

By re-arranging the terms, we arrive at

$$\|\nabla f(\boldsymbol{x})\|^2 \geq 2\mu\left[f(\boldsymbol{x}) - f^*\right], \tag{3.1.9}$$

which is valid for any $\boldsymbol{x} \in \mathbb{R}^n$. Using (3.1.9) together with (3.1.4) thus gives

$$f(\boldsymbol{w}_{k+1}) \leq f(\boldsymbol{w}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{w}_k)\|^2 \leq f(\boldsymbol{w}_k) - \frac{\mu}{L}(f(\boldsymbol{w}_k) - f^*).$$

This leads to

$$f(\boldsymbol{w}_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)(f(\boldsymbol{w}_k) - f^*),$$

which we can iterate in order to obtain

$$f(\boldsymbol{w}_K) - f^* \leq \left(1 - \frac{\mu}{L}\right)^K (f(\boldsymbol{w}_0) - f^*).$$

It then suffices to note that the bound is also valid for $K = 0$. □

Equivalently, we can show a worst-case complexity result: the method computes $\boldsymbol{w}_k$ such that $f(\boldsymbol{w}_k) - f^* \leq \epsilon$ in at most $\mathcal{O}(\frac{L}{\mu}\ln(\frac{1}{\epsilon}))$ iterations.

Similar results can be shown for the criterion $\|\boldsymbol{w}_k - \boldsymbol{w}^*\|$: in other words, the distance between the current iterate and the (unique) global optimum decreases at a rate $\mathcal{O}\left((1 - \frac{\mu}{L})^k\right)$.

Remark 3.1.1 *Proofs of convergence rates are typically more technical for convex and strongly convex problems: in order to obtain better bounds than in the nonconvex setting, one must make careful use of the (strong) convexity inequalities. In this course, we do not focus on these aspects, but rather draw insights from the final complexity bounds or convergence rates.*

3.2 Acceleration

3.2.1 Introduction: the momentum principle

In Section 3.1.3, we derive complexity bounds for the gradient descent algorithm, and we saw in particular that assuming that the function was convex (respectively, strongly convex) improved the complexity. These results are called *upper* complexity bounds, in the sense that they reflect the worst

[2] This is because the gradient of the quadratic function on the right-hand side with respect to $\boldsymbol{y}$ is $\nabla f(\boldsymbol{x}) + \mu(\boldsymbol{y} - \boldsymbol{x})$. This vector is equal to $\mathbf{0}$ if and only if $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{\mu}\nabla f(\boldsymbol{x})$

possible convergence rate that this algorithm could exhibit on a given problem. The issue of *lower* bounds, that show a rate that cannot be improved upon, has been the subject to a lot of attention, particularly in the convex optimization community.

For nonconvex optimization, it is known that there exists a function for which gradient descent converges exactly at the $\mathcal{O}(\frac{1}{\sqrt{K}})$ rate: in this case, the lower bound matches the upper bound. On the contrary, for convex functions, the lower bound is actually $\mathcal{O}(\frac{1}{K^2})$, which is a sensible improvement over the bound in $\mathcal{O}(\frac{1}{K})$ of Theorem 3.1.2. There are methods that can achieve this bound, thanks to an algorithmic technique called **acceleration**.

The underlying idea of acceleration is that, at a given iteration and given the available information from previous iterations (in particular, the latest displacement), one can move along a better step than that given by the current gradient.

3.2.2 Nesterov's accelerated gradient method

Among the existing methods based on acceleration, the accelerated gradient algorithm proposed by Yurii Nesterov in 1983 is the most famous, to the point that it has been termed "Nesterov's algorithm".

Algorithm 3: Accelerated gradient method.

Initialization: $\boldsymbol{w}_0 \in \mathbb{R}^d$, $\boldsymbol{w}_{-1} = \boldsymbol{w}_0$.

for $k = 0, 1, \ldots$ **do**

1. Compute a steplength $\alpha_k > 0$ and a parameter $\beta_k > 0$.
2. Compute the new iterate as

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f\left(\boldsymbol{w}_k + \beta_k(\boldsymbol{w}_k - \boldsymbol{w}_{k-1})\right) + \beta_k(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}). \tag{3.2.1}$$

end

Algorithm 3 provides a description of the method. Like the gradient descent method of Section 3.1, it requires a single gradient calculation per iteration; however, unlike in gradient descent, the gradient is not evaluated at the current iterate $\boldsymbol{w}_k$, but at a combination of this iterate with the previous step $\boldsymbol{w}_k - \boldsymbol{w}_{k-1}$: this term is called the **momentum term**, and is key to the performance of accelerated gradient techniques.

Another view of the accelerated gradient descent is that of a two-loop recursion: given $\boldsymbol{w}_0$ and $\boldsymbol{z}_0 = \boldsymbol{w}_0$, the update (3.2.1) can be rewritten as

$$\begin{cases} \boldsymbol{w}_{k+1} &= \boldsymbol{z}_k - \alpha_k \nabla f(\boldsymbol{z}_k) \\ \boldsymbol{z}_{k+1} &= \boldsymbol{w}_{k+1} + \beta_{k+1}(\boldsymbol{w}_{k+1} - \boldsymbol{w}_k). \end{cases} \tag{3.2.2}$$

This formulation decouples the two steps behind the accelerated gradient update: a gradient step on $\boldsymbol{z}_k$, combined with a momentum step on $\boldsymbol{w}_{k+1}$.

Choosing the parameters We now comment on the choice of the stepsize α_k and the momentum parameter β_k. The same techniques than those presented in Section 3.1.2 can be considered for the choice of α_k (stepsize parameter). As in the gradient descent case, the choice $\alpha_k = \frac{1}{L}$ is a standard one.

The choice of β_k is most crucial to obtaining the improved complexity bound. The standard values proposed by Nesterov depend on the nature of the objective function:

- If f is a μ-strongly convex, we set

$$\beta_k = \beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}} \tag{3.2.3}$$

  for every k. Note that this requires the knowledge of both the Lipschitz constant of the gradient and the strong convexity constant.

- For a general convex function f, β_k is computed in an adaptive way using two sequences, as follows:

$$t_{k+1} = \frac{1}{2}(1+\sqrt{1+4t_k^2}), t_0 = 0, \quad \beta_k = \frac{t_k - 1}{t_{k+1}}. \tag{3.2.4}$$

The following informal theorem summarizes the complexity results that can be proven for Algorithm 3.

Theorem 3.2.1 *Consider Algorithm 3 applied to a convex function f satisfying Assumption 3.0.1, with $\alpha_k = \frac{1}{L}$, and let $\epsilon > 0$. Then, for any $K \geq 1$, the iterate $\boldsymbol{w}_K$ computed by Algorithm 3 satisfies*

i) $f(\boldsymbol{w}_K) - f^* \leq \mathcal{O}(\frac{1}{K^2})$ for a generic convex function if β_k is set according to the adaptive rule (3.2.4);

ii) At most $f(\boldsymbol{w}_K) - f^* \leq \left((1-\sqrt{\frac{\mu}{L}})^K\right)$ for a μ-strongly convex function, provided β_k is set to the constant value given by (3.2.3).

Note that we can also derive worst-case complexity bounds for the accelerated gradient method, that show the same improvement. For instance, for strongly convex functions, we can establish that $f(\boldsymbol{w}_k) - f^* \leq \epsilon$ after at most $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\ln(\epsilon^{-1})\right)$ iterations, which represents an improvement over the $\mathcal{O}\left(\frac{L}{\mu}\ln(\epsilon^{-1})\right)$ complexity over gradient descent. Here the improvement is in terms of problem-dependent constants.

3.2.3 Other accelerated methods

Heavy ball method The heavy ball method is a precursor of the accelerated gradient algorithm, that was proposed by Boris T. Polyak in 1964. Its k-th iteration can be written as

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha\nabla f(\boldsymbol{w}_k) + \beta(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}),$$

where the stepsize and momentum parameters are chosen to be constant values. The key difference between this iteration and Nesterov's lies in the gradient evaluation, which the heavy ball method performs at the current point: in that sense, the heavy ball method performs first the gradient update, then the momentum step, while Nesterov's method adopts the inverse approach. This method achieves the optimal rate of convergence on strongly convex quadratic functions, but can fail on general strongly convex functions.

Conjugate gradient The (linear) conjugate gradient method, proposed by Hestenes and Stiefel in 1952, has remained to this day one of the preferred methods to solve linear systems of equations and strongly convex quadratic minimization problems. Unlike Polyak's method, the conjugate gradient algorithm does not require knowledge of the Lipschitz constant L nor the parameter μ, because it exploits knowledge from the past iterations. The k-th iteration of conjugate gradient can be written as:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \alpha_k p_k, \quad p_k = -\nabla f(x_k) + \beta_k p_{k-1}.$$

In a standard conjugate gradient algorithm, α_k and β_k are computed using formulas tailored to the problem: this contributes to their convergence rate analysis, which leads to a rate similar to that of accelerated gradient. However, unlike accelerated gradient, the conjugate gradient is guaranteed to terminate after d iterations on a d-dimensional problem. When d is very large, the bound for conjugate gradient matches that of the other methods, and in that sense does not depend on the problem dimension.

Example 3.2.1 (Strongly convex quadratic minimization) *A strongly convex quadratic minimization problem is an optimization problem of the form*

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \; q(\boldsymbol{w}) := \tfrac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{w}$$

where $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix and $\boldsymbol{b} \in \mathbb{R}^d$. This problem is smooth (because the objective is polynomial in all of the decision variables) and $\nabla^2 f(\boldsymbol{w}) \succ \mathbf{0}$ for every $\boldsymbol{w}$, meaning that the problem is μ-strongly convex with μ denoting the minimum eigenvalue of $\boldsymbol{A}$. As a result, there exist a unique global minimum given by the solution of $\nabla q(\boldsymbol{w}) = \boldsymbol{A}\boldsymbol{w} - \boldsymbol{b} = 0$. This equation is a linear system but the cost of inverting this system and computing a solution can be prohibitive. For this reason, one can replace the exact solve by an iterative, gradient-based approach, and apply Algorithm 1 or Algorithm 3. Note that $q \in \mathcal{C}^{1,1}_{\|\boldsymbol{A}\|}(\mathbb{R}^d)$, hence the choice of steplength 3.1.3 is a valid one.

If gradient descent is applied, then an ϵ-accuracy in the objective value can be reached in at most $\mathcal{O}\left(\frac{L}{\mu}\ln(\frac{1}{\epsilon})\right)$ iterations, while if one applies the accelerated gradient or the heavy ball method with appropriately chosen parameters, this bound improves to $\mathcal{O}\left(\frac{L}{\mu}\ln(\frac{1}{\epsilon}\right)$. Finally, if we aim at using conjugate gradient, the result bound will be in $\mathcal{O}\left(\min\{d, \frac{L}{\mu}\ln(\frac{1}{\epsilon})\}\right)$.

Chapter 4

Stochastic gradient techniques

4.1 Introduction

In this chapter, we will develop method that fully exploit the structure of data science problems. To this end, we assume that we are given a dataset of n examples under the form $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ and $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$ are obtained from a certain data distribution. As in the linear regression example from the previous chapter, we seek a model function $h : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ such that $h(\boldsymbol{x}_i) \approx \boldsymbol{y}_i$ for any $i = 1, \dots, n$. We will parameterize this model by a vector $\boldsymbol{w} \in \mathbb{R}^d$ ($h(\boldsymbol{x}_i) = h(\boldsymbol{x}_i; \boldsymbol{w})$) so that knowing the vector $\boldsymbol{w}$ suffices to be able to evaluate the model.

In order to quantify the model's ability to represent the data, we define a cost function (or loss function) of the form $\ell : (h, y) \mapsto \ell(h, y)$: this function aims at penalizing values (h, y) such that $h \neq y$. The mean-square or ℓ_2 loss defined by $(h, y) \mapsto \|h - y\|^2$ is a canonical example of a loss. We then define the loss at a given example by $\ell(h(\boldsymbol{w}; \boldsymbol{x}_i), y_i)$.

We wish to compute the best model according to our entire dataset, leading to average the losses over all examples. This finally leads to the following optimization problem.

Definition 4.1.1 (Finite sum problem) *Consider a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ and $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$, a model class $\{h(\boldsymbol{w}; \cdot) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}\}_{\boldsymbol{w} \in \mathbb{R}^d}$ and a cost function ℓ. The* **finite-sum problem** *associated with this setup is defined by:*

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^n \ell(h(\boldsymbol{w}; \boldsymbol{x}_i), \boldsymbol{y}_i) = \frac{1}{n} \sum_{i=1}^n f_i(\boldsymbol{w}). \tag{4.1.1}$$

Provided the f_is are all $\mathcal{C}^1$ functions, the function f is also $\mathcal{C}^1$ and we can apply gradient descent to problem (4.1.1). The k-th iteration of this method is:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k) = \boldsymbol{w}_k - \frac{\alpha_k}{n} \sum_{i=1}^n \nabla f_i(\boldsymbol{w}_k).$$

From the iteration, it is clear that every gradient descent iteration requires to access**the entire dataset** for a single gradient calculation. In a big data setting, the number of examples n can be extremely large, and gradient descent can be too expensive to be used in practice.

Remark 4.1.1 *In a purely stochastic or "online", examples may only be available in a streaming fashion: in this context, it may be impossible to compute a finite average over all samples, and we*

consider instead the stochastic optimization problem:

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\left[f_{(\boldsymbol{x},\boldsymbol{y})}(\boldsymbol{w})\right]. \tag{4.1.2}$$

Even though the gradient of the objective may exist, it can be quite difficult to compute This formulation is often closer to the real goal of machine learning (i.e. have a good model for all possible examples in the distribution), yet it is often replaced by (4.1.1) to account for the empirical setting.

4.2 Stochastic gradient method

4.2.1 Algorithm

The idea behind stochastic gradient is remarkably simple. It considers the problem (4.1.1) under the assumption that every component function f_i possesses a gradient: every iteration then consists in choosing a random index i_k and taking a step in the direction of the negative gradient of the function f_{i_k}. Algorithm 4 details this process.

Algorithm 4: Stochastic gradient method

Initialization: $\boldsymbol{w}_0 \in \mathbb{R}^d$.

for $k = 0, 1, \ldots$ **do**

1. Define a stepsize $\alpha_k > 0$.
2. Draw an index $i_k \in \{1, \ldots, n\}$.
3. Compute the new iterate

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f_{i_k}(\boldsymbol{w}_k). \tag{4.2.1}$$

end

The vector $\nabla f_{i_k}(\boldsymbol{w}_k)$ is called a **stochastic gradient** for $\boldsymbol{w}_k$. The key property of iteration (4.2.1) is that it only requires one example from the dataset: as a result, **its cost in terms of accessing the data is n times lower compared to an iteration of gradient descent**. As a result, the comparison between the two methods is often conducted in terms of **epochs**.

Definition 4.2.1 (Epoch) *For problem (4.1.1), an epoch represents n calculations of a sample gradient ∇f_i.*

As a result, the cost of one iteration of gradient descent is that of one epoch, an iteration of Algorithm 4 only costs a fraction $(\frac{1}{n})$ of an epoch.

To conclude this presentation of stochastic gradient, we make several key remarks on its behavior.

Remark 4.2.1 *In general, it is not possible to guarantee convergence of gradient descent. Consider the two-dimensional problem:*

$$\underset{w\in\mathbb{R}}{\text{minimize}}\, \frac{1}{2}(f_1(w) + f_2(w))$$

with $f_1(w) = 2w^2$ and $f_2 = -w^2$. If $w_0 > 0$ and $i_k = 2$ for any k, then the method will diverge.

It is easy to build problems and draws of indices for which stochastic gradient does not converge. In practice, however, finite-sum problems from data science involve component functions that are quite similar, corresponding to data samples that follow a similar distribution. As a result, improving the loss specific to a given sample may very well improve other losses with respect to similar samples. This observation partly explains the success of stochastic gradient methods in data-related settings.

Remark 4.2.2 *Algorithm 4 is often called* **Stochastic Gradient Descent**, *or SGD. This denomination is inaccurate, in that one cannot guarantee that the stochastic gradient method will behave like a descent method (i.e. that it will decrease the function value at every iteration). For this reason, we will call this method* **stochastic gradient** *in these notes, thereby following other authors' terminology [1]. However, we point out that many librairies refer to their implementation of Algorithm 4 as SGD.*

4.2.2 Convergence rate analysis

In this section, we describe the key steps to deriving convergence rates (and complexity bounds) for stochastic gradient, under a slightly altered version of Assumption 3.0.1.

Assumption 4.2.1 *The objective function $f = \frac{1}{n}\sum_{i=1}^n f_i$ of problem (4.1.1) is $\mathcal{C}_L^{1,1}(\mathbb{R}^d)$ for some $L > 0$, and bounded below by $f_{low} \in \mathbb{R}$. Moreover, every function f_i is of class $\mathcal{C}^1$.*

The key argument for analyzing gradient descent is the result of Proposition 3.1.1: in particular, the inequality

$$f(\boldsymbol{w}_{k+1}) \leq f(\boldsymbol{w}_k) - \left(\alpha_k - \frac{L}{2}\alpha_k^2\right)\|\nabla f(\boldsymbol{w}_k)\|^2$$

leads to decrease guarantees for a sufficiently small step size.

It is possible to obtain a similar inequality in the case of stochastic gradient (in a probabilistic sense). To this end, assumptions on the random indices (and thus the stochastic gradients used throughout the algorithm) are necessary.

Assumption 4.2.2 (Stochastic gradients' properties) *For every iteration k of Algorithm 4, the random index i_k is drawn so that*

i) i_k *is independent of* $i_0, \ldots, i_{k-1}$;

ii) $\mathbb{E}_{i_k}\left[\nabla f_{i_k}(\boldsymbol{w}_k)\right] = \nabla f(\boldsymbol{w}_k)$;

iii) $\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(\boldsymbol{w}_k)\|^2\right] \leq \|\nabla f(\boldsymbol{w}_k)\|^2 + \sigma^2$ *with* $\sigma^2 \in (0, \infty)$.

The second property of Assumption 4.2.2 implies that the stochastic gradient $\nabla f_{i_k}(\boldsymbol{w}_k)$ is an unbiased estimator of the true gradient $\nabla f(\boldsymbol{w}_k)$. The third property guarantees that the norm of this stochastic estimator cannot deviate too much from the true norm, which is critical in guaranteeing some form of convergence. The most classical strategy to satisfy these assumptions is given below.

Example 4.2.1 (Uniform sampling) *Suppose that $\{i_k\}_k$ is a sequence of independent, identically distributed indices uniformly drawn in $\{1, \ldots, n\}$. Then, Algorithm 4 satisfies Assumption 4.2.2.*

Using Assumption 4.2.2, one can establish a useful inequality for analyzing Algorithm 4.

Proposition 4.2.1 *Let Assumptions 3.0.1 and 4.2.2 hold, and consider the k-th iteration of Algorithm 4. Then,*

$$\mathbb{E}_{i_k}\left[f(\boldsymbol{w}_{k+1})\right] - f(\boldsymbol{w}_k) \leq \nabla f(\boldsymbol{w}_k)^{\mathrm{T}}\, \mathbb{E}_{i_k}\left[\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\right] + \frac{L}{2}\, \mathbb{E}_{i_k}\left[\|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\|^2\right]. \tag{4.2.2}$$

As shown by (4.2.2), one can guarantee decrease over an iteration of stochastic gradient **in expectation** (provided the right-hand side of (4.2.2) is negative). This property is sufficient to obtain convergence rates (or complexity bounds) for stochastic gradient applied to strongly convex, convex and nonconvex problems. These results heavily depend on the choice for the stepsize sequence $\{\alpha_k\}_k$: in fact, tuning the stepsize is a major issue in machine learning (in which the stepsize is often referred to as a learning rate).

In the rest of this section, we mainly focus on strongly convex problems, and make the following assumption.

Assumption 4.2.3 *The objective function f of problem (4.1.1) is continuous and μ-strongly convex. It possesses a unique global minimum $\boldsymbol{w}^*$, and we let $f^* = f(\boldsymbol{w}^*)$.*

We first consider the case of a constant stepsize, for which we can establish the following result.

Theorem 4.2.1 (Stochastic gradient with constant stepsize) *Let Assumptions 3.0.1, 4.2.2 and 4.2.3 hold. Suppose that we apply Algorithm 4 with a constant stepsize*

$$\alpha_k = \alpha \in (0, \tfrac{1}{L}] \quad \forall k.$$

Then, for every $K \in \mathbb{N}$,

$$\mathbb{E}\left[f(\boldsymbol{w}_K) - f^*\right] \leq \frac{\alpha L \sigma^2}{2\mu} + (1 - \alpha\mu)^K \left[f(\boldsymbol{w}_0) - f^* - \frac{\alpha L \sigma^2}{2\mu}\right], \tag{4.2.3}$$

where the expected value is taken on all randomness up to iteration K.

The result of Theorem 4.2.1 shows a linear rate of convergence for stochastic gradient in expectation, which is comparable to that of gradient descent. However, this rate only guarantees that $\{f(\boldsymbol{w}_K)\}$ converges to a value in $\left[f^*, f^* + \frac{\alpha L \sigma^2}{2\mu}\right]$: even on average, stochastic gradient can only be guaranteed to converge towards a neighborhood of the optimal value. Indeed, the result (4.2.3) involves a constant term on its right-hand side, showing a possible gap to the optimum in $\frac{\alpha L \sigma^2}{4\mu}$. It is therefore not possible to guarantee that $\mathbb{E}\left[f(\boldsymbol{w}_k) - f^*\right] \leq \epsilon$ for any $\epsilon > 0$, but only for sufficiently large values of ϵ. However, the use of a constant stepsize guarantees a linear rate of convergence.

In the original version of stochastic gradient (proposed by Robbins and Monro in 1951), the sequence of stepsizes was required to satisfy

$$\sum_{k=0}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty,$$

which lead to asymptotic convergence of the method. In order to satisfy these requirements, the sequence α_k must go to 0 as k goes to infinity. Therefore, we now describe a convergence result based on using a decreasing stepsize.

Theorem 4.2.2 (Stochastic gradient with decreasing stepsize) *Under Assumptions 3.0.1, 4.2.2 and 4.2.3, consider Algorithm 4 applied with a decreasing stepsize sequence of the form*

$$\alpha_k = \frac{\beta}{k+\gamma},$$

with $\beta > \frac{1}{\mu}$ and $\gamma > 0$ chosen so that $\alpha_0 = \frac{\beta}{\gamma} \leq \frac{1}{L}$. Then, for any $K \in \mathbb{N}$, we have

$$\mathbb{E}\left[f(\boldsymbol{w}_K) - f^*\right] \leq \frac{\nu}{\gamma + K}, \tag{4.2.4}$$

where

$$\nu := \max\left\{\frac{\beta^2 L \sigma^2}{2(\beta\mu - 1)}, \gamma(f(\boldsymbol{w}_0) - f^*)\right\}.$$

As for gradient descent, using a decreasing stepsize sequence poses the risk of producing tiny steps early in the algorithmic run. However, note that it guarantees convergence of $\{f(\boldsymbol{w}_k)\}$ to f^* in expectation, which is a strongly result than that of Theorem 4.2.1. Still, the convergence rate in $\mathcal{O}(1/K)$ is worse than the one we established for gradient descent, in $\mathcal{O}((1-\mu/L)^K)$. Note however that this comparison is related to the number of iterations: if we include the per-iteration cost (in terms of accesses to the data), the stochastic gradient method has a convergence rate of $\mathcal{O}(1/K)$ (1 access per iteration) while the gradient descent method yields a rate in $\mathcal{O}(n(1-\mu/L)^K)$. The latter can be significantly worse than the former for large n.

Remark 4.2.3 (A hybrid approach) *A common stepsize strategy used in deep learning consists in adopting a learning rate schedule. In this hybrid approach, one runs first stochastic gradient with a constant value α upon a certain point, then replaces α by $\alpha' < \alpha$: the process is repeated until a suitable point has been found, or the budget of epochs is exhausted.*

In the strongly convex setting, the result of Theorem 4.2.1 provides a useful criterion to reduce the stepsize. Indeed, starting with a stepsize α, one can guarantee that a neighborhood of size $\frac{\alpha L \sigma^2}{\mu}$ can be reached in a bounded number of iterations (or epochs). Therefore, once that point is reached, one can reduce α. This approach guarantees to converge to the true optimal value, at a sublinear rate: for any $\epsilon > 0$,

$$\mathbb{E}\left[f(\boldsymbol{w}_k) - f^*\right] \leq \epsilon \quad \textit{after} \quad \mathcal{O}(1/\epsilon) \textit{ iterations.}$$

In practice, knowing how the neighborhood is reached may not be possible: other criteria, such as detecting stalling in the algorithm (lack of progress in the iterates, in the function values, etc), can help in designing strategies: several learning rate schedules are implemented in modern libraries.

Stepsize choice and nonconvex optimization Stochastic gradient and its variants are the most common methods for training deep neural networks: the associated optimization problem is highly nonconvex, and the above analysis cannot be applied. Still, it is possible to derive convergence rates for the nonconvex case by studying the following quantities:

- $\mathbb{E}\left[\frac{1}{K}\sum_{i=1}^K \|\nabla f(\boldsymbol{w}_k)\|^2\right]$ for the variants of Algorithm 4 based on constant stepsizes.
- $\mathbb{E}\left[\frac{1}{\sum_{i=1}^K \alpha_k}\sum_{i=1}^K \alpha_k \|\nabla f(\boldsymbol{w}_k)\|^2\right]$ for the variants using a decreasing stepsize sequence.

As in the strongly convex case, the iteration results for stochastic gradient will be worse than that of gradient descent. For instance, applying 4 with a constant stepsize guarantees that

$$\mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\|\nabla f(\boldsymbol{w}_k)\|^2\right] \leq \epsilon$$

in at most $\mathcal{O}(\epsilon^{-4})$ iterations. This bound is worse than the $\mathcal{O}(\epsilon^{-2})$ bound for gradient descent. In addition, this result is only valid for sufficiently large ϵ because of the stochastic nature of the method.

Remark 4.2.4 (Advanced stochastic gradient techniques) *The most common stochastic gradient variants for deep learning (*SGD WITH MOMENTUM, ADAM, ADAGRAD, RMSPROP*) build on the stochastic gradient principle by adding a momentum term to the stochastic gradient and/or a scaling to every component of the step so as to reduce the need for intensive stepsize tuning. Analyzing these methods is significantly more complicated than in the deterministic setting, and theoretical results have yet to be reconciled with practical behavior.*

4.3 Variance reduction

The theory of Section 4.2.2 relies heavily on Assumption 4.2.2, and more precisely on a control over the norm of the stochastic gradient through the quantity σ^2. A higher value of σ leads to looser neighborhoods of the solution and, as a result, worse theoretical guarantees. Numerically, this translates into high variability of the method's output.

Variance reduction techniques have been designed to reduce the "variance" of gradient estimates used in stochastic gradient. The goal of this section is to summarize key approaches to reducing variance.

4.3.1 Batch methods

As explained in the previous section, the iteration of stochastic gradient is

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f_{i_k}(\boldsymbol{w}_k),$$

where i_k is a random index drawn within $\{1, \ldots, n\}$. This method relies on a **single** example to build a stochastic gradient estimate, and this estimate comes with a certain variance (illustrated by the parameter σ^2 in Assumption 4.2.2). On the contrary, gradient descent relies on an exact gradient computed using all the samples, which leads to a deterministic execution and no variance in the (exact) gradient estimate.

To improve the variance of this method, it appears natural to consider stochastic gradient estimates based on several samples: this is the underlying principle of **batch stochastic gradient methods**.

At every iteration of a batch method, a (random) set of indices S_k with indices between 1 and n is drawn, and the following update is performed:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \frac{1}{|S_k|}\sum_{i\in S_k}\nabla f_i(\boldsymbol{w}_k) \tag{4.3.1}$$

When $|S_k| = 1$, we recover the classical stochastic gradient formula. When $|S_k| = n$ and the indices are drawn without replacement, we obtain $S_k = \{1, \ldots, n\}$, and iteration (4.3.1) corresponds to that of gradient descent.

More broadly, we can identify two regimes of batch sizes:

- $|S_k| \approx n$: the per-iteration cost of such a variant is close to that of gradient descent, hence these batch methods often exhibit a behavior comparable to that of gradient descent.
- $1 < |S_k| = n_b << n$: this regime, called mini-batching, is often thought as a good way to reduce variance while keeping the per-iteration cost to an affordable level.

Assuming that the batch size is constant over all iterations, i.e. that $|\mathcal{S}_k| = n_b \; \forall k$, it is possible to show that a mini-batch variant reaches a closer neighborhood than stochastic gradient. Every iteration of the mini-batch method is of course more expensive than an iteration of stochastic gradient (in fact, an epoch corresponds to n/n_b iterations of a batch stochastic gradient method with a fixed batch size of n_b). Note that the cost of batch methods can be mitigated by exploiting parallel computing resources. Indeed, in a distributed data setup, stochastic gradients can often be evaluated in parallel, which reduces the effective cost of batch techniques.

Proposition 4.3.1 *Under Assumptions 3.0.1 and 4.2.2, we have*

$$\mathbb{E}_{S_k}\left[\frac{1}{|S_k|}\sum_{i\in S_k}\nabla f_i(\boldsymbol{w}_k)\right] = \nabla f(\boldsymbol{w}_k)$$

and

$$\mathbb{E}_{S_k}\left[\left\|\frac{1}{|S_k|}\sum_{i\in S_k}\nabla f_i(\boldsymbol{w}_k)\right\|_2^2\right] \leq \frac{\sigma^2}{n_b}.$$

Finally, we mention that (mini)-batch approaches remain more expensive than "vanilla" stochastic gradient, while being more sensitive to redundancies in the data, and introducing an new hyperparameter to be tuned (the batch size). For this reason, the classical stochastic gradient approach can still be observed to be more efficient on some example, and it remains a popular algorithmic choice in practice.

4.3.2 Other variance reduction techniques

Gradient aggregation techniques are variance reduction approaches with well-established convergence rates, that are provably better than that of stochastic gradient. They have attracted significant interest from the academic machine learning and optimization community, though their use in modern data science settings like deep learning has not been successful (these techniques are still efficient on simple models, and are part of librairies like `scikit-learn`). They essentially consist in maintaing a **full gradient estimate** throughout the iterations, which requires at least one full gradient calculation: the gradient estimate used to make a step then combines the full gradient estimator with a new stochastic gradient, leading to a corrected step, and a method with provably smaller variance.

Iterate averaging is another way to reduce the variance in stochastic gradient at no additional cost in terms of accesses to data points. The idea consists in analyzing the properties of a running

average $\frac{1}{K}\sum_{k=0}^{K-1} \boldsymbol{w}_k$. In certain cases (in particular, when $\alpha = \frac{1}{\mu(k+1)}$ and f is μ-strongly convex), this sequence possesses favorable properties and its behavior is less variable. However, computing this average either requires to store all iterates (which becomes expensive in terms of storage) or necessitates to maintain a running average (subject to numerical errors).

4.4 Stochastic gradient methods for deep learning

In this section, we review the main stochastic gradient techniques that are used to train deep learning models. Our focus remains on finite-sum problems of the form (4.1.1) under Assumption 4.2.1. Our goal is to study several variants on the iterative scheme

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha \boldsymbol{g}_k, \tag{4.4.1}$$

where $\alpha > 0$ is a fixed stepsize (or learning rate), and $\boldsymbol{g}_k$ is a stochastic gradient estimator, that may correspond to taking a single term from the finite sum (as in stochastic gradient) or to a batch of indices.

To encompass all the variants of interest, we consider a more general iteration of the form

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha \boldsymbol{m}_k \oslash \boldsymbol{v}_k, \tag{4.4.2}$$

where $\alpha > 0$, $\boldsymbol{m}_k, \boldsymbol{v}_k \in \mathbb{R}^d$ and $\oslash$ denotes the componentwise division operator:

$$\boldsymbol{m}_k \oslash \boldsymbol{v}_k := \left[\frac{[\boldsymbol{m}_k]_i}{[\boldsymbol{v}_k]_i}\right]_{i=1,\dots,d}.$$

To see that (4.4.2) generalizes (4.4.1), set $\boldsymbol{m}_k = \boldsymbol{g}_k$ and $\boldsymbol{v}_k = \mathbf{1}_{\mathbb{R}^d}$. The iteration (4.4.2) is particularly convenient to express the popular methods in deep learning using a single format.

4.4.1 Stochastic gradient with momentum

Most practical implementations of stochastic gradient combine the basic step (4.4.1) together with a momentum term, similarly to the accelerated methods described in Chapter 3. An iteration of gradient descent with momentum reads

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha(1-\beta)\boldsymbol{g}_k + \alpha\beta\left(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\right), \tag{4.4.3}$$

where $\beta \in (0,1)$ is a momentum parameter (when $\beta = 0$ we again obtain the standard stochastic gradient iteration). Iteration (4.4.3) can be seen as a version of Polyak's method where the gradient-type step is combined with the previous direction. As for the heavy-ball method, the idea is to incorporate information from the previous step through momentum. In practice, the iteration (4.4.3) often leads to accumulation of good steps (in terms of optimization), whereas bad directions and bad steps tend to cancel out.

The basic method (4.4.3) can be recovered from (4.4.2), by setting $\boldsymbol{v}_k = \mathbf{1}_{\mathbb{R}^d}$ and defining $\boldsymbol{m}_k$ in a recursive manner through $\boldsymbol{m}_{-1} = \mathbf{0}_{\mathbb{R}^d}$ and

$$\boldsymbol{m}_k = (1-\beta)\boldsymbol{g}_k - \beta \boldsymbol{m}_{k-1} \quad \forall k \in \mathbb{N}.$$

where $\beta \in (0,1)$.

Stochastic gradient with momentum is implemented in most deep learning librairies such as PyTorch. It has shown great success in training deep neural networks on computer vision problems, and is partly responsible for the rise of deep learning in the early 2010s.

Remark 4.4.1 *Theoretical guarantees for the iteration (4.4.3) are much more difficult to obtain than for accelerated gradient (in particular, it is not well understood whether such a method can be provably faster than SGD). Nevertheless, stochastic gradient techniques with momentum are widely used in practice, even on nonconvex problems such as neural network training.*

4.4.2 AdaGrad

The adaptive gradient method, or AdaGrad, was proposed in 2011 to address the issue of setting the learning rate α in stochastic gradient. Rather than using costly procedures such as line search, AdaGrad scales every coordinate of the stochastic gradient using information from the values of that coordinate in the previous iterations. Mathematically, the method maintains a sequence $\{\boldsymbol{r}_k\}_k$ defined by

$$\forall i = 1, \ldots, d, \quad \left\{ \begin{array}{l} [\boldsymbol{r}_{-1}]_i = 0 \\ [\boldsymbol{r}_k]_i = [\boldsymbol{r}_{k-1}]_i + [\boldsymbol{g}_k]_i^2 \quad \forall k \geq 0, \end{array} \right. \tag{4.4.4}$$

The AdaGrad iteration can then be written as

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha \boldsymbol{g}_k \oslash \sqrt{\boldsymbol{r}_k}, \tag{4.4.5}$$

where the square root is applied componentwise to $\boldsymbol{r}_k$. This iteration is a special cas of (4.4.2), where $\boldsymbol{m}_k = \boldsymbol{g}_k$ and $\boldsymbol{v}_k = \sqrt{\boldsymbol{r}_k}$. The novelty in AdaGrad does not lie in the use of momentum, but in the use of one stepsize per coordinate. The stepsize sequence thus has the form

$$\left\{ \left[\frac{\alpha}{\sqrt{[\boldsymbol{r}_k]_i}} \right]_{i=1}^{d} \right\}_k .$$

The resulting *diagonal scaling* on the coordinates of $\boldsymbol{g}_k$ leads to stepsizes that adapt to coordinates that can vary by orders of magnitude (which would require a careful choice of α in basic stochastic gradient). On the other hand, the accumulation process at work in the definition of $\boldsymbol{r}_k$ results in stepsizes that are monotonically decreasing, and that often converge quickly towards 0.

Remark 4.4.2 *In practice, $\boldsymbol{r}_k$ is replaced by $\boldsymbol{r}_k + \eta \mathbf{1}_{\mathbb{R}^d}$ where $\eta > 0$ is a small value that helps with numerical stability.*

The AdaGrad method is particularly interesting for problems with sparse gradients, in which stochastic gradients tend to have many zero coordinates. In this situation, using $\boldsymbol{r}_k$ will only modify the stepsizes corresponding to nonzero gradient coordinates. Many problems in recommendation systems have a sparse structure, and AdaGrad is considered to be an efficient method for this class of problems.

4.4.3 RMSProp

The RMSProp (*Root Mean Square Propagation*) algorithm is similar in spirit to AdaGrad, in that it scales the gradient components at every step. This method relies on a sequence $\{\boldsymbol{r}_k\}_k$ defined by

$$\forall i = 1, \ldots, d, \quad \left\{ \begin{array}{l} [\boldsymbol{r}_{-1}]_i = 0 \\ [\boldsymbol{r}_k]_i = (1-\lambda)[\boldsymbol{r}_{k-1}]_i + \lambda [\boldsymbol{g}_k]_i^2 \quad \forall k \geq 0, \end{array} \right. \tag{4.4.6}$$

where $\lambda \in (0,1)$. The value of λ is used to put more weight on the current gradient coordinates than on the coordinates from past iterations (this information being contained in $\boldsymbol{r}_{k-1}$). This simple idea slows down the decrease of the stepsizes to 0, compared to the stepsizes of AdaGrad.

With the definition (4.4.6), the RMSProp iteration corresponds has the same form as that of AdaGrad, that is, a special case of (4.4.2) with $\boldsymbol{m}_k = \boldsymbol{g}_k$ and $\boldsymbol{v}_k = \sqrt{\boldsymbol{r}_k}$.

Remark 4.4.3 *As for* AdaGrad, *standard practice replaces* $\boldsymbol{r}_k$ *by* $\boldsymbol{r}_k + \eta \mathbf{1}_{\mathbb{R}^d}$, *where* $\eta > 0$ *is a small quantity.*

The RMSProp method has been found quite successful for training very deep neural networks.

4.4.4 Adam

The Adam optimization method was proposed in 2013. This method can be thought as combining the idea of momentum (used in the stochastic gradient method of Section 4.4.1) with the diagonal scaling procedure on which both AdaGrad and RMSProp are based. An iteration of Adam falls into the generic scheme (4.4.2) by setting

$$\boldsymbol{m}_k = \frac{(1-\beta_1)\sum_{j=0}^{k} \beta_1^{k-j} \boldsymbol{g}_j}{1-\beta_1^{k+1}} \tag{4.4.7}$$

and

$$\boldsymbol{v}_k = \sqrt{\frac{(1-\beta_2)\sum_{j=0}^{k} \beta_2^{k-j} \boldsymbol{g}_j \odot \boldsymbol{g}_j}{1-\beta_2^{k+1}}}. \tag{4.4.8}$$

Here $\beta_1, \beta_2 \in (0,1)$, and $\odot$ denotes the Hadamard or componentwise product

$$\boldsymbol{g}_k \odot \boldsymbol{g}_k = \left[[\boldsymbol{g}_k]_i^2\right]_{i=1}^d.$$

Remark 4.4.4 *In practice,* $\boldsymbol{v}_k + \eta \mathbf{1}_{\mathbb{R}^d}$ *(with small* $\eta > 0$*) is used in lieu of* $\boldsymbol{v}_k$.

The above formulas describe the two components of Adam. On one hand, a weighted combination of the previous steps that puts the emphasis on the most recent steps (and the current stochastic gradient) defines the direction of the next step. On the other hand, a diagonal scaling is applied to the coordinates of this direction, again according to a weighted average of the coordinates from the previous iterations. This important feature, that has a statistical motivation, appears to be responsible for the success of Adam in practice. The impressive performance of Adam on training neural networks has contributed to its popularity, and it remains the preferred method today in numerous applications. In particular, Adam and its variant AdamW (based on regularization principle) are quite efficient on natural language processing models.

4.5 Conclusion

Stochastic gradient methods rely on partial gradient information selected in a random fashion: as such, they are not guaranteed to succeed deterministically, and their convergence guarantees are typically weaker than that of gradient descent. Nevertheless, they have proven very successful in data-driven problems, wherein computing the gradient involves accessing a massive amount of data

points. In this setting, stochastic gradient approaches have an attractive, low per-iteration cost, and typically make fast, significant progress early on in the algorithmic run. The nature of the data is key to the performance of stochastic gradient: on standard datasets from machine and deep learning, it appears beneficial to resort to such techniques.

Multiple variants of stochastic gradient have been developed in academic and industrial contexts. Batch stochastic gradient techniques are among the most popular, as they allow to incorporate more than one sample into the gradient approximation: this leads to variance reduction, a concept that is driving some of the latest advances in stochastic gradient.

Chapter 5

Nonsmooth optimization and regularization

The purpose of this chapter is to address two common characteristics of data science problems: the possible lack of differentiability of the optimization problem on one hand, and the desire to produce models with a specific structure on the other hand. We illustrate these issues using a classical learning paradigm, then describe the underlying optimization concepts.

5.1 Introductory example: The perceptron method

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n vectors of $\mathbb{R}^d$, and suppose each vector $\boldsymbol{x}_i$ is given a label $y_i \in \{-1, 1\}$. We wish to design a linear model $\boldsymbol{x} \mapsto \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}$ that correctly classifies the data. Since this classification is binary and corresponds to finding a sign (positive or negative. Therefore, we will consider that the model correctly classifies an input if $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} \gg 1$ and $y_i = 1$, or $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} \ll 1$ and $y_i = -1$, and such models should have a small loss with respect to the ith data point. Conversely, models such that $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} \gg 1$ and $y_i = -1$ or $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} \ll 1$ and $y_i = 1$ should have a very large loss with respect to the ith data point. Finally, models for which $|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}| \leq 1$ should also be penalized, as they are likely to correctly (or incorrectly) classify data points by a small margin: we would rather correctly classify by a large margin.

To this end, we will use the following loss function, called the hinge loss:

$$\ell(h, y) := \max\{1 - yh, 0\}. \tag{5.1.1}$$

Then, one can translate the binary classification problem into the following optimization problem:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{1 - y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}, 0\} + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \tag{5.1.2}$$

where $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ represents the dataset, and $\lambda > 0$. The solution to this problem belongs to the family of **support vector machine** models, or SVMs.

One of the earliest methods that was proposed to solve this algorithm is the **perceptron algorithm**, given in Algorithm 5.

In its basic form, the preceptron algorithm is quite similar to stochastic gradient with a constant step size, in that it selects a single sample at every iteration and performs an update based on this

Algorithm 5: Perceptron algorithm for problem 5.1.3.

Initialization: $\boldsymbol{w}_0 \in \mathbb{R}^d$, $\alpha > 0$.
for $k = 0, 1, \ldots$ **do**

1. Draw an index $i_k \in \{1, \ldots, n\}$ at random.
2. Compute the new iterate as

$$\boldsymbol{w}_{k+1} = \left(1 - \frac{\alpha\lambda}{n}\right)\boldsymbol{w}_k + \begin{cases} \alpha y_{i_k}\boldsymbol{x}_{i_k} & \text{if } 1 - y_{i_k}\boldsymbol{x}_{i_k}^{\mathrm{T}}\boldsymbol{w}_k > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.1.3}$$

end

value. However, the *hinge loss* is a **nonsmooth** function, i.e. the gradient does not exist at every point. Therefore, we cannot consider that the perceptron algorithm is the stochastic gradient method *stricto sensu*. Still, for structured functions such as the hinge loss, it is possible to define quantities that act as a proxy for the gradient, and can thus drive the optimization process: we detail these aspects in Section 5.2.

Another interesting property of the problem (5.1.2) is that the objective function involves two terms: the hinge loss term, which depends on the data and possesses the finite-sum structure we already saw in Chapter 4, and a regularizing term, **which does not depend on the data** and serves to enforce structural properties on the solution. We will address this topic and the associated algorithms in Section 5.3.

5.2 Nonsmooth optimization

5.2.1 From nonsmooth functions to nonsmooth problems

Problems such as (5.1.2), that involve a function possibly not differentiable, are termed *nonsmooth problems*. They involve functions that we will call nonsmooth (by opposition with smooth) : for the purpose of these notes, we will define nonsmooth functions as follows.

Definition 5.2.1 (Nonsmooth functions) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is called* **nonsmooth** *if it is not differentiable everywhere.*

Remark 5.2.1 *A nonsmooth function can be continuous (this is the case for the hinge loss above).*

Example 5.2.1 *Examples of nonsmooth functions*

- $w \mapsto |w|$ *from* $\mathbb{R}$ *to* $\mathbb{R}$;
- $w \mapsto \|w\|_1$ *from* $\mathbb{R}^d$ *to* $\mathbb{R}$;
- *ReLU:* $w \mapsto \max\{w, 0\}$ *from* $\mathbb{R}^d$ *to* $\mathbb{R}$.

Since nonsmooth functions are not differentiable everywhere, optimization problems that involve nonsmooth functions may be impossible to solve via gradient-based methods. Still, several approaches can be used to tackle these problems.

One useful technique consists in reformulating a nonsmooth problem as a smooth one when possible. For instance, the problem $\min_{w\in\mathbb{R}} |w|$ is equivalent to

$$\min_{w,t^+,t^-\in\mathbb{R}} t^+ + t^- \quad \text{s. t.} \quad w = t^+ - t^-, t^+ \geq 0, t^- \geq 0.$$

This reformulation is a smooth problem involving only linear objective and constraints, which is easily solvable by smooth solvers.

Another technique, frequently employed in practice, consists in working with functions that are nonsmooth but Lipschitz continuous (denoted by $\mathcal{C}_L^{0,0}$, by analogy with $\mathcal{C}_L^{1,1}$) and using a gradient-based scheme. This approach is motivated by the following property.

Theorem 5.2.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous function. Then it is differentiable at almost every point in $\mathbb{R}^d$.*

For instance, the ReLU function is Lipschitz continuous (not differentiable at 0) thus most constructions involving ReLU (such as neural networks) would not be differentiable everywhere. However, most algorithms will operate under the assumption that the function is indeed differentiable. This is the case for most points (in fact, almost every point), but nonsmooth functions are likely to be non-differentiable at their minima, should they possess one.

5.2.2 Subgradient methods

In the case of convex functions, one can define a proxy for the gradient called the subgradient.

Definition 5.2.2 (Subgradient and subdifferential) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. A vector $\boldsymbol{g} \in \mathbb{R}^d$ is called a subgradient of f at $\boldsymbol{w} \in \mathbb{R}^d$ if*

$$\forall \boldsymbol{z} \in \mathbb{R}^n, \qquad f(\boldsymbol{z}) \;\geq\; f(\boldsymbol{w}) + \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{z} - \boldsymbol{w}).$$

The set of all subgradients of f at $\boldsymbol{w}$ is called the subdifferential of f at $\boldsymbol{w}$, and denoted by $\partial f(\boldsymbol{w})$.

Note that when the function f is differentiable at $\boldsymbol{w}$, we have $\partial f(\boldsymbol{w}) = \{\nabla f(\boldsymbol{w})\}$, thus the notion of subdifferential matches that of the gradient for differentiable functions.

The interest of subgradients is further illustrated by the following result.

Theorem 5.2.2 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and $\boldsymbol{w} \in \mathbb{R}^d$.*

$$\mathbf{0} \in \partial f(\boldsymbol{w}) \quad \Leftrightarrow \quad \boldsymbol{w} \text{ minimum of } f.$$

Example 5.2.2 *Let $f : \mathbb{R} \to \mathbb{R}$, $f(w) = |w|$.*

$$\partial f(w) \;=\; \begin{cases} -1 & \text{if } w < 0 \\ 1 & \text{if } w > 0 \\ [-1, 1] & \text{if } w = 0. \end{cases}$$

The set $[-1, 1]$ contains 0, which confirms that $w^* = 0$ is the minimum of f.

Algorithm 6: Subgradient descent method.

Initialization: $\boldsymbol{w}_0 \in \mathbb{R}^d$.
for $k = 0, 1, \ldots$ **do**
1. Compute a subgradient $\boldsymbol{g}_k \in \partial f(\boldsymbol{w}_k)$.
2. Compute a steplength $\alpha_k > 0$.
3. Set $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \boldsymbol{g}_k$.

end

Remark 5.2.2 *Subgradients can also be defined for nonconvex functions, however in that case the subdifferential may be empty (typically at local maxima of the function).*

By analogy with gradient descent, we can design a subgradient method, as shown by Algorithm 6.

Such a method offers a flexibility in choosing the subgradient, which can be an issue. Moreover, choosing the stepsize is more difficult than for gradient descent, due to the nonsmooth nature of the problem. In fact, a subgradient can lead to increase in the function value for any stepsize, hence the choice of subgradient is critical to the success of this method.

Variants of subgradient method Based on the existing variants on the gradient descent paradigm, one can build algorithms that incorporate momentum and/or stochastic aspects; however, their analysis is also more intricate.

5.3 Regularization

5.3.1 Regularized problems

As we mentioned in introduction, a common practice in machine learning problems consists in enforcing a specific structure of the machine learning model through the objective function. Such regularized problems have the following form :

$$\min_{\boldsymbol{w} \in R^d} \underbrace{f(\boldsymbol{w})}_{loss\ function} + \underbrace{\lambda \Omega(\boldsymbol{w})}_{regularization\ term}.$$

where $\lambda > 0$ is called a regularization parameter.

Example 5.3.1 (Ridge regularization) *A problem with ridge regularization has the following form:*

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2.$$

The ridge regularizer $\boldsymbol{w} \mapsto \frac{1}{2}\|\boldsymbol{w}\|^2$ has several interpretations. It effectively penalizes $\boldsymbol{w}$s with large components, and can be shown to be equivalent to a constraint on the squared norm $\|\boldsymbol{w}\|^2$. In addition, a ridge regularizer has the effect to reduce the variance of the problem solution with respect to the data. Finally, when the regularizer $\lambda > 0$ is big enough, this often turns the objective function into a strongly convex one, with the positive implications in terms of convergence speed and uniqueness of the (global) minimum.

5.3.2 Sparsity-inducing regularizers

While computing a model to explain some data, we might want to compute a model that explains the data using as few features as possible[1]. Mathematically speaking, if our model is parameterized by a vector $\boldsymbol{w} \in \mathbb{R}^d$, our goal is to compute a vector that explains the data with as few nonzero coordinates as possible.

There exists a regularizer that penalized vectors with nonzero components (not just large as opposed to the ridge regularizer), called the ℓ_0 norm [2]. An ℓ_0-regularized problem has the form

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_0, \quad \|\boldsymbol{v}\|_0 = |\{i | [\boldsymbol{v}]_i \neq 0\}|.$$

However, this function is nonsmooth and discontinuous; its combinatorial nature also introduces more complexity to the original problem. As a result, researchers have turned to an intermediate regularization term, the ℓ_1 norm defined by

$$\|\boldsymbol{w}\|_1 = \sum_{i=1}^{d} |w_i|. \tag{5.3.1}$$

This function is continuous and convex; moreover, it is a norm function, which endows it with many desirable properties.

An illustration of this method is given below.

Example 5.3.2 *LASSO (Least Absolute Shrinkage and Selection Operator) Consider the setting of linear regression with data $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$. With an ℓ_1 regularizer, the problem becomes:*

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{w}\|_1.$$

The solution of this problem is known to possess fewer nonzero elements than the un-regularized, least-squares solution.

5.3.3 Proximal methods

Following our introduction of regularized problems in the previous section, we now describe optimization algorithms tailored to such formulations.

We begin by describing our problem class of interest.

Definition 5.3.1 (Composite optimization) *A composite optimization problem is of the form:*

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) + \lambda \Omega(\boldsymbol{w}),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth, $\mathcal{C}^{1,1}$ function, $\lambda > 0$ and $\Omega : \mathbb{R}^d \to \mathbb{R}$ is a convex, nonsmooth regularizer.

The **proximal approach** follows a classical optimization paradigm, in which a given problem is replaced by a sequence of easier problems called subproblems (note that all methods that we covered in these notes implicitly rely on these techniques). In the case of proximal methods, one aims at

Algorithm 7: Proximal gradient method.

Initialization: $\boldsymbol{w}_0 \in \mathbb{R}^d$.
for $k = 0, 1, \ldots$ **do**

1. Compute the gradient of the smooth part $\nabla f(\boldsymbol{w}_k)$.
2. Compute a steplength $\alpha_k > 0$.
3. Compute $\boldsymbol{w}_{k+1}$ such that

$$\boldsymbol{w}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{w}_k) + \tfrac{1}{2\alpha_k}\|\boldsymbol{w} - \boldsymbol{w}_k\|_2^2 + \lambda\Omega(\boldsymbol{w}) \right\}. \tag{5.3.2}$$

end

exploiting the smoothness of f to obtain easier problems, while using the structure of Ω directly into the subproblems.

Algorithm 7 gives a sketch of a proximal gradient method. The cost of an iteration of this algorithm is clearly more than that of other methods we have seen so far, given that it includes a gradient calculation as well as solving an auxiliary optimization problem (5.3.2), called the proximal subproblem.

Remark 5.3.1 *If $\Omega \equiv 0$ (i. e. Ω is the zero function and the problem is un-regularized), one can show that the solution of (5.3.2) is given by*

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k).$$

We thus recognize the gradient iteration of Algorithm 1.

Proximal gradient methods can be designed using most of the tools that can be applied to gradient descent : this includes stepsize choices, acceleration as well as stochastic aspects. Moreover, complexity results exist for nonconvex and convex f, though the latter has attracted more attention in the literature.

Example of proximal method: ISTA We end this section on proximal methods by a instance of Algorithm 7 that has proven successful in signal and image processing. This method is dedicated to solving problems with an ℓ_1 regularization term, of the form:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1.$$

Unlike for general regularizers, one can obtain a closed-form solution of the subproblem (5.3.2). Indeed, the proximal subproblem, given by

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{w}_k) + \tfrac{1}{2\alpha_k}\|\boldsymbol{w} - \boldsymbol{w}_k\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \right\},$$

[1] The goal of this process is *feature selection*.
[2] Though technically this function defines a semi-norm.

has a unique solution. To obtain it, one computes the usual gradient step $\boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)$, then one applies the **soft-thresholding function** $s_{\alpha_k \lambda}(\bullet)$ to each component, where this function is given by

$$\forall \mu > 0, \forall t \in \mathbb{R}, \qquad s_\mu(t) = \begin{cases} t + \mu & \text{if } t < -\mu \\ t - \mu & \text{if } t > \mu \\ 0 & \text{otherwise.} \end{cases}$$

As a result, the solution of the proximal subproblem is defined component-wise according to the components of the gradient step. The resulting update is at the heart of the corresponding proximal algorithm, called ISTA (Iterative Soft-Thresholding Algorithm): a description of ISTA is given in Algorithm 8.

Algorithm 8: ISTA: Iterative Soft-Thresholding Algorithm.

Initialization: $\boldsymbol{w}_0 \in \mathbb{R}^d$.
for $k = 0, 1, \ldots$ **do**

1. Compute the gradient of the smooth par $\nabla f(\boldsymbol{w}_k)$.
2. Compute a steplength $\alpha_k > 0$.
3. Compute $\boldsymbol{w}_{k+1}$ component-wise through the following rule

$$[\boldsymbol{w}_{k+1}]_i = \begin{cases} [\boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)]_i + \alpha_k \lambda & \text{if } [\boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)]_i < -\alpha_k \lambda \\ [\boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)]_i - \alpha_k \lambda & \text{if } [\boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)]_i > \alpha_k \lambda \\ 0 & \text{if } [\boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)]_i \in [-\alpha_k \lambda, \alpha_k \lambda]. \end{cases} \tag{5.3.3}$$

end

It can be shown that the use of the soft-thresholding function does promote zero components in the new iterates, which results in sparser solutions at the end of the algorithmic run.

Remark 5.3.2 *A notable improvement on ISTA was the inclusion of momentum, which resulted in a new algorithm called FISTA (Fast ISTA): this method is now the most widely used instance of ISTA.*

5.4 Conclusion

Nonsmoothness is a very common property in optimization, that can lead to mild or major challenges in implementing algorithms to minimize nonsmooth functions. In certain cases, the structure and the impact of nonsmoothness are well understood; in other cases, generalized notions of derivative such as subgradients may have to be used in order to design optimization algorithms.

Nonsmoothness frequently arises in regularized problem, where the goal is to enforce structural properties for a model, that do not depend on the data. The optimization schemes of choice for these problems are proximal gradient methods, that proceed by solving subproblems involving the regularizer. For instance, for ℓ_1 regularization, that promotes sparsity of the solution, the proximal

gradient algorithm can be written in an explicit form (ISTA). Note that a regularizer need not be nonsmooth: for smooth problems, proximal gradient is equivalent to gradient descent. This is for instance the case with the ℓ_2 regularizer, that aims at reducing variance with respect to the data.

Bibliography

[1] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Rev.*, 60:223–311, 2018.

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

[3] S. Boyd and L. Vandenberghe. *Introduction to Applied Linear Algebra - Vectors, Matrices and Least Squares*. Cambridge University Press, Cambridge, United Kingdom, 2018.

[4] M. C. Ferris, O. L. Mangasarian, and S. J. Wright. *Linear programming with MATLAB*. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2007.

[5] S. J. Wright. Optimization algorithms for data analysis. In 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.

Appendix A

Notations and mathematical tools

A.1 Notations

A.1.1 Generic 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 : $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{C}}$.
- 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.

A.1.2 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 $\mathbb{R}^d$ is used for the set of vectors with $d \in \mathbb{N}$ real components; although we do not explicitly indicate it in the rest of these notes, we always assume that $d \geq 1$.

- A vector $\boldsymbol{w} \in \mathbb{R}^d$ is thought as a column vector, with $w_i \in \mathbb{R}$ denoting its i-th coordinate in the canonical basis of $\mathbb{R}^d$. We thus write $\boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$, or, in a compact form, $\boldsymbol{w} = [w_i]_{1 \leq \imath \leq d}$.

- Given a column vector $\boldsymbol{w} \in \mathbb{R}^d$, the corresponding row vector is denoted by $\boldsymbol{w}^{\mathrm{T}}$, so that $\boldsymbol{w}^{\mathrm{T}} = [w_1 \ \cdots \ w_d]$ and $[\boldsymbol{w}^{\mathrm{T}}]^{\mathrm{T}} = \boldsymbol{w}$.

- For any integer $d \geq 1$, the vectors $\mathbf{0}_d$ and $\mathbf{1}_d$ correspond to the vectors of $\mathbb{R}^d$ for which all elements are 0 or 1, respectively. For simplicity, we may write $\boldsymbol{w} \geq 0$ to indicate that all components of $\boldsymbol{w}$ are nonnegative.

A.1.3 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).

A.2 Mathematical tools

Optimization has mathematical roots in real analysis, mostly through differential calculus. Linear algebra structures also play a major role in optimization (and data science). In this section, we list the basic results that will be used in the course.

For a deeper dive into these notions, the following links are recommended.

- For linear algebra:
  - https://www.ceremade.dauphine.fr/∼carlier/polyalgebre.pdf (in Frnech);
  - http://vmls-book.stanford.edu/vmls.pdf (Chapters 1 to 3, in English).
- For differential calculus:
  - https://www.ceremade.dauphine.fr/∼bouin/ens1819/Cours_Bolley.pdf (in French);
  - https://sebastianraschka.com/pdf/books/dlb/appendix_d_calculus.pdf (in English).

A.2.1 Vector linear algebra

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

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

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

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

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

Definition A.2.1 *A set $\mathcal{S} \subseteq \mathbb{R}^d$ satisfying the conditions*

1. $\mathbf{0}_d \in \mathcal{S}$;
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 $\mathbb{R}^d$.*

Definition A.2.2 *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ be p vectors in $\mathbb{R}^d$. 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 subspace of $\mathbb{R}^d$ 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 A.2.3
- *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 subspace of $\mathbb{R}^n$ has dimension at most n.

Example A.2.1 *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* i*th vector of the canonical basis (with a* 1 *in the* i*th 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 A.2.4 *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 A.2.1 *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. $\|\boldsymbol{x}\| = 0 \ \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}_{\mathbb{R}^n}$*;*
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 A.2.5 *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 A.2.1 *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 A.2.2 *The last inequality is a key result in both linear algebra and analysis. In this course, it will play a major role in deriving Taylor-type inequalities.*

A.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 A.2.3 *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 subspaces are of interest.

Definition A.2.6 (Matrix subspaces) *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.*

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

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

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

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

  The dimension of this 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 A.2.1 (Rank-nullity theorem) *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then,*

$$\dim(\ker(\boldsymbol{A})) + \operatorname{rang}(\boldsymbol{A}) = n.$$

Definition A.2.7 (Matrix norms) *Consider the space $\mathbb{R}^{m\times n}$. The operator norm $\|\cdot\|$ and the Frobenius norm $\|\cdot\|_F$ are defined by*

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

Definition A.2.8 (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 A.2.9 (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 $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}_n$ (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 A.2.10 (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 $\boldsymbol{A} \succ 0$.

Definition A.2.11 (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 A.2.1 *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 A.2.12 (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 A.2.2 *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 A.2.2 (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 A.2.4 *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*).

A.2.3 Calculus

Note: This section gives additional background to the concepts and properties introduced in Chapter 1.

Definition A.2.13 (Continuity) *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called* **continuous in $\boldsymbol{x} \in \mathbb{R}^n$** *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 A.2.5 *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 A.2.14 (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 A.2.2 *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 A.2.15 (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*

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

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

Corollary A.2.1

- *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:*

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

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

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 A.2.3 (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 $h : \mathbb{R}^n \mapsto \mathbb{R}^p$ 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 A.2.6 *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 A.2.4 (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 A.2.16 (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 A.2.5 (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 A.2.17 (Lipschitz continuity) *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is L-**Lipschitz continuous** on $\mathcal{A} \subset \mathbb{R}^n$ if*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}^2, \quad \|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

Proposition A.2.3 *Any Lipschitz continuous function on a set is continuous on this set.*

Definition A.2.18 (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 $p \in \mathbb{N}$.*
- *We say that f is $\mathcal{C}_L^{p,p}(\mathbb{R}^d)$ (or simply $\mathcal{C}_L^{p,p}$) if it is differentiable p times and its pth-order derivative is L-Lipschitz continuous.*

Theorem A.2.6 (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*

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

Moreover, if $f \in \mathcal{C}_L^{1,1}(\mathbb{R}^d)$, then

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{A.2.1}$$

Theorem A.2.7 (Taylor expansion of order 2) *Let $f \in \mathcal{C}^2(\mathbb{R}^d)$. For any vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of $\mathbb{R}^d$, one has*

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}\int_0^1 (\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}} \nabla^2 f(\boldsymbol{x} + t\,(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})\, dt.$$

Moreover, if $f \in \mathcal{C}_L^{2,2}(\mathbb{R}^d)$, then

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}} \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^3. \tag{A.2.2}$$

A.3 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 random variables and random vectors defined as follows:

- random variables z defined on a probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ by
$$\forall B \in \mathcal{B}(\mathbb{R}), \quad \mathbb{P}(z \in B) = \mathbb{P}(B);$$
- random vectors $\boldsymbol{z} = \begin{bmatrix} z_1 \\ \cdots \\ z_d \end{bmatrix}$ of size d, defined on the probability space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P})$.

In both case, the set of events will be the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$.

A.3.1 Random variables

A random variable is a function from a probability space onto another space that induces a probability measure in the latter. This notion is often restricted to the scalar case (in which case the arrival space is $\mathbb{R}$). Random vectors, that correspond to multidimensional outputs, will be addressed in a subsequent section.

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

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

- *A discrete set of possible values $\mathcal{Z} = \{z_i\} \subset \mathbb{R}$;*
- *An associated set of probabilities $p = \{p_i\}$ such that $p_i \geq 0$, $\sum_i p_i = 1$ and*
$$\forall \mathcal{S} \subset \mathcal{Z}, \quad \mathbb{P}(z \in \mathcal{S}) = \sum_{z_i \in \mathcal{S}} p_i.$$

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

- *A continuous set of possible values $\mathcal{Z} \subset \mathbb{R}$;*

- *An associated probability density* $p : \mathcal{Z} \to \mathbb{R}^+$ *such that* $\int_{\mathbb{R}} p(z)\, dz = 1$ *and*

$$\forall \mathcal{S} \subset \mathcal{Z}, \quad \mathbb{P}\left(z \in \mathcal{S}\right) \; = \; \int_{z \in \mathcal{S}} p(z)\, dz.$$

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

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

$$\mathbb{E}\left[z\right] = \mathbb{E}_z\left[z\right] = \begin{cases} \sum_{z_i \in \mathcal{Z}} z_i\, p(z = z_i) & \textit{(discrete case)} \\ \\ \int_{\mathcal{Z}} z\, p(z)\, dz & \textit{(continuous case).} \end{cases}$$

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

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

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

The expected

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

- *The variance of* z *is defined by*

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

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

Lemma A.3.1

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

A.3.2 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 A.3.5 (Joint distribution (discrete case)) *Let* z *and* w *be two discrete random variables taking values in* $\mathcal{Z} = \{z_i\}$ *and* $\mathcal{W} = \{w_j\}$*, respectively. The distribution of the pair of random variables* (z, w) *is defined by*

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

$$p_{i,j} = \mathbb{P}\left(z = z_i, w = w_j\right).$$

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

- *The set of possible values* $\mathcal{Z} \times \mathcal{W}$;
- *The continuous probability density* $p : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$ *such that*

$$\int_z \int_w p(z, w)\, dz\, dw \;=\; 1.$$

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 A.3.7 (Marginal laws (discrete case)) *Let* z *and* w *be two discrete random variables taking values in* $\mathcal{Z} = \{z_i\}$ *and* $\mathcal{W} = \{w_j\}$, *respectively. Let* $\{p_{i,j}\}$ *be the joint distribution of* (z, w).

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

$$p_{i\bullet} := \mathbb{P}\left(z = z_i\right) = \sum_{j | w_j \in \mathcal{W}} \mathbb{P}\left(z = z_i,\ w = w_j\right) = \sum_j p_{i,j}.$$

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

$$p_{\bullet j} := \mathbb{P}\left(w = w_j\right) = \sum_{i | z_i \in \mathcal{Z}} \mathbb{P}\left(z = z_i,\ w = w_j\right) = \sum_i p_{i,j}.$$

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

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

$$\forall z \in \mathcal{Z}, \quad p_z(z) = \int_{\mathcal{W}} p(z, w)\, dw.$$

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

$$\forall w \in \mathcal{W}, \quad p_w(w) = \int_{\mathcal{Z}} p(z, w)\, dz.$$

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

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

The correlation of z *and* w *is*

$$\mathrm{Corr}\left[z, w\right] = \frac{\mathrm{Cov}\left[z, w\right]}{\sqrt{\mathrm{Var}_z\left[z\right]}\sqrt{\mathrm{Var}_w\left[w\right]}}.$$

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

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

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

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

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

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

A.3.3 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 A.3.11 (Law of a random vector) *Let $\boldsymbol{z} = [z_i]_i$ be a random vector in $\mathbb{R}^n$: the law (or the distribution) of $\boldsymbol{z}$ is given by the joint distribution of its components. In particular, we define the following moments of this distribution:*

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

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

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

- *the* **covariance** *matrix of $\boldsymbol{z}$, denoted by $\mathrm{Var}[\boldsymbol{z}]$ or $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ is the matrix of the covariances between each component*

$$\forall 1 \leq i, j \leq n, \quad [\boldsymbol{\Sigma}_{\boldsymbol{z}}]_{i,j} := \mathbb{E}[(z_i - \mathbb{E}[z_i])(z_j - \mathbb{E}[z_j])].$$

Note that the covariance matrix can be written as

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

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