Lecture Notes

Optimization Theory and Methods

Course held by

Prof. Giancarlo Bigi

Notes written by

Francesco Paolo Maiale

Department of Mathematics

Pisa University

November 23, 2017

Disclaimer

These notes (still in progress) comes out of the *Optimization Theory and Methods* course, held by Professor Giancarlo Bigi in the first semester of the academic year 2017/2018.

N.B. Send an email to francescopaolo (dot) maiale (at) gmail (dot) com to report any kind of mistake.

Contents

Chapter 1

Introduction

Optimization problems arise from every discipline, and it is thus fundamental to develop a theory which aims to deal with it, and methods to find a suitable solution.

In this first chapter, we shall introduce the terminology of the optimization theory, and we shall also describe the type of problems we will be dealing with in this course.

1.1 Mathematical Formulation

Optimization is the minimization (or maximization) of a real-valued function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, eventually subject to constraints.

Mathematically speaking, given a (eventually proper) subset $X \subseteq \mathbb{R}^n$, we want to solve the minimization problems

$$\min_{x \in X} f(x) := \min\{f(x) \, : \, x \in X\},$$

and find the set of all $\bar{x} \in X$ realizing the minimum above.

Clearly, a generic function f does not admit a minimum on a generic subset X; hence we will be often interested in the solution of the infimum problem:

$$\inf_{x \in X} f(x) := \inf\{f(x) \, : \, x \in X\}. \tag{1.1}$$

It may happen that we need to find the maximum (or supremum) of a function f subject to constraints, but we can restrict our theory to minimization problems since

$$\min_{x \in X} f(x) = \max_{x \in X}(-f)(x) \quad \text{and} \quad \inf_{x \in X} f(x) = \sup_{x \in X}(-f)(x).$$

We use the following common notation for a problem of the form (1.1):

- The function f is the *objective function*, and it will often belong to a class of function satisfying particular properties that are useful in minimization problems (e.g., convex).
- The region X defining the constraints is called *feasible region*, while the complement $\mathbb{R}^n \setminus X$ is called *unfeasible region*.
- A solution $\bar{x} \in X$ to the minimization problem (1.1) is called *optimal value*.

Discrete Optimization. In some optimization problems it makes sense to consider only integer-valued variables (or even binary-valued variables $(x_1, \ldots, x_n) \in \{0, 1\}^n$). Therefore, the mathematical formulation (1.1) needs to be slightly modified:

$$\inf_{x \in X} f(x) := \inf\{f(x) \,:\, x \in X \cap \mathbb{Z}^n\} \quad \text{or} \quad \inf_{x \in X} f(x) := \inf\{f(x) \,:\, x \in X \cap \{0, 1\}^n\}.$$

Problems of this type are called *integer programming* problems, and they form a subset of *discrete optimization problems*.

The defining property of discrete problems is that the variables x_i takes value in a large, but finite, set. These are discussed in a different course, therefore we will only deal with *continuous*, possibly nonlinear, problems here.

Global/Local. Optimization algorithm are often able to locate *local* solutions (local minimum or maximum) to a minimization problem, but, in general, it is hard to write a converging algorithm, which is also optimal and robust, able to locate a global solution in the feasible region.

Chapter 2

Convex Functions Theory

The goal of this chapter is to introduce the most important class of functions we will be dealing with in minimization problems: *convex* functions.

More precisely, we shall study the fundamental properties of (strictly) convex functions, and prove that every local minimum for f is also a global minimum (solving, at least partially, the issues with the algorithms that tends to be trapped in local minima while looking for a global one.)

2.1 Minimization Problems: Definitions of Solution

In this brief section, we recall the main definitions of minimum point we will be dealing with in this course, and show some basic examples.

Let $X \subseteq \mathbb{R}^n$ be a subset, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function.

Definition 2.1 (Local Minimum)**.** A point $\bar{x} \in \mathbb{R}^n$ is a *local minimum* (LM) for the function f, subject to the constraint X, if the following properties are satisfied:

1. The point $\bar{x}$ belongs to X.
2. There exists $\epsilon > 0$ such that

$$f(\bar{x}) \leq f(x) \quad \text{for every } x \in B(\bar{x}, \epsilon).$$

Definition 2.2 (Strict Local Minimum)**.** A point $\bar{x} \in \mathbb{R}^n$ is a *strict local minimum* for the function f subject to the constraint X if the following properties are satisfied:

1. The point $\bar{x}$ belongs to X.
2. There exists $\epsilon > 0$ such that

$$f(\bar{x}) < f(x) \quad \text{for every } x \in B(\bar{x}, \epsilon) \setminus \{\bar{x}\}.$$

Definition 2.3 (Global Minimum)**.** A point $\bar{x} \in \mathbb{R}^n$ is a *global minimum* (GM) for the function f subject to the constraint X if the following properties are satisfied:

1. The point $\bar{x}$ belongs to X.
2. For every $x \in X$ it turns out that $f(\bar{x}) \leq f(x)$.

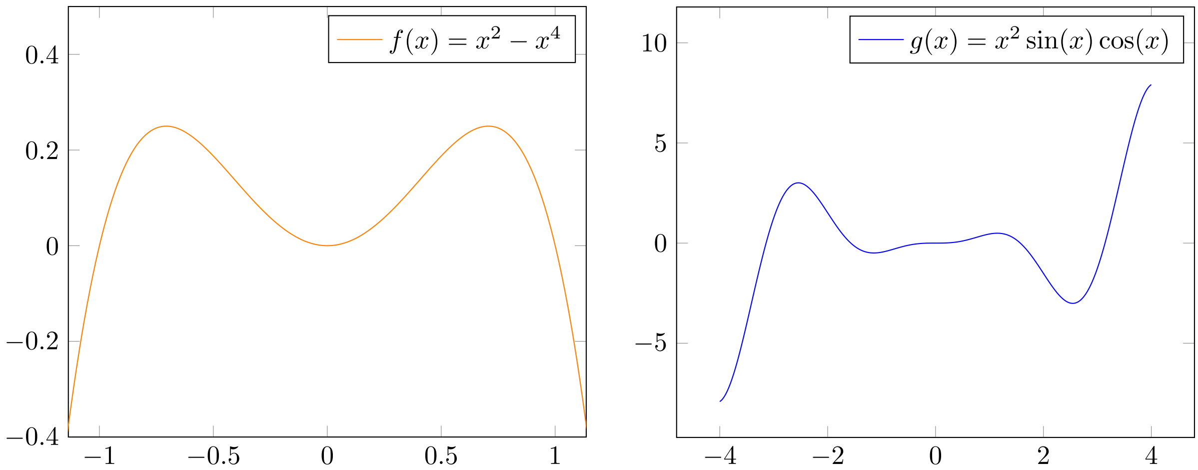

Figure 2.1: The function $f(x)$ admits a strict local minimum at $x = 0$. The function $g(x)$ admits an infinite number of strict local minima whose value is strictly decreasing.

Definition 2.4 (Strict Global Minimum)**.** A point $\bar{x} \in \mathbb{R}^n$ is a *strict global minimum* for the function f subject to the constraint X if the following properties are satisfied:

1. The point $\bar{x}$ belongs to X.
2. For every $x \in X \setminus \{\bar{x}\}$ it turns out that $f(\bar{x}) \leq f(x)$.

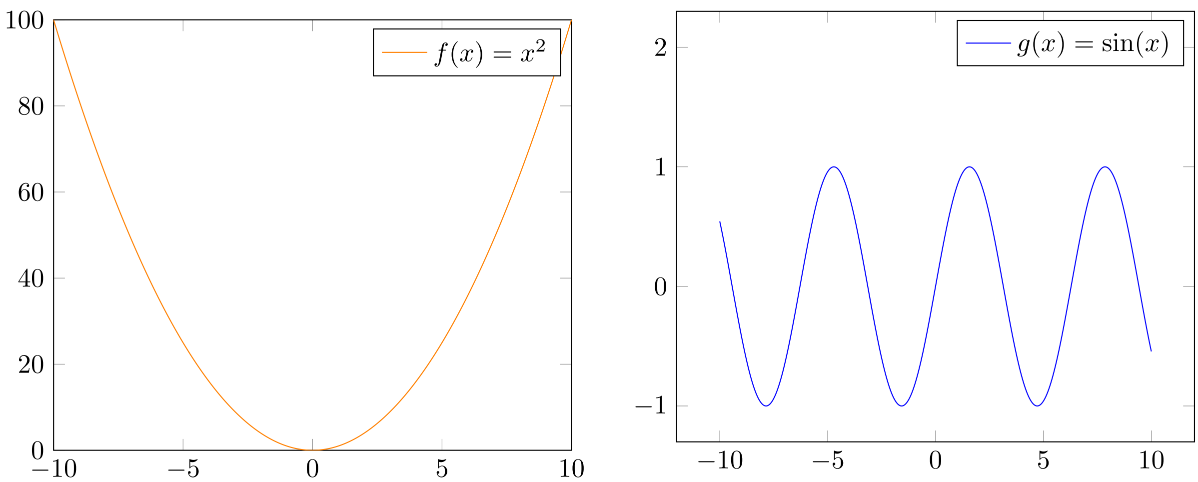

Figure 2.2: The function $f(x)$ admits a strict global minimum at $x = 0$. The function $g(x)$ admits an infinite number of global minima (-1) at $x_k := 3/2\pi + 2k\pi$.

2.2 Convex Functions

In this section, we study the fundamental properties of *convex* (resp. *concave*) functions, and we describe their role in minimization (maximization) problems.

2.2.1 Convex Sets

We briefly recall the definition of convex sets, together with the central properties of closure with respect to certain operations.

Definition 2.5 (Convex Set)**.** A set $C \subseteq \mathbb{R}^n$ is *convex* if

$$\lambda x + (1-\lambda)y \in C \quad \text{for every } x,\, y \in C \text{ and } \lambda \in [0,\, 1].$$

In particular, a set $C \subseteq \mathbb{R}^n$ is convex if and only if it contains every segment with extremal points in C. A simple example is shown in the figure below.

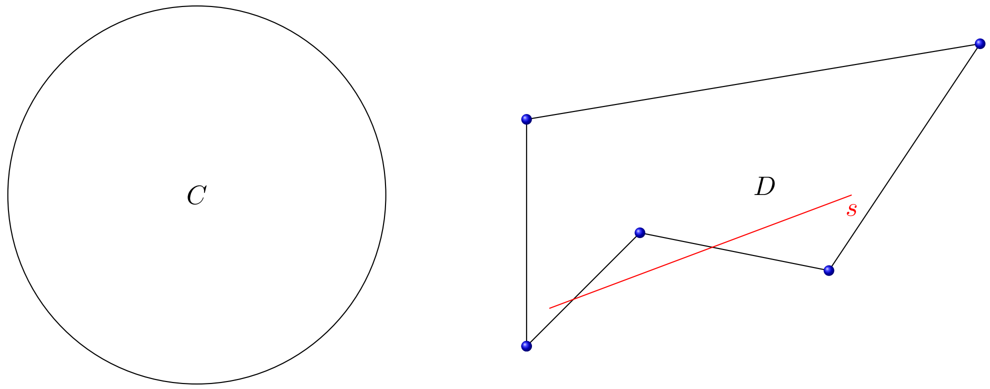

Figure 2.3: The set on the left is convex. The set on the right, on the other hand, is not convex because it does not contain entirely the segment s.

Lemma 2.6 (Closure Properties of Convex Sets)**.**

(1) An arbitrary (not necessarily countable!) intersection of convex sets is convex, that is,

$$\{C_i\}_{i\in\mathcal{I}} \text{ convex collection } \implies \bigcap_{i\in\mathcal{I}} C_i \text{ is convex.}$$

(2) A finite sum of convex sets is convex, that is,

$$\{C_i\}_{i=1,\,\dots,\,k} \text{ convex collection } \implies \sum_{i=1}^{k} C_i \text{ is convex.}$$

(3) The scaling of a convex set is convex, that is,

$$C \text{ convex set } \implies \mu \cdot C \text{ is convex for any } \mu \in \mathbb{R}.$$

Proof.

(1) Let $x,\, y \in C := \cap_{i \in \mathcal{I}} C_i$ be two points in the intersection. By definition of intersection x and y belong to C_i for every $i \in \mathcal{I}$, and by convexity we infer that

$$\lambda x + (1-\lambda)y \in C_i \quad \text{for every } i \in \mathcal{I} \implies \lambda x + (1-\lambda)y \in C.$$

(2) Let $x,\, y \in C := \sum_{i=1}^k C_i$ be two points in the finite sum. There are $x_i,\, y_i \in C_i$ such that

$$x = \sum_{i=1}^k x_i \quad \text{and} \quad y = \sum_{i=1}^k y_i,$$

and therefore, for any $\lambda \in [0,\, 1]$, it turns out that

$$\lambda x + (1-\lambda)y = \lambda \sum_{i=1}^k x_i + (1-\lambda) \sum_{i=1}^k y_i = \sum_{i=1}^k \left[\lambda x_i + (1-\lambda) y_i\right].$$

By convexity $\lambda x_i + (1-\lambda)y_i \in C_i$ for every $i = 1,\, \dots,\, k$, and thus $\lambda x + (1-\lambda)y$ necessarily belongs to the sum C.

(3) Suppose $\mu \neq 0$ and let $x,\, y \in \mu \cdot C$ be two point. For any $\lambda \in [0,\, 1]$ it turns out that

$$\lambda x + (1-\lambda)y = \mu \underbrace{\left[\lambda \frac{x}{\mu} + (1-\lambda)\frac{y}{\mu}\right]}_{\in C} \implies \lambda x + (1-\lambda)y \in \mu \cdot C,$$

which is exactly what we wanted to prove.

□

2.2.2 Convex and Concave Functions

We are finally ready to define a *convex* function, which is, intuitively, a function f whose graph between $f(x)$ and $f(y)$ is under the segment $\lambda x + (1-\lambda)y$ for $\lambda \in [0,\, 1]$.

Definition 2.7. Let $C \subseteq \mathbb{R}^n$ be a convex set. A function $f : C \longrightarrow \mathbb{R}$ is *convex* if and only if

$$f\left(\lambda x + (1-\lambda)y\right) \leq \lambda f(x) + (1-\lambda) f(y) \quad \text{for every } x,\, y \in C \text{ and } \lambda \in [0,\, 1]. \tag{2.1}$$

Furthermore, the function f is *strictly convex* if and only if

$$f\left(\lambda x + (1-\lambda)y\right) < \lambda f(x) + (1-\lambda) f(y) \quad \text{for every } x,\, y \in C \text{ and } \lambda \in (0,\, 1). \tag{2.2}$$

There is an alternative definition of *convex* function that is a little more abstract. The *epigraph* of a function f is "what lies above the graph of f", that is,

$$\mathrm{epi}(f) := \{(x,\, y) \in \mathbb{R}^n \times \mathbb{R} \,:\, y \geq f(x)\} \cap C,$$

where C is the domain of f. The following criterion states that it is enough to check the convexity of the epigraph to infer that f is convex.

Lemma 2.8. *Let C be a convex set. A function $f : C \longrightarrow \mathbb{R}$ is convex if and only if* epi(f) *is a convex subset of $\mathbb{R}^n$.*

Proof.

" $\Longrightarrow$ " Let $(x,\, y)$ and $(x',\, y')$ be two points of the epigraph, and let $\lambda \in (0,\, 1)$ be a parameter: we need to prove that

$$\lambda y + (1-\lambda) y' \overset{?}{\geq} f(\lambda x + (1-\lambda) x'),$$

that is, $\lambda(x,\, y) + (1-\lambda)(x',\, y')$ belongs to $\operatorname{epi}(f)$. By assumption f is a convex function, and therefore

$$\begin{aligned} f(\lambda x + (1-\lambda) x') &\leq \lambda f(x) + (1-\lambda) f(x') \leq \\ &\leq \lambda y + (1-\lambda) y', \end{aligned}$$

where the last inequality follows from the fact that both $(x,\, y)$ and $(x',\, y')$ belong to $\operatorname{epi}(f)$.

" $\Longrightarrow$ " Suppose that $\operatorname{epi}(f)$ is convex, and let $x,\, y \in C$ and $\lambda \in (0,\, 1)$. We need to prove that f is convex, or, equivalently, that (2.1) holds. By convexity of the epigraph, it turns out that

$$\lambda y + (1-\lambda) y' \overset{?}{\geq} f(\lambda x + (1-\lambda) x') \quad \text{for every } y \geq f(x),$$

and thus also for $y = f(x)$. $\square$

Definition 2.9. Let $C \subseteq \mathbb{R}^n$ be a convex set. A function $f : C \longrightarrow \mathbb{R}$ is *concave* if and only if

$$f\left(\lambda x + (1-\lambda) y\right) \geq \lambda f(x) + (1-\lambda) f(y) \quad \text{for every } x,\, y \in C \text{ and } \lambda \in [0,\, 1].$$

Furthermore, the function f is *strictly concave* if and only if

$$f\left(\lambda x + (1-\lambda) y\right) > \lambda f(x) + (1-\lambda) f(y) \quad \text{for every } x,\, y \in C \text{ and } \lambda \in (0,\, 1).$$

Remark 2.1. It is enough to develop the theory of convex function since one can easily prove that if $f : C \longrightarrow \mathbb{R}$ is a concave function, then $-f : C \longrightarrow \mathbb{R}$ is a convex function, and vice versa.

Remark 2.2. Let $f : C \longrightarrow \mathbb{R}$ be a convex function. By induction we can easily generalize the definition (2.1) to finite convex combinations, that is,

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i)$$

for every k-tuple of points $x_1,\, \dots,\, x_k \in C$ and coefficients $\lambda_i \geq 0$ satisfying the constraints

$$\sum_{i=1}^{k} \lambda_i = 1.$$

Definition 2.10 (Convex Hull)**.** The convex hull of a finite sequence of points $x_1,\, \dots,\, x_k \in \mathbb{R}^n$ is the set of all the possible convex combinations, that is,

$$\operatorname{Conv}(x_1,\, \dots,\, x_k) := \left\{ \sum_{i=1}^{k} \lambda_i x_i \;:\; \lambda_i \geq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

Similarly, if $S \subset \mathbb{R}^n$ is a subset of $\mathbb{R}^n$, the convex hull is defined by the set of all finite convex combinations, that is,

$$\operatorname{Conv}(S) := \left\{ \sum_{i=1}^{k} \lambda_i x_i \;:\; x_1,\, \dots,\, x_k \in S,\, \lambda_i \geq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

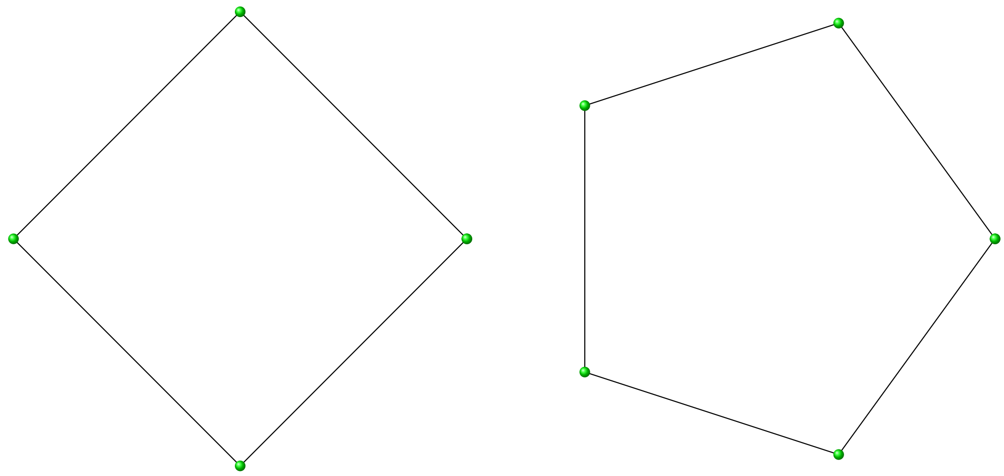

Figure 2.4: **Left:** Convex hull of 4 points. **Right:** Convex hull of 5 points.

In particular, one can easily prove that

$$\mathrm{Conv}\,(S) = \bigcup_{F \subset S} \mathrm{Conv}\,(x_1, \ldots, x_k)\,,$$

where F ranges among all the finite subsets $\{x_1, \ldots, x_k\}$ of S.

Lemma 2.11 (Carathéodory)**.** *Let $A \subseteq \mathbb{R}^n$ be a subset. For every $x \in \mathrm{Conv}(A)$ there are nonnegative coefficients $\lambda_1, \ldots, \lambda_{n+1} \in \mathbb{R}_{\geq 0}$ such that*

$$\sum_{i=1}^{n+1} \lambda_i = 1 \quad \textit{and} \quad x = \sum_{i=1}^{n+1} \lambda_i x_i,$$

where $x_i \in A$ for every i.

Proof. Let $x \in \mathrm{Conv}(A)$. Then x is a convex combination of a finite number of points in A, that is, there are $\lambda_1, \ldots, \lambda_k \in \mathbb{R}_{>0}$ and $x_1, \ldots, x_k \in A$ such that

$$\sum_{i=1}^{k} \lambda_i = 1 \quad \text{and} \quad x = \sum_{i=1}^{k} \lambda_i x_i,$$

Suppose $k > n+1$ (otherwise, there is nothing to prove). Then $x_2 - x_1, \ldots, x_k - x_1$ are linearly dependent vectors in $\mathbb{R}^n$, which means that

$$\sum_{j=2}^{k} \mu_j (x_j - x_1) = 0 \quad \text{for some } (\mu_2, \ldots, \mu_k) \in \mathbb{R}^{k-1} \setminus \{(0, \ldots, 0)\}.$$

In particular, if $\mu_1 := -\sum_{j=2}^{k} \mu_j$, then the formula above proves that

$$\sum_{j=1}^{k} \mu_j x_j = 0 \quad \text{and} \quad \sum_{j=1}^{k} \mu_j = 0,$$

and thus there is at least $\mu_{\bar{j}} > 0$. Then

$$x = \sum_{j=1}^{k} (\lambda_j - \alpha\mu_j)\, x_j$$

for every $\alpha \in \mathbb{R}$. In particular, we can choose

$$\alpha := \min_{1\le j\le k} \left\{ \frac{\lambda_j}{\mu_j} \;:\; \mu_j > 0 \right\} =: \frac{\lambda_i}{\mu_i}.$$

Note that $\alpha > 0$, and for every $j \in \{1, \ldots, k\}$ we have $\lambda_j - \alpha\mu_j \geq 0$. In particular, we have $\lambda_i - \alpha\mu_i = 0$, which means that

$$\sum_{j=1}^{k} (\lambda_j - \alpha\mu_j) = 1 \quad \text{and} \quad x = \sum_{j=1}^{k} (\lambda_j - \alpha\mu_j)\, x_j,$$

and the sum is over $k-1$ nonzero coefficients. We iterate the process until x is represented as a convex combination of $n+1$ points (and we cannot go any furthers since $x_i - x_1$ for $i = 2, \ldots, n+1$ can be linearly independent). ☐

Definition 2.12 (Sublevels)**.** Let $f : X \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$ be a function. The M-sublevel of f, for $M \in \mathbb{R}$, is defined by

$$f^M := \{x \in X \;:\; f(x) \leq M\}\,.$$

Remark 2.3. If $f : C \longrightarrow \mathbb{R}$ is convex, then the sublevel f^M is convex, as a subset of $\mathbb{R}^n$, for every $M \in \mathbb{R}$. Indeed, for any $x, y \in f^M$ and $\lambda \in [0, 1]$ it turns out that

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \leq \lambda M + (1-\lambda)M = M.$$

The converse, on the other hand, is generally **false**. For example, the sublevels of the function

$$\mathbb{R} \ni x \longmapsto x^3 \in \mathbb{R}$$

are half-lines, thus convex, but the function is not convex by any means. A non-convex function with all sublevels convex is called *quasi-convex*.

Lemma 2.13 (Closure Properties of Convex Functions)**.**

(a) *The supremum of a collection of convex functions is convex, that is,*

$$\{C_i\}_{i\in\mathcal{I}} \text{ convex functions} \implies f := \sup_{i\in\mathcal{I}} f_i \text{ is convex.}$$

(b) *A finite sum of convex functions is convex, that is,*

$$\{f_i\}_{i=1,\ldots,k} \text{ convex functions} \implies f := \sum_{i=1}^{k} f_i \text{ is convex.}$$

(c) *The scaling of a convex function is convex, that is,*

$$f \text{ convex function} \implies \lambda \cdot f \text{ is convex for any } \lambda \in \mathbb{R}.$$

(d) *Let $h : \mathbb{R} \longrightarrow \mathbb{R}$ and $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be two real-valued functions. The composition $h \circ f$ is convex if one of the following hold:*

(1) The function f is convex and the function h is convex and nondecreasing.

(2) The function f is concave and the function h is convex and nonincreasing.

Proof. The proof of **(a)**, **(b)** and **(c)** follows immediately from Lemma 2.6 and the characterization of convexity in terms of the epigraph. In fact, one can easily check that

$$f = \sup_i f_i \implies \operatorname{epi}(f) = \bigcap_{i \in \mathcal{I}} \operatorname{epi}(f_i),$$

$$f = \sum_{i=1}^{k} f_i \implies \operatorname{epi}(f) = \sum_{i \in \mathcal{I}} \operatorname{epi}(f_i),$$

$$g = \mu \cdot f \implies \operatorname{epi}(g) = \lambda \cdot \operatorname{epi}(f).$$

We now prove the assertion (1) of the property **(d)**; the second one is similar and left to the reader as a simple exercise.

Notice that the assumption h nondecreasing is sufficient, and it comes out naturally if both f and h are functions of class C^2 by computing the second derivative of $h \circ f$.

Proof of (d). Let $x,\, y \in \mathbb{R}^n$ be any two point, and let $\lambda \in (0,\,1)$ be a parameter. We have

$$\begin{aligned}(h \circ f)(\lambda x + (1-\lambda)y) = h\left(f(\lambda x + (1-\lambda)y)\right) &\leq \\ \leq h\left(\lambda f(x) + (1-\lambda)f(y)\right) &\leq \\ \leq \lambda (h \circ f)(x) + (1-\lambda)(h \circ f)(y)&,\end{aligned}$$

which is exactly what we wanted to prove. The red inequality holds if h is a nondecreasing function since, in general, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \;\not\Longrightarrow\; h(f(\lambda x + (1-\lambda)y)) \leq h(\lambda f(x) + (1-\lambda)f(y)).$$

□

Corollary 2.14. *Let $A \in M(n, \mathbb{R})$ be a $n \times n$ matrix with real coefficients, and let $b \in \mathbb{R}^n$ be a vector. If $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a convex function, then the composition*

$$x \longmapsto f(Ax + b)$$

is also a convex function.

Proof. The composition is clearly convex because $x \longmapsto Ax + b$ is a linear function, which means that the inequality above is, actually, an equality:

$$\lambda x + (1-\lambda)y \longmapsto \lambda\, Ax + (1-\lambda)Ay + b = \lambda(Ax + b) + (1-\lambda)(Ay + b).$$

□

2.2.3 Examples

Exercise 2.1. Prove that the norm $\|\cdot\| : \mathbb{R}^n \longrightarrow \mathbb{R}_+$ and the squared norm $\|\cdot\|^2 : \mathbb{R}^n \longrightarrow \mathbb{R}_+$ are both convex functions.

Solution. Recall that the norm $\|\cdot\|$ is homogeneous and satisfies the triangular inequality, that is,

$$\|\lambda x\| = |\lambda| \|x\| \quad \text{and} \quad \|x+y\| \leq \|x\| + \|y\|$$

for any $x, y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$. Therefore, it easily follows that

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = \lambda\|x\| + (1-\lambda)\|y\|,$$

which means that $\|\cdot\|$ is convex.

The function $x^2 : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is convex and nondecreasing (the domain is the positive half-line), and hence the composition $\|\cdot\|^2 = x^2 \circ \|\cdot\|$ is convex by **(d)** of Lemma 2.13. $\square$

Exercise 2.2. Let Q be a $n \times n$ matrix, and let $b \in \mathbb{R}^n$ be a (translation) vector. Prove that the function

$$\varphi_Q : \mathbb{R}^n \ni x \longrightarrow \frac{1}{2} x^T Q x + b^T x \in \mathbb{R}$$

is convex if and only if Q is semidefinite positive.

Solution. The Hessian criterion (see Lemma 2.36) allows us to infer that

$$\varphi_Q \text{ is convex} \iff \nabla^2(\varphi_Q(x)) \geq 0,$$

and therefore it suffices to notice that the Hessian of φ_Q is constantly equal to Q. It follows that

$$\varphi_Q \text{ is convex} \iff Q \text{ is semidefinite positive},$$

which is exactly what we wanted to prove. $\square$

Definition 2.15 (Support Function)**.** Let $C \subseteq \mathbb{R}^n$ be a set. The *support function* of C is defined by setting

$$\sigma_C(x) := \sup \{\langle x, y\rangle \ : \ y \in C\},$$

where $\langle \cdot, -\rangle$ denotes the 2-norm scalar product.

Lemma 2.16. *Let $C \subseteq \mathbb{R}^n$ be a set. The support function σ_C is convex*[1]*.*

Proof. The function $x \longmapsto \langle x, y\rangle$ is linear for every fixed y, which means that σ_C is given by the supremum of an arbitrary family of convex function. By Lemma 2.13 we infer that σ_C is convex. $\square$

2.2.4 Convex Functions in Minimization Problems

We are finally ready to state and prove a result which explain, at least partially, why the class of convex (concave) functions is so important in minimization (maximization) problems.

More precisely, we shall prove that local minima and global minima coincide for convex functions, and we will also show that a particular subclass of the convex functions admits one and only one minimum point.

[1] Here we need to be careful. We defined convex functions $f : C \longrightarrow \mathbb{R}$, but, a priori, the function σ_C may attain $+\infty$ in some points. We will not do it, but with reasonable changes it is possible to develop a theory of convex functions with values in $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$.

Theorem 2.17. *Let X be a convex set in $\mathbb{R}^n$, and let $f : X \longrightarrow \mathbb{R}$ be a convex function. Then the following assertions hold:*

(a) *If $\bar{x}$ is a local minimum for f on X, then $\bar{x}$ is also a global minimum for f in X.*

(b) *The set of minima is convex, that is,*

$$\operatorname{argmin}_X(f) := \{x \in X \ : \ x \text{ is a minimum point for } f \text{ in } X\} \text{ is convex.}$$

(c) *Furthermore, if f is strictly convex, then f admits at most one minimum point.*

Proof.

(a) We argue by contradiction. Suppose that f admits a local minimum $\bar{x} \in X$ that is not a global minimum (i.e., there exists $x \in X$ such that $f(x) < f(\bar{x})$). Then, for every $\lambda \in [0, 1]$ it follows from (2.1) that

$$f(\lambda x + (1-\lambda)\bar{x}) \leq \lambda f(x) + (1-\lambda) f(\bar{x}) < \lambda f(\bar{x}) + (1-\lambda) f(\bar{x}) = f(\bar{x}),$$

which means that $\bar{x}$ cannot be a local minimum since the segment $\lambda x + (1-\lambda)\bar{x}$ is arbitrarily close to $\bar{x}$ (i.e., it intersects every ball $B(\bar{x}, \epsilon)$ for any positive ϵ).

(b) Let x and y be two local/global minima for f on X. Since $f(x) = f(y)$, it follows from (2.1) that

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) = f(x),$$

which means that the whole segment of extremal points x and y is also contained in $\operatorname{argmin}_X(f)$.

(c) We argue by contradiction. Suppose that f admits two global/local minima $x, y \in X$, and notice that $f(x) = f(y)$ by **(a)**. Then, for every $\lambda \in [0, 1]$ it follows from (2.2) that

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y) = f(x) = f(y),$$

which contradicts the assumption that x is a global minimum.

□

2.3 Strongly Convex Functions

Suppose that we are interested in the minimization problem

$$\min\{f(x) \ : \ x \in X\},$$

where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a function and $X \subseteq \mathbb{R}^n$ a subset. By Weierstrass theorem, the minimization problem admits a solution if we assume, for example, that f is continuous on X, and X is compact. Unfortunately, we will often deal with problems lacking the compactness we need; therefore, we need to introduce a class of function that can, in a certain way, replace it.

Definition 2.18 (Coercive)**.** A function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is coercive in $\mathbb{R}^n$ if and only if

$$\lim_{\|x\| \to +\infty} f(x) = +\infty.$$

If f is a coercive function, then the sublevel $f^M := \{x \in \mathbb{R}^n \,:\, f(x) \le M\}$ is bounded (and eventually empty) for every $M \in \mathbb{R}$. Indeed, we notice that

$$f \text{ is coercive } \implies \forall M \in \mathbb{R}\, \exists R > 0 \,:\, f(x) \ge M \quad \text{for any } x \in \{x \in \mathbb{R}^n \,:\, \|x\| \ge R\},$$

and, in particular, we have $f^M \subset B(0,\, R)$ for some $R > 0$ depending solely on M. It follows that the initial minimization problem is equivalent to

$$\min\{f(x) \,:\, x \in X \cap f^M\}$$

for any choice of $M \in \mathbb{R}$ such that $f^M \neq \varnothing$. Moreover, if the function f is continuous (or lower semi-continuous=closed function), then one can prove that f^M is closed (and thus compact). In conclusion, if X is closed and f is continuous (or lower semi-continuous), then we simply need to solve the minimization problem

$$\min\{f(x) \,:\, x \in X \cap f^M\},$$

which admits a solution by Weierstrass theorem, as the intersection $X \cap f^M$ is compact.

The primary goal of this section is to introduce a subclass of strictly convex functions that are also coercive. Furthermore, we shall prove that every convex function is continuous on the interior of its domain, which will allow us to infer the existence of one and only one solution to the minimization problem (1.1).

Definition 2.19. Let $C \subseteq \mathbb{R}^n$ be a convex set. A function $f : C \longrightarrow \mathbb{R}$ is *strongly convex* if there exists a positive constant μ such that

$$f\left(\lambda x + (1-\lambda)y\right) \le \lambda f(x) + (1-\lambda) f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|y - x\|^2 \tag{2.3}$$

for every $x,\, y \in C$ and $\lambda \in [0,\, 1]$. The constant μ is known as *modulus of strong convexity.*

Lemma 2.20. *A function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is μ-strongly convex if and only if the function*

$$g(x) := f(x) - \frac{\mu}{2}\|x\|^2$$

is convex. In particular, if f is strongly convex, then f is also strictly convex.

Proof.

$\underline{\implies}$: Let $x,\, y \in \mathbb{R}^n$ and $\lambda \in (0,\, 1)$. The definition (2.3) yields to the following inequality

$$g(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) - \frac{\mu}{2}\left[\lambda(1-\lambda)\|y - x\|^2 + \|\lambda x + (1-\lambda)y\|^2\right],$$

which means that it is enough to prove that

$$\lambda(1-\lambda)\|y - x\|^2 + \|\lambda x + (1-\lambda)y\|^2 = \lambda\|x\|^2 + (1-\lambda)\|y\|^2.$$

The thesis follows immediately from a straightforward computation:

$$\begin{aligned}
&\lambda(1-\lambda)\|y - x\|^2 + \|\lambda x + (1-\lambda)y\|^2 = \\
&= \lambda(1-\lambda)\left[\|x\|^2 + \|y\|^2 - 2\langle x,\, y\rangle\right] + \lambda^2\|x\|^2 + (1-\lambda)^2\|y\|^2 + 2\lambda(1-\lambda)\langle x,\, y\rangle = \\
&= \lambda\|x\|^2 + (1-\lambda)\|y\|^2.
\end{aligned}$$

$\underline{\Longleftarrow}$: It suffices to go backward on the proof of the opposite implication. The details are left to the reader as an easy exercise. $\square$

Lemma 2.21. *Let $f : C \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. Then there exist $x^* \in \mathbb{R}^n \setminus \{0\}$ and $\beta \in \mathbb{R}$ such that*

$$f(x) \geq \langle x,\, x^* \rangle + \beta \quad \textit{for every } x \in C. \tag{2.4}$$

The proof of this result follows from a finite-dimensional statement of the geometrical Hahn-Banach theorem. Namely, if $A \subset \mathbb{R}^n$ is a nonempty closed convex set and $B \subset \mathbb{R}^n$ is a nonempty compact convex set, and if $A \cap B = \varnothing$, then there exist $\gamma_1 < \gamma_2 \in \mathbb{R}$ and $v \in \mathbb{R}^n \setminus \{0\}$ such that

$$\langle v,\, a \rangle \leq \gamma_1 \quad \text{for every } a \in A,$$

$$\langle v,\, b \rangle \geq \gamma_2 \quad \text{for every } b \in B.$$

We shall prove this result later, with more sophisticated tools. Now we explain why the statement of Lemma 2.21 follows immediately from the Hahn-Banach theorem.

Proof. Assume (wlog) that $C = \mathbb{R}^n$, and let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be the convex function. We utilize the separation theorem (see Theorem 2.42) with[2]

$$A := \overline{\operatorname{epi}(f)} \quad \text{and} \quad B := \{(x_a,\, y_a) \in \mathbb{R}^n \times \mathbb{R}\},$$

where $(x_a,\, y_a)$ is a point that does not belong to the closure of the epigraph of f. Then there exist $(x^*,\, y^*) \in \mathbb{R}^{n+1} \setminus \{0\}$ and $\gamma \in \mathbb{R}$ such that

$$\langle x^*,\, x_a \rangle + y^* y_a = \langle x^* \oplus y^*,\, x_a \oplus y_a \rangle < \gamma < \langle x^* \oplus y^*,\, x \oplus y \rangle = \langle x^*,\, x \rangle + y^* y,$$

and now the conclusion follows easily. $\square$

Proposition 2.22. *A μ-strongly convex function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is coercive.*

Proof. By Lemma 2.20, if f is μ-strongly convex, then

$$g(x) := f(x) - \frac{\mu}{2}\|x\|^2$$

is convex, and by (2.4) it follows that there are $\beta \in \mathbb{R}$ and $x^* \in \mathbb{R}^n \setminus \{0\}$ such that

$$f(x) \geq \langle x,\, x^* \rangle + \beta + \frac{\mu}{2}\|x\|^2 \quad \text{for every } x \in \mathbb{R}^n.$$

By Cauchy-Schwartz inequality, we have

$$f(x) \geq \beta - \|x\|\|x^*\| + \frac{\mu}{2}\|x\|^2,$$

which means that

$$\lim_{\|x\| \to +\infty} f(x) \sim \lim_{\|x\| \to +\infty} \frac{\mu}{2}\|x\|^2 = +\infty.$$

$\square$

Remark 2.4. We notice that, in particular, a μ-strongly convex function is supercoercive (=superlinear), that is,

$$\lim_{\|x\| \to +\infty} \frac{f(x)}{\|x\|} = +\infty.$$

[2]We need to be careful here. The epigraph of a convex function is a convex set by definition, but we need to prove also that the closure of a convex function is convex. This property is left as an exercise for the reader.

Corollary 2.23. *Let $f : X \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ be a μ-strongly convex function. Then f admits one and only one minimum point in X.*

Proof. The function f is coercive by Proposition 2.22, and f^M is closed by continuity (see Theorem 2.24). Therefore, Weierstrass theorem assures the existence of at least one minimum point for f, which is unique by Theorem 2.17 **(c)**, since strongly convex implies strictly convex. □

2.4 Differentiable Convex Functions

In this section, we prove that a convex function is continuous in the interior of its domain and, also, almost everywhere differentiable (with respect to the Lebesgue measure).

2.4.1 Continuity of Convex Functions

Let $\Omega \subseteq \mathbb{R}^n$ be an open subset. Recall that a function $f : \Omega \longrightarrow \mathbb{R}$ is *Lipschitz-continuous* if there exists a positive constant $L > 0$ such that

$$|f(x) - f(y)| \leq L \cdot \|x - y\| \quad \text{for every } x,\, y \in \Omega. \tag{2.5}$$

Similarly, a function $f : \Omega \longrightarrow \mathbb{R}$ is *locally Lipschitz* if for every compact set $K \subset \Omega$ there exists a positive constant $L := L(K) > 0$, depending on the set K, such that

$$|f(x) - f(y)| \leq L \cdot \|x - y\| \quad \text{for every } x,\, y \in K. \tag{2.6}$$

Remark 2.5. A locally Lipschitz function $f : \Omega \longrightarrow \mathbb{R}$ is continuous on Ω.

We are finally ready to state the main result of this section, that is a convex function is continuous on the interior part of its domain.

Theorem 2.24. *Let $\Omega \subseteq \mathbb{R}^n$ be a convex (not necessarily open) set. A convex function $f : \Omega \longrightarrow \mathbb{R}$ is locally Lipschitz on the interior* $\mathrm{Int}(\Omega)$.

Proof.

Step 1. We claim that f is locally bounded at every interior point $x \in \mathrm{Int}\,\Omega$. Indeed, let us consider the vertexes of a regular convex polygon with center x, i.e.

$$x_i := x + \delta e_i \quad \text{and} \quad x_{i+n} := x - \delta e_i$$

for $i = 1, \ldots, n$, where $\{e_1, \ldots, e_n\}$ is the usual orthonormal basis of $\mathbb{R}^n$. Define

$$S := \mathrm{Conv}\,(x_1, \ldots, x_n, x_{n+1}, x_{2n}),$$

and notice that, for every $y \in S$, by convexity we have

$$f(y) \leq \sum_{i=1}^{2n} \lambda_i f(x_i) \leq \max_{i=1,\ldots,2n} f(x_i) \underbrace{\sum_{i=1}^{2n} \lambda_i}_{=1} =: M,$$

which gives us a local upper bound. Similarly, given $y \in S$, we consider the symmetric point $\hat{y} \in S$ with respect to x in such a way that

$$x = \frac{y + \hat{y}}{2} \implies f(x) \leq \frac{1}{2} f(y) + \frac{1}{2} f(\hat{y}).$$

It follows that

$$f(y) \geq 2f(x) - \frac{1}{2} f(\hat{y}) \geq 2f(x) - M,$$

which gives us a local lower bound.

Step 2. The function f is locally bounded around $\bar{x} \in \operatorname{Int}\Omega$, hence there exist $\epsilon > 0$ and $M \in \mathbb{R}_{>0}$ such that

$$|f(y)| \le M \qquad \text{for every } y \in B(\bar{x},\, 2\epsilon)\ .$$

Let $x,\, y \in B(\bar{x},\, \epsilon)$, and set $\alpha := \|x - y\| \neq 0$. Clearly

$$z := x + \frac{\epsilon}{\alpha}(x - y) \in B(\bar{x},\, 2\epsilon) \implies x = \frac{\alpha}{\alpha + \epsilon} z + \frac{\epsilon}{\alpha + \epsilon} y,$$

and thus by convexity of f we have

$$f(x) \le \frac{\alpha}{\alpha + \epsilon} f(z) + \frac{\epsilon}{\alpha + \epsilon} f(y).$$

It follows that

$$\begin{aligned} f(x) - f(y) &\le \frac{\alpha}{\alpha + \epsilon}\,[f(z) - f(y)] \le \\ &\le \frac{\alpha}{\alpha + \epsilon}\,|f(z) - f(y)| \le \\ &\le \frac{\alpha}{\epsilon}\,|f(z) - f(y)| \le \\ &\le \frac{2M}{\epsilon}\alpha = \frac{2M}{\epsilon}\|x - y\|, \end{aligned}$$

which means that $f(x) - f(y) \le c \cdot \|x - y\|$. The argument is symmetric with respect to x and y, and therefore we also have that $f(y) - f(x) \le c \cdot \|x - y\|$. □

Proof 2. We divide the proof into two steps: the 1-dimensional case, and next we reduce the general case to it with a simple trick.

Step 1. Suppose that $f : \mathbb{R} \longrightarrow \mathbb{R}$. We prove here that for every compact set $K := [a,\, b] \subset \mathbb{R}^n$, the restriction $f\big|_K$ is Lipschitz with constant

$$L(K) := 2 \max\{f(x)\ :\ x \in [a - 1,\, b + 1]\}.$$

The graph of the function $f\big|_{\mathbb{R}\setminus K}$ is always above the secant line $r(\cdot)$ passing through x_1 and x_2, for any $x_1,\, x_2 \in [a,\, b]$. Therefore, we have the inequalities

$$r(a - 1) \le f(a - 1) \quad \text{and} \quad r(b + 1) \le f(b + 1).$$

The slope of the line r is constant (by definition), and hence

$$r' = \frac{r(b + 1) - r(x)}{b + 1 - x} \le \frac{f(b + 1) - f(x)}{1} \le L(K)$$

for any $x \in K$. In a similar fashion, it turns out that

$$r' = \frac{r(x) - r(a - 1)}{x - (a - 1)} \ge \frac{f(x) - f(a - 1)}{1} \ge -L(K),$$

which means that $r' \in [-L(K),\, L(K)]$. In conclusion, we notice that

$$|f(x) - f(y)| \le L(K)\|x - y\| \quad \text{for every } x,\, y \in K.$$

Step 2. The ball of radius R, denoted by B_R, is given by the union of the segments of length R starting at the origin. More precisely, if we set

$$L_\theta(R) := \{t\theta \,:\, t \in [0,\, R]\} \quad \text{for } \theta \in S^{n-1},$$

then we easily obtain the following identity:

$$B_R = \bigcup_{\theta \in S^{n-1}} L_\theta(R).$$

If $M := \sup_{B_R} |f(x)|$, then f is a M-Lipschitz function on each segment L_θ, and thus on their union B_R. The proof of the theorem follows immediately since we can always cover a compact set $K \subset \Omega$ with a finite union of balls $B(x_i,\, R_i)$ for $i = 1,\, \dots,\, N$. □

In particular, a convex function $f : \Omega \longrightarrow \mathbb{R}$ defined on a convex set Ω is continuous on the interior, but, a priori, it could be discontinuous on the boundary $\partial\Omega$.

Example 2.1. In the 1-dimensional setting, we can easily find a convex function discontinuous at the boundary. For example, let $I := (-\infty,\, 5] \subset \mathbb{R}$ and consider the function

$$f(x) := \begin{cases} x^2 & \text{if } x < 5, \\[2ex] 30 & \text{if } x = 5. \end{cases}$$

The function f is convex because the epigraph is convex, but it is upper semi-continuous only at $x = 5$ since we have that

$$\limsup_{x \to 5^-} f(x) = 25 < 30 = f(5).$$

Proposition 2.25. *A convex function $f : I \subset \mathbb{R} \longrightarrow \mathbb{R}$ is continuous on the interior of I, and upper semi-continuous on the boundary ∂I.*

Example 2.2. A convex function is always upper semi-continuous in dimension one only. Set

$$\Omega := \big\{(x,\, y) \in \mathbb{R}^2 \,:\, y < 0\big\} \cup \{(0,\, 0)\} \quad \text{and} \quad f(x,\, y) := \begin{cases} -\frac{x^2}{2y} & \text{if } (x,\, y) \neq (0,\, 0), \\[2ex] 0 & \text{if } (x,\, y) = (0,\, 0). \end{cases}$$

One can easily check that f is a convex function. The parabola $(t,\, -1/2t^2)$ intersects the boundary of Ω at $(0,\, 0)$, and we have that

$$f(t,\, -\frac{1}{2}t^2) = 1 \quad \text{for every } t > 0,$$

and therefore

$$\liminf_{t \to 0^+} f(t,\, -\frac{1}{2}t^2) = 1 > 0 = f(0,\, 0),$$

which means that f is not upper semi-continuous at $(0,\, 0)$.

2.4.2 Differentiability of Convex Functions

In this section, we investigate the Lusin property of Lipschitz maps (from $\mathbb{R}^n$ to $\mathbb{R}^m$) with C^1 maps and, in particular, we prove that Lipschitz maps are $\mathcal{L}^n$-almost everywhere differentiable.

Theorem 2.26 (Lusin Property)**.** *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a Lipschitz map. Then, for every $\epsilon > 0$ there exists a function $g_\epsilon \in C^1(\mathbb{R}^n;\, \mathbb{R})$ satisfying the following properties:*

(1) The two maps coincide up to a set of measure at most ϵ, that is,

$$\mathcal{L}^n\left(\{x \in \mathbb{R}^n \mid f(x) \neq g_\epsilon(x)\}\right) \leq \epsilon.$$

(2) The Lipschitz constant is bounded from above, that is,

$$\mathrm{Lip}(g_\epsilon) \leq \mathrm{Lip}(f).$$

We will not prove this theorem entirely as we will only sketch the proof of a "local" statement. On the other hand, at the end of the section we briefly show how to obtain the global statement via a *partition of unity* (but this is a highly nontrivial result!)

Regularity. First, we prove that Lipschitz maps from $\mathbb{R}^n$ to $\mathbb{R}$ (and thus to $\mathbb{R}^m$) are almost everywhere differentiable with respect to the Lebesgue measure.

Theorem 2.27 (Rademacher)**.** *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a Lipschitz map. Then f is $\mathcal{L}^n$-almost everywhere differentiable.*

Remark 2.6. The extension property of Lipschitz functions in $\mathbb{R}^n$ automatically gives us the Rademacher theorem for Lipschitz maps defined on a subset $E \subset \mathbb{R}^n$.

Remark 2.7. The statement of the Rademacher theorem presented here is not the most general possible. Indeed, the proof we are about to give works for Lipschitz maps with values in *any finite-dimensional normed space.*

On the other hand, for infinite-dimensional Banach spaces, the statement is usually false. There is a particular class of Banach spaces for which the theorem holds, usually referred in the literature as *Banach spaces with the Lusin property*).

Lemma 2.28. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a Lipschitz map. Then f belongs to $W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^n)$, that is, the distributional gradient Df is essentially bounded.*

Sketch of the Proof. Given a locally summable $f \in L^1_{\mathrm{loc}}(\mathbb{R}^n)$, and given a direction $v \in \mathbb{R}^n$, it is a well-known fact that the distributional directional derivative is given by

$$\frac{\partial f}{\partial v} = \lim_{h \to 0} \frac{f - \tau_{hv} f}{h}.$$

Since f is a Lipschitz map, we easily infer that

$$\sup_{x \in \mathbb{R}^n} \left| \frac{f(x) - \tau_{hv} f(x)}{h} \right| \leq \mathrm{Lip}(f)|v|,$$

that is, the directional distributional derivative belongs to $L^\infty(\mathbb{R}^n)$, and so also the distributional gradient. $\square$

Remark 2.8. For every $p \in [1, +\infty)$ it turns out that

$$W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^n) \subset W^{1,p}_{\mathrm{loc}}(\mathbb{R}^n).$$

Theorem 2.29. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a continuous Sobolev function, that is $f \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^n)$ for some $p \in (n, +\infty)$. Then f is $\mathcal{L}^n$-almost everywhere differentiable, and the pointwise gradient agrees with the distributional gradient for $\mathcal{L}^n$-almost every $x \in \mathbb{R}^n$.*

Proof. Fix a ball $B := B\left(\underline{x},\, r\right)$.

Step 1. We claim that

$$\omega(f,\,B) \le C(n)\cdot r\left(\fint_B |\nabla f(x)|^p\, \mathrm{d}x\right)^{1/p}, \tag{2.7}$$

where $C(n)$ is a universal constant[3] and ω is oscillation function, that is,

$$\omega(f,\,B) := \sup\left\{|f(x) - f(\underline{x})| \mid |x - \underline{x}| < r\right\}.$$

Step 2. In order to prove (2.7), we notice that we can always reduce to the case $B = B(0,\,1)$ by composing with the transformation

$$x \longmapsto \frac{x - \underline{x}}{r}.$$

The Poincaré inequality gives the inequality[4]

$$\left|\int_{B(0,\,1)} f(x)\,\mathrm{d}x - \fint_{B(0,\,1)} f(x)\,\mathrm{d}x\right| \le c(n)\|\nabla f\|_{L^p(B)},$$

from which we infer that

$$\Phi(f) = \|\nabla f\|_{L^p(B(0,\,1))} + \left|\fint_{B(0,\,1)} f(x)\,\mathrm{d}x\right|$$

is equivalent to the usual $W^{1,\,p}(B(0,\,1))$ norm. Therefore, we apply Sobolev embedding theorem and obtain the inequality

$$\|f\|_{\infty,\,B(0,\,1)} \le c_1(n)\left[\|\nabla f\|_{L^p(B(0,\,1))} + \left|\fint_{B(0,\,1)} f(x)\,\mathrm{d}x\right|\right].$$

In conclusion, it turns out that

$$\omega(f,\,B) \le 2\,\|f\|_\infty \le 2\,c_1(n)\left[\|\nabla\, f\|_{L^p(B(0,\,1))} + \left|\fint_{B(0,\,1)} f(x)\,\mathrm{d}x\right|\right],$$

and, by replacing f with $f - \mathrm{av}(f)$, we infer that (2.7) holds true since the left-hand side does not change if we add a constant to the function.

Step 3. For every $h \in \mathbb{R}^n$ it turns out that

$$|f(\underline{x} + h) - f(\underline{x})| \le C(n)|h|\left(\fint_B |\nabla f(x)|^p\, \mathrm{d}x\right)^{1/p}.$$

Therefore, if L is a linear application, it turns out that

$$|f(\underline{x} + h) - f(\underline{x}) - Lh| \le C(n)|h|\left(\fint_B |\nabla f(x) - L|^p\, \mathrm{d}x\right)^{1/p},$$

and this concludes the proof. Indeed, if $\underline{x}$ is a point of L^p-approximate continuity of $\nabla f \in \mathrm{Ł}^p_{\mathrm{loc}}(\mathbb{R}^n)$, and if we take $L := \nabla f(\underline{x})$, then it turns out that

$$|f(\underline{x} + h) - f(\underline{x}) - \nabla f(\underline{x})h| \le C(n)|h|\mathcal{O}(1) = C(n)\,\mathcal{O}(\|h\|).$$

□

[3] We say that a constant is universal if it depends only on the dimension of the ambient space

[4] The constant depends only on the dimension n since the domain is the unitary ball!

The Rademacher theorem for Lipschitz maps follows easily combining the lemma and the theorem above. We are finally ready to state the local version of the Lusin property theorem and to give, at least, the main ideas behind it.

Theorem 2.30. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a $\mathcal{L}^n$-almost everywhere differentiable map, and let Ω be an open bounded subset of $\mathbb{R}^n$. For every $\epsilon > 0$ there exist a compact subset $K_\epsilon \subset \Omega$ and a function $g_\epsilon \in C^1(\mathbb{R}^n; \mathbb{R})$ satisfying the following properties:*

(1) The two maps coincide up to a set of measure at most ϵ, that is,

$$f\big|_{K_\epsilon} = g_\epsilon\big|_{K_\epsilon} \qquad \textit{and} \qquad \mathcal{L}^n\left(\Omega \setminus K_\epsilon\right) \leq \epsilon.$$

(2) The Lipschitz constant is bounded from above, that is,

$$\mathrm{Lip}(g_\epsilon) \leq \mathrm{Lip}(f) + \epsilon$$

provided that f is Lipschitz.

Proof. Let

$$D := \{x \in \mathbb{R}^n \mid f \text{ is differentiable at } x\}$$

be the set of points where f is differentiable. Then, for any $x \in D$, it turns out that

$$\omega(x, r) := \sup_{|h|<r} \left| \frac{f(x+h) - \nabla f(x)h}{h} \right| \to 0$$

decreasingly as $r \to 0^+$. For every $\epsilon > 0$ we can find a compact set $K_\epsilon \subset \Omega$ such that

(a) $\mathcal{L}^n\left(\Omega \setminus K_\epsilon\right) \leq \epsilon$;

(b) ∇f is continuous at each point of K_ϵ (Lusin);

(c) $\omega_r(x) \searrow 0$ as $r \to 0^+$, uniformly with respect to $x \in K_\epsilon$.

The reader may check these assertions as an exercise[5]. Let $A = \mathbb{R}^n \setminus K_\epsilon$ be the complement of K_ϵ in Ω; for any integer $i \in \mathbb{Z}$ we consider the set

$$A_i := \left\{ x \in A \;\middle|\; \frac{1}{2^{i+1}} < d(x, K_\epsilon) < \frac{1}{2^{i-1}} \right\}.$$

Clearly $\{A_i\}_{i\in\mathbb{Z}}$ is a covering of A; hence there exists a smooth partition of unity $\{\sigma_i\}_{i\in\mathbb{Z}}$ subordinated to it, with the additional property

$$|\nabla \sigma_i| \leq 2^{i+3}.$$

Let ρ be a regularizing kernel with support contained in the unitary ball, and assume also that $\int \rho = 1$ and $\int x\rho = 0$. If we let $\rho_i := \rho_{\frac{1}{2^{i+1}}}$ be the rescaling, then we can set

$$g(x) := \begin{cases} f(x) & \text{if } x \in K_\epsilon, \\ \displaystyle\sum_{i\in\mathbb{Z}} \left(f(x)\sigma_i(x)\right) * \rho_i(x) & \text{if } x \notin K_\epsilon. \end{cases}$$

To conclude the proof the reader may check that

(a) $f \equiv g$ on K_ϵ;

(b) g is smooth on A;

(c) g is differentiable at every $x \in K_\epsilon$ and $\nabla g = \nabla f$ on A.

□

[5]**Hint.** The second assertion follows from the Lusin theorem, while the third one follows from the Egorov theorem.

2.4.3 Criteria for Differentiable Functions

In this section, we shall always assume that $f : \Omega \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$ is a differentiable function defined on an open set Ω. The goal is to describe criteria, using the differentiability, that characterize a convex function completely.

Lemma 2.31. *The function f is convex if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x\rangle \quad \textit{for any } x,\, y \in \Omega. \tag{2.8}$$

The function f is strictly convex if and only if

$$f(y) > f(x) + \langle \nabla f(x),\, y - x\rangle \quad \textit{for any } x \neq y \in \Omega. \tag{2.9}$$

Proof. The two statements looks very much similar, but the proof is slightly different since a strict inequality is, in general, not preserved under a limit.

Step 1. Suppose that f is convex. A simple rewrite of (2.1) yields to

$$f(y) - f(x) \geq \frac{f(x + \lambda(y - x)) - f(x)}{\lambda},$$

and hence it suffices to take the limit of the right-hand side for $\lambda \to 0$ to obtain (2.8). Vice versa, we apply (2.8) twice with $(\lambda x + (1 - \lambda)y,\, y)$ and $(\lambda x + (1 - \lambda)y,\, x)$, and we find that

$$\begin{cases} f(y) \geq f(\lambda x + (1 - \lambda)y) + \langle \nabla f(\lambda x + (1 - \lambda)y),\, \lambda(y - x)\rangle\,, \\ f(x) \geq f(\lambda x + (1 - \lambda)y) + \langle \nabla f(\lambda x + (1 - \lambda)y),\, (1 - \lambda)(x - y)\rangle\,. \end{cases}$$

We multiply the first inequality by $(1 - \lambda)$ and the second inequality by λ, and we subtract them obtaining

$$\lambda f(x) + (1 - \lambda)f(y) \geq (\lambda + 1 - \lambda)f(\lambda x + (1 - \lambda)y) = f(\lambda x + (1 - \lambda)y)$$

for any $\lambda \in (0,\, 1)$ and for any $x,\, y \in \Omega$.

Step 2. The "vice versa" proof presented above works, without any change (except $\geq \to >$), for the arrow "(2.9) implies f strictly convex". Suppose now that f is strictly convex, and notice that the previous step implies that

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x\rangle \quad \text{for any } x,\, y \in \Omega.$$

The idea is to argue by contradiction. Suppose that there are $x,\, y \in \Omega$ such that

$$f(y) = f(x) + \langle \nabla f(x),\, y - x\rangle,$$

and recall that, by strict convexity (2.2), we have

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y) \quad \text{for all } \lambda \in [0,\, 1].$$

Combining these two relations, we obtain a strict inequality

$$f(\lambda x + (1 - \lambda)y) < f(x) + (1 - \lambda)\langle \nabla f(x),\, y - x\rangle \quad \text{for all } \lambda \in [0,\, 1],$$

which gives us a contradiction $(f(x) < f(x))$ by choosing $\lambda = 1$. ☐

Corollary 2.32. *The function f is strongly convex if and only if*

$$f(y) \geq f(x) + \frac{\mu}{2}\left[\|y\|^2 - \|x\|^2\right] + \langle \nabla f(x),\, y-x\rangle - \mu\|x\|\sum_{i=1}^{n}(y_i - x_i) \tag{2.10}$$

for all $x,\, y \in \Omega$.

Lemma 2.33. *The function f is convex if and only if the gradient is a monotone operator, that is,*

$$\langle \nabla f(y-x),\, y-x\rangle \geq 0 \quad \textit{for all } x,\, y \in \Omega. \tag{2.11}$$

The function f is strictly convex if and only if the gradient is a strictly monotone operator, that is,

$$\langle \nabla f(y-x),\, y-x\rangle > 0 \quad \textit{for all } x \neq y \in \Omega. \tag{2.12}$$

Proof. The argument for f convex works for f strictly convex, and hence we only prove (2.11).

$\underline{\Longrightarrow}$: If f is convex, we apply (2.8) twice and obtain the following system of inequalities:

$$\begin{cases} f(y) - f(x) \geq \langle \nabla f(x),\, y-x\rangle, \\ f(x) - f(y) \geq \langle \nabla f(y),\, x-y\rangle. \end{cases}$$

The gradient is a linear operator, and therefore the difference between the two inequalities yields to

$$\langle \nabla f(x),\, y-x\rangle - \langle \nabla f(y),\, x-y\rangle \leq 0 \implies \langle \nabla f(y-x),\, y-x\rangle \geq 0,$$

which is exactly what we wanted to prove.

$\underline{\Longleftarrow}$: Vice versa, consider the function $g(\lambda) := f(x + \lambda(y-x))$. Then

$$g'(\lambda) = \langle \nabla f(x+\lambda(y-x),\, y-x\rangle\,,$$

and by assumption we have $g'(1) \geq g'(0)$. The fundamental theorem of integral calculus implies that

$$f(y) = g(1) = g(0) + \int_0^1 g'(t)\,\mathrm{d}t = \underbrace{g(0)}_{=f(x)} + \int_0^1 g'(t),$$

which means that by (2.8) it suffices to prove that

$$\int_0^1 g'(t)\,\mathrm{d}t \geq \langle \nabla f(x),\, y-x\rangle = g'(0).$$

These inequality follows immediately from (2.11) by choosing suitable x and y. □

Corollary 2.34. *The function f is strongly convex if and only if*

$$\langle \nabla f(y-x),\, y-x\rangle \geq \frac{\mu}{2}\|y-x\|\sum_{i=1}^{n}(y_i - x_i) \quad \textit{for all } x,\, y \in \Omega. \tag{2.13}$$

To conclude this section, we prove a criterion easy to check, but that requires more regularity to hold true. In particular, we assume that $f : \Omega \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$ is a twice-differentiable function.

Definition 2.35. A $n \times n$ matrix M is *positive-definite* if

$$\langle Mx, x\rangle > 0 \quad \text{for all } x \in \mathbb{R}^n.$$

A $n \times n$ matrix M is *positive-semidefinite* if

$$\langle Mx, x\rangle \geq 0 \quad \text{for all } x \in \mathbb{R}^n.$$

Lemma 2.36. *The function f is convex if and only if the Hessian matrix $H(f)(x)$ is positive-semidefinite for all $x \in \Omega$.*

Proof.

$\underline{\Longrightarrow}$: Let $x, y \in \Omega$ and $\lambda \in [0, 1]$. By Taylor's formula with Lagrange remainder we have

$$f(x + \lambda y) = f(x) + \lambda\langle \nabla f(x), y\rangle + \frac{1}{2}\lambda^2\langle H(f)(x), y\rangle + r(\lambda y),$$

and applying (2.8) we find that

$$\frac{1}{2}\lambda^2\langle H(f)(x), y\rangle + r(\lambda y) \geq 0.$$

If $y \neq 0$ (otherwise we just switch the roles of x and y), then

$$\frac{1}{2}\langle H(f)(x), y\rangle + \frac{r(\lambda y)}{\|\lambda y\|^2}\|y\|^2 \geq 0,$$

and the thesis follows by taking the limit as $\lambda \to 0$.

$\underline{\Longleftarrow}$: Vice versa, let $x, y \in \Omega$. By Taylor's formula there exists $\lambda \in [0, 1]$ such that

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\langle H(f)(x + \lambda(y - x)), y - x)\rangle.$$

The matrix $H(f)(x + \lambda(y - x))$ is positive-semidefinite, and therefore

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle,$$

which implies f convex by (2.8). $\square$

Lemma 2.37. *The function f is strictly convex if the Hessian matrix $H(f)(x)$ is positive-definite for all $x \in \Omega$.*

Remark 2.9. The opposite implication in Lemma 2.37 is not true. Indeed, the function $f(x) = x^4$ is strictly convex, but its Hessian $12x^2$ vanishes in 0.

Corollary 2.38. *The function f is strongly convex if and only if $H(f)(x) - \mu\mathrm{Id}$ is positive-semidefinite for all $x \in \Omega$ if and only if*

$$\min\{\lambda : \lambda \textit{ eigenvalue of } H(f)(x)\} \geq \mu.$$

Remark 2.10. A quadratic function $f(x) = \frac{1}{2}x^TQx + b^Tx + c$ is strictly convex if and only if it is strongly convex.

2.5 Subgradient and Subdifferential

In this section, we introduce a notion that generalizes the use of the gradient in Section 2.4.3 for the set of points where a convex functions is (eventually) not differentiable.

2.5.1 Directional Derivative of a Convex Function

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a function, and let $v \in \mathbb{R}^n$ be a vector. The derivative along v of f at x, known as directional derivative, is defined by setting

$$\partial_v f(x) := \lim_{t \searrow 0} \frac{f(x+tv) - f(x)}{t}.$$

Our first goal is to prove that a convex function admits a directional derivative at any point $x \in \operatorname{Int}\Omega$ and for any direction v. We remark that for function f differentiable at x we have

$$\partial_v f(x) = \langle \nabla f(x), \, v \rangle,$$

where $\langle \cdot, \, \cdot \rangle$ denotes the usual scalar product.

Theorem 2.39. *Let $\Omega \subseteq \mathbb{R}^n$ be a convex set. A convex function $f : \Omega \longrightarrow \mathbb{R}$ has a directional derivative $\partial_v f(x)$ for every $v \in \mathbb{R}^n$ and $x \in \operatorname{Int}\Omega$.*

In particular, a convex function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ has directional derivatives at any point.

Proof. Fix $x \in \operatorname{Int}\Omega$ and a direction $v \in \mathbb{R}^n$. It is enough to prove that the function

$$g(t) := \frac{f(x+tv) - f(x)}{t}$$

is bounded and monotone for $t \searrow 0$.

Bounded. The function f is locally Lipschitz, which means that for $t > 0$ small enough we have

$$|g(t)| \leq \frac{1}{t} |f(x+tv) - f(x)| \leq 2M \frac{t\|v\|}{t} = 2M\|v\|.$$

Monotone. We claim that g is monotone decreasing on compact sets, that is,

$$0 < t_2 < t_1 \implies g(t_2) < g(t_1).$$

Let $t_1 > 0$ be small enough for $x + t_1 v$ to belong to Ω (which is always possible because x is an interior point). For every $0 < t_2 < t_1$ we have that

$$x + t_2 v = \left(1 - \frac{t_2}{t_1}\right) x + \frac{t_2}{t_1} (x + t_1 v) \implies x + t_2 v \in \Omega,$$

and the convexity of f yields to

$$f(x + t_2 v) \leq \left(1 - \frac{t_2}{t_1}\right) f(x) + \frac{t_2}{t_1} f(x + t_1 v).$$

Therefore

$$g(t_2) = \frac{f(x+t_2 v) - f(x)}{t_2} \leq \frac{1}{t_1} f(x + t_1 v) - \frac{1}{t_1} f(x) = g(t_1),$$

which means that g is monotone decreasing on $[0, \, T]$ for every $T > 0$.

Conclusion. The function g is bounded at $t = 0$ and monotone decreasing; hence the limit exists and it coincides with the infimum over $t > 0$, that is,

$$\partial_v f(x) = \inf_{t>0} g(t) = \lim_{t\searrow 0} g(t).$$

$\square$

2.5.2 Subgradient and Subdifferential of a Convex Function

Definition 2.40. Let $f : \Omega \longrightarrow \mathbb{R}$ be a function. A vector $s \in \mathbb{R}^n$ is a *subgradient* of f at $x \in \Omega$ if

$$f(y) \geq f(x) + \langle s,\, y - x\rangle \quad \text{for every } y \in \Omega. \tag{2.14}$$

The *subdifferential* of f at $x \in \Omega$ is the set of all its subgradients at x, that is,

$$\partial f(x) := \{ s \in \mathbb{R}^n \mid s \text{ satisfies } (2.14) \}.$$

We first need to verify that the new definition actually generalizes the notion of derivative, that is, we need to check if the subdifferential $\partial f(x)$ of a differentiable function is a singlet $\{\nabla f(x)\}$.

Proposition 2.41. *Let $f : \Omega \longrightarrow \mathbb{R}$ be a function. The subdifferential of f at x consists of $\nabla f(x)$ only, i.e.*

$$\partial f(x) = \{\nabla f(x)\}.$$

Proof. The inclusion $\supseteq$ is obvious, hence we only need to prove that the difference between a vector $s \in \partial f(x)$ and $\nabla f(x)$ is necessarily equal to 0. It follows from (2.14) that

$$f(x + tv) \geq f(x) + \langle s,\, tv\rangle, \tag{2.15}$$

while using the Taylor's formula we find that

$$f(x + tv) = f(x) + t\langle \nabla f(x),\, v\rangle + o(t\|v\|). \tag{2.16}$$

If we plug (2.16) into (2.15), we find that

$$\langle s - \nabla f(x),\, v\rangle + \frac{o(t\|v\|)}{t} \leq 0 \implies \langle s - \nabla f(x),\, v\rangle \leq 0$$

by taking the limit as t tends to 0. Finally, we notice that v is arbitrary, and thus we can take $v = s - \nabla f(x)$, which means that

$$\|s - \nabla f(x)\|^2 \leq 0 \implies s - \nabla f(x) = 0.$$

$\square$

The next result will describe the main topological properties of the subdifferential of a convex function at any interior point. To prove that it is not empty, we need a slightly stronger separation theorem (which can be obtained as "limit" of the Hahn-Banach geometric theorem, see Theorem 3.9).

Theorem 2.42. *Let $A \subseteq \mathbb{R}^n$ be a convex nonempty set, and let $\bar{z} \in \partial A$ be a point on its boundary. There exists an hyperplane H thorough $\bar{z}$ such that A is on one side of it, that is,*

$$\exists\, z^* \in \mathbb{R}^n \setminus \{0\} \,:\, \langle z^*,\, z\rangle \leq \langle z^*,\, \bar{z}\rangle \quad \textit{for every } z \in A.$$

Theorem 2.43. *Let $f : \Omega \longrightarrow \mathbb{R}$ be a convex function. The subdifferential $\partial f(x)$ is nonempty, convex and compact at every internal point $x \in \operatorname{Int}\Omega$.*

Proof.

Step 1. We first prove the convexity of $\partial f(x)$. Let $s_1, s_2 \in \partial f(x)$, and let $\lambda \in (0, 1)$. By (2.14) we have

$$f(y) \geq f(x) + \langle s_1, y - x\rangle \quad \text{and} \quad f(y) \geq f(x) + \langle s_2, y - x\rangle$$

for every $y \in \Omega$. We multiply the first inequality by λ, the second by $(1 - \lambda)$, and we sum them to find that

$$f(y) \geq f(x) + \langle \lambda s_1 + (1 - \lambda)s_2, y - x\rangle,$$

which means that $\partial f(x)$ is convex.

Step 2. We now prove that $\partial f(x)$ is bounded and closed. Let $(s_k)_{k\in\mathbb{N}} \subset \partial f(x)$ be a converging sequence, and let $s := \lim_{k\to+\infty} s_k$. By (2.14) we have that

$$f(y) \geq f(x) + \langle s_k, y - x\rangle \quad \text{for every } y \in \Omega,$$

and taking the limit for $k \to +\infty$ yields to

$$f(y) \geq f(x) + \langle s, y - x\rangle \quad \text{for every } y \in \Omega,$$

which means that $\partial f(x)$ is closed. To prove the boundedness, recall that the function f is locally Lipschitz, hence for $y \in B(x, \epsilon)$ we have that

$$\langle s, y - x\rangle \leq f(y) - f(x) \leq M \cdot \|y - x\|.$$

Let $y := x + \epsilon\frac{s}{\|s\|}$. Clearly $y \in B(x, \epsilon)$, and we have that

$$\epsilon\|s\| = \langle s, \epsilon\frac{s}{\|s\|}\rangle \leq M \cdot \left\|\epsilon\frac{s}{\|s\|}\right\| \implies \|s\| \leq M,$$

which means that $\partial f(x)$ is bounded, and thus compact.

Step 3. Let $A := \mathrm{epi}(f)$ and $(\bar{x}, f(\bar{x})) \in \partial A$. By Theorem 2.42 there exists $(\mu^*, t^*) \in \mathbb{R}^n \times \mathbb{R} \setminus \{\mathbf{0}\}$ such that

$$\langle \mu^*, x\rangle + t^*t \leq \langle \mu^*, \bar{x}\rangle + t^*f(\bar{x}),$$

from which it follows that

$$\langle \mu^*, x\rangle + t^*f(x) \leq \langle \mu^*, x\rangle + t^*t \leq$$
$$\leq \langle \mu^*, \bar{x}\rangle + t^*f(\bar{x}).$$

From the previous inequality we easily deduce that $t^* \leq 0$, and therefore

$$(-t^*)f(\bar{x}) \geq (-t^*)f(x) + \langle \mu^*, x - \bar{x}\rangle \quad \text{for every } x \in \Omega,$$

which means that to conclude the proof it is enough to show that t^* cannot be equal to 0. We argue by contradiction. If $t^* = 0$, the inequality above yields to

$$\langle \mu^*, x - \bar{x}\rangle \leq 0 \quad \text{for every } x \in \Omega,$$

and, in particular, for every $x \in B(\bar{x}, \epsilon)$. Therefore

$$\langle \mu^*, y\rangle \leq 0 \quad \text{for every } y \in B(0, \epsilon),$$

and this is enough to infer that $\mu^* = 0$, which gives a contradiction since (μ^*, t^*) is not the zero vector in $\mathbb{R}^{n+1}$. □

2.6 Sublinear Functions

In this section, we introduce the notion of *sublinear* function and we show the connection with convex functions, directional derivatives and subgradients.

2.6.1 Main Properties and Support Functions

Definition 2.44. A function $p : \mathbb{R}^n \longrightarrow \mathbb{R}$ is *sublinear* if the following properties are satisfied:

(a) The function p is positively homogeneous, that is,

$$p(tv) = tp(v) \quad \text{for every } t > 0 \text{ and } v \in \mathbb{R}^n.$$

(b) The function p is subadditive, that is,

$$p(v_1 + v_2) \leq p(v_1) + p(v_2) \quad \text{for every } v_1,\, v_2 \in \mathbb{R}^n.$$

Proposition 2.45. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. The directional derivative is a sublinear function of the direction, that is,*

$$v \longrightarrow \partial_v f(x)$$

is sublinear for all fixed $x \in \mathbb{R}^n$.

Proof. Let $t > 0$ be a positive real number, and let $v \in \mathbb{R}^n$ be a direction. By definition

$$\begin{aligned} \partial_{tv} f(x) &= \lim_{h \searrow 0} \frac{f(x + htv) - f(x)}{h} = \\ &= t \lim_{h \searrow 0} \frac{f(x + htv) - f(x)}{ht} = \\ &= t \lim_{\ell \searrow 0} \frac{f(x + \ell v) - f(x)}{\ell} = t\partial_v f(x), \end{aligned}$$

where $\ell := ht$ decreases to zero because t is positive and fixed, and $h \searrow 0$. To prove the subadditivity property, we notice that

$$x + t(v_1 + v_2) = \frac{1}{2}(x + 2tv_1) + \frac{1}{2}(x + 2tv_2).$$

By convexity, it turns out that

$$f(x + t(v_1 + v_2)) \leq \frac{1}{2} f(x + 2tv_1) + \frac{1}{2} f(x + 2tv_2),$$

and this inequality yields to

$$\begin{aligned} \partial_{v_1 + v_2} f(x) &= \lim_{t \searrow 0} \frac{f(x + t(v_1 + v_2)) - f(x)}{h} \leq \\ &\leq \lim_{t \searrow 0} \left[\frac{f(x + 2tv_1) - f(x)}{2t} + \frac{f(x + 2tv_2) - f(x)}{2t} \right] = \\ &= \partial_{v_1} f(x) + \partial_{v_2} f(x), \end{aligned}$$

which is exactly what we wanted to prove. $\square$

Theorem 2.46. *A function $p : \mathbb{R}^n \longrightarrow \mathbb{R}$ is sublinear if and only if p is convex and positively homogeneous.*

Proof. A positively homogeneous subadditive function is trivially convex because

$$p(\lambda x + (1-\lambda)y) \leq p(\lambda x) + p((1-\lambda)y) = \lambda p(x) + (1-\lambda)p(y)$$

for every $x,\, y \in \mathbb{R}^n$ and every $\lambda \in [0,\, 1]$.

Vice versa, suppose that p is convex and positively homogeneous, and let $x,\, y \in \mathbb{R}^n$ be any two points. A straightforward computation shows that

$$\frac{1}{2}p(x+y) = p\left(\frac{x}{2} + \frac{y}{2}\right) \leq \frac{1}{2}(p(x) + p(y)),$$

from which we infer that p is subadditive. □

Corollary 2.47. *The epigraph of a sublinear function $p : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a convex cone*[6].

Proof. As a consequence of Theorem 2.46, it is enough to show that $\mathrm{epi}(f)$ is a cone. But this follows immediately from the definition:

$$(x,\, s) \in \mathrm{epi}(f) \implies s \geq p(x) \implies ts \geq \underbrace{tp(x)}_{=p(tx)} \implies (tx,\, ts) \in \mathrm{epi}(f).$$

□

Definition 2.48 (Support Function)**.** Let $C \subseteq \mathbb{R}^n$ be a set. The *support function* of C is defined by setting

$$\sigma_C(x) := \sup\left\{\langle x,\, y\rangle \ :\ y \in C\right\},$$

where $\langle \cdot,\, - \rangle$ denotes, e.g., the 2-norm scalar product.

Recall that the support function σ_C is convex, as it is the supremum of a collection of linear mappings $x \longmapsto \langle x,\, y\rangle$. The next goal is to prove that, for a convex function f, the directional derivative map $v \longrightarrow \partial_v f(x)$ is the support function of a nonempty convex compact set, and vice versa. We need a fundamental result in function analysis.

Theorem 2.49 (Hahn-Banach, analytic form)**.** *Let X be a real vector space and let $p : X \longrightarrow \mathbb{R}$ be a function satisfying the following properties:*

1) $p(\lambda x) = \lambda p(x)$ *for every $x \in X$ and every positive $\lambda \in \mathbb{R}$.*

2) $p(x+y) \leq p(x) + p(y)$ *for every $x,\, y \in X$.*

Let $G \subset X$ be a linear subspace of X, and let $g : G \longrightarrow \mathbb{R}$ be a linear functional such that

$$g(x) \leq p(x), \qquad \forall\, x \in G.$$

Then there exists a linear functional $f : X \longrightarrow \mathbb{R}$, which extends g, satisfying the following property:

$$f(x) \leq p(x) \qquad \forall\, x \in X.$$

[6]**Definition.** A subspace $\mathcal{C}$ of a vector space V is a *cone* if for every $x \in \mathcal{C}$ and $t > 0$, we have $tx \in \mathcal{C}$.

Proof. Let us consider the family

$$\mathcal{F} := \left\{ h : Y \longrightarrow \mathbb{R} \;\middle|\; \begin{array}{c} Y \text{ is a linear subspace of } X, \\ h \text{ is linear, } G \subseteq Y, \, h \text{ extends } g, \\ \text{and } h(x) \leq p(x) \text{ for every } x \in Y \end{array} \right\}.$$

A standard argument proves that

$$h_1 \leq h_2 \iff Y_1 \subset Y_2 \text{ and } h_2 \text{ extends } h_1$$

is a partial order relation on $\mathcal{F}$. The reader may easily check that the family $\mathcal{F}$ is nonempty (since $(g, Y) \in \mathcal{F}$) and that it is *inductive*. Indeed, let $\{h_i\}_{i \in I} \subset P$ be a totally ordered chain. If we set

$$Y = \bigcup_{i \in I} Y_i, \qquad h(x) = h_i(x) \quad \text{if } x \in Y_i \text{ for some } i,$$

then it is easy to prove that (h, Y) is an upper bound for the chain. By Zorn's lemma, there is a maximal element $(f, Y) \in \mathcal{F}$: If we can prove that Y is equal to X, then the functional $f : X \longrightarrow \mathbb{R}$ will be the sought one.

We argue by contradiction. Suppose that there exists an element $x_0 \in X \setminus Y$, and let us consider the linear space $Z := Y \oplus \mathbb{R}x_0$. For every $x \in Y$, we also define

$$h(x) = f(x) + t\alpha, \qquad t \in \mathbb{R},$$

where the constant $\alpha \in \mathbb{R}$ needs to be chosen in such a way that $(h, Z) \in \mathcal{F}$. Let α be any real number satisfying the inequalities

$$\sup_{y \in Y} \left[f(y) - p(y - x_0) \right] \leq \alpha \leq \inf_{x \in Y} \left[p(x + x_0) - f(x) \right].$$

It is easy to prove that such an α must exist. Indeed, the subadditivity of p yields to

$$f(u) + f(v) \leq p(u + v + x_0 - x_0) \leq p(u - x_0) + p(v + x_0), \qquad \forall\, u,\, v \in Y,$$

which, in turn, implies

$$f(u) - p(u - x_0) \leq -f(v) + p(v + x_0), \qquad \forall\, u,\, v \in Y.$$

It follows that for any $x \in Y$ we have

$$\begin{cases} f(x) + \alpha \leq p(x + x_0) \\ \\ f(x) - \alpha \leq p(x - x_0), \end{cases}$$

and, multiplying both by $t > 0$, if we set $y = tx$, then we obtain

$$\begin{cases} f(y) + \alpha y \leq p(y + tx_0) \\ \\ f(y) - t\alpha \leq p(y - tx_0). \end{cases}$$

In conclusion, for every $t \in \mathbb{R}$, the inequalities above implies that

$$f(x) + t\alpha \leq p(x + tx_0),$$

and hence we can always take $h(x_0) = \alpha$. In particular, the map h extends f and it satisfies the inequality $h \leq p$; hence $(h, Z) \in \mathcal{F}$, which is in contradiction with the maximality of f. $\square$

Remark 2.11. The Hahn-Banach theorem holds in every vector space, but in this course, we will use it mainly when $X := \mathbb{R}^n$. Note also that, if $L \subset \mathbb{R}^n$ is a linear proper subspace, then a linear functional

$$L \ni x \longmapsto \langle x, s \rangle \in \mathbb{R} \quad \text{for some } s \in \mathbb{R}^n$$

can be extended by Hahn-Banach to a linear functional

$$\mathbb{R}^n \ni x \longmapsto \langle x, \hat{s} \rangle \in \mathbb{R} \quad \text{for some } \hat{s} \in \mathbb{R}^n,$$

but, in general, $\hat{s}$ is not equal to s. Consider, for example, the linear subspace $L := \mathbb{R} \times \{0\} \subset \mathbb{R}^2$, the linear functional $(x, y) \longmapsto x + y$ and the sublinear function $p := \| \cdot \|_2$.

We now introduce a technical result that gives us an alternative definition of the subdifferential of f at a point $x \in \mathbb{R}^n$. Recall that, for a given f and $x \in \mathrm{dom}(f)$, we have

$$\partial f(x) := \{ s \in \mathbb{R}^n \mid s \text{ satisfies } (2.14) \} .$$

Lemma 2.50. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a function. Then*

$$\partial f(x) = \{ s \in \mathbb{R}^n \mid \partial_v f(x) \geq \langle s, v \rangle \text{ for all } v \in \mathbb{R}^n \} .$$

Proof.

" $\subseteq$ " Let $s \in \partial f(x)$. By (2.14) we have that

$$f(y) \geq f(x) + \langle s, y - x \rangle,$$

and hence we can chose $y := x + tv$. It follows that

$$f(x + tv) - f(x) \geq \langle s, tv \rangle \implies \frac{f(x + tv) - f(x)}{t} \geq \langle s, v \rangle,$$

which proves the first inclusion by taking the limit for $t \searrow 0$.

" $\supseteq$ " Let $s \in \{ s \in \mathbb{R}^n \mid \partial_v f(x) \geq \langle s, v \rangle \text{ for all } v \in \mathbb{R}^n \}$. The incremental ratio is monotonically decreasing when $t \searrow 0$, hence

$$\langle s, v \rangle \leq \inf_{t>0} \frac{f(x + tv) - f(x)}{t} \leq$$

$$\leq f(x + v) - f(x),$$

and, since v is arbitrary, we see that (2.14) is satisfied by setting $y := x + v$. □

Theorem 2.51. *Let $x \in \mathbb{R}^n$ and let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. Then the directional derivative $v \longmapsto \partial_v f(x)$ is the support function of the nonempty convex compact set $\partial f(x)$.*

Proof. Note that we can equivalently prove that

$$\partial_v f(x) = \sup \{ \langle v, s \rangle \, : \, s \in \partial f(x) \} = \max \{ \langle v, s \rangle \, : \, s \in \partial f(x) \} ,$$

where the last equality follows from the compactness of $\partial f(x)$.

" $\geq$ " By Lemma 2.50 we have that

$$s \in \partial f(x) \implies \partial_v f(x) \geq \langle s, v \rangle,$$

and hence by taking the sup we infer that

$$\partial_v f(x) \geq \max \{ \langle s, v \rangle \ : \ s \in \mathbb{R}^n \} .$$

" $\leq$ " We argue by contradiction. Suppose that there exists $\bar{v} \in \mathbb{R}^n$ (wlog $\|v\| = 1$) such that

$$\partial_{\bar{v}} f(x) > \max \{ \langle s, \bar{v} \rangle \ : \ s \in \mathbb{R}^n \} = \langle \bar{s}, \bar{v} \rangle,$$

for some $\bar{s} \in \partial f(x)$, and let

$$\bar{\gamma} := \partial_{\bar{v}} f(x) - \langle \bar{s}, \bar{v} \rangle > 0.$$

Let $\hat{s} = \bar{s} + \bar{\gamma}\bar{v}$, and notice that

$$\langle \hat{s}, \bar{v} \rangle = \langle \bar{s}, \bar{v} \rangle + \bar{\gamma} = \partial_{\bar{v}} f(x),$$

$$\langle \hat{s}, -\bar{v} \rangle = -\langle \bar{s}, \bar{v} \rangle - \bar{\gamma} = -\partial_{\bar{v}} f(x).$$

Furthermore, the directional derivative is sublinear with respect to the direction, which yields to the inequality

$$\partial_{\bar{v}} f(x) + \partial_{-\bar{v}} f(x) \geq \partial_0 f(x) = 0 \implies \partial_{-\bar{v}} f(x) \geq -\partial_{\bar{v}} f(x).$$

Therefore

$$\langle \hat{s}, -\bar{v} \rangle = -\partial_{\bar{v}} f(x) \leq \partial_{-\bar{v}} f(x),$$

and thus it follows from the positive homogeneity that

$$\partial_{\gamma\bar{v}} f(x) \geq \langle \hat{s}, \gamma\bar{v} \rangle \quad \text{for every } \gamma \in \mathbb{R}. \tag{2.17}$$

The linear function $\langle \hat{s}, \bar{v} \rangle$ is defined on the one-dimensional linear subspace Span $< \bar{v} > := \{ \gamma\bar{v} \ : \ \gamma \in \mathbb{R} \}$, and it satisfies (2.17); hence by the Hahn-Banach Theorem we can find $s \in \mathbb{R}^n$ such that

$$\langle s, \gamma\bar{v} \rangle = \langle \hat{s}, \gamma\bar{v} \rangle \quad \text{and} \quad \langle s, v \rangle \leq \partial_v f(x)$$

for every $v \in \mathbb{R}^n$. It follows that $s \in \partial f(x)$, and

$$\langle s, \bar{v} \rangle = \langle \hat{s}, \bar{v} \rangle > \langle \bar{s}, \bar{v} \rangle,$$

which is absurd since we assumed $\langle \bar{s}, \bar{v} \rangle$ to be the maximum of the scalar product on $\partial f(x)$. □

The same argument presented here can be used to prove a more general version of Theorem 2.52 valid for all sublinear functions. Indeed, if p is a sublinear function, then one can easily prove that p is the support function of the subdifferential $\partial p(0)$. One inequality is trivial since

$$p(0) = 0 \implies p(v) \geq p(0) + \langle s, v \rangle = \langle s, v \rangle \implies p(v) \geq \max \{ \langle s, v \rangle \ : \ s \in \partial p(0) \} .$$

The opposite inequality follows arguing by contradiction with the very same proof, replacing $\partial f(x)$ with $\partial p(0)$.

Theorem 2.52. *A function $p : \mathbb{R}^n \longrightarrow \mathbb{R}$ is sublinear if and only if p is the support function of a nonempty convex compact set. Furthermore, it turns out that $p \equiv \sigma_{\partial p(0)}$.*

We now use Theorem 2.52 to show that there is a one-to-one correspondence between convex compact sets of $\mathbb{R}^n$ and sublinear functions.

Lemma 2.53. *Let $A, B \subseteq \mathbb{R}^n$ be nonempty convex compact sets. Then*

$$A = B \iff \sigma_A \equiv \sigma_B.$$

Proof. Note that by the thesis is symmetric with respect to A and B. Therefore, it is enough to prove that

$$A \subseteq B \iff \sigma_A \leq \sigma_B.$$

" $\implies$ " This implication follows immediately from the fact that $\sigma_B(v)$ is defined as the maximum over the set B of the function $s \longmapsto \langle s, v\rangle$, which is bigger than the maximum over the smaller set A.

" $\impliedby$ " We argue by contradiction. Suppose that there exists $a \in B \notin A$. The geometrical form of the Hahn-Banach theorem (see Theorem 3.8) asserts that there exist $\bar{s} \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$ such that

$$\langle \bar{s}, a\rangle > \gamma \geq \langle \bar{s}, b\rangle \quad \text{for every } b \in B.$$

We take the supremum (or maximum) over $b \in B$, and this yields to the inequality

$$\sigma_A(\bar{s}) \geq \langle \bar{s}, a\rangle > \sigma_B(\bar{s}),$$

and this is a contradiction since $\sigma_A \leq \sigma_B$ by assumption. ☐

Corollary 2.54. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. Then*

$$\partial\left(v \longmapsto \partial_v f(x)\right) = \partial f(x).$$

2.6.2 Topological Properties of the Subdifferential

We now employ the theory of sublinear functions developed in the previous section to investigate the main topological properties of the subdifferential.

Lemma 2.55. *Let $C_1, \dots, C_k \subseteq \mathbb{R}^n$ be nonempty convex compact sets. Then the convex hull of the union $C := \mathrm{Conv}(\cup_{i=1}^k C_i)$ is convex and compact.*

Proof. A finite union of compact sets is compact, which means that it is enough to prove that $\mathrm{Conv}(K)$ is bounded and closed, whenever K is a compact set in $\mathbb{R}^n$.

Step 1. The idea is to show that a sequence in $\mathrm{Conv}(K)$ has a convergent subsequence whose limit belongs to $\mathrm{Conv}(K)$. By Carathéodory's Theorem a sequence in the convex hull of K can be expressed as

$$\left(\sum_{i=1}^{n+1} \lambda_i^k x_i^k\right)_{k\in\mathbb{N}},$$

where $\sum_i \lambda_i^k x_i^k$ is a convex combination of $n+1$ points of A for every fixed k.

Step 2. The sequence $\big((x_1^k, \dots, x_{n+1}^k)\big)_{k\in\mathbb{N}}$ has a converging subsequence to a point

$$(x_1, \dots, x_{n+1}) \in K^{\otimes n+1},$$

as a consequence of the compactness of the set $K^{\otimes n+1} := K \times \cdots \times K$. Similarly, the constraints on the coefficients

$$\lambda_i^k \geq 0 \quad \text{and} \quad \sum_{i=1}^{n+1} \lambda_i^k = 1$$

implies that the sequence $\big((\lambda_1^k, \dots, \lambda_{n+1}^k)\big)_{k\in\mathbb{N}}$ has a converging subsequence to a point

$$(\lambda_1, \dots, \lambda_{n+1}) \in \mathbb{R}^{n+1}$$

satisfying the same constraints

$$\lambda_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{n+1} \lambda_i = 1.$$

In particular, the sequence of points in the convex hull of K converges to a point

$$x = \sum_{i=1}^{n+1} \lambda_i x_i,$$

which belongs to $\mathrm{Conv}(K)$ as a consequence of the properties discussed above. $\square$

Proposition 2.56. *Let $f_1, \dots, f_k : \mathbb{R}^n \longrightarrow \mathbb{R}$ be convex functions. The following assertions hold:*

(1) *For every $t > 0$ it turns out that*

$$\partial(t \cdot f)(x) = t \cdot \partial f(x).$$

(2) *The subdifferential of a sum is the sum of the subdifferentials*[7]*, that is,*

$$\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x).$$

(3) *Let $f(x) := \max_{i=1,\dots,k} f_i(x)$ be the pointwise maximum. Then*

$$\partial f(x) = \mathrm{Conv}\left(\bigcup_{i\in I_{f(x)}} \partial f_i(x)\right) \quad \text{where } I_{f(x)} = \{i \in \{1, \dots, k\} \ : \ f_i(x) = f(x)\}.$$

Proof.

(1) The reader can easily check that $\partial(t \cdot f)(x) = t \cdot \partial f(x)$ for every $t > 0$ using the definition (2.14) of subdifferential.

[7] Here the reader needs to be careful. If f_1 and f_2 are convex on $\Omega_1, \Omega_2 \subset \mathbb{R}^n$, then we need additional assumptions for the thesis to hold.

(2) By Lemma 2.53 it suffices to show that the support functions coincide. A straightforward computation yields to

$$
\begin{aligned}
\sigma_{\partial(f_1+f_2)(x)}(v) &= \partial_v(f_1+f_2)(x) = \\
&= \partial_v f_1(x) + \partial_v f_2(x) = \\
&= \max\{\langle s_1,\, v\rangle \,:\, s_1 \in \partial f_1(x)\} + \max\{\langle s_2,\, v\rangle \,:\, s_1 \in \partial f_2(x)\} = \\
&= \max\{\langle s_1 + s_2,\, v\rangle \,:\, s_1 \in \partial f_1(x),\, s_2 \in \partial f_2(x)\} = \\
&= \max\{\langle s,\, v\rangle \,:\, s \in \partial f_1(x) + \partial f_2(x)\} = \sigma_{\partial f_1(x)+\partial f_2(x)}(v),
\end{aligned}
$$

which is exactly what we wanted to prove.

(3) The proof of this statement is rather technical, and we shall not present it here. The main idea is to use Lemma 2.55 and prove that, in general,

$$
\sigma_S(v) = \max\{\sigma_{S_i}(v) \,:\, i \in \{1, \ldots, k\}\},
$$

where $S := \mathrm{Conv}(\cup_{i=1}^k S_i)$ and S_i convex compact sets.

□

2.7 Convexity and Tangent Cone

In this final section, we introduce the notions of *polar cone* and of *tangent cone* to a set X at a point $\bar{x}$. We prove that, for a convex set, the tangent cone is given by the polarization of the *normal cone* to X at $\bar{x}$, and we derive a nice characterization of the subdifferential and the directional derivative of a convex function.

Definition 2.57 (Cone)**.** A set $X \subset \mathbb{R}^n$ is a *cone* centered at 0 if it contains all the half-lines originating from 0, that is,

$$
x \in X \implies tx \in X \quad \text{for every } t > 0.
$$

Definition 2.58 (Polar Cone)**.** Let $X \subset \mathbb{R}^n$ be a set. The *polar cone* of X, denoted by X°, is the set of all the vectors $z \in \mathbb{R}^n$ with negative scalar product against each element of X, that is,

$$
X^\circ := \{z \in \mathbb{R}^n \,:\, \langle z,\, x\rangle \leq 0 \quad \text{for every } x \in X\}.
$$

Definition 2.59 (Dual Cone)**.** Let $X \subset \mathbb{R}^n$ be a set. The *dual cone* of X, denoted by X^*, is the set of all the vectors $z \in \mathbb{R}^n$ with positive scalar product against each element of X, that is,

$$
X^* := (-X)^\circ.
$$

Definition 2.60 (Normal Cone)**.** Let $X \subset \mathbb{R}^n$ be a set, and let $\bar{x} \in \mathrm{cl}\, X$. The *normal cone* to X at $\bar{x}$, denoted by $N_X(\bar{x})$, is given by

$$
N_X(\bar{x}) := \{z \in \mathbb{R}^n \,:\, \langle z,\, x - \bar{x}\rangle \leq 0 \quad \text{for every } x \in X\}.
$$

Definition 2.61 (Tangent Cone)**.** Let $X \subset \mathbb{R}^n$ be a set, and let $\bar{x} \in \mathrm{cl}\, X$. The *tangent cone* to X at $\bar{x}$, denoted by $T_X(\bar{x})$, is given by

$$
T_X(\bar{x}) := \left\{v \in \mathbb{R}^n \,:\, \exists\, t_k \searrow 0 \text{ and } \exists\, v^k \to v \text{ such that } \bar{x} + t_k v^k \in X \text{ for every } k \in \mathbb{N}\right\}.
$$

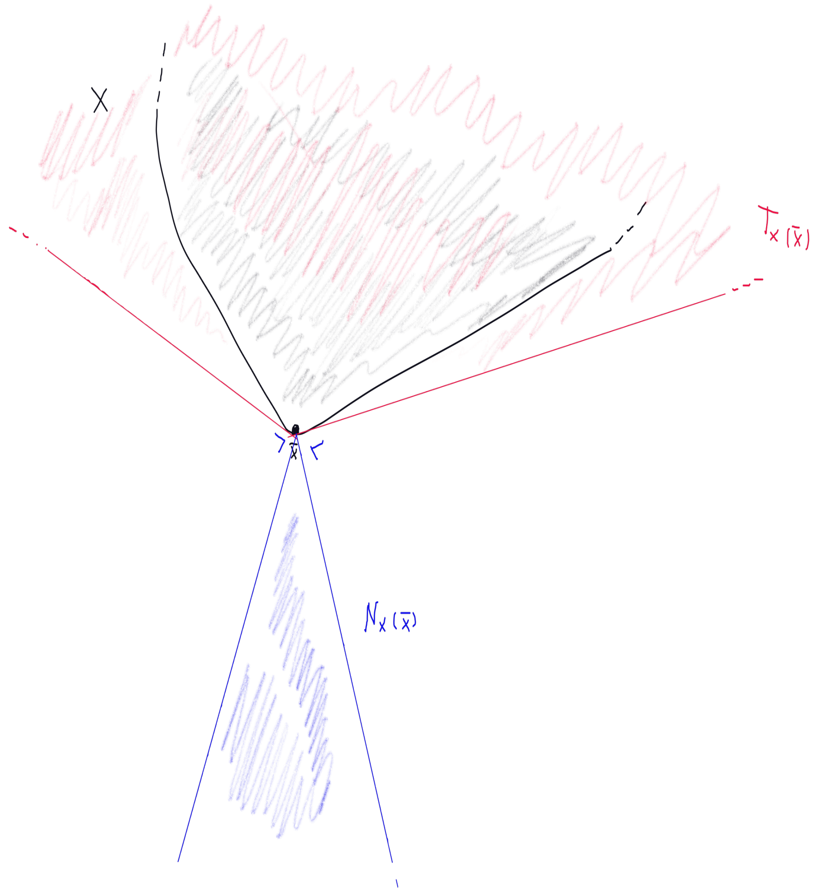

Figure 2.5: In this picture, we show the difference between the tangent cone and the normal cone at a point $\bar{x}$ on the boundary of a set X.

Remark 2.12. The reader can easily check that X°, X^*, $N_X(\bar{x})$ and $T_X(\bar{x})$ are all nonempty (the zero vector) convex closed cones (but they are by no means compact).

Proof. We prove that X° satisfies the property enlisted above; for the other three cones, it suffices to run a similar argument. First, we notice that for a given $t > 0$ we have

$$z \in X^\circ \implies \langle z,\, x\rangle \le 0 \quad \forall\, x \in X \implies \underbrace{t\langle z,\, x\rangle}_{=\langle tz,\, x\rangle} \le 0 \quad \forall\, x \in X \implies tz \in X^\circ.$$

Similarly, for any $\lambda \in (0,\, 1)$ and any couple of points $z_1,\, z_2 \in X^\circ$, it is easy to see that

$$\langle \lambda z_1 + (1-\lambda) z_2,\, x\rangle = \lambda\langle z_1,\, x\rangle + (1-\lambda)\langle z_2,\, y\rangle \le 0 \implies \lambda z_1 + (1-\lambda) z_2 \in X^\circ,$$

which means that X° is a nonempty convex cone. To prove that it is also closed, let $(z_k)_{k\in\mathbb{N}}$ be a sequence of points in X° converging to some $\bar{z} \in \mathbb{R}^n$. It turns out that

$$\langle z_k,\, x\rangle \le 0 \implies \langle \lim_{k\to+\infty} z_k,\, x\rangle = \lim_{k\to+\infty} \langle z_k,\, x\rangle \le 0 \implies \lim_{k\to+\infty} z_k = \bar{z} \in X^\circ,$$

and therefore X° is also closed. $\square$

Theorem 2.62 (Bipolarity Theorem)**.** *Let $\mathcal{C}$ be a nonempty closed convex cone. Then*

$$(\mathcal{C}^\circ)^\circ = \mathcal{C} \quad \text{and} \quad (\mathcal{C}^*)^* = \mathcal{C}.$$

Proof. The inclusion $(\mathcal{C}^\circ)^\circ \supseteq \mathcal{C}$ is always satisfied, even if $\mathcal{C}$ is a set of $\mathbb{R}^n$ with no additional properties. Therefore, it suffices to prove that

$$c \in (\mathcal{C}^\circ)^\circ \implies c \in \mathcal{C}.$$

We argue by contradiction. Suppose that there exists $\bar{c} \in (\mathcal{C}^\circ)^\circ \setminus \mathcal{C}$; the (strong) geometric form of the Hahn-Banach theorem (see Theorem 2.42) implies that the nonempty closed convex cone $\mathcal{C}$ and the compact singlet $\{\bar{c}\}$ may be separated via a hyperplane. Precisely, there exist $z \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ and $\gamma \in \mathbb{R}$ such that

$$\langle z, \bar{c}\rangle > \gamma > \langle z, c\rangle \quad \text{for every } c \in \mathcal{C}.$$

By definition, the origin $\mathbf{0}$ belongs to the cone $\mathcal{C}$, and thus γ is strictly greater than 0. On the other hand, by positive homogeneity of the cone we have that ☐

$$\exists\, c \in \mathcal{C} \; : \; \langle z, c\rangle > 0 \implies \langle z, tc\rangle = t\langle z, c\rangle \xrightarrow{t \to +\infty} +\infty,$$

and this is absurd since γ needs to be finite. In particular, γ needs to be strictly bigger than 0 and also less or equal than 0, which means that we have derived the sought contradiction.

Corollary 2.63. *Let $\mathcal{C}_1$, $\mathcal{C}_2$ be a nonempty closed convex cones. Then*

$$\mathcal{C}_1 \subseteq \mathcal{C}_2 \iff \mathcal{C}_2^\circ \subseteq \mathcal{C}_1^\circ \quad \text{and} \quad \mathcal{C}_1 \subseteq \mathcal{C}_2 \iff \mathcal{C}_2^* \subseteq \mathcal{C}_1^*.$$

We are now ready to derive some of the fundamental properties of the tangent cone, which will come in handy shortly to investigate the relationship with the normal cone.

Proposition 2.64. *Let $X \subset \mathbb{R}^n$.*

(a) *If x is an internal point of X, then $T_X(x) = \mathbb{R}^n$.*

(b) *The tangent cone is a local notion, that is,*

$$T_{X \cap B(x,\, r)}(x) = T_X(x) \quad \text{for every } r > 0.$$

Furthermore, if X is convex the following properties hold:

(c) *For any $\bar{x}$ it turns out that $X \subseteq T_X(\bar{x}) + \bar{x}$.*

(d) *The tangent cone is convex at every $\bar{x} \in X$.*

Proof. The proof of **(a)** and **(b)** is a straightforward application of the definition via sequences (since any internal point is the center of a small ball, entirely contained in X.)

(c) The proof of this assertion is left as an exercise to the reader.

(d) Let v_1, $v_2 \in T_X(\bar{x})$ and $\lambda \in (0, 1)$. By definition there are $(t_k^1)_{k\in\mathbb{N}}$ and $(t_k^2)_{k\in\mathbb{N}}$ sequences of positive real numbers decreasing to 0, and sequences of vectors $(v_1^k)_{k\in\mathbb{N}}$ and $(v_2^k)_{k\in\mathbb{N}}$ converging respectively to v_1 and v_2 such that

$$\bar{x} + t_k^i v_i^k \in X \quad \text{for every } k \in \mathbb{N} \text{ and } i = 1, 2.$$

We choose $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ sequences in such a way that

$$t_k := a_k t_k^1 + b_k t_k^2 \searrow 0 \quad \text{and} \quad \lambda v_1^k + (1-\lambda) v_2^k \xrightarrow{k\to+\infty} \lambda v_1 + (1-\lambda) v_2,$$

an we also require that

$$\bar{x} + \left(a_k t_k^1 + b_k t_k^2\right)\left(\lambda v_1^k + (1-\lambda) v_2^k\right) = \lambda(\bar{x} + t_k^1 v_1^k) + (1-\lambda)(\bar{x} + t_k^2 v_2^k).$$

The set X is convex by assumption; hence $\bar{x} + t_k(\lambda v_1^k + (1-\lambda)v_2^k) \in X$, and this is exactly what we need to infer that $T_X(\bar{x})$ is a convex cone.

□

Proposition 2.65. *Let $X \subset \mathbb{R}^n$. The normal cone is a subset of the polarization of the tangent cone (and vice versa), i.e.*

$$N_X(\bar{x}) \subseteq T_X(\bar{x})^\circ.$$

Moreover, if X is a convex set, then

$$T_X(\bar{x}) = N_X(\bar{x})^\circ \quad \textit{and} \quad T_X(\bar{x})^\circ = N_X(\bar{x}).$$

Proof. Let $z \in N_X(\bar{x})$. The inclusion is proved if we can show that

$$\langle z,\, v\rangle \leq 0 \quad \text{for every } v \in T_x(\bar{x}). \tag{2.18}$$

Let v be a given vector in the tangent cone $T_X(\bar{x})$. By definition there is a sequence of positive real numbers $t_k \searrow 0$ and a sequence of vectors v^k converging to v such that

$$\bar{x} + t_k v^k \in X \quad \text{for all } k \in \mathbb{N}.$$

Since z is a point in the normal cone, we have that

$$\langle z,\, x - \bar{x}\rangle \leq 0 \quad \text{for every } x \in X,$$

and therefore we can take $x := \bar{x} + t_k v^k$. It turns out that

$$\langle z,\, t_k v^k\rangle = t_k \langle z,\, v^k\rangle \leq 0 \quad \text{for every } k \in \mathbb{N},$$

and dividing by t_k, which is positive for k fixed, we infer that

$$\langle z,\, v^k\rangle \leq 0 \quad \text{for every } k \in \mathbb{N}.$$

Taking the limit as k approaches $+\infty$, we use the fact that $v^k \to v$ to conclude that (2.18) holds.

Convexity. Let $z \in T_X(\bar{x})^\circ$. The set X is convex, hence by Proposition 2.64 **(c)** we have the following inclusion inclusion:

$$X \subseteq \bar{x} + T_X(\bar{x}) \implies x - \bar{x} \in T_X(\bar{x}) \quad \text{for all } x \in X.$$

Therefore, the scalar product $\langle z,\, x - \bar{x}\rangle$ is less than or equal to zero, and this is enough to conclude that z belongs to $N_X(\bar{x})$. □

2.7.1 Characterization of the Subdifferential

We are finally ready to conclude the chapter on the theory of convex functions with a particular characterization of the subdifferential in terms of the normal cone and the directional derivative in terms of the tangent cone.

Theorem 2.66. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function on $\mathbb{R}^n$. For any $\bar{x} \in X$ it turns out that*

$$\partial f(\bar{x}) = \left\{ s \in \mathbb{R}^n \, : \, (s, \, -1) \in N_{\mathrm{epi}(f)} \left((\bar{x}, \, f(\bar{x})) \right) \right\}. \tag{2.19}$$

Moreover, for any direction $v \in \mathbb{R}^n$ the directional derivative is given by

$$\partial_v f(\bar{x}) = \min \left\{ r \in \mathbb{R} \, : \, (v, \, r) \in T_{\mathrm{epi}(f)} \left((\bar{x}, \, f(\bar{x})) \right) \right\}. \tag{2.20}$$

Proof.

Step 1. Recall that

$$s \in \partial f(\bar{x}) \iff f(x) \geq f(\bar{x}) + \langle s, \, x - \bar{x} \rangle \quad \text{for all } x \in X.$$

In particular, from the definition of epigraph we have that

$$s \in \partial f(\bar{x}) \iff t \geq f(\bar{x}) + \langle s, \, x - \bar{x} \rangle \quad \text{for all } (x, \, t) \in \operatorname{epi} f,$$

which, in turn, is equivalent to

$$\langle \begin{pmatrix} s \\ -1 \end{pmatrix}, \begin{pmatrix} x \\ t \end{pmatrix} - \begin{pmatrix} \bar{x} \\ f(\bar{x}) \end{pmatrix} \rangle \leq 0 \quad \text{for all } (x, \, t) \in \operatorname{epi} f,$$

and this holds if and only if $(s, \, -1) \in N_{\mathrm{epi}(f)} \left((\bar{x}, \, f(\bar{x})) \right)$.

Step 2. Let $v \in T_{\mathrm{epi}(f)} \left((\bar{x}, \, f(\bar{x})) \right)$. The Bipolarity Theorem 2.62 implies that $v \in N_{\mathrm{epi}(f)} \left((\bar{x}, \, f(\bar{x})) \right)^{\circ}$, and thus the characterization (2.21) proves that

$$\langle \begin{pmatrix} v \\ r \end{pmatrix}, \begin{pmatrix} \gamma s \\ -\gamma \end{pmatrix} \rangle \leq 0 \quad \text{for all } \gamma \geq 0 \text{ and } s \in \partial f(\bar{x}).$$

If we evaluate the scalar product, dividing by γ, we find that

$$r \geq \langle v, \, s \rangle \quad \text{for all } s \in \partial f(\bar{x}),$$

which means that

$$r \geq \max \left\{ \langle v, \, s \rangle \, : \, s \in \partial f(\bar{x}) \right\} = \partial_v f(\bar{x})$$

By compactness of $\partial f(\bar{x})$, there exists $\bar{s} \in \partial f(\bar{x})$ such that

$$\langle v, \, \bar{s} \rangle = \max \left\{ \langle v, \, s \rangle \, : \, s \in \partial f(\bar{x}) \right\},$$

and therefore $\langle v, \, \bar{s} \rangle \geq \langle v, \, s \rangle$ for all $s \in \partial f(\bar{x})$, which means that $(v, \, \langle \bar{s}, \, v \rangle) \in T_{\mathrm{epi}(f)} \left((\bar{x}, \, f(\bar{x})) \right)$. ☐

Corollary 2.67. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function on $\mathbb{R}^n$. For any $\bar{x} \in X$ it turns out that*

$$N_{\mathrm{epi}(f)} \left((\bar{x}, \, f(\bar{x})) \right) = \left\{ (\gamma s, \, -\gamma) \, : \, \gamma \geq 0, \, s \in \partial f(\bar{x}) \right\}. \tag{2.21}$$

The reader may fill in the details of the proof of this corollary, but it is, mostly, a straightforward consequence of the characterization (2.19).

2.8 Convex Conjugate

In this section, we introduce the notion of *convex conjugate* of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, and we prove that for a convex closed function f there is a nice characterization of its subdifferential in terms of the subdifferential of the conjugate.

Motivating Example. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a differentiable function, and let $s \in \mathbb{R}^n$ be a vector. Our goal is to find a vector $x \in \mathbb{R}^n$ such that

$$\nabla f(x) = s,$$

that is, we want to investigate the inverse (if any) of the gradient of f. In other words, we are looking for the stationary points of the function $f(\cdot) - \langle s, \cdot \rangle$, i.e.

$$\nabla f(x) = s \iff x \in \operatorname{argmin} \{ f(y) - \langle s, y \rangle \ : \ y \in \mathbb{R}^n \}.$$

The main goal of this section is to deal with a similar problem (regarding the subdifferential) in the setting of convex, but not necessarily differentiable, functions.

Definition 2.68 (Convex Conjugate)**.** The *convex conjugate* of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is the function defined by setting

$$f^*(x^*) := \sup \{ \langle x^*, x \rangle - f(x) \ : \ x \in \mathbb{R}^n \}.$$

Remark 2.13. The convex conjugate is given by the supremum of an arbitrary collection of affine functions

$$\langle \cdot, x \rangle - f(x),$$

and hence (see Lemma 2.13) f^* is convex, even if f is not.

The convex conjugate is, a priori, a function with values in the extended real line $\mathbb{R} \cup \{+\infty\}$ since the supremum may not be a maximum. For example, if f attains the value $-\infty$ at some point, then one can easily prove that

$$f^* \equiv +\infty.$$

Definition 2.69 (Proper)**.** The domain of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is the set of all points where f attains a finite value, that is,

$$\operatorname{dom} f = \{ x \in \mathbb{R}^n \mid f(x) < +\infty \}.$$

A function is *proper* if and only if the domain $\operatorname{dom} f$ is not empty.

Hereinafter we shall assume that f is a function with values in $\mathbb{R}$, which is enough to assure that the convex conjugate is a proper function (i.e., there exists $x \in \mathbb{R}^n$ such that $f^*(x^*) < +\infty$).

Proposition 2.70. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. Then the convex conjugate is a (convex) closed (=lower semi-continuous) and proper function.*

Proof. It suffices to show that f^* is a closed function (=the epigraph is closed). Consider a sequence $(x_k^*, r_k)_{k \in \mathbb{N}} \subset \operatorname{epi} f^*$ converging to an element $(x^*, r) \in \mathbb{R}^n \times \mathbb{R}$. The convex conjugate is a supremum, and thus by definition we obtain

$$r_k \geq f^*(x_k^*) \geq \langle x_k^*, x \rangle - f(x) \quad \text{for all } k \in \mathbb{N}.$$

If we take the limit for $k \to +\infty$, it turns out that

$$r = \lim_{k \to +\infty} r_k \geq \lim_{k \to +\infty} \langle x_k^*, x \rangle - f(x) = \langle x^*, x \rangle - f(x),$$

and the thesis follows taking the supremum of the right-hand side with respect to $x \in \mathbb{R}^n$. $\square$

The convex conjugate of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is defined in such a way that the Fenchel's inequality holds, that is,

$$f^*(x^*) + f(x) \geq \langle x^*, x\rangle \quad \text{for all } x,\, x^* \in \mathbb{R}^n. \tag{2.22}$$

For a convex proper function the points $(x, x^*) \in \mathbb{R}^n \times \mathbb{R}^n$ such that (2.22) is an equality can be easily characterized via the subdifferential of f (and vice versa).

Proposition 2.71. *Let f be a convex proper function. Then*

$$f^*(x^*) + f(x) = \langle x^*, x\rangle \iff x^* \in \partial f(x).$$

Proof. Let $x^* \in \partial f(x)$. By definition

$$f(y) \geq f(x) + \langle x^*, y - x\rangle \quad \text{for all } y \in \mathbb{R}^n,$$

which is equivalent to

$$\langle x^*, x\rangle \geq f(x) + \langle x^*, y\rangle - f(y) \quad \text{for all } y \in \mathbb{R}^n.$$

Taking the supremum over $y \in \mathbb{R}^n$ of the right-hand side, we find that

$$f^*(x^*) + f(x) \leq \langle x^*, x\rangle$$

which, combined with (2.22), gives us the sought equality. The "only-if" part of the proof is given by the same argument, read backwards. $\square$

Before we give a characterization of the subdifferential of f in terms of the subdifferential of f^* (and vice versa), we need to to introduce the notion of convex biconjugate.

Definition 2.72 (Convex Biconjugate)**.** The *convex biconjugate* of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is the function defined by setting

$$f^{**}(x^*) := \sup \left\{ \langle x^*, x\rangle - f^*(x) \;:\; x \in \mathbb{R}^n \right\}.$$

Theorem 2.73. *Let f be a proper closed convex function. Then*

$$f^{**} \equiv f.$$

Proof. We assume, for simplicity, that f is a function defined on the whole $\mathbb{R}^n$ with values in $\mathbb{R}$. A well-known fact in convex functions theory asserts that

$$f(x) = \sup \left\{ \langle s, x\rangle - b \;:\; (s, b) \in \Sigma_f \right\},$$

where

$$\Sigma_f := \left\{ (s, b) \in \mathbb{R}^n \times \mathbb{R} \mid f(y) \geq \langle s, y\rangle - b \text{ for all } y \in \mathbb{R}^n \right\}.$$

Let $(x^*, b) \in \Sigma_f$. Then $b \geq \sup_y \left\{ \langle x^*, y\rangle - f(y) \right\} = f^*(x^*)$, and, in particular, we infer that a function f is proper if and only if there is an affine function below. It follows that

$$\begin{aligned} f(x) &= \sup \left\{ \langle s, x\rangle - b \;:\; (s, b) \in \Sigma_f \right\} = \\ &= \sup \left\{ \langle x^*, x\rangle - b \;:\; b \geq f^*(x^*) \right\} = \\ &= \sup \left\{ \langle x^*, x\rangle - f^*(x^*) \;:\; x^* \in \mathbb{R}^n \right\} = f^{**}(x), \end{aligned}$$

which is exactly what we wanted to prove. $\square$

Corollary 2.74. *Let f be a proper closed convex function. Then*

$$x^* \in \partial f(x) \iff x \in \partial f^*(x^*).$$

Example 2.3. Let C be a nonempty convex subset of $\mathbb{R}^n$, and let

$$\delta_C(x) := \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{if } x \notin C, \end{cases}$$

be its indicator function. We can easily prove that δ_C is a convex proper closed function, and its convex conjugate is given by the support function, i.e.,

$$\delta_C^*(x) = \sigma_C(x).$$

Moreover, one can prove that the subdifferential of the indicator function is given by

$$\partial \delta_C(x) = N_C(x),$$

and, similarly, we have

$$\partial \sigma_C(x) = ?$$

Chapter 3

Convex Optimization Theory

In this chapter, we shall apply the theory of convex functions developed in the previous lessons to study a particular class of minimum problems with either a differentiable or a convex function defined on a subset X of $\mathbb{R}^n$.

3.1 Optimality Conditions

Let $X \subseteq \mathbb{R}^n$ be a subset, and let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a function. We are interested in solving (at least locally) the minimum problem

$$\min_{x \in X} f(x) := \min\{f(x) \,:\, x \in X\}, \tag{3.1}$$

and find the set of all the (local) minimizers $\bar{x} \in X$.

3.1.1 Admissible Directions Cone

Definition 3.1 (Admissible Direction)**.** A vector $v \in \mathbb{R}^n$ is an *admissible direction* for X at $\bar{x} \in X$ if there exists $T > 0$ such that

$$x + \tau v \in X \quad \text{for all } \tau \in [0,\, T].$$

Definition 3.2 (Cone)**.** For a given point $\bar{x} \in X$, the *admissible direction cone* is the set of all admissible directions, that is,

$$F_X(\bar{x}) = \{v \in \mathbb{R}^n \,:\, \exists T > 0 \text{ such that } x + \tau v \in X \text{ for all } \tau \in [0,\, T]\}\,.$$

Proposition 3.3. *Let X be given as above. Then:*

(a) *The admissible direction cone is a subset (eventually proper) of the tangent cone, i.e.*

$$F_X(\bar{x}) \subseteq T_X(\bar{x}).$$

(b) *If $\bar{x}$ is an internal point of X, then $F_X(\bar{x}) = \mathbb{R}^n$.*

Furthermore, if X is convex the following properties hold:

(c) *The tangent cone is the closure of the admissible direction cone, that is,*

$$T_X(\bar{x}) = \operatorname{cl} F_X(\bar{x}).$$

(d) *The admissible directions are all of the form $\gamma(x - \bar{x})$, that is,*

$$F_X(\bar{x}) = \{v \in \mathbb{R}^n \,:\, v = \gamma(x - \bar{x}) \textit{ for some } x \in X\}\,.$$

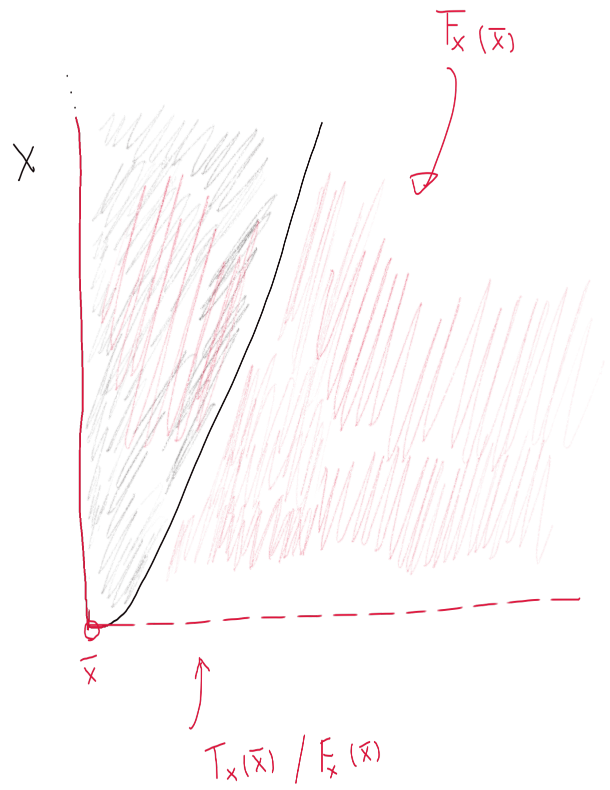

Figure 3.1: In this picture, we show the difference between the admissible directions cone and the tangent cone at a point $\bar{x}$ on the boundary of a convex set X.

3.1.2 Optimality for Differentiable Functions

In this section, we shall always assume that $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is differentiable at $\bar{x}$. First, we derive necessary conditions for a point $\bar{x}$ to be optimal and, as we will see shortly, we will also recover a fundamental result in one-variable analysis.

Theorem 3.4. *Let $\bar{x} \in X$ be a local minimum for f.*

(i) *The gradient at $\bar{x}$ has positive scalar product against every vector v in the tangent cone, that is,*

$$\langle \nabla f(\bar{x}), v \rangle \geq 0 \quad \text{for all } v \in T_X(\bar{x}). \tag{3.2}$$

(ii) *Suppose that f is twice-differentiable at $\bar{x}$. For all $v \in F_X(\bar{x})$ such that $\langle \nabla f(\bar{x}), v \rangle = 0$ it turns out that*

$$\langle \nabla^2 f(\bar{x})v, v \rangle \geq 0, \tag{3.3}$$

where $\nabla^2 f(\bar{x})$ is the Hessian of f at $\bar{x}$.

In particular, if $\bar{x}$ is an internal point of X, then the tangent cone is the whole $\mathbb{R}^n$, which means that (3.2) is equivalent to

$$\langle \nabla f(\bar{x}),\, v\rangle \geq 0 \quad \text{for all } v \in \mathbb{R}^n.$$

The unique vector with positive scalar product against **every** other vector in $\mathbb{R}^n$ is the zero vector, which means that, if $\bar{x} \in \operatorname{int} X$ local minimum, then $\nabla f(\bar{x}) = 0$.

Remark 3.1. It follows from (3.1) that $-\nabla f(\bar{x}) \in T_X(\bar{x})^\circ$ and, if X is a convex set, then the polarity theorem implies that $-\nabla f(\bar{x}) \in N_X(\bar{x})$. It follows from the definition that

$$(3.1) \iff \langle \nabla f(\bar{x}),\, x - \bar{x}\rangle \geq 0 \quad \text{for all } x \in X,$$

and this equivalent form is to be preferred in optimization theory because X is involved instead of the cone, which is usually hard to compute.

Proof. The main idea here is to apply the Taylor's formula up to the first (resp. second) order to prove the first (resp. second) assertion.

(i) Let $v \in T_X(\bar{x})$ be given. By definition there exist a decreasing sequence $t_k \searrow 0$ and a sequence of vectors v^k converging to v such that

$$\bar{x} + t_k v^k \in X \quad \text{for every } k \in K.$$

For any $\epsilon > 0$ there exists $N \in \mathbb{N}$ big enough such that

$$\bar{x} + t_k v^k \in X \cap B(\bar{x},\, \epsilon) \quad \text{for every } k \geq N,$$

and by definition of local minimum we have that

$$0 \leq f(\bar{x} + t_k v^k) - f(\bar{x}) = t_k \langle \nabla f(\bar{x}),\, v^k\rangle + o(t_k v_k) \quad \text{for every } k \geq N$$

as a consequence of the Taylor's formula up to the first order. If we divide the right-hand side by t_k and take the limit as $k \to +\infty$, we find that

$$\langle \nabla f(\bar{x}),\, v\rangle = \lim_{k\to+\infty} \langle \nabla f(\bar{x}),\, v^k\rangle \geq 0,$$

which is exactly what we wanted to prove.

(ii) Let $v \in F_X(\bar{x})$ be given. By definition there exists $\tau > 0$ such that

$$\bar{x} + tv \in X \quad \text{for every } t \in [0,\, \tau].$$

For any $\epsilon > 0$ there exists $\tau_0 \in (0,\, \tau)$ small enough such that

$$\bar{x} + tv \in X \cap B(\bar{x},\, \epsilon) \quad \text{for every } t \in [0,\, \tau_0],$$

and by definition of local minimum we have that

$$0 \leq f(\bar{x} + tv) - f(\bar{x}) = t\langle \nabla f(\bar{x}),\, v\rangle + \frac{1}{2} t^2 \langle \nabla^2 f(\bar{x}) v,\, v\rangle + o(|tv|^2)$$

for all $t \in [0,\, \tau_0]$ as a consequence of the Taylor's formula up to the second order. By assumption the first-order term $\langle \nabla f(\bar{x}),\, v\rangle$ is zero; hence we have

$$0 \leq \frac{1}{2} t^2 \langle \nabla^2 f(\bar{x}) v,\, v\rangle + o(|tv|^2).$$

We divide the right-hand side by t^2 and take the limit as t approaches 0. The remainder goes to zero, and therefore we obtain the sought inequality:

$$\langle \nabla^2 f(\bar{x}) v,\, v\rangle \geq 0.$$

$\square$

We are now interested in whether or not we can replace the admissible directions cone $F_X(\bar{x})$ with the tangent cone $T_X(\bar{x})$ and still have the assertion **(ii)** to hold.

Example 3.1. Consider the function

$$f(x_1, x_2) := -x_1^2 + x_1x_2 + x_2$$

subject to the constraint

$$X := \{(x_1, x_2) \in \mathbb{R}^2 \,:\, x_2 \geq x_1^2,\, x_1 \geq 0\}.$$

The point $(0, 0) \in X$ is a global minimum for f in X, as the reader can easily check. Moreover, we have already proved that the tangent cone is

$$T_X((0, 0)) = \{(d_1, d_2) \in \mathbb{R}^2 \,:\, d_1 \geq 0,\, d_2 \geq 0\},$$

and the admissible directions cone is

$$F_X((0, 0)) = \{(d_1, d_2) \in \mathbb{R}^2 \,:\, d_1 \geq 0,\, d_2 > 0\} \cup \{(0, 0)\}.$$

A straightforward computation shows that

$$\nabla f(0, 0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \nabla^2 f(0, 0) = \begin{pmatrix} -2 & 1 \\ 1 & 0 \end{pmatrix}.$$

We now notice that

$$d \in F_X((0, 0)) \quad \text{and} \quad \langle \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \rangle = 0 \implies d = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$d \in T_X((0, 0)) \quad \text{and} \quad \langle \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \rangle = 0 \implies d = \begin{pmatrix} d_1 \\ 0 \end{pmatrix}$$

for all $d_1 \geq 0$, which means that the necessary condition **(ii)** is satisfies with the admissible directions cone. On the other hand, if d belongs to the tangent cone we have

$$\langle \begin{pmatrix} -2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ 0 \end{pmatrix}, \begin{pmatrix} d_1 \\ 0 \end{pmatrix} \rangle = \langle \begin{pmatrix} -2d_1 \\ d_1 \end{pmatrix}, \begin{pmatrix} d_1 \\ 0 \end{pmatrix} \rangle = -2d_1^2,$$

and this is stricter less than zero for all $d_1 > 0$.

In particular, the question raised above has a negative answer. We are now ready to give a sufficient condition for a differentiable function to have a local minimum at $\bar{x} \in X$.

Theorem 3.5. *Suppose that f is differentiable at $\bar{x} \in X$, and*

$$\langle \nabla f(\bar{x}), d \rangle > 0 \quad \text{for all } d \in T_X(\bar{x}) \setminus \{\mathbf{0}\}. \tag{3.4}$$

Then $\bar{x}$ is a local minimum for f in X.

Proof. We argue by contradiction. If $\bar{x}$ is not a local minimum, we can find a sequence $(x_k)_{k\in\mathbb{N}} \subset X$ such that $x_k \to \bar{x}$ and $f(x_k) < f(\bar{x})$. We can rewrite x_k as follows:

$$x_k = \bar{x} + \|x_k - \bar{x}\| \frac{x_k - \bar{x}}{\|x_k - \bar{x}\|} =: \bar{x} + t_k d^k.$$

Note that the vectors d^k are unitary (norm equal to one), and therefore they converge (up to subsequences) to some $d \in \mathbb{R}^n$. Furthermore, $t_k \searrow 0$ because $x_k \to \bar{x}$, and hence we infer that $d \in T_X(\bar{x})$. It follows that

$$0 > f(x_k) - f(\bar{x}) = t_k\langle \nabla f(\bar{x}),\, d^k\rangle + o(|t_k d^k|),$$

and dividing by t_k and taking the limit for $k \to +\infty$ we find that

$$0 > \langle \nabla f(\bar{x}),\, d\rangle,$$

which contradicts the assumption (3.4). □

We are now interested in showing that the assertion **(ii)** of the necessary conditions is not sufficient, not even if we replace admissible directions cone $F_X(\bar{x})$ with the tangent cone $T_X(\bar{x})$ and require

$$\langle \nabla^2 f(\bar{x})v,\, v\rangle > 0 \quad \text{for all } v \in T_X(\bar{x}) \text{ such that } \langle \nabla f(\bar{x}),\, v\rangle = 0.$$

Example 3.2. Consider the function

$$f(x_1,\, x_2) := x_1 + x_2^2$$

subject to the constraint

$$X := \{(x_1,\, x_2) \in \mathbb{R}^2 \ :\ x_2^3 \geq x_1^2\}.$$

The tangent cone at (0, 0) coincides with the admissible directions cone, and they are given by

$$T_X((0,\, 0)) = F_X((0,\, 0)) = \{(d_1,\, d_2) \in \mathbb{R}^2 \ :\ d_1 = 0,\, d_2 \geq 0\}.$$

A straightforward computation shows that

$$\nabla f(0,\, 0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \nabla^2 f(0,\, 0) = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}.$$

We now notice that

$$d \in T_X((0,\, 0)) \quad \text{and} \quad \langle \begin{pmatrix} 1 \\ 0 \end{pmatrix},\, \begin{pmatrix} 0 \\ d_2 \end{pmatrix} \rangle,$$

which means every $d \in T_X((0,\, 0))$ satisfies the requirement $\langle \nabla f(\bar{x}),\, v\rangle = 0$. On the other hand, if d belongs to the tangent cone we have

$$\langle \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ d_2 \end{pmatrix},\, \begin{pmatrix} 0 \\ d_2 \end{pmatrix} \rangle = 2d_2^2,$$

and this is stricter bigger than zero for all $d_2 > 0$, i.e. all the tangent vectors different from the null vector. This condition fail to be sufficient because, as the reader can easily check, the point (0, 0) is not even a local minimum for f in X. To prove this conclusive assertion, consider the path

$$x(t) := (t,\, \sqrt[3]{t^2}) \in X,$$

and show that $f(x(t)) < 0$ for all negatives t small enough.

We can easily find a second-order sufficient condition for the free problem, that is, if we assume that $X = \mathbb{R}^n$.

Theorem 3.6. *Suppose that f is differentiable at $\bar{x} \in \mathbb{R}^n$, and suppose that*

$$\nabla f(\bar{x}) = 0 \quad \textit{and} \quad \nabla^2 f(\bar{x}) \textit{ is a positive-definite matrix.} \tag{3.5}$$

Then $\bar{x}$ is a global minimum for f in $\mathbb{R}^n$. Furthermore, there exists a positive constant $\gamma \in \mathbb{R}_{>0}$ such that

$$f(x) \geq f(\bar{x} + \gamma\|x - \bar{x}\|^2 \quad \textit{for any } x \in B(\bar{x},\, \epsilon).$$

3.1.3 Optimality for Convex Functions

In this section, we shall always assume that $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex, and subject to the convex constraint X. The necessary condition $\langle \nabla f(\bar{x}),\, v \rangle \geq 0$ for all $v \in T_X(\bar{x})$ investigated above is here replaced by

$$\partial_v f(\bar{x}) \geq 0 \quad \text{for all } v \in T_X(\bar{x}), \tag{3.6}$$

where $\partial_v f(\bar{x})$ denotes the directional derivative at $\bar{x}$ with direction v.

The characterization of the directional derivative in terms of the subdifferential (see Theorem 2.52) implies that (3.6) is equivalent to

$$\inf_{v \in T_X(\bar{x})} \max_{s \in \partial f(\bar{x})} \langle s,\, v \rangle \geq 0. \tag{3.7}$$

Suppose now that we can invert the infimum and the maximum in the inequality (3.7). It turns out that

$$\max_{s \in \partial f(\bar{x})} \inf_{v \in T_X(\bar{x})} \langle s,\, v \rangle \geq 0,$$

which means that there exists $\bar{s} \in \partial f(\bar{x})$ such that

$$\inf_{v \in T_X(\bar{x})} \langle \bar{s},\, v \rangle \geq 0 \iff \langle \bar{s},\, v \rangle \geq 0 \quad \text{for all } v \in T_X(\bar{x}).$$

The set X is convex; thus the Bipolarity Theorem implies that $-\bar{s}$ belongs to the normal cone $N_X(\bar{x})$, which means that

$$(3.6) \iff 0 \in N_X(\bar{x}) + \partial f(\bar{x}).$$

We shall not follow this approach in this section and, in particular, we will not prove that the maximum and the infimum in (3.7) can be inverted (although, it is true).

Theorem 3.7. *A point $\bar{x} \in X$ is a minimum point if and only if $0 \in N_X(\bar{x}) + \partial f(\bar{x})$.*

Proof.

"$\underline{\Longleftarrow}$" Let $s \in \partial f(\bar{x})$ be the subgradient such that $-s \in N_X(\bar{x})$. By definition of subgradient

$$f(x) \geq f(\bar{x}) + \underbrace{\langle s,\, x - \bar{x} \rangle}_{\geq 0} \implies f(\bar{x}) \leq f(x),$$

which is exactly what we wanted to prove.

"$\underline{\Longrightarrow}$" Note that

$$\bar{x} \in X \text{ minimum of } f \iff (X \times (-\infty,\, f(\bar{x}))) \cap \operatorname{epi} f = \varnothing.$$

Both are convex sets (because f and X are convex by assumption). Therefore, we can apply the usual separation result (see Theorem 3.10)to find $(s^*,\, \mu^*) \in \mathbb{R}^{n+1} \setminus \{\mathbf{0}\}$ and $\gamma \in \mathbb{R}$ such that

$$\langle s^*,\, x \rangle + \mu^* t \geq \gamma \geq \langle s^*,\, y \rangle + \mu^* r \quad \text{for all } (x,\, t) \in \operatorname{epi} f \text{ and } (y,\, r) \in X \times (-\infty,\, f(\bar{x})).$$

The inequalities are preserved if we take the left limit for $r \to f(\bar{x})$, and hence we have that

$$\langle s^*,\, x \rangle + \mu^* t \geq \gamma \geq \langle s^*,\, y \rangle + \mu^* r \quad \text{for all } (x,\, t) \in \operatorname{epi} f \text{ and } (y,\, r) \in X \times (-\infty,\, f(\bar{x})].$$

Suppose now that $\mu^* = 0$. It turns out that

$$\langle s^*,\, x \rangle = \gamma \quad \text{for all } x \in X,$$

and this is enough to conclude that also $s^* = \mathbf{0}$, which yields to a contradiction. In a similar fashion, we notice that μ^* cannot be negative, and thus the vector $s := \frac{1}{\mu^*} s^* \in \mathbb{R}^n$ is well-defined and the inequality does not change when we divide by μ^*. Namely, we have

$$\langle s,\, x\rangle + t \geq \gamma' \geq \langle s,\, y\rangle + r \quad \text{for all } (x,\, t) \in \operatorname{epi} f \text{ and } (y,\, r) \in X \times (-\infty,\, f(\bar{x})].$$

If we choose $y := \bar{x}$, then the inequality above proves that

$$f(x) \geq f(\bar{x}) + \underbrace{\langle s,\, \bar{x} - x\rangle}_{=\langle -s,\, x - \bar{x}\rangle} \quad \text{for all } x \in X,$$

which means that $-s$ is a subgradient. Similarly, choosing $x := \bar{x}$ and $t = f(\bar{x})$ we obtain

$$\langle s,\, y - \bar{x}\rangle \leq 0 \quad \text{for all } y \in X,$$

which means that s belongs to the normal cone at $\bar{x}$. □

Remark 3.2. If $X := \mathbb{R}^n$ (or $\bar{x}$ is an internal point of X), we know that the normal cone at $\bar{x}$ collapses to the null vector only, which means that

$$\bar{x} \in X \text{ minimum of } f \iff 0 \in \partial f(\bar{x}).$$

In particular, if $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a convex differentiable function, we automatically find a known result in Calculus:

$$\bar{x} \in X \text{ minimum of } f \iff \nabla f(\bar{x}) = 0.$$

3.2 Projection Problem

3.2.1 Introduction

Let $E \subseteq \mathbb{R}^n$ be a subset, and let $y \in \mathbb{R}^n$ be a point. The projection problem consists in finding a well-defined involution ($P^2 = P$) such that $P(y)$ is the unique minimum of the following problem:

$$\inf_{x \in E} \|y - x\|.$$

It is not hard to see that E needs to be convex, closed, and nonempty for the projection to be well-defined. Furthermore, the norm is always positive, and therefore it is equivalent to solve the infimum problem

$$\inf_{x \in E} \frac{1}{2}\|y - x\|^2.$$

The function $f(x) = \frac{1}{2}\|x - y\|^2$ is quadratic and its Hessian is given by $\nabla^2 f(x) = \mathrm{id}_{n \times n}$, which means that it is positive-definite. It follows that f is strongly convex (see Remark 2.10), and therefore the minimization problem

$$\inf \left\{ \frac{1}{2}\|y - x\|^2 \,:\, x \in E \right\}$$

admits one and only one solution for any given $y \in \mathbb{R}^n$. We define a projection by sending a point $y \in \mathbb{R}^n$ to the point of minimum distance $d(y,\, E)$, that is,

$$P : \mathbb{R}^n \longrightarrow E, \qquad y \longmapsto \operatorname{argmin}\left(\left\{ \frac{1}{2}\|y - x\|^2 \,:\, x \in E \right\} \right).$$

In particular, as a consequence of the necessary and sufficient criterion (see Theorem 3.7), we have the following characterization of the projection:

$$P_C(y) = \bar{x} \iff -\nabla f(\bar{x}) \in N_X(\bar{x}) \iff \langle y - \bar{x},\, x - \bar{x}\rangle \leq 0 \quad \text{for all } x \in \mathbb{R}^n,$$

3.2.2 Separation Theorems

In this section, we finally employ the construction of the projection onto convex closed sets to prove the separation results we have been using extensively in the previous chapters.

Theorem 3.8 (Strong Separation)**.** *Let $K \subset \mathbb{R}^n$ and $C \subset \mathbb{R}^n$ be two nonempty convex subsets such that $K \cap C = \varnothing$. Assume that K is compact and C is closed. Then there exist $s^* \in \mathbb{R}^n \setminus \{0\}$ and $\alpha < \beta \in \mathbb{R}$ such that*

$$\langle s^*, x\rangle \leq \alpha < \beta \leq \langle s^*, y\rangle, \quad \textit{for all } y \in K \textit{ and for all } x \in C.$$

More precisely, there exists a closed hyperplane H that strictly separates K and C.

We prove the (strong) separation theorem in a slightly less general form, i.e. we assume that K is a single point $\{y\}$ that does not belong to C.

Proof. Let $\bar{x} := P_C(y)$, where $P_C(\cdot)$ is the projection onto the convex set C. If we set $s^* := y - \bar{x}$ and $\gamma := \langle s^*, \bar{x}\rangle$, then one can easily check that

$$\langle s^*, x\rangle \leq \gamma \quad \text{for all } x \in C,$$

and

$$\langle s^*, y\rangle - \gamma = \|y - \bar{x}\|^2 > 0,$$

which is exactly what we wanted to prove. □

Theorem 3.9 (Support Hyperplane)**.** *Let $C \subset \mathbb{R}^n$ be a nonempty convex subsets, and let $\hat{y} \notin \operatorname{Int} C$. Then there exists $s^* \in \mathbb{R}^n \setminus \{0\}$ such that*

$$\langle s^*, x\rangle \leq \langle s^*, \hat{y}\rangle, \quad \textit{for all } x \in C.$$

More precisely, there exists a closed hyperplane H that passes through $\hat{y}$ such that C is entirely contained in one side.

Proof. We may assume without loss of generality that $\hat{y}$ belongs to the boundary ∂C, otherwise we can apply the previous theorem. Therefore we can take a sequence of points

$$y^k \in \partial B_{\frac{1}{k}}(\hat{y})$$

converging to $\hat{y}$, and apply the previous separation result to $\overline{C}$ and $\{y^k\}$. It turns out that there are a vector $s^k \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ and a real number $\gamma_k \in \mathbb{R}$ such that

$$\langle s^k, x\rangle < \gamma_k \leq \langle s^k, y^k\rangle, \quad \text{for all } x \in \overline{C}.$$

The sequence $(s^k/\|s^k\|)_{k\in\mathbb{N}}$ converges (up to subsequences) to a unitary vector $s^* \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, and thus we can take the limit for $k \to +\infty$ of the inequality above:

$$\langle s^*, x\rangle \leq \langle s^*, \hat{y}\rangle, \quad \text{for all } x \in \overline{C}.$$

□

Theorem 3.10 (Weak Separation)**.** *Let $C_1 \subset \mathbb{R}^n$ and $C_2 \subset \mathbb{R}^n$ be two nonempty convex subsets such that $C_1 \cap C_2 = \varnothing$. Then there exist $s^* \in \mathbb{R}^n \setminus \{0\}$ and $\gamma \in \mathbb{R}$ such that*

$$\langle s^*, x\rangle \leq \gamma \leq \langle s^*, y\rangle, \quad \textit{for all } x \in C_1 \textit{ and for all } y \in C_2.$$

More precisely, there exists a closed hyperplane H that weakly separates C_1 and C_2.

Proof. The disjointness implies that $\mathbf{0}$ does not belong to the set difference

$$C_1 - C_2 := \{x - y \mid x \in C_1,\, y \in C_2\}.$$

We apply the previous separation result between $\mathbf{0}$ and $C_1 - C_2$ to find the sought result. □

3.2.3 Motzkin Alternative Theorem

In this brief section, we show how one can use the separation theorems enlisted above to find a fundamental result in optimization theory, known as the *Motzkin alternative theorem.*

Theorem 3.11 (Motzkin)**.** *Let $A \in \mathrm{M}(n \times n)$ be a nonzero matrix, and let $B \in \mathrm{M}(p \times n)$ and $C \in \mathrm{M}(q \times n)$ be arbitrary matrices. The system*

$$\begin{cases} Ad < 0, \\ Bd \leq 0, \\ Cd = 0, \end{cases} \tag{3.8}$$

does not admit a unique solution $d \in \mathbb{R}^n$ if and only if there are $\theta \in \mathbb{R}^n_{\geq 0} \setminus \{\mathbf{0}\}$, $\lambda \in \mathbb{R}^p_{\geq 0}$, and $\mu \in \mathbb{R}^q$ such that

$$A^T\theta + B^T\lambda + C^T\mu = 0. \tag{3.9}$$

Proof. We present here a sketch of the proof. The interested reader can easily apply the mentioned separation theorem and fill in the missing details.

"$\Longleftarrow$" We argue by contradiction. Suppose that there exists a unique solution $d \in \mathbb{R}^n$ of the system (3.8), and notice that

$$\langle A^T\theta + B^T\lambda + C^T\mu,\, d\rangle = \langle\theta,\, Ad\rangle + \langle\lambda,\, Bd\rangle + \langle\mu,\, Cd\rangle =$$

$$= \langle\theta,\, \underbrace{Ad}_{<0}\rangle + \langle\lambda,\, \underbrace{Bd}_{\leq 0}\rangle < 0.$$

This is a contradiction with the assumption (3.9), and thus the thesis follows immediately.

"$\Longrightarrow$" Consider the closed convex set

$$D := \left\{ A^T\theta + B^T\lambda + C^T\mu \;\middle|\; \theta \in \mathbb{R}^n_{\geq 0} \setminus \{\mathbf{0}\},\ \lambda \in \mathbb{R}^p_{\geq 0},\ \text{and } \mu \in \mathbb{R}^q \text{ and } \sum_{i=1}^n \theta_i = 1 \right\}.$$

It suffices to prove that $\mathbf{0}$ belongs to D. We argue by contradiction, and assume that $\mathbf{0}$ does not belong to D. The strong separation theorem (see Theorem 3.8) proves the existence of a nonzero vector $d \in \mathbb{R}^n$ such that

$$\langle A^T\theta + B^T\lambda + C^T\mu,\, d\rangle > 0 \quad \text{for all } A^T\theta + B^T\lambda + C^T\mu \in D.$$

The reader may conclude the proof by showing that d is a solution of the system (3.8) by choosing θ, λ and μ in an appropriate way. $\square$

Chapter 4

Optimization Algorithms

In this chapter, we start employing the theory developed so far to produce different type of algorithms and investigate its properties of convergence.

4.1 Introduction

Let us consider the minimization problem

$$\min \{f(x) \,:\, x \in X\}. \tag{4.1}$$

The primary goal of this chapter is to produce different algorithms that allows us to find a solution to (4.1), starting from a sequence of admissible points $x^k \in X$. We shall mainly see two kinds of algorithms:

(1) Given a function $\Phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ that measures the distance between a point and a solution of (4.1), find an admissible sequence of points $(x^k)_{k\in\mathbb{N}} \subset X$ such that

$$\Phi(x^{k+1}) < \Phi(x^k) \quad \text{for every } k \in \mathbb{N}.$$

(2) Find a sequence of functions $f_k : \mathbb{R}^n \longrightarrow \mathbb{R}$ and a sequence of subsets $X_k \subset \mathbb{R}^n$ such that

$$f_k \xrightarrow{k\to+\infty} f \quad \text{and} \quad X_k \xrightarrow{k\to+\infty} X$$

with respect to a certain notion of convergence. Consider a sequence of admissible points

$$x^{k+1} \in \operatorname{argmin} \{f_k(x) \,:\, x \in X_k\},$$

and determine whether or not the limit, if any, of the sequence is a minimum point for the initial problem (4.1).

Typical results we will be able to determine by studying, from a theoretical point of view, an algorithm are the following ones:

(a) The sequence $(x^k) \subset X$ converges to a stationary point $\bar{x} \in X$ for the function f.

(b) Every cluster point of the sequence $(x^k) \subset X$ is a stationary point for the function f.

(c) There exists a cluster point of the sequence $(x^k) \subset X$ that is also a stationary point for the function f.

4.2 Gradient Method with Constant Step

In this section, we consider the minimization problem with no constraint at all (i.e., we assume $X = \mathbb{R}^n$). In this particular setting, to solve (4.1) it suffices to find all the stationary points $x \in \mathbb{R}^n$ for the function f, which means that

$$\bar{x} \in \mathbb{R}^n \;:\; \nabla f(\bar{x}) = 0 \iff \bar{x} \text{ solution of } (4.1). \tag{4.2}$$

Let $\alpha > 0$ be a positive real number, and set $G_\alpha(x) := x - \alpha \nabla f(x)$. The problem (4.2) is clearly equivalent to finding the fixed points of G_α for any $\alpha > 0$, i.e.,

$$\bar{x} \in \mathbb{R}^n \;:\; G_\alpha(\bar{x}) = \bar{x} \iff \nabla \bar{x} \in \mathbb{R}^n \;:\; f(\bar{x}) = 0. \tag{4.3}$$

More precisely, we consider the following first-order problem with constant step α defined by setting

$$\begin{cases} x^{k+1} := G_\alpha(x^k) = x^k - \alpha \nabla f(x^k), \\ x^0 \in \mathbb{R}^n. \end{cases} \tag{4.4}$$

We now briefly recall a fundamental notion in functional analysis, that of contraction, which yields to a surprisingly easy result for fixed-point problems.

Definition 4.1 (Contraction)**.** Let $(\mathscr{X}, d)$ be metric space. A mapping $T : \mathscr{X} \longrightarrow \mathscr{X}$ is a *contraction* if and only if there exists a constant $k \in [0, 1)$ such that

$$d(T(x), T(y)) \leq k d(x, y) \quad \text{for all } x, y \in \mathscr{X}.$$

Theorem 4.2 (Banach)**.** *Let $(\mathscr{X}, d)$ be a nonempty complete metric space with a contraction mapping $T : \mathscr{X} \longrightarrow \mathscr{X}$. Then T admits a unique fixed-point $\bar{x} \in \mathscr{X}$. Furthermore, $\bar{x}$ can be found as follows: Start with an arbitrary element $x^0 \in \mathscr{X}$ and obtain $\bar{x}$ as a limit of the sequence*

$$x^n := T(x^{n-1}) \quad \textit{for all } n \in \mathbb{N}.$$

It follows that, if we are able to find a value of $\alpha > 0$ for which G_α is a contraction, then we can find the unique fixed point $\bar{x} \in X$ as a limit of the sequence (4.4) for any choice of the initial point $x_0 \in \mathbb{R}^n$. We now prove that such an $\alpha > 0$ exists if f satisfies certain assumptions.

Proposition 4.3. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a quadratic function*

$$f(x) = \frac{1}{2}\langle x, Qx \rangle + \langle b, x \rangle + c,$$

and assume that Q is symmetric and positive-definite. Then there exists $\bar{\alpha} > 0$ such that $G_{\bar{\alpha}}$ is a contraction in $\mathscr{X} := \mathbb{R}^n$.

Proof. First, we notice that a simple computation yields to

$$\begin{aligned} \|G_\alpha(x) - G_\alpha(y)\|_2 &= \|(x - y) - \alpha(Qx - Qy)\|_2 = \\ &= \|(\mathrm{id} - \alpha Q)(x - y)\|_2 \leq \\ &\leq \|\mathrm{id} - \alpha Q\|_2 \|x - y\|_2, \end{aligned}$$

and therefore we only need to prove that the 2 norm of the matrix $\mathrm{id} - \alpha Q$ is less than one. On the other hand, it is a symmetric matrix and thus

$$\|\mathrm{id} - \alpha Q\|_2 = \max\{|1 - \alpha\lambda_i| \;:\; \lambda_i \text{ eigenvalue of } Q\} = \max\{|1 - \alpha\lambda_{max}|, |1 - \alpha\lambda_{min}|\}.$$

If we see the right-hand side as a function of α, say $g(\alpha)$, then we can easily find the minimum point

$$\bar{\alpha} := \frac{2}{\lambda_{min} + \lambda_{max}}.$$

It turns out that

$$g(\bar{\alpha}) = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{min} + \lambda_{max}} < 1,$$

and therefore we can finally infer that $G_{\bar{\alpha}}$ is a contraction:

$$\|G_{\bar{\alpha}}(x) - G_{\bar{\alpha}}\|_2 \leq g(\bar{\alpha})\|x - y\|_2.$$

□

The same result is true under even more general assumptions on f, but we will not give the proof here because it is rather messy and with a lot of computations.

Proposition 4.4. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a strongly convex function, and assume that the gradient is a Lipschitz function, that is, there exists $L > 0$ such that*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for every } x,\, y \in \mathbb{R}^n.$$

Then there exists $\bar{\alpha} > 0$ such that $G_{\bar{\alpha}}$ is a contraction in $\mathscr{X} := \mathbb{R}^n$.

If we give up the strong convexity of the function f, we can prove a very similar result for differentiable functions with Lipschitz gradient. First, we need a technical result that follows from the well-known Taylor's formula.

Lemma 4.5. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a differentiable function, and suppose that its gradient is Lipschitz of constant $L > 0$. Then*

$$f(x + y) \leq f(x) + \langle \nabla f(x),\, y\rangle + \frac{L}{2}\|y\|^2 \tag{4.5}$$

for every x and y in $\mathbb{R}^n$.

Proof. Let $g(t) := f(x + ty)$, and notice that

$$g'(t) = \langle \nabla f(x + ty),\, y\rangle \implies g'(0) = \langle \nabla f(x),\, y\rangle.$$

A straightforward computation concludes the proof:

$$\begin{aligned} f(x + y) - f(x) = g(1) - g(0) &= \\ &= \int_0^1 g'(t)\,\mathrm{d}t = \\ &= \int_0^1 \langle \nabla f(x + ty),\, y\rangle\,\mathrm{d}t = \\ &= \int_0^1 \langle \nabla f(x + ty) \pm \nabla f(x),\, y\rangle\,\mathrm{d}t = \\ &= \langle \nabla f(x),\, y\rangle + \int_0^1 \langle \nabla f(x + ty) - \nabla f(x),\, y\rangle\,\mathrm{d}t \leq \\ &\leq \langle \nabla f(x),\, y\rangle + \frac{L}{2}\|y\|^2. \end{aligned}$$

□

Proposition 4.6. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a differentiable function, and assume that the gradient is a Lipschitz function, that is, there exists $L > 0$ such that*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for every } x,\, y \in \mathbb{R}^n.$$

Suppose that $\alpha \in (0,\, 2/L)$. Then every cluster point of the sequence (4.4) is stationary for f.

Proof. First, we want to show that the method is decreasing, i.e.,

$$\alpha \in (0,\, \frac{2}{L}) \iff f(x^{k+1}) < f(x^k) \quad \text{for all } k \in \mathbb{N}.$$

A simple computation shows that

$$\begin{aligned} f(x^{k+1}) &= f(x^k - \alpha \nabla f(x^k)) \leq \\ &\leq f(x^k) + \langle \nabla f(x^k),\, -\alpha \nabla f(x^k) \rangle + \frac{L}{2}\|\alpha \nabla f(x^k)\|^2 = \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L}{2}\alpha^2 \|\nabla f(x^k)\|^2 = \\ &= f(x^k) + \left(\frac{L}{2}\alpha - 1\right)\alpha \|\nabla f(x^k)\|^2 < f(x^k), \end{aligned}$$

provided that $\alpha \in (0,\, 2/L)$. Let $\bar{x}$ be a cluster point of the sequence $(x^k)_{k\in\mathbb{N}}$, and let $(x^{k_\ell})_{\ell\in\mathbb{N}}$ be a subsequence converging to $\bar{x}$. Then

$$f(x^{k_{\ell+1}}) < f(x^{k_\ell + 1}) < f(x^{k_\ell}) + \left(\frac{L}{2}\alpha - 1\right)\alpha \|\nabla f(x^{k_\ell})\|^2 \quad \text{for all } \ell \in \mathbb{N},$$

and, by taking the limit for $\ell \to +\infty$, we find that

$$f(\bar{x}) \leq f(\bar{x}) + \left(\frac{L}{2}\alpha - 1\right)\alpha \|\nabla f(\bar{x})\|^2 \implies \|\nabla f(\bar{x})\|^2 \leq 0 \implies \nabla f(\bar{x}) = 0.$$

□

In particular, we can produce a straightforward algorithm that follows what we have proved so far. Notice that the algorithm has the disadvantage that one needs to compute/estimate the Lipschitz constant of the gradient, and L could be arbitrarily big (and thus α small).

Algorithm 1 Gradient Method with Constant Step

1: Select $x^0 \in \mathbb{R}^n$.
2: *loop*:
3: **if** $\nabla f(x^k) = 0$ **then stop**.
4: $x^{k+1} = x^k - \alpha \nabla f(x^k)$.
5: $k \leftarrow k + 1$.
6: **goto** *loop*.

4.3 Gradient Method with Exact/Inexact Research

Let us consider the *search function* defined by setting

$$\varphi_k(t) := f(x^k - t\nabla f(x^k)) \quad \text{for all } k \in \mathbb{N}.$$

The *exact search* consists in the following first-order problem with a nonconstant step

$$\begin{cases} x^{k+1} := x^k - t_k \nabla f(x^k), \\ x^0 \in \mathbb{R}^n, \end{cases} \tag{4.6}$$

where

$$t_k \in \operatorname{argmin} \{\varphi_k(t) \ : \ t \geq 0\}.$$

Notice that it is not restrictive to require $t \geq 0$ because

$$\varphi_k'(t) = \langle \nabla f(x^k - t\nabla f(x^k)), -\nabla f(x^k)\rangle \implies \varphi_k'(0) = -\|\nabla f(x^k)\| < 0 \implies t_k > 0.$$

The problem with this method is that the algorithm may end up going around the minimum point indefinitely, which means that it may take more time than necessary.

Algorithm 2 Gradient Method with Exact Line Research

1: Select $x^0 \in \mathbb{R}^n$ and set $k = 0$.
2: *loop*:
3: **if** $\nabla f(x^k) = 0$ **then stop**.
4: Compute $t_k \in \operatorname{argmin} \{\varphi_k(t) \ : \ t \geq 0\}$.
5: $x^{k+1} = x^k - t_k \nabla f(x^k)$.
6: $k \leftarrow k + 1$.
7: **goto** *loop*.

4.3.1 Armijo Procedure

The inexact search method is very similar to the exact search, but we do not take t_k as a solution to the minimum problem given by φ_k. First, we consider a generalization of the search function, which is given by

$$\varphi_k(t) := f(x^k + td^k) \quad \text{for all } k \in \mathbb{N}.$$

Armijo Condition. Fix $c_1 \in (0, 1)$. For every $k \in \mathbb{N}$ we choose t_k in such a way that

$$f(x^k + t_k d^k) \leq f(x^k) + c_1 t_k \langle \nabla f(x^k), d^k\rangle. \tag{4.7}$$

Notice that, if φ_k is the search function introduced above, the Armijo condition (4.7) is equivalent to choosing t_k in such a way that

$$\varphi_k(t_k) \leq \varphi_k(0) + c_1 t_k \varphi_k'(0). \tag{4.8}$$

Remark 4.1. If $d^k = -\nabla f(x^k)$, then φ_k is the search function introduced in the previous section, and (4.7) can be rewritten as follows:

$$f(x^{k+1}) \leq f(x^k) - c_1 t_k \|\nabla f(x^k)\|^2. \tag{4.9}$$

Proposition 4.7. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a differentiable function, and let $c_1 \in (0, 1)$. Then there exists $\tau_k > 0$ such that (4.7) is satisfied for every $t_k \in [0, \tau_k]$.*

Proof. Recall that d^k is chosen in such a way that the method is descent, which means that

$$\langle \nabla f(x^k), d^k\rangle \leq 0 \quad \text{for all } k \in \mathbb{N}.$$

Therefore, we simply notice that

$$\lim_{t \searrow 0} \frac{f(x^k + td^k) - f(x^k)}{t} = \langle \nabla f(x^k), d^k\rangle < c_1\langle \nabla f(x^k), d^k\rangle,$$

and this is enough to conclude the proof since for t small enough the inequality (4.7) holds. ☐

Armijo Rule. Let $\bar{t} > 0$ and $\gamma \in (0, 1)$ be given. Take $t_k = \bar{t}\gamma^n$, where $n \in \mathbb{N}$ is the smallest natural number such that t_k satisfies (4.7).

Algorithm 3 Gradient Method with Exact/Inexact Research

Select $x^0 \in \mathbb{R}^n$ and set $k = 0$.
loop:
if $\nabla f(x^k) = 0$ **then stop**.
Select $d^k \in \mathbb{R}^n$ such that $\langle \nabla f(x^k), d^k\rangle < 0$.
Compute t_k via exact search/Armijo rule.
$x^{k+1} = x^k - t_k\nabla f(x^k)$.
$k \leftarrow k + 1$.
goto *loop*.

Theorem 4.8. *Suppose that the GRI algorithm (3) generates an infinite sequence of points $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$. Suppose that f is bounded from below, i.e.*

$$f(x) \geq -M \quad \text{for every } x \in \mathbb{R}^n \tag{4.10}$$

and also that the angle between $\nabla f(x^k)$ and d^k, denoted by θ_k, satisfies the following inequality

$$\theta_k \geq \bar{\theta} + \frac{\pi}{2} \tag{4.11}$$

for a given $\bar{\theta} \in (0, \pi/2)$. Then every cluster point $\bar{x}$ of the sequence $(x_k)_{k\in\mathbb{N}}$ is a stationary point for f.

Proof. We divide the proof into two steps to ease the notation a little.

Step 1. First, we notice that the Armijo condition (4.7) implies that

$$0 \leq -c_1 t_k\langle \nabla f(x^k), d^k\rangle = -c_1\|\nabla f(x^k)\|_2\|d^k\|_2 t_k \cos\theta_k \leq f(x^k) - f(x^{k+1}),$$

which means that the sequence $(f(x^k))_{k\in\mathbb{N}}$ is monotonically decreasing. In particular, since f is also bounded from below, we infer that the sequence converges, and thus

$$f(x^k) - f(x^{k+1}) \xrightarrow{k\to+\infty} 0.$$

Furthermore, we notice that

$$\cos\theta_k \leq \cos\left(\frac{\pi}{2} + \bar{\theta}\right) = \sin\bar{\theta},$$

and therefore we have that

$$t_k\|\nabla f(x^k)\|_2\|d^k\|_2 \xrightarrow{k\to+\infty} 0.$$

Step 2. Let $\bar{x} \in \mathbb{R}^n$ be a cluster point, and denote by $(x_k)_{k\in\mathbb{N}}$ the subsequence converging to $\bar{x}$.

Step 2.1. If

$$\tau := \limsup_{k\to+\infty} t_k \|d^k\| > 0,$$

then $t_{k_\ell}\|d^{k_\ell}\| \to \tau$ for a suitable subsequence and $\|\nabla f(x^{k_\ell})\| \to 0$. Since we also have that

$$\|\nabla f(x^{k_\ell})\| \to \|\nabla f(\bar{x})\|,$$

we immediately infer that $\|\nabla f(\bar{x})\| = 0$, which means that $\nabla f(\bar{x}) = 0$.

Step 2.2. Suppose now that

$$\tau := \limsup_{k\to+\infty} t_k \|d^k\| = 0,$$

and thus $t_k\|d^k\| \to 0$. The Armijo procedure implies that

$$f(x^k + \frac{t_k}{\gamma} d^k) > f(x^k) + c_1 \frac{t_k}{\gamma} \langle \nabla f(x^k),\, d^k \rangle,$$

and therefore, if we set $\hat{d}^k := d^k/\|d^k\|$, then

$$f(x^k + \frac{t_k}{\gamma}\|d^k\|\hat{d}^k) - f(x^k) > c_1 \frac{t_k\|d^k\|}{\gamma} \langle \nabla f(x^k),\, \hat{d}^k \rangle.$$

The left-hand side, using the medium value theorem, is equal to

$$\frac{t_k\|d^k\|}{\gamma} \langle f(x^k + \gamma_k d^k),\, \hat{d}^k \rangle \quad \text{where } \gamma_k \in [0, \frac{t_k\|d^k\|}{\gamma}] \text{ and } \gamma_k \searrow 0,$$

and therefore

$$\langle \nabla f(x^k + \gamma_k d^k),\, \hat{d}^k \rangle > c_1 \langle \nabla f(x^k),\, \hat{d}^k \rangle.$$

If we take the limit for $k \to +\infty$, and we denote by $\hat{d}$ the limit of a subsequence of $\hat{d}^k$, then we infer that

$$\langle \nabla f(x^k),\, \hat{d} \rangle \geq c_1 \langle \nabla f(x^k),\, \hat{d} \rangle \implies \langle \nabla f(x^k),\, \hat{d} \rangle \geq 0.$$

Furthermore, we also have that

$$\langle \nabla f(x^k),\, \hat{d}^k \rangle < 0 \implies \langle \nabla f(x^k),\, \hat{d} \rangle \leq 0 \implies \langle \nabla f(x^k),\, \hat{d} \rangle = 0.$$

It turns out that

$$\sin\bar{\theta}\|\nabla f(x^k)\| \leq -\cos\theta_k \|\nabla f(x^k)\|\|\hat{d}^k\| = \langle \nabla f(x^k),\, \hat{d}^k \rangle \to 0,$$

from which it follows that $\|\nabla f(x^k)\| \to 0$, i.e. $\nabla f(\bar{x}) = 0$. □

Remark 4.2. The same proof works for the GRE Algorithm (3). Indeed, if we consider $\bar{x}^{k+1}$ and $\bar{t}_k$ given by the Armijo rule, then one can easily check that

$$f(x^k) - f(\bar{x}^{k+1}) < f(x^k) - f(x^{k+1}),$$

and proceed in an analogous way.

Remark 4.3. If the step-size of (4.7) is derived with a different procedure than the Armijo rule, then we can obtain a similar convergence result, provided that the step-size satisfies the following additional condition:

$$\langle \nabla f(x^k + t_k d^k),\, d^k \rangle \geq c_2 \langle \nabla f(x^k),\, d^k \rangle, \tag{4.12}$$

where $c_1 < c_2 < 1$ is called curvature. Notice that, in terms of the search function, we are asking that

$$\varphi_k'(t_k) \geq c_2 \varphi_k'(0). \tag{4.13}$$

4.4 Conjugate Gradient Methods

In this section, we introduce a new family of methods that provides an alternative to the GRI by choosing a suitable descent direction, without losing track of the directions that have already been chosen in all of the previous iterations.

4.4.1 The Linear Case

The linear conjugate gradient method was originally designed to solve the linear system

$$Ax = b \quad \text{for } A \text{ positive-definite } n \times n \text{ matrix and } b \in \mathbb{R}^n,$$

via the minimization of the strictly convex function

$$f(x) = \frac{1}{2}\langle x,\, Ax\rangle - \langle b,\, x\rangle.$$

Algorithm 4 Linear Conjugate Gradient Method

---

1. Select $x^0 \in \mathbb{R}^n$ and set $k = 0$.
2. *loop*:
3. **if** $r^k := b - Ax^k = 0$ **then stop**.
4. Set $\beta_k := \frac{\langle r^k,\, r^k\rangle}{\langle r^{k-1},\, r^{k-1}\rangle}$ for $k \geq 1$.
5. Set $d^k := r^k + \beta_k d^{k-1}$ for $k \geq 1$, otherwise $d^0 = r^0$.
6. Compute $t_k = \frac{\langle r^k,\, r^k\rangle}{\langle d^k,\, Ad^k\rangle}$.
7. $x^{k+1} = x^k + t_k d^k$.
8. $k \leftarrow k + 1$.
9. **goto** *loop*.

Since $r^k = -\nabla f(x^k) = b - Ax^k$, the first iteration is the same of the GRE line method, and next the search direction is modified in such a way that convergence can be achieved in a **finite** number of iterations.

Proposition 4.9. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a quadratic function of the form $f(x) = \frac{1}{2}\langle x,\, Ax\rangle - \langle b,\, x\rangle$, and suppose that there exists $\bar{k} \in \mathbb{N}$ such that the Algorithm (4) generates a sequence $(r^k)_{k\in\mathbb{N}}$ with $r^k \neq 0$ for all $k < \bar{k}$. Then the following identities hold:*

(**1**) *The r^ks are orthogonal, that is, $\langle r^k,\, r^j\rangle = 0$ for all $j < k$ and $k \leq \bar{k}$.*

(**2**) *The Ad^js are orthogonal to d^k, that is, $\langle d^k,\, Ad^j\rangle = 0$ for all $j < k$ and $k \leq \bar{k}$.*

(**3**) *The d^js are orthogonal to r^k, that is, $\langle r^k,\, d^j\rangle = 0$ for all $j < k$ and $k \leq \bar{k}$.*

(**4**) *For all $k \leq \bar{k}$ it turns out that $\langle d^k,\, r^0\rangle = \langle r^k,\, r^k\rangle$.*

Remark 4.4. The property (**3**) allows us to conclude that the Algorithm (4) is a descent method since we have the following chain of equalities:

$$\langle \nabla f(x^k),\, d^k\rangle = -\langle r^k,\, r^k\rangle - \beta_k \underbrace{\langle r^k,\, d^{k-1}\rangle}_{=0} = -\langle r^k,\, r^k\rangle = -\|r^k\|_2^2 < 0.$$

Remark 4.5. Notice that in the Algorithm (4) t_k is chosen in such a way that it minimizes the search function

$$\varphi_k(t) := f(x^k + td^k).$$

Indeed, one can easily prove that $t_k > 0$ and

$$\begin{aligned}\varphi_k'(t) &= \langle Ax^k + t_k Ad^k - b,\, d^k\rangle = \\ &= \langle t_k Ad^k - r^k,\, d^k\rangle = \\ &= t_k\langle d^k,\, Ad^k\rangle - \langle r^k,\, r^k + \beta_k d^{k-1}\rangle = \\ &= t_k\langle d^k,\, Ad^k\rangle - \langle r^k,\, r^k\rangle = 0.\end{aligned}$$

Remark 4.6. The property (**1**) guarantees that the Algorithm (4) stops after at most n iterations. In fact, if $r^k \neq 0$ for all $k = 0, \ldots, n-1$, then $r_0, \ldots, r^{n-1}$ are necessarily linearly independent, and thus we cannot have another nonzero linear independent vector $r^n \in \mathbb{R}^n$.

Remark 4.7. The property (**2**), on the other hand, proves that the vectors $d^0, \ldots, d^k$ are all linearly independent for all $k < n$. Indeed, if we have

$$d^k = \alpha_0 d^0 + \cdots + \alpha_{k-1} d^{k-1}$$

for some $\alpha_i \in \mathbb{R}$, then

$$\langle Ad^k,\, d^k\rangle = \alpha_0\langle Ad^k,\, d^0\rangle + \cdots + \alpha_{k-1}\langle Ad^k,\, d^{k-1}\rangle = 0,$$

and therefore $d^k = 0$ since A is positive-definite. It turns out that

$$0 = \alpha_0 d^0 + \cdots + \alpha_{k-1} d^{k-1},$$

and this is enough to infer that $\alpha_0 = \cdots = \alpha_{k-1} = 0$ since the d^i, for $i = 0, \ldots, k-1$, are linearly independent by inductive assumption.

Theorem 4.10. *Let $(x_k)_{k\in\mathbb{N}}$ be the sequence produced by the Algorithm (4). Then*

$$f(x^k) = \min\left\{f(x) \,:\, x - x^0 \in S_k\right\},$$

where S_k denotes the linear span of the vectors $d^0, \ldots, d^{k-1}$.

Proof. Let us consider the function

$$\psi_k(\alpha_0, \ldots, \alpha_{k-1}) := f(x^0 + \alpha_0 d^0 + \cdots + \alpha_{k-1} d^{k-1}),$$

in such a way that the minimization problem

$$\min\left\{f(x) \,:\, x - x^0 \in S_k\right\},$$

can be easily restated as the following unconstrained problem on the real line:

$$\min\left\{\psi_k(\alpha_0, \ldots, \alpha_{k-1}) \,:\, \alpha_0, \ldots, \alpha_{k-1} \in \mathbb{R}\right\}. \tag{4.14}$$

Furthermore, the function ψ_k is strictly convex and quadratic since f is quadratic and strictly convex. Therefore, the unique minimum point of the problem (4.14) is the unique solution $(\bar\alpha_0, \ldots, \bar\alpha_{k-1}) \in \mathbb{R}^k$ of the linear system of equations

$$\nabla\psi_k(\alpha_0, \ldots, \alpha_{k-1}) = 0.$$

In particular, we notice that since both

$$0 = \frac{\partial \psi_k}{\partial \alpha_i}(\bar{\alpha}_0, \ldots, \bar{\alpha}_{k-1}) = \langle \nabla f(x^0 + \bar{\alpha}_0 d^0 + \cdots + \bar{\alpha}_{k-1} d^{k-1}), d^i \rangle$$

and

$$\langle \nabla f(x^k), d^i \rangle = -\langle r^k, d^i \rangle = 0$$

hold for every $i < k$, the uniqueness of the solution implies

$$x^k = x^0 + \bar{\alpha}_0 d^0 + \cdots + \bar{\alpha}_{k-1} d^{k-1}),$$

which is exactly what we wanted to prove. □

In particular, since $S_1 \subset S_2 \subset \cdots \subset S_n = \mathbb{R}^n$, the convergence in a finite number of steps follows immediately from Theorem 4.10 as well.

4.4.2 The Nonlinear Case

Suppose now that $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a nonlinear function. The Algorithm (4) does not work anymore because, for example, in the 6th step we consider the solution t_k of the exact search, and this is not always possible to compute in the general case. Furthermore, the scalar product

$$\langle \nabla f(x^k), d^k \rangle = -\|\nabla f(x^k)\|^2 + \beta_k \langle \nabla f(x^k), d^{k-1} \rangle$$

is less than zero for sure if the second term is zero, but this happens if we are able to find the solution of the exact search. In particular, the Algorithm (4) for a nonlinear function may not be a descent method at all steps. Let

$$\varphi_k(t) := f(x^k + t d^k),$$

and let us require the following additional property:

$$|\varphi_k'(t_k)| \leq c_2 |\varphi_k'(0)| \quad \text{for some } c_2 < c_1 \in (0, 1). \tag{4.15}$$

If (4.15) holds and f is bounded from below, one can easily show that

$$-\frac{\|\nabla f(x^k)\|^2}{1 - c_2} \leq \langle \nabla f(x^k), d^k \rangle \leq \frac{2c_2 - 1}{1 - c_2} \|\nabla f(x^k)\|^2, \tag{4.16}$$

which means that the method descent for all $c_2 < c_1$ and $c_2 \in (0, 1/2)$. The algorithm above needs to be slightly modified as follows:

Algorithm 5 Linear Conjugate Gradient Method - Nonlinear Case

1: Select $x^0 \in \mathbb{R}^n$ and set $k = 0$.
2: *loop*:
3: **if** $r^k := b - Ax^k = 0$ **then stop**.
4: Set $\beta_k := \frac{\langle r^k, r^k \rangle}{\langle r^{k-1}, r^{k-1} \rangle}$ for $k \geq 1$.
5: Set $d^k := r^k + \beta_k d^{k-1}$ for $k \geq 1$, otherwise $d^0 = r^0$.
6: Compute $t_k > 0$ in such a way that t_k satisfies (4.15) and the Armijo rule.
7: $x^{k+1} = x^k + t_k d^k$.
8: $k \leftarrow k + 1$.
9: **goto** *loop*.

The convergence result deriving from this general algorithm is, as expected, not as good as it was in the linear case.

Theorem 4.11. *Let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence generated by the Algorithm (5). Suppose that $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is bounded from below, and suppose that $\nabla f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is Lipschitz. Then there exists a subsequence $(k_\ell)_{\ell\in\mathbb{N}}$ such that*

$$\|\nabla f(x^{k_\ell})\| \xrightarrow{\ell\to+\infty} 0.$$

Furthermore, if the sublevels of f are compact, then there exists a sub-subsequence $(k_{\ell_j})_{j\in\mathbb{N}}$ converging to a stationary point.

In order to understand why the convergence of this method is, theoretically, not satisfactory, we now present a simple non-rigorous argument. Suppose that $\theta_k \approx \pi/2$, and notice that

$$(4.16) \implies \frac{2c_2 - 1}{1 - c_2}\|\nabla f(x^k)\|^2 < 0.$$

For a suitable value of c_2, this is enough to infer that $\|\nabla f(x^k)\| \ll \|d^k\|$. Therefore, it is **possible** that the step t_k is almost zero, and it is also possible that $x^{k+1} \approx x^k$ and $\nabla f(x^{k+1}) \approx \nabla f(x^k)$. It follows that $\beta_k \approx 1$, $d^{k+1} \approx d^k$, and hence

$$\theta^{k+1} = \frac{\nabla f(x^k),\, d^k\rangle}{\|\nabla f(x^k)\|\|d^k\|} \approx \theta^k \implies \theta^{k+1} \approx \frac{\theta}{2}.$$

This is the reason why the Algorithm presented above may get stuck in some points, but, obviously, the argument is not a rigorous proof of this fact.

4.4.3 Polak-Ribiere Method

In the linear case, we have that

$$\beta_k = \frac{\langle\nabla f(x^k),\, \nabla f(x^k) - \nabla f(x^{k-1})\rangle}{\|\nabla f(x^{k-1})\|^2}.$$

In the general case, we note that, if the angle $\theta_k \approx \pi/2$ and $\nabla f(x^k) \approx \nabla f(x^{k-1})$, then $\beta_k \approx 0$. In particular, we do not risk to get stuck again, but the method may not be descent (unless there is some form of strong convexity and exact search). A possible solution is to take

$$\beta_k = \max\{0, \frac{\langle\nabla f(x^k),\, \nabla f(x^k) - \nabla f(x^{k-1})\rangle}{\|\nabla f(x^{k-1})\|^2}\},$$

and in this case one can prove descent convergence results.

4.5 Newton-Raphson Method

The Newton-Raphson (NR) method searches for the zeros of a function $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ satisfying certain regularity properties. More precisely, if we assume $F \in C^1(\mathbb{R}^n;\, \mathbb{R}^n)$ and the Jacobian $J_F(\cdot)$ invertible at all $x \in \mathbb{R}^n$, then we can define the kth iteration as follows:

$$x^k = x^{k-1} - \left[J_F(x^{k-1})\right]^{-1} F(x^{k-1}). \tag{4.17}$$

We now prove that the NR method is **locally** convergent to a stationary point, even though the radius of the ball of convergence can be very small.

Theorem 4.12. *Let $\bar{x} \in \mathbb{R}^n$ be a zero of F. Suppose that the following hold:*

(a) *The function is continuously differentiable, that is,* $F \in C^1(\mathbb{R}^n;\, \mathbb{R}^n)$.

(b) *The Jacobian* $J_F(\cdot)$ *is invertible at all* $x \in \mathbb{R}^n$.

(c) *The function* $x \longmapsto J_F(x)$ *is Lipschitz in a neighborhood of* $\bar{x}$.

Then there are positive constants δ *and* M *such that* (4.17) *is well-defined for all* $x^0 \in B(\bar{x},\, \delta)$, *and the generated sequence* $(x^k)_{k\in\mathbb{N}}$ *converges superlinearly, i.e.,*

$$\|x^{k+1} - \bar{x}\| \leq M\|x^k - \bar{x}\|^2. \tag{4.18}$$

Remark 4.8. The estimate (4.18) implies the convergence of the generated sequence $(x^k)_{k\in\mathbb{N}}$ to the zero $\bar{x}$. Indeed, it is enough to take

$$x^0 \in B(\bar{x},\, \epsilon) \quad \text{and} \quad \epsilon < \min\{\delta,\, \frac{1}{M}\}$$

to infer that

$$\|x^{k+1} - \bar{x}\| \xrightarrow{k\to+\infty} 0.$$

Proof. The Jacobian is locally Lipschitz in a neighborhood of $\bar{x}$, and therefore there are positive constants L and δ_1 such that

$$\|J_F(x) - J_F(y)\| \leq L\|x - y\| \quad \text{for all } x,\, y \in B(\bar{x},\, \delta_1).$$

Furthermore, from the assumption (**b**) and the local inverse theorem, we infer that there exists a positive constant δ_2 such that

$$F\,\big|_{B(\bar{x},\,\delta_2)} : B(\bar{x},\, \delta_2) \longrightarrow F\left(B(\bar{x},\, \delta_2)\right)$$

is invertible, with inverse of class $C^1(\mathbb{R}^n;\, \mathbb{R}^n)$, and satisfying the identity

$$J_{F^{-1}}(F(x)) = [J_F(x)]^{-1} \quad \text{for all } x \in B(\bar{x},\, \delta_2).$$

By continuity, there exists a positive constant $K > 0$ such that

$$\|[J_F(x)]^{-1}\| \leq K \quad \text{for all } x \in B(\bar{x},\, \delta_2).$$

The final premise follows from the mean value theorem for integrals, since we can always write

$$F(x^k) = F(x^k) - F(\bar{x}) = \int_0^1 J_F(x^k + t(\bar{x} - x^k))(x^k - \bar{x})\, \mathrm{d}t.$$

Step 1. Let $\delta < \min\{\delta_1,\, \delta_2,\, \frac{1}{KL}\}$. Then

$$\begin{aligned}
x^{k+1} - \bar{x} &= (x^k - \bar{x}) - \left[J_F(x^k)\right]^{-1} F(x^k) = \\
&= \left[J_F(x^k)\right]^{-1} \left[J_F(x^k)(x^k - \bar{x}) - F(x^k)\right] = \\
&= \left[J_F(x^k)\right]^{-1} \left[J_F(x^k)(x^k - \bar{x}) - \int_0^1 J_F(x^k + t(\bar{x} - x^k))(x^k - \bar{x})\, \mathrm{d}t\right] = \\
&= \left[J_F(x^k)\right]^{-1} \left[\int_0^1 \left(J_F(x^k) - J_F(x^k + t(\bar{x} - x^k))\right)(x^k - \bar{x})\, \mathrm{d}t\right].
\end{aligned}$$

Step 2. From the identity proved in the first step, it follows that

$$\|x^{k+1}-\bar{x}\| \leq \left\| \left[J_F(x^k)\right]^{-1} \right\| \left[\int_0^1 \left\| J_F(x^k) - J_F(x^k + t(\bar{x}-x^k)) \right\| \|x^k - \bar{x}\| \, \mathrm{d}t \right] \leq$$

$$\leq KL \int_0^1 \|x^k - \bar{x}\|^2 t \, \mathrm{d}t =$$

$$= \frac{KL}{2} \|x^k - \bar{x}\|^2,$$

and therefore the M of the thesis is given by $\frac{KL}{2}$. In conclusion, the NR method is well-defined because

$$x^k \in B(\bar{x}, \delta) \implies \|x^{k+1} - \bar{x}\| \leq \frac{KL}{2}\delta^2 < \delta \implies x^{k+1} \in B(\bar{x}, \delta).$$

$\square$

Remark 4.9. Suppose that $F(x) = \nabla f(x)$ for some function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. Then the Jacobian of F is the Hessian matrix of f, which means that (4.17) can be rewritten as

$$x^{k+1} = x^k - \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k). \tag{4.19}$$

This method corresponds to step-size $t_k = 1$ for all $k \in \mathbb{N}$, and direction $d^k = -\left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$. Therefore, a simple computation shows that

$$\left[\nabla^2 f(x^k)\right]^{-1} \text{ positive-definite} \implies \langle \nabla f(x^k), \, d^k \rangle < 0,$$

which means that the method NR is descent if f is, for example, strongly convex.

Alternative Derivation. There is another way to derive the same result when $F(x) = \nabla f(x)$, where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a sufficiently regular function. Using the Taylor's formula up to the order two, we obtain the identity

$$f(x+d) = f(x) + \langle \nabla f(x), \, d \rangle + \frac{1}{2} \langle d, \, \nabla^2 f(x) d \rangle + r(d).$$

If we forget about the remainder, we found a quadratic function with respect to the variable d, and we denote it by

$$n(d) := \frac{1}{2} \langle d, \, \nabla^2 f(x) d \rangle + \langle \nabla f(x), \, d \rangle + f(x).$$

The gradient of $n(d)$ is given by

$$\nabla n(d) = \nabla f(x) + \nabla^2 f(x) d,$$

and (assuming that $\nabla^2 f(x)$ is invertible) it is equal to zero if and only if

$$d = -\left[\nabla^2 f(x)\right]^{-1} \nabla f(x).$$

4.6 Ordinary Least Squares

Let A be a $n \times m$ real matrix (with $m \ll n$), and let $b \in \mathbb{R}^n$ be a vector. The idea is to find a solution $x \in \mathbb{R}^m$ of the problem

$$Ax = b$$

via the minimization of the square of the norm of the residual, that is,

$$\min\{\frac{1}{2}\|Ax - b\|_2^2 \ : \ x \in \mathbb{R}^m\}.$$

Notice that, if we set

$$f(x) := \frac{1}{2}\|Ax - b\|_2^2,$$

then f is a quadratic function with gradient $\nabla f(x) = A^T Ax - A^T b$. In particular, any stationary point of the function f is given by a solution to the normal equations

$$A^T Ax - A^T b = 0. \tag{4.20}$$

We now want to generalize the main ideas for a function of the form

$$f(x) = \frac{1}{2}\sum_{j=1}^{n} r_j^2(x),$$

where the r_j's are called residua of f. Suppose that $r_j \in C^2(\mathbb{R}^n;\ \mathbb{R})$ for all $j = 1, \ldots, n$, and set

$$R(x) := \begin{pmatrix} r_1(x) \\ \vdots \\ r_n(x) \end{pmatrix}.$$

Then a straightforward computations shows that the gradient of f is given by

$$\nabla f(x) = J_R(x)^T R(x),$$

and the Hessian by

$$\nabla^2 f(x) = J_R(x)^T J_R(x) + \sum_{j=1}^{n} r_j(x)\nabla^2 r_j(x).$$

The Gauss-Newton (GN) method is defined as follows. Suppose that, for some reason, we have

$$\nabla^2 f(x) \approx J_R(x)^T J_R(x),$$

and let us introduce the following notation

$$J_k := J_R(x^k) \quad \text{and} \quad r_k := R(x^k)$$

so that we can write

$$\nabla f(x^k) = J_k^T r_k \quad \text{and} \quad \nabla^2 f(x) \approx J_k^T J_k.$$

The $(k+1)$th iteration is thus defined by

$$x^{k+1} = x^k + t_k d_{GN}^k, \tag{4.21}$$

where d_{GN}^k is the solution of the problem

$$J_k^T J_k d_{GN}^k = -J_k^T r_k. \tag{4.22}$$

Notice that (4.22) are the normal equations associated to the linear system $J_k d = -r_k$, whose minimum problem is given by

$$\min\{\frac{1}{2}\|J_k d + r_k\|_2^2 \,:\, d \in \mathbb{R}^n\}.$$

Proposition 4.13. *The Gauss-Newton direction d_{GN}^k is a descent direction for all $k \in \mathbb{N}$, provided that x^k is not a stationary point for f.*

Proof. We use the identity (4.22), and we find that

$$\begin{aligned}\langle \nabla f(x^k),\, d_{GN}^k \rangle &= \langle J_k^T r_k,\, d_{GN}^k \rangle = \\ &= -\langle J_k^T J_k d_{GN}^k,\, d_{GN}^k \rangle = \\ &= -\langle J_k d_{GN}^k,\, J_k d_{GN}^k \rangle = -\|J_k d_{GN}^k\|^2,\end{aligned}$$

and this last term is necessarily less than or equal to zero. On the other hand, notice that

$$\|J_k d_{GN}^k\|^2 = 0 \implies J_k d_{GN}^k = 0 \overset{(4.22)}{\implies} J_k^T r_k = \nabla f(x^k) = 0,$$

and this is a contradiction with the assumptions. □

To conclude this section, we present the convergence result (without any proof) of the Gauss-Newton method applied to a function of the form $\sum_j r_j^2(x)$ satisfying certain properties.

Theorem 4.14. *Let $f(x)$ be a function of the form $\sum_{j=1}^n r_j^2(x)$, and suppose that the following properties hold:*

(a) *The function r_j belongs to $C^1(\mathbb{R}^n)$ for all $j = 1, \ldots, n$.*

(b) *The sublevel $f^{f(x^0)} := \{x \in \mathbb{R}^n \,:\, f(x) \leq f(x^0)\}$ is compact.*

Then every cluster point $\bar{x}$ of the sequence $\{x^k\}_{k\in\mathbb{N}}$ (generated by the GN method) is necessarily a stationary point for f, provided that $J_R(\bar{x})$ has maximal rank.

4.7 Convex Optimization

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. The idea is to modify (4.6) into a different formula that takes into account the fact that the gradient is replaced by the notion of subgradient. More precisely, we consider an iterative algorithm defined by setting

$$x^{k+1} := x^k - t_k s^k \tag{4.23}$$

where $s^k \in \partial f(x^k)$ is a subgradient chosen appropriately.

Remark 4.10. The vector s^k may be, a priori, not a descent direction. Therefore, when we define an algorithm, we also need to introduce a condition that can control how far s^k is from a descent direction.

Example 4.1. Let $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be the function defined by

$$f(x) := \max\{f_1(x),\, f_2(x)\},$$

where

$$f_1(x_1,\, x_2) = \frac{1}{2}\left[(x_1 - 1)^2 + (x_2 + 1)^2\right] \quad \text{and} \quad f_2(x_1,\, x_2) = f_1(-x_1,\, -x_2).$$

The function f is not differentiable at $\bar{x} = (1,\, 1)$ since one can easily show that

$$\nabla f_1(\bar{x}) = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \quad \text{and} \quad \nabla f_2(\bar{x}) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

It follows that the subdifferential of f at $\bar{x}$ is the convex hull generated by these two vectors, that is,

$$\partial f(\bar{x}) = \operatorname{conv}\left(\nabla f_1(\bar{x}),\, \nabla f_2(\bar{x})\right) = \left\{ \begin{pmatrix} 2 - 2\lambda \\ 2\lambda \end{pmatrix} \,:\, \lambda \in [0,\, 1] \right\},$$

and we set $s := \nabla f_2(\bar{x})$. If $x(t) := \bar{x} - ts$, then

$$f_1(x(t)) = 2(t^2 + 1) \quad \text{and} \quad f_2(x(t)) = f_1(x(t)) - 4t < f_1(x(t))$$

for $t > 0$, which means that

$$f(x(t)) = f_1(x(t)) \quad \text{for all } t > 0.$$

It follows that $f(x(t)) > f(\bar{x})$ for every $t > 0$, and this is equivalent to saying that $-s$ is not a descent direction.

We now want to present a result of convergence of the method (4.23), provided that some assumptions are satisfied. Denote by x^k_{best} the partial minimum, that is,

$$f(x^k_{\text{best}}) = \min\{f(x^1),\, \dots,\, f(x^k)\} =: f^k_{\text{best}},$$

and denote by f^* the minimum of the problem, that is,

$$f^* = \min\{f(x) \,:\, x \in \mathbb{R}^n\}.$$

Theorem 4.15. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function, and let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by the algorithm (4.23). Suppose that the following properties hold:*

(1) There exists $x^* \in \mathbb{R}^n$ such that $f(x^*) = f^*$.

(2) The function f is globally (or locally) Lipschitz of constant $L > 0$.

(3) There exists[1] $R > 0$ such that $\|x^1 - x^*\| \leq R$.

Then $f^k_{\text{best}} \to f^*$.

Proof. The idea is to estimate the 2-norm of the difference between a point of the sequence and the minimum point x^*. First, we notice that

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - t_k s^k - x^*\|_2^2 =$$

$$= \langle x^k - x^*,\, x^k - x^* \rangle + t_k^2 \langle s^k,\, s^k \rangle - 2t_k \langle x^k - x^*,\, s^k \rangle.$$

[1]The value of the constant R is important in the practical applications of this convergence result since it gives us the rate of convergence.

On the other hand, by definition of subgradient we have that

$$f(x^*) \geq f(x^k) - \langle s^k,\, x^k - x^* \rangle \implies f(x^k) - f(x^*) \leq \langle s^k,\, x^k - x^* \rangle,$$

and therefore

$$\|x^{k+1} - x^*\|_2^2 \leq \langle x^k - x^*,\, x^k - x^* \rangle + t_k^2 \langle s^k,\, s^k \rangle - 2t_k(f(x^k) - f(x^*)). \tag{4.24}$$

If we iterate this estimate k times, then we find that

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^1 - x^*\|^2 + \sum_{i=1}^k t_i^2 \|s^i\|^2 - 2\sum_{i=1}^k t_i(f(x^i) - f(x^*)),$$

and therefore

$$2\sum_{i=1}^k t_i(f_{\text{best}}^k - f(x^*)) \leq 2\sum_{i=1}^k t_i(f(x^i) - f(x^*)) \leq R^2 + \sum_{i=1}^k t_i^2 \|s^i\|^2,$$

which proves that

$$f_{\text{best}}^k - f^* \leq \frac{R^2 + \sum_{i=1}^k t_i^2 \|s^i\|^2}{2\sum_{i=1}^k t_i}.$$

To conclude it suffices to recall that f is L-Lipschitz, which means that $\|s^i\| \leq L$, and therefore we simply need to chose the sequence $(t_i)_{i\in\mathbb{N}}$ in such a way that it is square summable but not summable (which is always possible). □

Remark 4.11. One can prove the convergence under less restrictive assumptions, e.g., one can choose a sequence $(t_i)_{i\in\mathbb{N}}$ that is not square-summable, but satisfying reasonable properties (see Cesaro's theorem).

Rate of Convergence. The function

$$g(t_1,\, \dots,\, t_k) := \frac{R^2 + L^2 \sum_{i=1}^k t_i^2}{2\sum_{i=1}^k t_i}$$

is convex, and it is easy to prove that the minimum point must be of the form $(t,\, \dots,\, t)$. Then

$$h(t) := g(t,\, \dots,\, t) = \frac{R^2 + kL^2t^2}{2kt}$$

is a one-variable function, and therefore we can easily find that the minimum is attained at the point $\bar{t} = \frac{R\sqrt{k}}{L}$ and is given by

$$h(\bar{t}) = \frac{RL}{\sqrt{k}}.$$

In particular, it turns out that

$$f_{\text{best}}^k - f^* \leq \epsilon \iff k \geq \left(\frac{RL}{\epsilon}\right)^2,$$

which means that the method can be really slow to compute the minimum accurately, even if it does not require a lot of memory.

Ideal Case. If we also know the value of the minimum f^*, then we could, ideally, attain the equality

$$k = \left(\frac{RL}{\epsilon}\right)^2$$

by setting

$$t_i := \frac{f(x^i) - f^*}{\|s^i\|^2},$$

which is clearly the minimum point of the quadratic function (4.24).

Index