Introduction to Optimization

J. Peypouquet and D. Cortild

Last Updated: June 26, 2025

These notes describe the material of the course "Introduction to Optimization", which is part of the BSc Applied Mathematics program at the University of Groningen. They will continuously be updated. Please do not hesitate to communicate any mistakes or typos.

Contents

1 Introduction

1.1 Optimization problems and related terminology

An *optimization problem* consists in finding the minimal or maximal value that a function can have on a given set and the points at which it attains that value, whenever they exist.

More precisely, consider a set X, along with a subset $C \subset X$. A *minimization problem* consists in finding a point $x_{\min} \in C$ such that

$$f(x_{\min}) \leq f(x)$$

for all $x \in C$, where $f : X \to \mathbb{R}$ is the *objective function*, while C is the *feasible set*, whose elements are the *feasible points*. The problem is *constrained* if $C \neq X$, and *unconstrained* otherwise. We refer to the point $x_{\min}$ above as a *(global) minimizer* of f on C, or a *solution* to the minimization problem

$$\min_C(f) = \min_{x \in C} f(x) = \min\{f(x) : x \in C\}. \tag{1}$$

The set of such solutions is denoted by $\operatorname{argmin}_C(f) = \operatorname{argmin}\{f(x) : x \in C\}$. If $C = X$, we just write $\operatorname{argmin}(f)$. If a function has no minimizers, we may still be interested in finding the value

$$\inf_C(f) = \inf_{x \in C} f(x) = \inf\{f(x) : x \in C\}.$$

In any case (empty or not), the set of minimizers can be described as

$$\operatorname{argmin}_C(f) = \bigcap_{\gamma > \inf_C(f)} [f \leq \gamma],$$

where

$$[f \leq \gamma] = \{x \in X : f(x) \leq \gamma\}$$

is the γ-*sublevel set* of f.

A *maximization problem*, in turn, consists in finding a point $x_{\max} \in C$ such that $f(x_{\max}) \geq f(x)$ for every $x \in C$, which we call a *(global) maximizer* of f on C. The rest of the notation is adapted accordingly. Now, since

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

one can transform any maximization problem into a minimization one. Therefore, we shall restrict ourselves to minimization throughout this course, without any loss of generality.

1.2 Two simple examples in $\mathbb{R}$ that actually tell us a lot

In solving the two problems that follow, we shall encounter several theoretical tools that will allow us to determine the existence of solutions, and provide characterizations to compute them.

1.2.1 Cost-efficient cylindrical containers

Suppose one intends to produce a cylindrical container of minimal cost, containing a given volume. It is composed of two flat circular surfaces at the bottom and top, and the curved (but otherwise rectangular) surface forming the walls. These different components may not be equally costly, in view of the materials used, their resistance and other requirements.

In terms of the radius r and height h of the can, the cost can be expressed as

$$C(r,h) = 2a\pi r^2 + 2b\pi rh,$$

where a and b are the costs of producing one unit of the flat and curved surfaces of the can, respectively. Now, if the can is to hold a volume V, the dimensions must satisfy

$$V = \pi r^2 h.$$

We can solve for h, and substitute in the cost to obtain an expression depending only on r:

$$\mathcal{C}(r) = 2a\pi r^2 + \frac{2bV}{r}.$$

The function $\mathcal{C} : (0,\infty) \to (0,\infty)$ is twice (actually infinitely) differentiable, with

$$\mathcal{C}'(r) = 4a\pi r - \frac{2bV}{r^2} \qquad \text{and} \qquad \mathcal{C}''(r) = 4a\pi + \frac{4bV}{r^2} > 0.$$

Observe that $\mathcal{C}'(\bar{r}) = 0$ if, and only if, $\bar{r}^3 = \frac{bV}{2a\pi}$, which is equivalent to $\bar{r} = \sqrt[3]{\frac{bV}{2a\pi}}$. Since $\mathcal{C}''$ is always strictly positive, $\mathcal{C}'$ is strictly increasing. Therefore, $\mathcal{C}'$ is negative on $(0,\bar{r})$ and positive on $(\bar{r},\infty)$, which means that $\mathcal{C}$ decreases strictly on $(0,\bar{r})$ and increases strictly on $(\bar{r},\infty)$. We conclude that $\mathcal{C}(\bar{r}) < \mathcal{C}(r)$ whenever $0 < r \neq \bar{r}$. To obtain the value of the corresponding $\bar{h}$, let us multiply the formula for the volume by $\bar{r}$, to get

$$\bar{r}V = \pi \bar{r}^3 \bar{h} = \frac{\pi b V \bar{h}}{2a\pi} = \frac{bV\bar{h}}{2a},$$

from which we deduce that

$$\bar{h} = 2\bar{r} \cdot \frac{a}{b}.$$

In other words, the ratio between the height and the diameter is the same as that between the unit costs of the materials involved. For example, if the cost of the flat surfaces is higher, the can will be taller. The arguments above can be easily generalized, in order to establish the following:

Proposition 1.1. *Let I be an open interval, and let $f : I \to \mathbb{R}$ be twice differentiable, with $f''(x) > 0$ for all $x \in I$. If there is $\bar{x} \in I$ such that $f'(\bar{x}) = 0$, then $f(\bar{x}) < f(x)$ for every $x \in I \setminus \{\bar{x}\}$.*

Actually, the converse is also true, under more general conditions.

Proposition 1.2. *Let I be an open interval, and let $f : I \to \mathbb{R}$ be differentiable. If there is $\bar{x} \in I$ such that $f(\bar{x}) \leq f(x)$ for every $x \in I$, then $f'(\bar{x}) = 0$.*

Proof. From the hypotheses, $f(\bar{x}) \leq f(\bar{x} \pm h)$ for every sufficiently small $h > 0$ (small enough so that $\bar{x} \pm h \in I$). For such h, we have

$$\frac{f(\bar{x}-h) - f(\bar{x})}{-h} \leq 0 \leq \frac{f(\bar{x}+h) - f(\bar{x})}{h}.$$

Since f is differentiable, we may let $h \downarrow 0$ to deduce that $f'(\bar{x}) \leq 0 \leq f'(\bar{x})$, which proves the statement. □

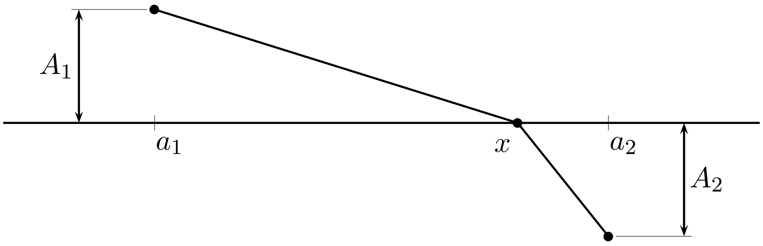
Exercise 1.3. What happens if the interval is not open?

Exercise 1.4. Among all the cones that can be inscribed inside the unit sphere, find the one with the largest volume.

1.2.2 A *broken* fastest path

An object is to be transported from one point (a_1, A_1) in the plane to another point (a_2, A_2) along the fastest route possible. If the object travels at a constant speed, the answer is to follow a straight path. Now, let $A_1 A_2 < 0$, so the object must cross the horizontal axis, and suppose that it travels at speed c_1 above the said axis, and at speed c_2 underneath it. Will it follow a straight line?

In order to answer this question, let x be the point in the horizontal axis where the object crosses.



The time it takes the object to travel from (a_1, A_1) to (a_2, A_2), as a function of x, is

$$T(x) = \frac{\sqrt{A_1^2 + (x - a_1)^2}}{c_1} + \frac{\sqrt{A_2^2 + (x - a_2)^2}}{c_2}.$$

The function T is twice differentiable on all of $\mathbb{R}$, and we have

$$\begin{aligned} T'(x) &= \frac{x - a_1}{c_1 \sqrt{A_1^2 + (x - a_1)^2}} + \frac{x - a_2}{c_2 \sqrt{A_2^2 + (x - a_2)^2}} \\ T''(x) &= \frac{A_1^2}{c_1 \left(A_1^2 + (x - a_1)^2\right)^{3/2}} + \frac{A_2^2}{c_1 \left(A_2^2 + (x - a_2)^2\right)^{3/2}}. \end{aligned}$$

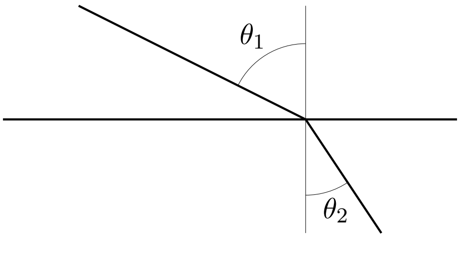
As before, $T''(x) > 0$ for all $x \in \mathbb{R}$. The unique zero $\bar{x}$ of T' lies in (a_1, a_2) and satisfies

$$\frac{|\bar{x} - a_1|}{c_1 \sqrt{A_1^2 + (\bar{x} - a_1)^2}} = \frac{|\bar{x} - a_2|}{c_2 \sqrt{A_2^2 + (\bar{x} - a_2)^2}}.$$

In other words,

$$\frac{\sin(\theta_1)}{c_1} = \frac{\sin(\theta_2)}{c_2},$$

where θ_1 and θ_2 are known as the *incidence* and *refraction* angles, respectively.



The solution *is* a straight line either if $c_1 = c_2$ or if $a_1 = a_2$ (in which case $\theta_1 = \theta_2 = 0$).

Fermat's *Principle of Least Time* states that the path taken by a ray of light between two given points is the path that can be traveled in the shortest time. As a consequence, we obtain the following:

Proposition 1.5 (Law of Refraction)**.** *For a given pair of media, the ratio of the sines of the angles of incidence and refraction equals the ratio of the speed of light in the corresponding media.*

1.3 Examples from data analysis

Optimization problems arise in a number of disciplines, frequently motivated by data. In this section, we briefly mention a few examples.

Consider a *data set*

$$\mathcal{D} = \{(a_j, y_j) \in X \times Y : j = 1, 2, \dots, J\},$$

where the a_j's are certain *features* (within set X) and the y_j's are *observations* or *labels* (in set Y), corresponding to a *sample* or *individual* $j \in \{1, \dots, J\}$. The aim is to *discover* or *learn* a function $\phi : X \to Y$ such that

$$\phi(a_j) \simeq y_j$$

for, let us say, many j's. Typically, the ϕ will belong to a *family* of functions, each of which is characterized by some *parameter* $\theta \in \Theta \subset \mathbb{R}^N$. The function ϕ can be used to make predictions about future samples. In fact, given an unseen data feature a, we predict the label y of a to be $y = \phi(a)$.

Example 1.6. When Y is a finite set, we speak of *classification problems.* For instance, the vector a_j could contain all the pixels of an image of a hand-written digit, and y_j could be its corresponding value, between 0 and 9. The *binary classification problem*, which corresponds to the case $Y = \{-1, 1\}$ (or any two-element set, for that matter), is particularly common. For example, based on the results of a medical examination, one would like to predict whether or not a patient has a given disease.

Among the many strategies to tackle these problems, we briefly comment on two, namely the *loss function* approach, and the *support vector machines.*

1.3.1 The loss function approach

A common approach is to define a *loss function*

$$\mathcal{L}_{\mathcal{D}}(\theta) = \frac{1}{J} \sum_{j=1}^{J} \ell(a_j, y_j, \theta),$$

in such a way that the parameters *better adapted to the purpose* that the function ϕ is intended to serve are the ones that give the lowest values of $\mathcal{L}_{\mathcal{D}}$.
One of the most typical problems in data analysis is the linear least squares problem. Here we solve

$$\min_{\theta \in \mathbb{R}^N} \frac{1}{2J} \sum_{j=1}^{J} (\theta^T a_j - y_j)^2 = \min_{\theta \in \mathbb{R}^N} \frac{1}{2J} \|A\theta - y\|^2,$$

where A is the matrix whose rows are a_j^T and $y = (y_1, \dots, y_J)^T$. According to the notation above, we seek $\phi : \mathbb{R}^N \to \mathbb{R}$ linear, namely $y = \phi(a) = \phi_\theta(a) = \theta^T a$, for which we use a quadratic loss function, whose terms are of the form $\ell(a, y, \theta) = (\theta^T a - y)^2$.

If additionally to approximating the data (a_j, y_j) with a linear structure, we want to impose a structure on θ, we can add a *regularization* term to the problem. For instance, *Ridge Regression* adds the term $\lambda\|\theta\|^2$, where λ is a *regularization parameter*, such that the problem is

$$\min_{\theta \in \mathbb{R}^N} \frac{1}{2J}\|A\theta - y\|^2 + \lambda\|\theta\|^2.$$

This form of regression yields a solution that linearly approximates the data, and which is also less sensitive to perturbations. The parameter λ offers a balance between the fidelity to the data and the sensitivity to perturbations. Another type of regularization consists in adding the term $\lambda\|\theta\|_1$, which yields the *LASSO formulation*:

$$\min_{\theta \in \mathbb{R}^N} \frac{1}{2J}\|A\theta - y\|^2 + \lambda\|\theta\|_1.$$

This formulation yields a linear approximation to the data, also taking into account that one wants a sparse vector θ, namely that θ contains as many 0 elements as possible. If many elements are zero, it makes the output vector θ more understandable, as the features the output depends on are reduced. It is however worth noting that the LASSO formulation is not differentiable, which will cause problems later on.

1.3.2 Support vector machines

A *Support Vector Machine* (SVM) is an optimization approach to solving binary classification problems, where the labels y_j take only the values -1 and 1. One then seeks a vector $w \in \mathbb{R}^N$ and a scalar $\beta \in \mathbb{R}$ satisfying

$$\begin{cases} w^T a_j - \beta \le -1 & \text{if } y_j = -1 \\ w^T a_j - \beta \ge 1 & \text{if } y_j = 1. \end{cases}$$

In other words, we seek a *hyperplane* parametrized by w and β that separates the sets $\{a_j \colon y_j = -1\}$ and $\{a_j \colon y_j = 1\}$. We can formulate the optimization problem as

$$\min_{w,\beta} \left\{ \sum_{j=1}^{J} \max\left\{1 - y_j(w^T a_j - \beta), 0\right\} \ : \ (w, \beta) \in \mathbb{R}^N \times \mathbb{R} \right\}.$$

If, for a given (w, β), the separating condition is satisfied for all points, the objective function is 0. Otherwise, it is strictly positive. If a separating hyperplane does not exist, the minimum value will be strictly positive, and the problem will find the best fit amongst all possible hyperplanes, knowing no perfect separating hyperplane exists.

If such a hyperplane exists, we may additionally want to maximize the distance to the closest point on each side to obtain a more *robust* classification. This can be modelled as the pair (w, β) that minimizes $\|w\|^2$, amongst all such pairs that satisfies the above separating condition. We model the *degree of robustness* by the parameter λ, and model the problem as

$$\min_{w,\beta} \left\{ \sum_{j=1}^{J} \max\left\{1 - y_j(w^T a_j - \beta), 0\right\} + \lambda\|w\|^2 \ : \ (w, \beta) \in \mathbb{R}^N \times \mathbb{R} \right\}.$$

In either case (classical or robust SVM), the objective function is non differentiable, adding an additional layer of complexity.

The problems arising from the loss function approach and the support vector machines have several points in common:

- They can be formulated as minimizing functions of *one or more real variables* (sometimes in a huge number).
- The functions involved are continuous. They are either smooth, or they have *simple* or *structured* forms of nonsmoothness.
- The function to be minimized is often a sum of simple functions that depend either on few data points or involve few variables.
- Usually, the nonsmooth parts can be separated from the rest (additive structure).

2 Basic optimization theory

This chapter contains an introduction to the theoretical foundations of continuous optimization. Section 2.1 contains a summary of the elementary metric and topological properties of finite-dimensional Euclidean spaces—especially open, closed and compact sets. Most readers will already have a thorough understanding of the topics discussed here. Nevertheless, we recommend a careful reading in order to get acquainted with notation and terminology. These topics are treated in greater depth in the course on *Metric and Topological Spaces*. Continuity and its relationship with the solvability of optimization problems will be presented in Section 2.2. In Section 2.3, we recall the geometric properties of the dot product, in connection with the angles between vectors. These ideas are exploited in Section 2.5, where we discuss the notion of differentiability, and apply it to derive applicable recipes to identify solutions to optimization problems. Further commentary on differentiability is provided in Section 2.6, also building tools to be used in the analysis of numerical methods. Finally, we present second order optimality conditions in Section 2.7, based on the notion of Hessian.

2.1 Metric and topological structure of $\mathbb{R}^N$

Throughout this course, $\mathbb{R}^N$ is the N-dimensional (real) vector space of N-tuples of real numbers, which we usually write as *column* vectors. In other words,

$$x \in \mathbb{R}^N \qquad \Longleftrightarrow \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}, \quad x_1, \dots, x_N \in \mathbb{R}.$$

Remark 2.1. From a geometric perspective, we can think of a vector $x \in \mathbb{R}^N$ as a *point* in the N-dimensional space, which immediately gives us a notion of *distance*. They can also be interpreted as proper vectors in the physical sense: directed segments (arrows) that can be placed anywhere in the space, which naturally gives rise to the concept of *angle*. Finally, we can view them as $N \times 1$ matrices, which is convenient in order to carry out some algebraic operations (see Remark 2.24 below).

The *(Euclidean) norm* of $x \in \mathbb{R}^N$ is $\|x\| = \sqrt{x_1^2 + \cdots + x_N^2}$. The following properties can be easily verified, and are left as an exercise to the reader:

i) $\|0\| = 0$ and $\|x\| > 0$ for all $x \neq 0$.

ii) $\|\alpha x\| = |\alpha|\,\|x\|$ for all $x \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$.

iii) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^N$. This is known as the *triangle inequality*.

Remark 2.2. Any function satisfying the three properties above is called a *norm* (see examples in the exercise below). In this course, however, the notation $\|\cdot\|$ will always denote the *Euclidean* norm.

Exercise 2.3. The *1-norm* and the ∞*-norm* are the functions $\|\ \|_1 : \mathbb{R}^N \to \mathbb{R}$ and $\|\ \|_\infty : \mathbb{R}^N \to \mathbb{R}$, respectively defined by

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_N| \qquad \text{and} \qquad \|x\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_N|\},$$

for each $x \in \mathbb{R}^N$. Show that they both satisfy conditions i), ii) and iii) above, and that

$$\|x\|_\infty \leq \|x\| \leq \|x\|_1 \leq \sqrt{N}\|x\| \leq N\|x\|_\infty$$

for every $x \in \mathbb{R}^N$. Are all these inequalities sharp?

The *distance* between $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^N$ is $\operatorname{dist}(x,y) = \|x-y\|$.

Exercise 2.4. Show that $\operatorname{dist} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is a *metric*, which means that

i) $\operatorname{dist}(x,x) = 0$ for all $x \in \mathbb{R}^N$ and $\operatorname{dist}(x,y) > 0$ for all $y \neq x$.

ii) $\operatorname{dist}(x,y) = \operatorname{dist}(y,x)$ for every $x, y \in \mathbb{R}^N$.

iii) $\operatorname{dist}(x,y) \leq \operatorname{dist}(x,z) + \operatorname{dist}(z,y)$ for every $x, y, z \in \mathbb{R}^N$.

The *open ball* centered at $x \in \mathbb{R}^N$ with radius $r > 0$ is

$$B(x,r) = \{y \in \mathbb{R}^N \ : \ \|x-y\| < r\}.$$

The *closed ball* centered at $x \in \mathbb{R}^N$ with radius $r > 0$ is

$$\bar{B}(x,r) = \{y \in \mathbb{R}^N \ : \ \|x-y\| \leq r\}.$$

A subset $S \subset \mathbb{R}^N$ is *open* if, for each $x \in S$, there is $r > 0$ such that $B(x,r) \subset S$.

Example 2.5. Open balls are open sets. Indeed, let $B = B(x_0, r_0)$ and let $x \in B$. We shall prove that $B(x,r) \subset B$, where $r = r_0 - \|x - x_0\|$. First, since $x \in B$, $\|x - x_0\| < r_0$, whence $r > 0$. Now, we take any $y \in B(x,r)$. By the triangle inequality, we have

$$\|y - x_0\| \leq \|y - x\| + \|x - x_0\| < r + \|x - x_0\| = r_0.$$

This means that $y \in B$. We conclude that $B(x,r) \subset B$.

Exercise 2.6. Show that the intersection of a finite number of open sets is open, while any union (finite or infinite) of open sets is open.

The *interior* of a set $S \subset \mathbb{R}^N$, denoted by $\operatorname{int}(S)$, is the largest open set contained in S. Equivalently, it is the union of all open subsets of S. The elements of $\operatorname{int}(S)$ are called *interior points* of S. Observe that $x \in \operatorname{int}(S)$ if, and only if, there is $r > 0$ such that $B(x,r) \subset S$. In particular, S is open if, and only if, $\operatorname{int}(S) = S$.

A subset $S \subset \mathbb{R}^N$ is *closed* if its complement is open.

Exercise 2.7. Show that closed balls are closed sets.

A sequence (x_k) in $\mathbb{R}^N$ *converges* to $\bar{x} \in \mathbb{R}^N$ (as $k \to \infty$) if, for every $\varepsilon > 0$, there is $k_0 \in \mathbb{N}$ such that $x_k \in B(\bar{x}; \varepsilon)$ (or, equivalently, $\|x_k - \bar{x}\| < \varepsilon$) for every $k \geq k_0$. In this case, $\bar{x}$ is the *limit* of (x_k), and we will indistinctly write either $\lim\limits_{k \to \infty} x_k = \bar{x}$, or $x_k \to \bar{x}$ (as $k \to \infty$). The notion of closedness can then be expressed in terms of convergent sequences as follows:

Proposition 2.8. *A set $S \subset \mathbb{R}^N$ is closed if, and only if, every convergent sequence of points in S has its limit in S.*

The *closure* of a set $S \subset \mathbb{R}^N$, denoted by $\overline{S}$, or sometimes as $\operatorname{cl}(S)$, is the smallest closed set that contains S. The set $\overline{S}$ can also be characterized as the intersection of all closed sets containing S, or as the set of the limits of all convergent sequences of elements of S. A set $S \subset \mathbb{R}^N$ is closed if, and only if, $\overline{S} = S$.

A set $S \subset \mathbb{R}^N$ is *bounded* if it is contained in a ball.

Proposition 2.9 (Bolzano-Weierstrass)**.** *Every bounded sequence in $\mathbb{R}^N$ has a convergent subsequence.*

Finally, a set $C \subset \mathbb{R}^N$ is *compact* if it is both closed and bounded. A consequence of Propositions 2.8 and 2.9 is the following:

Corollary 2.10. *Every sequence in a compact set has a subsequence that converges to a point in the set.*

2.2 Continuous functions and existence of minimizers

Let $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$, let $\bar{x} \in \overline{D}$, and let $\ell \in \mathbb{R}$. If, for every $\varepsilon > 0$ there is $\delta > 0$ such that $|f(x) - \ell| < \varepsilon$ whenever $\|x - \bar{x}\| < \delta$, we say that $f(x)$ *converges* to ℓ as x tends to $\bar{x}$, and that ℓ is the *limit* of $f(x)$ as x tends to $\bar{x}$. In this case, we write $\ell = \lim\limits_{x \to \bar{x}} f(x)$.

A function $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ is *continuous at* a point $\bar{x} \in D$ if $\lim\limits_{x \to \bar{x}} f(x) = f(\bar{x})$. If f is continuous at every point of D, it is *continuous in* D, or just *continuous*.

Exercise 2.11. Let $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ and $\bar{x} \in D$. Prove that f is continuous at $\bar{x}$ if, and only if, for every sequence (x_k) converging to $\bar{x}$, we have $\lim\limits_{k \to \infty} f(x_k) = f(\bar{x})$.

Theorem 2.12 (Weierstrass)**.** *If $C \subset \mathbb{R}^N$ is compact and $f : C \to \mathbb{R}$ is continuous, there exist $x_{\min}, x_{\max} \in C$ such that $f(x_{\min}) \leq f(x) \leq f(x_{\max})$ for all $x \in C$.*

Proof. Let (x_k) be a sequence in C such that $\lim\limits_{k \to \infty} f(x_k) = \inf(f)$. By Corollary 2.10, there is a subsequence (x_{n_k}) that converges to a point in C, which we shall denote by $x_{\min}$. Since f is continuous, we have

$$f(x_{\min}) = \lim_{k \to \infty} f(x_{n_k}) = \lim_{k \to \infty} f(x_k) = \inf(f),$$

which implies $f(x_{\min}) \leq f(x)$ for every $x \in C$. The argument for $x_{\max}$ is similar. □

A function $f : \mathbb{R}^N \to \mathbb{R}$ is *coercive* if for every $R > 0$, there is $\rho > 0$ such that $f(x) > R$ for every $x \notin B(0, \rho)$. In this case, we shall often write $\lim\limits_{\|x\| \to \infty} f(x) = \infty$.

Example 2.13. Let A be a matrix of size $M \times N$ with $\ker(A) = \{0\}$, and define $f : \mathbb{R}^N \to \mathbb{R}$ by $f(x) = \|Ax\|$. We have

$$\|Ax\|^2 = (Ax)^T(Ax) = x^T(A^TAx).$$

The symmetric matrix A^TA is positive definite since $\ker(A^TA) = \ker(A) = \{0\}$. Therefore, $\|Ax\|^2 \geq \lambda\|x\|^2$, where $\lambda > 0$ is the smallest eigenvalue of A^TA. It follows that $f(x) \geq \sqrt{\lambda}\|x\|$, and so f is coercive.

Recall the γ*-sublevel set* of a function $f\colon \mathbb{R}^N \to \mathbb{R}$ is defined as $[f \leq \gamma] = \{x \in \mathbb{R}^N\colon f(x) \leq \gamma\}$. Given $\gamma \in \mathbb{R}$, the γ*-level set* of $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ is

$$[f = \gamma] = \{x \in D : f(x) = \gamma\}.$$

Exercise 2.14. Show that the level and sublevel sets of a continuous function are closed.

Proposition 2.15. *The sublevel sets of a coercive function must be bounded.*

Proof. Let $\gamma > \inf(f)$. Since f is coercive, there is $r > 0$ such that $f(x) > \gamma$ if $x \notin B(0, r)$. In other words, $[f \leq \gamma] \subset B(0, r)$. □

Exercise 2.16. Show that $f : \mathbb{R}^N \to \mathbb{R}$ is coercive if, and only if, for every sequence (x_k) in $\mathbb{R}^N$ such that $\lim\limits_{k\to\infty} \|x_k\| = \infty$, we have $\lim\limits_{k\to\infty} f(x_k) = \infty$.

Proposition 2.17. *If $f : \mathbb{R}^N \to \mathbb{R}$ is continuous and coercive, then* $\operatorname{argmin}(f) \neq \emptyset$*.*

Proof. If $R > \inf(f)$, then

$$\min\left\{f(x) : x \in \mathbb{R}^N\right\} = \min\left\{f(x) : f(x) \leq R\right\} = \min_{[f\leq R]}(f).$$

The set $[f \leq R]$ is closed by Exercise 2.14, since f is continuous, and bounded, in view of Proposition 2.15. The conclusion follows from Theorem 2.12. □

Example 2.18. Considering the robust SVM from Section 1.3.2, we note that the objective function is lower-bounded by $\lambda\|w\|^2$, which is coercive, and hence the objective function itself is coervice. Given the continuity of the objective function, it follows that the robust SVM has a nonempty set of solutions.

Example 2.19. Consider the function $f : \mathbb{R}^N \to \mathbb{R}$, defined by

$$f(x) = \frac{1}{2}\|Ax - b\|^2 + \rho\|x\|_1 = \frac{1}{2}\|Ax - b\|^2 + \rho\sum_{i=1}^{N}|x_i|,$$

where A is a real matrix of size $M \times N$, $b \in \mathbb{R}^M$ and $\rho > 0$. Minimizing this function has relevant applications in statistics, particularly in compressed sensing. Observe that f is coercive since $f(x) \geq \rho\|x\|_1 \geq \rho\|x\|$ by Exercise 2.3. Since f is continuous and coercive, $\operatorname{argmin}(f) \neq \emptyset$.

The application of Proposition 2.17 is not always straightforward, as shown in the following:

Example 2.20. If we set $\rho = 0$ in the function of Example 2.19, we obtain

$$f(x) = \frac{1}{2}\|Ax - b\|^2.$$

If $\ker(A) = \{0\}$, f is coercive (see Example 2.13) and Proposition 2.17 can be directly applied. Otherwise, f *is not* coercive. In that case, we decompose $\mathbb{R}^N = \ker(A) \oplus \left[\ker(A)\right]^\perp$ so that each $x \in \mathbb{R}^N$ can be written in a unique way as $x = x_\parallel + x_\perp$, where $x_\parallel \in \ker(A)$ and $x_\perp \in \left[\ker(A)\right]^\perp$. Clearly, $f(x) = f(x_\perp)$. Define $Y = \left[\ker(A)\right]^\perp$, $\tilde{A} = A|_Y$ and $\tilde{f} = f|_Y$. Then, $\ker(\tilde{A}) = 0$, $\tilde{f}$ is continuous and coercive, and $\operatorname{argmin}(\tilde{f}) \neq \emptyset$. We conclude that $\operatorname{argmin}(f) = \operatorname{argmin}(\tilde{f}) + \ker(A)$.

Finally, a function $f \colon D \subset \mathbb{R}^N \to \mathbb{R}$ is *lower-semicontinuous at* $x \in D$ if, for every $\varepsilon > 0$, there is $\delta > 0$ such that $f(y) \geq f(x) - \varepsilon$ for every $y \in B(x, \delta) \cap D$. If f is lower-semicontinuous at every point of D, it is *lower-semicontinuous in* D, or just *lower-semicontinuous*. Every continuous function is lower-semicontinuous.

Lower-semicontinuity can be characterized in terms of convergent sequences, as follows:

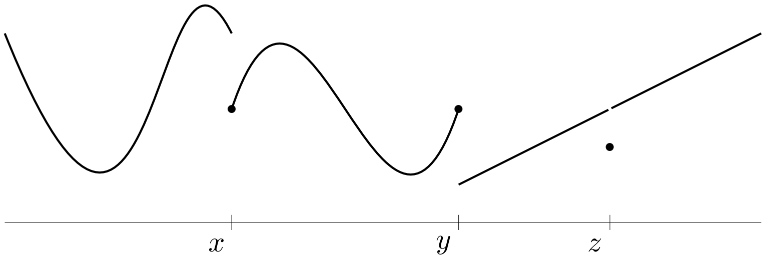
Proposition 2.21. *A function $f \colon D \subset \mathbb{R}^N \to \mathbb{R}$ is lower-semicontinuous at $x \in D$ if, and only if,*

$$f(x) \leq \liminf_{k\to\infty} f(x_k) \tag{2}$$

for every sequence (x_k) in D that converges to x.

Proof. Suppose f is lower-semicontinuous, and let $x_k \to x$. For every $\varepsilon > 0$, there is $\delta > 0$ such that $f(y) \geq f(x) - \varepsilon$ for every $y \in B(x, \delta) \cap D$. Since $x_k \to x$, there is $k_0 \in \mathbb{N}$ such that $x_k \in B(x, \delta)$ for every $k \geq k_0$. As a consequence, $f(x_k) \geq f(x) - \varepsilon$ for every $k \geq k_0$, and so $\liminf_{k\to\infty} f(x_k) \geq f(x) - \varepsilon$. Since ε was arbitrary, this implies (2). Conversely, assume f is not lower-semicontinuous. We shall find a sequence (x_k) in D such that $x_k \to x$ and $f(x) > \liminf_{k\to\infty} f(x_k)$. Indeed, if f is not lower-semicontinuous, there is $\varepsilon_0 > 0$ such that for every $\delta > 0$ there is $y_\delta \in B(x, \delta) \cap D$ such that $f(y_\delta) < f(x) - \varepsilon_0$. Taking $\delta_k \to 0$, and writing $x_k = y_{\delta_k}$, we observe that $x_k \to x$ and $f(x_k) < f(x) - \varepsilon_0$ for every k. It follows that $\liminf_{k\to\infty} f(x_k) \leq f(x) - \varepsilon_0 < f(x)$, as claimed. □

Exercise 2.22. The function depicted below is lower-semicontinuous at every point, *except* for y. Why?



Exercise 2.23. Let $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ be lower-semicontinuous, and let $C \subset D$ be closed. Suppose, moreover, that either C is bounded or f is coercive. Prove that $\operatorname{argmin}_C(f) \neq \emptyset$.

2.3 Geometry of $\mathbb{R}^N$

The *dot product* (or also *inner product*) of $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^N$ is

$$x \cdot y = x_1 y_1 + \cdots + x_N y_N.$$

Remark 2.24. Considering the elements in $\mathbb{R}^N$ as $N \times 1$ matrices, and using the matrix product, we get

$$x \cdot y = x^T y,$$

where x^T is the transpose of x. This identification will be frequently used in the sequel.

Example 2.25. Let A be a real matrix of size $M \times N$, and consider vectors $u \in \mathbb{R}^N$ and $v \in \mathbb{R}^M$. The identification described above, allows us to compute $Au \cdot v = (Au)^T v = u^T A^T v = u^T (A^T v) = u \cdot A^T v$.

Some of the basic algebraic properties of the dot product are:

i) $x \cdot x = \|x\|^2$ for all $x \in \mathbb{R}^N$.

ii) $x \cdot y = y \cdot x$ for all $x, y \in \mathbb{R}^N$.

iii) $(\alpha x + z) \cdot y = \alpha(x \cdot y) + z \cdot y$ for all $x, y, z \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$.

We also have the following:

Proposition 2.26 (Cauchy-Schwarz Inequality)**.** *For every $x, y \in \mathbb{R}^N$, we have $|x \cdot y| \leq \|x\|\,\|y\|$.*

Proof. The inequality is trivially satisfied if $y = 0$. If $y \neq 0$ and $t > 0$, then

$$0 \leq \|x \pm ty\|^2 = \|x\|^2 \pm 2\,t\,x \cdot y + t^2 \|y\|^2.$$

Therefore,

$$|x \cdot y| \leq \frac{1}{2t}\|x\|^2 + \frac{t}{2}\|y\|^2 \tag{3}$$

for each $t > 0$. The right-hand side is minimized when $t = \|x\|/\|y\|$, which gives precisely $\|x\|\|y\|$. □

Inequality (3), with $t = 1$ is known as *Young's Inequality*. For general $t > 0$, it is sometimes referred to as the *Peter-Paul Inequality*.

Another important consequence of the relationship between the inner product and the norm is the following:

Proposition 2.27 (Parallelogram Identity)**.** *For every $x, y \in \mathbb{R}^N$, we have*

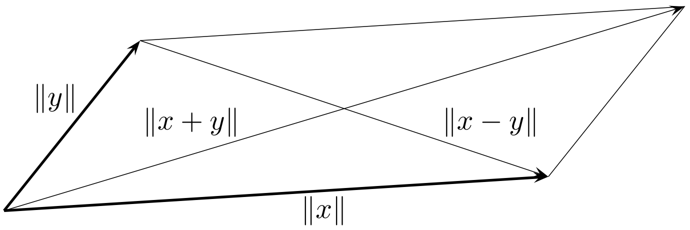
$$\|x + y\|^2 + \|x - y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right).$$

Proof. It suffices to add the following identities

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2\,x \cdot y \qquad \text{and} \qquad \|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2\,x \cdot y,$$

to cancel the terms containing the dot product. □

Proposition 2.27 shows the relationship between the length of the sides and the lengths of the diagonals in a parallelogram.



The vectors $x, y \in \mathbb{R}^N \setminus \{0\}$ are *perpendicular* or *orthogonal* if $x \cdot y = 0$. In that case, we write $x \perp y$. On the other hand, the vectors $x, y \in \mathbb{R}^N \setminus \{0\}$ are *parallel* if there is $\alpha \in \mathbb{R}$ such that $x = \alpha y$, in which case we write $x \parallel y$.

Exercise 2.28. Show that $x \parallel y$ if, and only if, $|x \cdot y| = \|x\|\,\|y\|$.

More generally, we can define the *angle* θ between $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^N$ as

$$\theta = \cos^{-1}\left(\frac{x \cdot y}{\|x\|\|y\|}\right).$$

Then, we have the following:

Proposition 2.29 (Law of Cosines and Pythagoras's Theorem)**.** *For every $x, y \in \mathbb{R}^N$, we have:*

i) $\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2\|x\|\|y\|\cos(\theta)$*; and*

ii) $x \perp y$ *if, and only if,* $\|x + y\|^2 = \|x\|^2 + \|y\|^2$.

Proof. It suffices to prove i), of which ii) is an immediate consequence. We have

$$\begin{aligned} \|x+y\|^2 &= \|x\|^2 + \|y\|^2 + 2\,x \cdot y \\ &= \|x\|^2 + \|y\|^2 + 2\|x\|\|y\|\cos(\theta), \end{aligned}$$

as stated. □

2.4 Tools from linear algebra

This subsection provides a concise overview of essential concepts from linear algebra that will be used throughout these notes. We focus on real vector spaces and their properties, particularly in the context of $\mathbb{R}^N$. For a more comprehensive treatment, we refer the interested reader to standard texts on linear algebra.

2.4.1 Real vector spaces

A *real vector space* is a set V equipped with two operations, vector addition and scalar multiplication by real numbers, satisfying certain axioms (associativity, commutativity, existence of identity elements and inverses for addition, distributivity of scalar multiplication over vector addition, compatibility of scalar multiplication). A subset $W \subseteq V$ is a *subspace* if it is itself a vector space under the same operations.

Example 2.30. There are many examples of real vector spaces, of which we elaborate a few:

- The most prominent example is $\mathbb{R}^N$, the set of all N-dimensional real column vectors.
- The set of all $M \times N$ real matrices, denoted $\mathbb{R}^{M\times N}$, forms a vector space.
- For any matrix $A \in \mathbb{R}^{M\times N}$, the space $W = \{Ax \colon x \in \mathbb{R}^N\} \subset \mathbb{R}^M$ for a vector space, which is a subspace of $\mathbb{R}^M$.
- The set of all real polynomials of degree at most n, denoted $\mathbb{P}_n[x]$, is also a vector space.

Given a set of vectors $\{v_1, \ldots, v_k\}$ in a vector space V, a *linear combination* is an expression of the form $c_1v_1 + \cdots + c_kv_k$, where $c_i \in \mathbb{R}$. The *span* of a set of vectors is the set of all possible linear combinations of those vectors, forming a subspace. A set of vectors is *linearly independent* if the only linear combination that equals the zero vector is the one where all scalars c_i are zero. The *dimension* of a vector space is the maximal number of linearly independent vectors it can contain. A *basis* for a vector space is a set of linearly independent vectors that span the entire space. Note that the cardinality of the basis of a vector space is equal to its dimension by definition. Every vector in V can be uniquely expressed as a linear combination of basis vectors, and the coefficients are called the *coordinates* of the vector with respect to that basis.

2.4.2 Linear transformations and matrices

A map $T \colon V \to W$ between two vector spaces V and W is a *linear transformation* if it preserves vector addition and scalar multiplication, namely if $T(u + v) = T(u) + T(v)$ and $T(cv) = cT(v)$ for all $u, v \in V$ and $c \in \mathbb{R}$.

For any linear transformation $T \colon \mathbb{R}^N \to \mathbb{R}^M$, there exists a unique real $M \times N$ matrix A such that $T(x) = Ax$ for all $x \in \mathbb{R}^N$. We often, but not always, identify linear transformations with their matrix representations.

Given a matrix $A \in \mathbb{R}^{M\times N}$ (or equivalently the associated linear transformation), we define

- the *range* of A as
$$\operatorname{ran}(A) = \{Ax : x \in \mathbb{R}^N\} \subset \mathbb{R}^M.$$
It represents all possible vectors that can be produced by applying A to vectors in $\mathbb{R}^N$.
- the *kernel* (also known as the null space) of A as
$$\ker(A) = \{x \in \mathbb{R}^N : Ax = 0\} \subset \mathbb{R}^N.$$
It consists of all vectors that are mapped to the zero vector by A.
- the *rank* of A as the dimension of its range, $\operatorname{rank}(A) = \dim(\operatorname{ran}(A))$.

The *transpose* of a matrix $A \in \mathbb{R}^{M\times N}$, denoted $A^T \in \mathbb{R}^{N\times M}$, is obtained by interchanging its rows and columns. In the context of linear transformations, the transpose corresponds to the *adjoint* operator in real vector spaces equipped with inner products. The adjoint T^* of a linear transformation T satisfies

$$T(x)\cdot y = x \cdot T^*(y)$$

for all vectors x and y in the respective vector spaces.

An important relationship between the kernel and range is given by the following theorem.

Theorem 2.31. *For any $A \in \mathbb{R}^{M\times N}$, it holds that*

$$\ker(A^TA) = \ker(A) \quad \textit{and} \quad \operatorname{ran}(AA^T) = \operatorname{ran}(A).$$

Furthermore, the kernel of A is the orthogonal complement of the range of A^T. This is implied by the fundamental decomposition of $\mathbb{R}^N$.

Theorem 2.32. *For any $A \in \mathbb{R}^{M\times N}$, it holds that*

$$\mathbb{R}^N = \ker(A) \oplus \operatorname{ran}(A^T) = \{x + y : x \in \ker(A), y \in \operatorname{ran}(A^T)\}.$$

As a consequence, $\dim(\ker(A)) + \operatorname{rank}(A) = N$ *and* $ker(A) = ran(A^T)^\perp$.

For a square matrix $S \in \mathbb{R}^{N\times N}$, a non-zero vector $v \in \mathbb{R}^N$ is an *eigenvector* of S if $Sv = \lambda v$ for some scalar $\lambda \in \mathbb{R}$, called the *eigenvalue* corresponding to v. The set of all eigenvectors corresponding to a particular eigenvalue, along with the zero vector, forms a subspace called the *eigenspace*.
A square matrix $S \in \mathbb{R}^{N\times N}$ is *symmetric* if $S = S^T$. Symmetric matrices have particularly useful properties.

Theorem 2.33. *Let $S \in \mathbb{R}^{N\times N}$ be a symmetric matrix. Then*

- *All eigenvalues of S are real.*
- *The matrix S admits an eigenvalue decomposition $S = \sum_{i=1}^N \lambda_i v_i v_i^T$, where λ_i are the eigenvalues of S, and $\{v_1, \dots, v_N\}$ is an orthonormal basis of eigenvectors.*

A symmetric matrix $P \in \mathbb{R}^{N \times N}$ is *positive semidefinite* if $x^T P x \geq 0$ for all $x \in \mathbb{R}^N$. It is *positive definite* if $x^T P x > 0$ for all non-zero $x \in \mathbb{R}^N$. These properties are directly linked to the eigenvalues.

Theorem 2.34. *Let $P \in \mathbb{R}^{N \times N}$ be a symmetric matrix. Then*

- *P is positive semidefinite if and only if all its eigenvalues are non-negative ($\lambda_i \geq 0$).*
- *P is positive definite if and only if all its eigenvalues are strictly positive ($\lambda_i > 0$).*

Consider the matrix $P = A^T A$, where $A \in \mathbb{R}^{M \times N}$. This matrix P is always symmetric and positive semidefinite. Let μ denote the smallest eigenvalue of P and L denote the largest eigenvalue of P. From the eigenvalue decomposition, we have that

$$\|Ax\|^2 = x^T A^T A x = x^T P x \geq \mu \|x\|^2 \quad \text{for all } x \in \mathbb{R}^N.$$

This inequality shows that the square of the norm of Ax is bounded below by a multiple of the square of the norm of x. From this we deduce that P is nonsingular (i.e., invertible) if, and only if, $\mu > 0$. This is equivalent to $\ker(P) = \{0\}$, which in turn is equivalent to $\ker(A) = \{0\}$.

Similarly, the eigenvalue decomposition also yields

$$\|Ax\|^2 = x^T P x \leq L \|x\|^2 \quad \text{for all } x \in \mathbb{R}^N.$$

This provides an upper bound on $\|Ax\|^2$. If P is invertible (i.e., $\mu > 0$), the ratio $\kappa = \frac{L}{\mu}$ is called the *condition number* of P. The condition number plays a crucial role in the sensitivity of solutions to linear systems and the convergence of optimization algorithms.

A square matrix $A \in \mathbb{R}^{N \times N}$ is *bijective* (or invertible) if its kernel is trivial ($\ker(A) = \{0\}$) and its range spans the entire space ($\operatorname{ran}(A) = \mathbb{R}^N$). In this case, there exists a unique inverse matrix A^{-1} such that $AA^{-1} = A^{-1}A = I$, where I is the identity matrix. When A is not square or not invertible, the *Moore-Penrose pseudoinverse*, denoted $A^\dagger$, provides a generalization of the inverse. For a matrix $A \in \mathbb{R}^{M \times N}$ such that $A^T A$ is invertible, the pseudoinverse is given by $A^\dagger = (A^T A)^{-1} A^T$.

2.5 Differentiable functions and a first order necessary condition for optimality

Let $D \subset \mathbb{R}^N$ be nonempty and open. A function $f : D \to \mathbb{R}$ is *differentiable* at $x \in D$ (*in the sense of Gâteaux*) if the *directional derivative*

$$f'(x; v) = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}$$

exists for all $v \in \mathbb{R}^N$, and there is $g(x) \in \mathbb{R}^N$ such that

$$g(x) \cdot v = f'(x; v)$$

for all $v \in \mathbb{R}^N$. In this case, the *gradient* of f at x is $\nabla f(x) = g(x)$. As usual, f is *differentiable* on D if it is so at every point of D.

Let f be differentiable at x, and let $g(x) = \nabla f(x)$ be its gradient at that point. If e_i denotes the i-th canonical vector in $\mathbb{R}^N$, then the i-th coordinate of $g(x)$, denoted by $g_i(x)$, satisfies

$$g_i(x) = \nabla f(x) \cdot e_i = f'(x; e_i) = \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t} = \frac{\partial f}{\partial x_i}(x).$$

In other words, the components of the gradient are the *partial derivatives.*

Example 2.35. Let us compute the gradient of the function $f : \mathbb{R}^N \to \mathbb{R}$, defined by

$$f(x) = \frac{1}{2}\|Ax - b\|^2,$$

where A is a real matrix of size $M \times N$ and $b \in \mathbb{R}^M$. For every $v \in \mathbb{R}^N$, we have

$$\frac{f(x+tv) - f(x)}{t} = \frac{\|A(x+tv) - b\|^2 - \|Ax - b\|^2}{2t} = \frac{2tAv \cdot (Lx - b) + t^2\|Av\|^2}{2t} = Ah \cdot (Ax - b) + \frac{t\|Av\|^2}{2}.$$

Letting $t \to 0$, we get $f'(x; v) = Ah \cdot (Ax - b)$. In order to identify the gradient $\nabla f(x)$, we must write the directional derivative in the form $g \cdot v$. Recalling, from Example 2.25, we see that

$$Av \cdot (Ax - b) = v \cdot A^T(Ax - b) = A^T(Ax - b) \cdot v,$$

whence $\nabla f(x) = A^T(Ax - b)$.

Proposition 2.36. *Let $D \subset \mathbb{R}^N$ be open, and let $f : D \to \mathbb{R}$ be differentiable at $x \in D$. If $\nabla f(x) \neq 0$, then $\nabla f(x)$ points in the direction of maximum ascent. More precisely,*

$$\frac{\nabla f(x)}{\|\nabla f(x)\|} \in \operatorname{argmax}\big\{f'(x; v) : \|v\| = 1\big\}.$$

Proof. By the definition of the gradient and the Cauchy-Schwarz inequality, for every $v \in \mathbb{R}^N$ such that $\|v\| = 1$, we have

$$f'(x; v) = \nabla f(x) \cdot v \le \|\nabla f(x)\|\|v\| = \|\nabla f(x)\|,$$

and so

$$\max_{\|v\|=1} f'(x; v) \le \|\nabla f(x)\|.$$

On the other hand, if we write

$$v_x = \frac{\nabla f(x)}{\|\nabla f(x)\|},$$

we have

$$\max_{\|v\|=1} f'(x; v) \ge f'(x; v_x) = \nabla f(x) \cdot v_x = \|\nabla f(x)\|,$$

which shows that the maximum is attained at v_x, as claimed. □

Remark 2.37. As a consequence, $-\nabla f(x)$ points in the direction of maximum descent of f at x. This fact will be useful in defining numerical methods to find minimizers of functions. If $\nabla f(x) \neq 0$, there is $\delta > 0$ such that $f\big(x - t\nabla f(x)\big) < f(x)$ for every $t \in (0, \delta)$. Indeed, let $\varepsilon = \frac{1}{2}\|\nabla f(x)\|^2$. According to the definition of gradient, there is $\delta > 0$ such that

$$\left| \frac{f\big(x - t\nabla f(x)\big) - f(x)}{t} - \nabla f(x) \cdot \big(-\nabla f(x)\big) \right| < \frac{1}{2}\|\nabla f(x)\|^2$$

for every $t \in (0, \delta)$. This implies that

$$f\big(x - t\nabla f(x)\big) < f(x) - \frac{t}{2}\|\nabla f(x)\|^2 < f(x)$$

for all such t, as claimed. As a consequence, one can reduce the values of the function by moving in the direction of $-\nabla f(x)$. The value of δ can be quantified in some cases, as explained in Remark 2.46. This will be highly relevant when we analyze certain numerical methods.

The gradient is orthogonal to the level sets of the function. More precisely, let $\sigma : \mathbb{R} \to [f = \gamma]$ be a differentiable function that describes a curve in the level set, with $\sigma(0) = x$ and $\sigma'(0) = v$. Then, $f \circ \sigma \equiv \gamma$, and so

$$0 = \frac{d}{dt} f\big(\sigma(t)\big) = \nabla f\big(\sigma(t)\big) \cdot \sigma'(t)$$

for every $t \in \mathbb{R}$. In particular, for $t = 0$, we get

$$\nabla f(x) \cdot v = 0.$$

Differentiability in the sense of Gâteaux does not imply continuity, as seen in the following:

Example 2.38. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} \dfrac{2x^4 y}{x^6 + y^3} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0). \end{cases}$$

A simple computation shows that $\nabla f(0,0) = (0,0)$. However, $\lim_{z \to 0} f(z, z^2) = 1 \neq f(0,0)$. Therefore, f is not continuous at $(0,0)$.

The following result allows us to characterize the minimizers of a differentiable function. It generalizes Proposition 1.2 to $\mathbb{R}^N$, and will be further generalized in the next section.

Theorem 2.39 (Fermat's Rule I)**.** *Let $D \subset \mathbb{R}^N$ be open, and let $f : D \to \mathbb{R}$ be differentiable. If $\hat{x} \in \operatorname{argmin}_D(f)$, then $\nabla f(\hat{x}) = 0$.*

Proof. Take $v \neq 0$ and $t > 0$ sufficiently small so that $\hat{x} + tv \in D$. Since $\hat{x} \in \operatorname{argmin}_A(f)$, $f(\hat{x}) \leq f(\hat{x} + tv)$. Therefore,

$$\nabla f(\hat{x}) \cdot v = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t} \geq 0.$$

Applying the same argument with $-v$ instead of v, we deduce that $\nabla f(\hat{x}) \cdot (-v) \geq 0$, which implies $\nabla f(\hat{x}) \cdot v \leq 0$. It follows that $\nabla f(\hat{x}) \cdot v = 0$ and, since v is arbitrary, $\nabla f(\hat{x}) = 0$. □

Example 2.40. Let $f : \mathbb{R}^N \to \mathbb{R}$ be defined by

$$f(x) = \frac{1}{2}\|Ax - b\|^2,$$

where A is a real matrix of size $M \times N$ and $b \in \mathbb{R}^M$. From Example 2.35, we know that $\nabla f(x) = A^T(Ax - b)$. Theorem 2.39 says that the minimizers of f are characterized by the equation $A^T A x = A^T b$, which has at least one solution $\hat{x}$, because $\operatorname{ran}(A^T A) = \operatorname{ran}(A^T)$. If $\ker(A^T A) = \ker(A) = \{0\}$, the unique solution may be found by computing $\hat{x} = (A^T A)^{-1} A^T b$, in which case $\operatorname{argmin}(f) = \{\hat{x}\}$. Otherwise, a solution is given by $\hat{x} = (A^T A)^\dagger A^T b$, where $X^\dagger$ denotes the *Moore-Penrose pseudoinverse* of the noninvertible matrix X. Since the difference of two solutions must be in $\ker(A^T A) = \ker(A)$, we conclude that

$$\operatorname{argmin}(f) = \{\hat{x}\} + \ker(A).$$

Remark 2.41. Theorem 2.39 remains true if D is not necessarily open and f is not necessarily differentiable in all of D, as long as $\hat{x} \in \operatorname{int}(D)$ and f is differentiable at $\hat{x}$.

2.6 Stronger first order regularity

Let $D \subset \mathbb{R}^N$ be nonempty and open. A function $f : D \to \mathbb{R}$ is *differentiable* at $x \in D$ *in the sense of Fréchet* if there is a vector $G(x) \in \mathbb{R}^N$ such that

$$\lim_{h \to 0} \frac{|f(x+h) - f(x) - G(x) \cdot h|}{\|h\|} = 0. \tag{4}$$

Exercise 2.42. Show that, if A is a matrix of size $M \times N$ and $b \in \mathbb{R}^M$, the function

$$f(x) = \frac{1}{2}\|Ax - b\|^2$$

is differentiable in the sense of Fréchet.

Proposition 2.43. *Let $D \subset \mathbb{R}^N$ be nonempty and open. If $f : D \to \mathbb{R}$ is differentiable at $x \in D$ in the sense of Fréchet, then f is continuous at x, and differentiable there in the sense of Gâteaux, with $G(x) = \nabla f(x)$.*

Proof. From (4), there is $\delta > 0$ such that

$$\frac{|f(x+h) - f(x) - G(x) \cdot h|}{\|h\|} < 1$$

for every $h \in B(0, \delta)$. This implies that

$$|f(x+h) - f(x)| < \|G(x)\|\|h\| + \|h\|,$$

and so $\lim_{h \to 0} f(x+h) = f(x)$ and f is continuous at x. It remains to show that

$$G(x) \cdot v = \lim_{t \to 0} \frac{f(x+tv) - f(x)}{t} \tag{5}$$

for every $v \in \mathbb{R}^N$. If $v = 0$, there is nothing to prove. Otherwise, write

$$\left| \frac{f(x+tv) - f(x)}{t} - G(x) \cdot v \right| = \frac{|f(x+tv) - f(x) - G(x) \cdot (tv)|}{t\|v\|}\|v\| = \frac{|f(x+h) - f(x) - G(x) \cdot h|}{\|h\|}\|v\|,$$

where we have written $h = tv$. As t tends to zero, so does h, and so does the right-hand side of the equality above, in view of (4). □

Example 2.44. The function in Example 2.38 is differentiable in the sense of Gâteaux but is not continuous. Therefore, it cannot be differentiable in the sense of Fréchet.

In all that follows, the term *differentiable* is meant in the sense of Gâteaux, unless otherwise stated.

A function $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ is *L-smooth* if it is differentiable and its gradient ∇f is *Lipschitz-continuous* with constant L, which means that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

for every $x, y \in A$.

Proposition 2.45. *Let $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ be L-smooth, and suppose the line segment joining x and y is contained in D. Then,*

$$\big|f(y) - f(x) - \nabla f(x) \cdot (y - x)\big| \leq \frac{L}{2}\|x - y\|^2. \tag{6}$$

Proof. Define $g : [0, 1] \to \mathbb{R}^N$ as $g(t) = f\big(x + t(y - x)\big)$, so that

$$g'(t) = \nabla f\big(x + t(y - x)\big) \cdot (y - x).$$

We have

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(s)\,ds = \int_0^1 \nabla f\big(x + s(y - x)\big) \cdot (y - x)\,ds,$$

and so

$$f(y) - f(x) - \nabla f(x) \cdot (y - x) = \int_0^1 \big[\nabla f\big(x + s(y - x)\big) - \nabla f(x)\big] \cdot (y - x)\,ds.$$

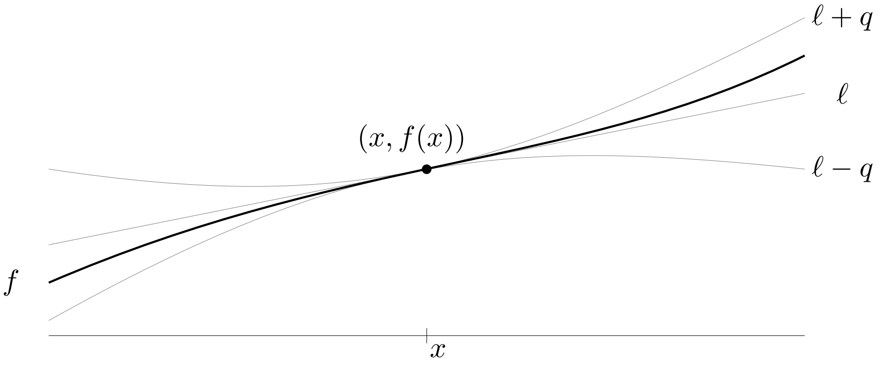
Using the monotonicity of the integral and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \big|f(y) - f(x) - \nabla f(x) \cdot (y - x)\big| &= \left|\int_0^1 \big[\nabla f\big(x + s(y - x)\big) - \nabla f(x)\big] \cdot (y - x)\,ds\right| \\ &\leq \int_0^1 \big\|\nabla f\big(x + s(y - x)\big) - \nabla f(x)\big\|\,\|y - x\|\,ds \\ &\leq L\|y - x\|^2 \int_0^1 s\,ds, \end{aligned}$$

which is exactly (6). □

The preceding result means that the difference between the values of a smooth function and those of its first order Taylor approximation can be bounded by a quadratic. More precisely, fix $x \in D$, and define $\ell(y) = f(x) + \nabla f(x) \cdot (y - x)$ and $q(x) = \frac{L}{2}\|y - x\|^2$. Then,

$$\ell - q \leq f \leq \ell + q.$$



Remark 2.46. As explained in Remark 2.37, for all sufficiently small t, we have $f\big(x - t\nabla f(x)\big) < f(x)$. If f is L-smooth, Proposition 2.45 allows us to quantify this decrease. Indeed, (6) implies that

$$f\big(x - t\nabla f(x)\big) - f(x) \leq -t\|\nabla f(x)\|^2 + \frac{Lt^2}{2}\|\nabla f(x)\|^2 = t\left[\frac{Lt}{2} - 1\right]\|\nabla f(x)\|^2.$$

If $t < 2/L$, the right-hand side is decreasing. Its smallest possible value is attained when $t = 1/L$, giving

$$f\left(x - \tfrac{1}{L}\nabla f(x)\right) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|^2.$$

This fact will be relevant in our analysis of the *gradient method*, a simple numerical procedure to minimize smooth functions without constraints, which is also one of the main building blocks in more sophisticated methods.

Smoothness is the strongest type of differentiability that we have studied in this chapter. Indeed, using Proposition 2.45, we can prove the following:

Proposition 2.47. *If $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ is L-smooth, then it is differentiable in the sense of Fréchet.*

Proof. From (6), we know that

$$\frac{|f(x+h) - f(x) - \nabla f(x)\cdot h|}{\|h\|} \leq \frac{L}{2}\|h\|,$$

and the right-hand side tends to zero with h. □

2.7 Hessian and second order optimality conditions

Let $D \subset \mathbb{R}^N$ be open, and let $f : D \to \mathbb{R}$ be differentiable. We say f is *twice differentiable* at $x \in D$ if there is a matrix H, of size $N \times N$, such that

$$Hv = \lim_{t\to 0}\frac{\nabla f(x+tv) - \nabla f(x)}{t}$$

for all $v \in \mathbb{R}^N$. The *Hessian* of f at x is $\nabla^2 f(x) = H$. If the function $x \mapsto \nabla^2 f(x)$ is continuous, we say f is *of class* $\mathcal{C}^2$, in which case, $\nabla^2 f(x)$ must be symmetric for each $x \in D$.

We have the following:

Proposition 2.48 (Second order Taylor approximation)**.** *Let $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ be of class $\mathcal{C}^2$, and let $x \in A$. For each $v \in \mathbb{R}^N$,*

$$\lim_{t\to 0}\frac{1}{t^2}\left|f(x+tv) - f(x) - t\nabla f(x)\cdot v - \frac{t^2}{2}\nabla^2 f(x)v\cdot v\right| = 0.$$

Proof. Define $\phi : I \subset R \to \mathbb{R}$ by $\phi(t) = f(x+tv)$, where I is a sufficiently small open interval around 0 such that $\phi(t)$ exists for all $t \in I$. It is easy to see that $\phi'(t) = \nabla f(x+tv)\cdot d$ and $\phi''(0) = (\nabla^2 f(x)v)\cdot v$. The second-order Taylor expansion for ϕ in $\mathbb{R}$ yields

$$\lim_{t\to 0}\frac{1}{t^2}\left|\phi(t) - \phi(0) - t\phi'(0) - \frac{t^2}{2}\phi''(0)\right| = 0,$$

which gives the result. □

The second order Taylor approximation discussed above is a *local* property, which allows us to understand the geometry of the function in a (possibly small) neighborhood of a point.

Theorem 2.49 (Second order optimality conditions)**.** *Let $D \subset \mathbb{R}^N$ be open, let $f : D \to \mathbb{R}$ be of class $\mathcal{C}^2$, and let $\hat{x} \in D$.*

i) If $\hat{x} \in \operatorname{argmin}(f)$, then $\nabla f(\hat{x}) = 0$ and $\nabla^2 f(\hat{x})$ is positive semidefinite.

ii) If $\nabla f(\hat{x}) = 0$ and $\nabla^2 f(\hat{x})$ is positive definite, there is $\varepsilon > 0$ such that $f(\hat{x}) < f(y)$ for every $y \in B(\hat{x}, \varepsilon)$.

Example 2.50. The Hessian of the quadratic function $f(x) = \frac{1}{2}\|Ax - b\|^2$ is constant, namely: $\nabla^2 f(x) = A^T A$. In agreement with i) of Theorem 2.49, it is positive semidefinite. As seen in Example 2.40, $\operatorname{argmin}(f) = \{\hat{x}\} + \ker(A)$, where $\hat{x} = (A^T A)^\dagger A^T b$, and $(A^T A)^\dagger$ is the Moore-Penrose pseudoinverse of $A^T A$. The set $\operatorname{argmin}(f)$ is reduced to a singleton when $\ker(A) = \{0\}$, which is equivalent to $A^T A$ being positive definite, as established in ii) of Theorem 2.49.

The converse of part i) of Theorem 2.49 is not true in general, even in $\mathbb{R}$ (when $N = 1$), as shown in the following:

Example 2.51. For the function $f(x) = x^3$, we have $\nabla f(0) = f'(0) = 0$ and $\nabla^2 f(0) = f''(0) \geq 0$, but 0 is not a minimizer of f.

Exercise 2.52. Determine $\operatorname{argmin}(f_a)$, where $f_a(x, y) = ax^4 + 8y^2 - 16xy$ and $a \in \mathbb{R}$.

It is also possible to prove the following, although its proof is postponed to a later section.

Proposition 2.53. *Let $D \subset \mathbb{R}^N$ be open and convex, let $f : D \to \mathbb{R}$ be of class $\mathcal{C}^2$, and let $\hat{x} \in D$. If $\nabla f(\hat{x}) = 0$ and $\nabla^2 f(y)$ is positive semidefinite for every $y \in D$, then $\hat{x} \in \operatorname{argmin}(f)$.*

3 Convexity

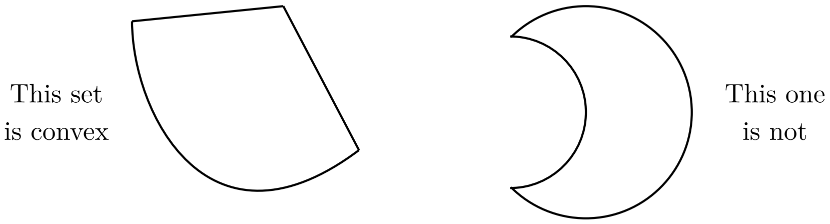
The concept of convexity plays a crucial role in the well-posedness of optimization problems, the characterization of minimizers, and the convergence and performance of computational methods. In Section 3.1, we introduce convex sets, along with many of their interesting properties. Their implication in the characterization of constrained minimizers is discussed in Section 3.2. Convex functions are introduced in Section 3.3, together with their properties, and some characterizations in the differentiable case. Finally, reinforced forms of convexity for functions are discussed in Section 3.4, including their connection with smoothness.

3.1 Convex sets

The *segment* joining $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^N$ is the set

$$\mathrm{seg}[x,y] = \big\{\lambda x + (1-\lambda)y : \lambda \in (0,1)\big\}.$$

A set $C \subset \mathbb{R}^N$ is *convex* if $\mathrm{seg}[x,y] \subset C$ whenever $x, y \in C$.



Example 3.1. Open balls are convex. Indeed, if $x, y \in B(z,r)$ and $\lambda \in (0,1)$, then

$$\begin{aligned} \|\lambda x + (1-\lambda)y - z\| &= \|\lambda(x-z) + (1-\lambda)(y-z)\| \\ &\leq \lambda\|x-z\| + (1-\lambda)\|y-z\| \\ &< \lambda r + (1-\lambda)r \\ &= r, \end{aligned}$$

and so $\lambda x + (1-\lambda)y \in B(z,r)$. By a similar argument, closed balls are convex.

Exercise 3.2. What happens if a ball contains *some* of its boundary elements?

Exercise 3.3. Let C be a convex set. Prove that, for any $\lambda_1, \ldots, \lambda_n \in [0,1]$ such that $\sum_{i=1}^n \lambda_i = 1$, and any $x_1 \ldots, x_n \in C$, it holds that

$$\lambda_1 x_1 + \cdots + \lambda_n x_n \in C.$$

A *box* in $\mathbb{R}^N$ is a set of the form

$$\prod_{i=1}^N [a_i, b_i],$$

where $a_i \leq b_i$ for $i = 1, \ldots N$. The *unit cube* is the box $[-1,1]^N$. Boxes are sometimes called N*-dimensional rectangles* or *parallelepipeds*.

An *affine subspace* of $\mathbb{R}^N$ is a set of the form

$$C = \{x_0\} + V = \{x_0 + v : v \in V\},$$

where x_0 is a point in $\mathbb{R}^N$, and V is a vector subspace of $\mathbb{R}^N$. Since this usually does not lead to confusion, one sometimes omits the braces in the singleton, and writes

$$C = x_0 + V.$$

The *dimension* of C is $\dim(C) = \dim(V)$.

Open and closed *halfspaces* are sets of the form

$$[v < \gamma] = \{x \in \mathbb{R}^N : v \cdot x < \gamma\}$$

and

$$[v \leq \gamma] = \{x \in \mathbb{R}^N : v \cdot x \leq \gamma\},$$

respectively. Although we have used the same notation as for sublevel sets, this should not lead to any confusion, since $[v \leq \gamma] = [f \leq \gamma]$ for $f(x) = v \cdot x$.

Exercise 3.4. Show that boxes, affine subspaces and halfspaces (both open and closed) are convex.

Example 3.5. The intersection of any (possibly infinite) collection of convex sets is convex. To see this, let $C = \bigcap\{C_j : j \in \mathcal{J}\}$, where C_j is convex for each $j \in \mathcal{J}$. Now, take $x, y \in C$. Then $x, y \in C_j$ for each $j \in \mathcal{J}$. By the convexity of C_j, $\mathrm{seg}[x, y] \subset C_j$. Since this is true for each C_j, we conclude that $\mathrm{seg}[x, y] \subset C$.

Exercise 3.6. When is the union of convex sets convex?

A *polyhedron* is the intersection of a finite number of halfspaces. Therefore, polyhedra are convex sets.

Exercise 3.7. Show that polyhedra are sets of the form $\{x \in \mathbb{R}^N \colon Ax \leq b\}$, with $A \in \mathbb{R}^{M\times N}$ and $b \in \mathbb{R}^M$.

Exercise 3.8. Consider a number $\alpha \in \mathbb{R}$, sets $C_1, C_2 \subset \mathbb{R}^N$ and $C_3 \subset \mathbb{R}^M$, and a linear function $A : \mathbb{R}^N \to \mathbb{R}^M$. Prove that, if C_1, C_2 and C_3 are convex, then so are $\alpha C_1 + C_2$, $C_1 \times C_3$, $A(C_1)$ and $A^{-1}(C_3)$.

Exercise 3.9. Let B_1 and B_2 be open (respectively, closed) balls, and let $\lambda \in (0, 1)$. Prove that the set $B = \lambda B_1 + (1 - \lambda)B_2$ is a ball. Compute its center and radius.

Exercise 3.10. Show that the interior and the closure of a convex set are convex.

Proposition 3.11. *Let C be convex, let $x \in \mathrm{int}(C)$, and let $y \in \overline{C}$. Then, $\lambda x + (1-\lambda)y \in \mathrm{int}(C)$ for every $\lambda \in (0, 1]$.*

Proof. Take $\lambda \in (0, 1]$ and write $z = \lambda x + (1-\lambda)y$. Since $x \in \mathrm{int}(C)$, there is $\rho > 0$ such that $B(x, \rho) \subset C$. On the other hand, since $y \in \overline{C}$, for every $\varepsilon > 0$, there is $p_\varepsilon \in \mathbb{R}^N$ such that $\|p_\varepsilon\| < \varepsilon$ and $y + p_\varepsilon \in C$. We shall pick an appropriate value for ε in a moment. Our purpose is to find $\delta > 0$ such that $z + \delta p \in C$ for all $p \in B(0, 1)$. We have

$$z + \delta p = \lambda x + (1 - \lambda)y + \delta p = \lambda \left[x + \frac{1-\lambda}{\lambda} p_\varepsilon + \frac{\delta}{\lambda} p\right] + (1 - \lambda)(y + p_\varepsilon).$$

Since $y + p_\varepsilon \in C$ and C is convex, it suffices to select ε and δ so that

$$x + \frac{1-\lambda}{\lambda} p_\varepsilon + \frac{\delta}{\lambda} p \in B(x, \rho) \subset C$$

(hence, $z + \delta p$ will be a convex combination of elements of C). In other words, it suffices that

$$\left\| x + \frac{1-\lambda}{\lambda} p_\varepsilon + \frac{\delta}{\lambda} p - x \right\| < \frac{\varepsilon(1-\lambda)}{\lambda} + \frac{\delta}{\lambda} \le \rho.$$

A suitable choice is $\delta = \varepsilon \le \frac{\lambda\rho}{2-\lambda}$. □

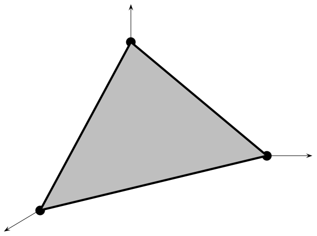
Exercise 3.12. Let C be a convex set with nonempty interior. Prove that $\overline{\mathrm{int}(C)} = \overline{C}$ and $\mathrm{int}(C) = \mathrm{int}(\overline{C}\,)$.

The *convex hull* of a set $S \subset R^N$ is

$$\mathrm{conv}(S) = \left\{ \sum_{j=1}^{J} \lambda_j x_j \ : \ x_1, \ldots, x_J \in S, \quad \lambda_j \in [0,1] \quad \text{and} \quad \sum_{j=1}^{J} \lambda_j = 1 \right\}.$$

The elements of $\mathrm{conv}(S)$ are the *convex combinations* of the elements of S.

Example 3.13. Let $e_1, \ldots, e_N$ denote the canonical vectors of $\mathbb{R}^N$. The set $\mathrm{conv}\big(\{e_1, \ldots, e_N\}\big)$ is usually known as the $(N-1)$-*dimensional simplex.* The picture on the right shows the 2-dimensional simplex in $\mathbb{R}^3$.



Proposition 3.14. *Let $S \subset \mathbb{R}^N$. The set $\mathrm{conv}(S)$ is convex and contains S. It is the intersection of all convex sets containing S. In particular, S is convex if, and only if, $S = \mathrm{conv}(S)$.*

Proof. It is clear that $S \subset \mathrm{conv}(S)$. Now, let us take $z_1, z_2 \in \mathrm{conv}(S)$ and $\nu \in [0,1]$, and prove that $\nu z_1 + (1-\nu) z_2 \in \mathrm{conv}(S)$. Since $z_1, z_2 \in \mathrm{conv}(S)$, there exist $x_1, y_1, x_2, y_2 \in S$ and $\lambda_1, \lambda_2 \in [0,1]$ such that $z_i = \lambda_i x_i + (1-\lambda_i) y_i$, for $i = 1, 2$. Therefore,

$$\nu z_1 + (1-\nu) z_2 = \nu\lambda_1 x_1 + \nu(1-\lambda_1) y_1 + (1-\nu)\lambda_2 x_2 + (1-\nu)(1-\lambda_2) y_2.$$

This point belongs to $\mathrm{conv}(S)$ because

$$\nu\lambda_1 + \nu(1-\lambda_1) + (1-\nu)\lambda_2 + (1-\nu)(1-\lambda_2) = \nu\big(\lambda_1 + 1 - \lambda_1\big) + (1-\nu)\big(\lambda_2 + 1 - \lambda_2\big) = 1,$$

and $x_1, y_1, x_2, y_2 \in S$. □

We finish this section with an important result about the existence of hyperplanes separating a convex set from points in their complement.

Proposition 3.15. *If C is a nonempty and convex subset of $\mathbb{R}^N$ not containing the origin, there is $v \in \mathbb{R}^N \setminus \{0\}$ such that $v \cdot x \le 0$ for each $x \in C$.*

Proof. Let $(x_n) \in C$ such that the set $\{x_n : n \ge 1\}$ is dense in C. Let C_n be the convex hull of the set $\{x_k : k = 1, \ldots, n\}$ and let p_n be the least-norm element of C_n. By convexity, for each $x \in C_n$ and $t \in (0,1)$, we have

$$\|p_n\|^2 \le \|p_n + t(x - p_n)\|^2 = \|p_n\|^2 + 2t\, p_n \cdot (x - p_n) + t^2 \|x - p_n\|^2.$$

Therefore,

$$0 \leq 2\|p_n\|^2 \leq 2\,p_n \cdot x + t\|x - p_n\|^2.$$

Letting $t \to 0$, we deduce that $p_n \cdot x \geq 0$ for all $x \in C_n$. Now write $v_n = -p_n/\|p_n\|$. The sequence (v_n) lies in the unit sphere, which is compact. We may extract a subsequence that converges to some $v \in \mathbb{R}^N$ with $\|v\| = 1$ (thus $v \neq 0$) and $v \cdot x \leq 0$ for all $x \in C$. □

Exercise 3.16. Let $C \subset \mathbb{R}^N$ be nonempty, open and convex. Show that, if $N \geq 2$ and $0 \notin C$, there is a nontrivial subspace of $\mathbb{R}^N$ that does not intersect C.

3.2 Optimization under convex constraints

Theorem 3.17 (Fermat's Rule II)**.** *Let $D \subset \mathbb{R}^N$ be open, and let $C \subset D$ be nonempty, closed and convex. Consider a differentiable function $f : D \to \mathbb{R}$ and a point $\hat{x} \in \operatorname{argmin}_C(f)$. Then,*

$$\nabla f(\hat{x}) \cdot (y - \hat{x}) \geq 0 \tag{7}$$

for every $y \in C$.

Proof. Take any $y \in C$ and $\lambda \in (0,1)$. Since C is convex and $\hat{x}$ also belongs to C, we have

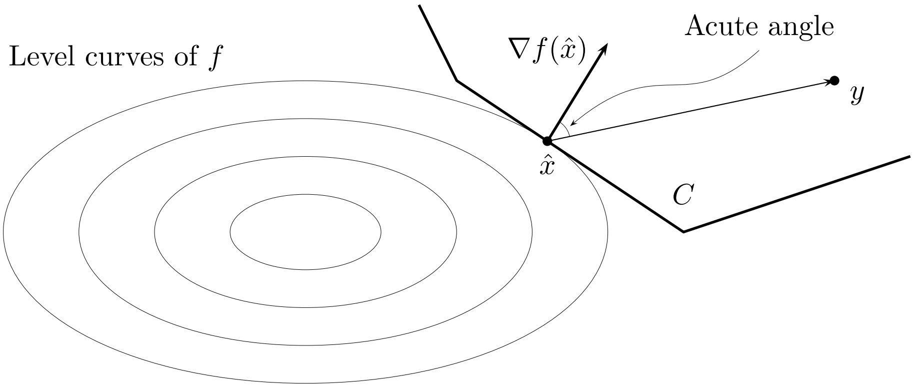
$$\lambda y + (1-\lambda)\hat{x} = \hat{x} + \lambda(y - \hat{x}) \in C.$$

Therefore, $f(\hat{x}) \leq f\big(\hat{x} + \lambda(y - \hat{x})\big)$, which implies that

$$\frac{f\big(\hat{x} + \lambda(y - \hat{x})\big) - f(\hat{x})}{\lambda} \geq 0.$$

Letting $\lambda \to 0$, and recalling the definition of gradient, we obtain precisely (7). □

This means that the gradient at a constrained minimizer forms an acute or straight angle with any vector pointing at any other element of C.



If C is an affine space, Theorem 3.17 gives:

Corollary 3.18. *Let $D \subset \mathbb{R}^N$ be open, let $f : D \to \mathbb{R}$, and let $C = \{x_0\} + V$ be an affine subspace of $\mathbb{R}^N$, contained in A. If $\hat{x} \in \operatorname{argmin}_C(f)$, then $\nabla f(\hat{x}) \perp V$.*

Example 3.19. Let $D \subset \mathbb{R}^N$ be open, let $f : D \to \mathbb{R}$ be differentiable, and let

$$C = \{x \in \mathbb{R}^N : Ax = b\},$$

where A is a real matrix of size $M \times N$ and $b \in \operatorname{ran}(A) \subset \mathbb{R}^M$. We can express C as

$$C = \{x_0\} + V,$$

where x_0 is any point in C and $V = \ker(A)$. If $\hat{x} \in \operatorname{argmin}_C(f)$, Corollary 3.18 states that

$$\nabla f(\hat{x}) \in V^{\perp} = \big(\ker(A)\big)^{\perp} = \operatorname{ran}(A^T).$$

Therefore, there exists $\bar{z} \in \mathbb{R}^M$ such that $\nabla f(\hat{x}) = A^T \hat{z}$. The elements of $\mathbb{R}^M$ are called the *dual variables* of the constrained minimization problem, and we say $\hat{z}$ is a *dual solution* of the problem. We shall discuss this in further detail later.

The following result shows that we can project onto closed convex sets.

Corollary 3.20. *Let $C \subset \mathbb{R}^N$ be nonempty, closed and convex. For each $x \in \mathbb{R}^N$, there is a unique point $P_C(x) \in C$ such that*

$$\big\|x - P_C(x)\big\| = \min\big\{\|x - z\| \ : \ z \in C\big\} = \operatorname{dist}(x, C). \tag{8}$$

Moreover, $P_C(x)$ is the only point in C that satisfies the inequality

$$\big(x - P_C(x)\big) \cdot \big(z - P_C(x)\big) \leq 0 \tag{9}$$

for all $z \in C$.

Proof. Fix $x \in \mathbb{R}^N$, and define $f : \mathbb{R}^N \to \mathbb{R}$ by $f(z) = \frac{1}{2}\|z - x\|^2$, so that (8) can be equivalently written as $\min\{f(z) : z \in C\}$. Since f is continuous (thus lower-semicontinuous) and coercive, and C is closed, Exercise 2.23 shows that $\operatorname{argmin}_C(f) \neq \emptyset$. On the other hand, since $\nabla f(z) = z - x$ for all $z \in \mathbb{R}^N$, Theorem 3.17 shows that every $\bar{y} \in \operatorname{argmin}_C(f)$ must satisfy

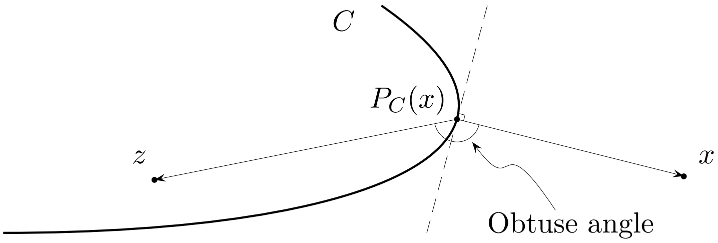
$$(x - \bar{y}) \cdot (z - \bar{y}) \leq 0 \tag{10}$$

for every $z \in C$. It remains to show that $\operatorname{argmin}_C(f)$ cannot have more than one element. Assume, on the contrary, that $\bar{y}_1, \bar{y}_2 \in \operatorname{argmin}_C(f)$. Since both must satisfy (10), we have

$$(x - \bar{y}_1) \cdot (\bar{y}_2 - \bar{y}_1) \leq 0 \qquad \text{and} \qquad (x - \bar{y}_2) \cdot (\bar{y}_1 - \bar{y}_2) \leq 0.$$

Combining these inequalities, we get $\|\bar{y}_1 - \bar{y}_2\|^2 \leq 0$, which implies $\bar{y}_1 = \bar{y}_2$. □

The point $P_C(x)$ is the *projection* of x onto C. Inequality (9) implies the angle condition depicted below.



Exercise 3.21. Show that, if $C = \{x_0\} + V$ is an affine subspace of $\mathbb{R}^N$, then $P_C(x) - x \perp V$.

Exercise 3.22. Let $C \subset \mathbb{R}^N$ be nonempty, closed and convex. Prove that $P_C : \mathbb{R}^N \to \mathbb{R}^N$ is a *nonexpansive* function, which means that $\|P_C(x) - P_C(y)\| \leq \|x - y\|$ for every $x, y \in \mathbb{R}^N$.

The term *orthogonal projection*, commonly used in the affine setting, is explained by Exercise 3.21. In the same context, if $C = \{x \in \mathbb{R}^N : Ax = b\}$, then

$$P_C(x) - x \in \big[\ker(A)\big]^\perp = \operatorname{ran}(A^T).$$

Example 3.23. Let us revisit Example 2.20 under the light of our last discoveries. On the one hand, we have

$$\min(f) = \min\left\{\frac{1}{2}\|Ax - b\|^2 : x \in \mathbb{R}^N\right\} = \frac{1}{2}\operatorname{dist}\big(b, \operatorname{ran}(A)\big)^2.$$

On the other hand, since $\operatorname{ran}(A)$ is nonempty, closed and convex, Corollary 3.20 says that

$$\operatorname{dist}\big(b, \operatorname{ran}(A)\big) = \|P_{\operatorname{ran}(A)}(b) - b\|.$$

The minimizers of f are thus the points that are mapped to $P_{\operatorname{ran}(A)}(b)$ by A. In other words,

$$\operatorname{argmin}(f) = A^{-1}\big(P_{\operatorname{ran}(A)}(b)\big).$$

How does this compare to Example 2.40?

Exercise 3.24. Let C be closed and convex. Show that there exists a family $(H_i)_{i\in\mathcal{I}}$ of closed halfspaces such that $C = \bigcap_{i\in\mathcal{I}} H_i$.

3.3 Convex functions

A function $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ is *convex* if D is a convex set and

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

for all $x, y \in D$ and all $\lambda \in (0,1)$.

Example 3.25. The function $f(x) = \|Ax - b\|^2$ is convex. Indeed,

$$\begin{aligned} f\big(\lambda x + (1-\lambda)y\big) &= \|\lambda(Ax-b) + (1-\lambda)(Ay-b)\|^2 \\ &= \lambda^2 f(x) + (1-\lambda)^2 f(y) + 2\lambda(1-\lambda)(Ax-b)\cdot(Ay-b). \end{aligned} \tag{11}$$

Using the identity $2\,u \cdot v = \|u\|^2 + \|v\|^2 - \|u - v\|^2$, with $u = Ax - b$ and $v = Ay - b$, we get

$$2\lambda(1-\lambda)(Ax-b)\cdot(Ay-b) = \lambda(1-\lambda)\big[f(x) + f(y) - \|A(x-y)\|^2\big]. \tag{12}$$

Summing (11) and (12), we obtain

$$f\big(\lambda x + (1-\lambda)y\big) = \lambda f(x) + (1-\lambda)f(y) - \lambda(1-\lambda)\|A(x-y)\|^2, \tag{13}$$

which implies that $f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda)f(y)$, and so f is convex.

Exercise 3.26. Let $f\colon \mathbb{R}^N \to \mathbb{R}$ be a convex function. Prove that, for all $k \geq 1$, for all $x_1, \dots, x_k \in \mathbb{R}^N$ and $\lambda_1, \dots, \lambda_k \in [0,1]$ satisfying $\lambda_1 + \cdots + \lambda_k = 1$, it holds that

$$f(\lambda_1 x_1 + \cdots + \lambda_k x_k) \leq \lambda_1 f(x_1) + \cdots + \lambda_k f(x_k).$$

Concerning the set of minimizers, we have the following:

Proposition 3.27. *Every sublevel set of a convex function is convex. In particular, if f is convex, then* $\operatorname{argmin}(f)$ *is convex.*

Proof. Let $\gamma \in \mathbb{R}$, let $x, y \in [f \leq \gamma]$, and let $\lambda \in (0,1)$. Then,

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

which shows that $\lambda x + (1-\lambda)y \in [f \leq \gamma]$. Since

$$\operatorname{argmin}(f) = \bigcap_{\gamma > \inf(f)} [f \leq \gamma],$$

Example 3.5 shows that $\operatorname{argmin}(f)$ is convex. □

Exercise 3.28. Let $f\colon \mathbb{R}^N \to \mathbb{R}$ be a convex function. Show that if f is bounded from above, then f must be constant. More generally, show that if $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is bounded from above on an affine subspace V of $\mathbb{R}^N$ (contained in D, of course), then f is constant on V.

Exercise 3.29. Let $f_i : D_i \subset \mathbb{R}^N \to \mathbb{R}$ be convex for $i \in \{1,2\}$, and let $a \geq 0$. Show that the function $f : D_1 \cap D_2 \to \mathbb{R}$, defined by $f = af_1 + f_2$, is convex.

Exercise 3.30. Let $f_i : D_i \subset \mathbb{R}^N \to \mathbb{R}$ be convex for each i in an index set $\mathcal{I}$, which may be finite or infinite. Set

$$D = \left\{ x \in \bigcap_{i \in \mathcal{I}} D_i : \sup_{i \in \mathcal{I}} f_i(x) < +\infty \right\}.$$

Prove that the function $f : D \to \mathbb{R}$, defined by $f(x) = \sup_{i \in \mathcal{I}} f_i(x)$, is convex.

Exercise 3.31. Let $D = [0, \infty)$, and consider a convex function $f : D \to \mathbb{R}$ such that $f(0) \leq 0$. Prove that f is *superadditive* on D, which means that, for every $x, y \in D$, one has

$$f(x) + f(y) \leq f(x+y).$$

Exercise 3.32. A function $f\colon \mathbb{R}^N \to \mathbb{R}$ is *midpoint-convex* if

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x) + f(y)}{2}$$

for all $x, y \in \mathbb{R}^N$. Prove that every continuous midpoint-convex function is convex.

The *epigraph* of a function $f : D \subset \mathbb{R}^N \to \mathbb{R}$ as

$$\operatorname{epi}(f) = \{(x,t) \in \mathbb{R}^N \times \mathbb{R} \colon t \geq f(x)\}.$$

Exercise 3.33. Show that $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is convex if, and only if, $\operatorname{epi}(f)$ is a convex subset of $\mathbb{R}^N \times \mathbb{R}$.

Proposition 3.34. *If $D \subset \mathbb{R}^N$ has nonempty interior, and f has at least one point of continuity, then* $\mathrm{epi}(f)$ *has nonempty interior.*

Proof. If f is continuous at $x_0 \in \mathrm{int}(D)$, there is $\delta > 0$ such that $|f(x)-f(x_0)| < 1$ for every $x \in B(x_0,\delta) \subset D$. Therefore, the open set $B(x_0,\delta) \times \big(f(x_0)+1,\infty\big)$ is nonempty and contained in $\mathrm{epi}(f)$. □

Exercise 3.35. Can you think of a convex function that is not continuous?

The previous exercise may result challenging in view of the following fact, whose proof is somewhat technical and will be omitted.

Proposition 3.36. *If $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ is convex, for each $x \in \mathrm{int}(D)$, there exist $L_x, r_x > 0$ such that*

$$|f(z) - f(y)| \le L_x \|z - y\|$$

for all $z, y \in B(x, r_x)$. In particular, f is continuous in $\mathrm{int}(D)$.

We invite the reader to revisit Exercise 3.35 under the light of Proposition 3.36.

The convexity of differentiable functions can be characterized in several ways, as shown in the following:

Proposition 3.37. *Let $D \subset \mathbb{R}^N$ be convex, and let $f\colon D \to \mathbb{R}$ be differentiable. The following are equivalent:*

i) *f is convex;*

ii) *for all $x, y \in D$, $f(y) \ge f(x) + \nabla f(x) \cdot (y - x)$;*

iii) *for all $x, y \in D$, $\big(\nabla f(y) - \nabla f(x)\big) \cdot (y - x) \ge 0$.*

If f is twice differentiable, the three statements above are equivalent to

iv) *for all $x \in D$, $\nabla^2 f(x)$ is positive semidefinite.*

Proof. By convexity,

$$f(\lambda y + (1-\lambda)x) \le \lambda f(y) + (1-\lambda) f(x)$$

for all $y \in X$ and all $\lambda \in (0,1)$. Rearranging the terms we get

$$\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \le f(y) - f(x).$$

As $\lambda \to 0$ we obtain *ii*). From *ii*), we immediately deduce *iii*). To prove that *iii*) ⇒ *i*), define $\phi : [0,1] \to \mathbb{R}$ by

$$\phi(\lambda) = f\left(\lambda x + (1-\lambda)y\right) - \lambda f(x) - (1-\lambda) f(y).$$

Then $\phi(0) = \phi(1) = 0$ and

$$\phi'(\lambda) = \left(\nabla f\left(\lambda x + (1-\lambda)y\right)\right) \cdot (x - y) - f(x) + f(y)$$

for $\lambda \in (0,1)$. Take $0 < \lambda_1 < \lambda_2 < 1$ and write $x_i = \lambda_i x + (1-\lambda_i)y$ for $i = 1,2$. A simple computation shows that

$$\phi'(\lambda_1) - \phi'(\lambda_2) = \frac{1}{\lambda_1 - \lambda_2} \left(\nabla f(x_1) - \nabla f(x_2)\right) \cdot (x_1 - x_2) \le 0.$$

In other words, ϕ' is nondecreasing. Since $\phi(0) = \phi(1) = 0$, there is $\bar{\lambda} \in (0,1)$ such that $\phi'(\bar{\lambda}) = 0$. Since ϕ' is nondecreasing, we must have $\phi' \leq 0$ (and so ϕ is nonincreasing) on $[0, \bar{\lambda}]$ and next $\phi' \geq 0$ (whence ϕ is nondecreasing) on $[\bar{\lambda}, 1]$. It follows that $\phi(\lambda) \leq 0$ on $[0,1]$, and so, f is convex.

Assume now that f is twice differentiable and let us prove that *iii)* $\Rightarrow$ *iv)* $\Rightarrow$ *i)*. For $t > 0$ and $v \in \mathbb{R}^N$, we have $(\nabla f(x + tv) - \nabla f(x)) \cdot (tv) \geq 0$. Dividing by t^2 and passing to the limit as $t \to 0$, we obtain $\big(\nabla^2 f(x)v\big) \cdot v \geq 0$. Finally, defining ϕ as above, we see that

$$\phi''(\lambda) = \left(\nabla^2 f\left(\lambda x + (1 - \lambda)y\right)(x - y)\right) \cdot (x - y) \geq 0.$$

It follows that ϕ' is nondecreasing and we conclude as before. □

The smoothness of convex functions can be characterized as follows:

Theorem 3.38 (Baillon-Haddad Theorem). *Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and differentiable, and let $L > 0$. The following statements are equivalent:*

i) *f is L-smooth;*

ii) *$f(z) \leq f(x) + \nabla f(x) \cdot (z - x) + \frac{L}{2}\|z - x\|^2$, for all $x, z \in \mathbb{R}^N$;*

iii) *$f(x) \geq f(y) + \nabla f(y) \cdot (x - y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$, for all $x, y \in \mathbb{R}^N$; and*

iv) *$(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$, for all $x, y \in \mathbb{R}^N$.*

Proof. We shall prove that *i)* $\Rightarrow$ *ii)* $\Rightarrow$ *iii)* $\Rightarrow$ *iv)* $\Rightarrow$ *i)*. The first implication is true, even if f is not convex (see Proposition 2.45). Suppose *ii)* holds. Subtract $\nabla f(y) \cdot z$ to both sides, write $h_y(z) = f(z) - \nabla f(y) \cdot z$ and rearrange the terms to obtain

$$h_y(z) \leq h_x(x) + (\nabla f(x) - \nabla f(y)) \cdot z + \frac{L}{2}\|z - x\|^2.$$

The function h_y is convex, differentiable and $\nabla h_y(y) = 0$. We deduce that $h_y(y) \leq h_y(z)$ for all $z \in \mathbb{R}^N$, and so

$$h_y(y) \leq h_x(x) + (\nabla f(x) - \nabla f(y)) \cdot z + \frac{L}{2}\|z - x\|^2. \tag{14}$$

Replace

$$z = x - \frac{1}{L}\big(\nabla f(x) - \nabla f(y)\big)$$

in (14) to obtain

$$h_y(y) \leq h_x(x) + (\nabla f(x) - \nabla f(y)) \cdot x - \frac{\|\nabla f(x) - \nabla f(y)\|^2}{2L},$$

which is precisely *iii)*. Interchanging the roles of x and y and adding the resulting inequality, we obtain *iv)*. The last implication is straightforward. □

Exercise 3.39. Let $f \colon \mathbb{R}^N \to \mathbb{R}$ be convex and L-smooth, and suppose $\operatorname{argmin}(f) \neq \emptyset$. Prove that, for all $x \in \mathbb{R}^N$, one has

$$f(x) - \min f \geq \frac{1}{2L}\|\nabla f(x)\|^2.$$

3.4 Strict and strong convexity

A function $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ is *strictly convex* if D is convex and

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

for every $\lambda \in (0,1)$ and every $x, y \in D$ such that $x \neq y$.

Exercise 3.40. State and prove a characterization result for strict convexity, in line with Proposition 3.37.

Proposition 3.41. *A strictly convex function can have, at most, one minimizer.*

Proof. Let f be strictly convex and let $x, y \in \operatorname{argmin}(f)$. If $x \neq y$, the strict convexity gives

$$f\left(\frac{1}{2}x + \frac{1}{2}y\right) < \frac{1}{2}f(x) + \frac{1}{2}f(y) = \frac{1}{2}\min(f) + \frac{1}{2}\min(f) = \min(f),$$

which is impossible. □

Given $\mu > 0$, a function $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is *strongly convex with parameter* μ, or μ*-strongly convex*, if D is convex and

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2$$

for every $x, y \in D$ and all $\lambda \in (0,1)$.

Exercise 3.42. Prove that f is μ-strongly convex if, and only if, $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

It is easy to see that, if f is convex and g is strictly (resp. strongly) convex, then $f + g$ is strictly (resp. strongly) convex.

Exercise 3.43. Show that strongly convex functions cannot be Lipschitz continuous.

Exercise 3.44. Show that the robust SVM from Section 1.3.2 has a unique solution.

The strong convexity of a function f can also be characterized in terms of properties of ∇f and $\nabla^2 f$, as in Proposition 3.37.

Proposition 3.45. *Let $f\colon D \subset \mathbb{R}^N \to \mathbb{R}$ be differentiable. The following are equivalent:*

i) f is μ-strongly convex;

ii) for all $x, y \in D$, $f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\mu}{2}\|x - y\|^2$;

iii) for all $x, y \in D$, $\big(\nabla f(y) - \nabla f(x)\big) \cdot (y - x) \geq \mu\|x - y\|^2$.

If f is twice differentiable, the three statements above are equivalent to

iv) for all $x \in D$, $\nabla^2 f(x)$ is positive definite, with $v \cdot \nabla^2 f(x) v \geq \mu\|v\|^2$ for every $v \in \mathbb{R}^N$.

Exercise 3.46. Prove Proposition 3.45.

Exercise 3.47. Determine the values of $p \in \mathbb{R}$ for which the expression $f(x) = x^p$ represents a strictly/strongly convex function on $(0, \infty)$.

Exercise 3.48. When is the function $f(x) = \frac{1}{2}\|Ax - b\|^2$ strictly/strongly convex?

Exercise 3.49. Give an example of a function that is convex, but not strictly convex; and of a function that is strictly convex, but not strongly convex.

Exercise 3.50. Show that every strongly convex function $f : \mathbb{R}^N \to \mathbb{R}$ is coercive.

For functions which are both smooth and strongly convex, we can obtain a reinforced version of Theorem 3.38, which is stronger than its simple combination with Proposition 3.45.

Proposition 3.51. *If $f : \mathbb{R}^N \to \mathbb{R}$ is L-smooth and μ-strongly convex, then*

$$(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \frac{\mu L}{L+\mu}\|x-y\|^2 + \frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|^2 \tag{15}$$

for all $x, y \in \mathbb{R}^N$.

Proof. Since f is μ-strongly convex, the function g, defined by $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex (see Exercise 3.42). We also know that $\nabla g(x) = \nabla f(x) - \mu x$. Next, we have

$$\begin{aligned} g(y) &= f(y) - \frac{\mu}{2}\|y\|^2 \\ &\leq f(x) + \nabla f(x) \cdot (y - x) + \frac{L}{2}\|y - x\|^2 - \frac{\mu}{2}\left(\|x\|^2 + \|y - x\|^2 + 2x \cdot (y - x)\right) \\ &= f(x) - \frac{\mu}{2}\|x\|^2 + (\nabla f(x) - \mu x) \cdot (y - x) + \frac{L-\mu}{2}\|y - x\|^2 \\ &= g(x) + \nabla g(x) \cdot (y - x) + \frac{L-\mu}{2}\|y - x\|^2. \end{aligned}$$

Since g is convex, we may use the fact that *ii*) $\Rightarrow$ *iv*) in Theorem 3.38, to conclude that

$$(\nabla g(x) - \nabla g(y)) \cdot (x - y) \geq \frac{1}{L-\mu}\|\nabla g(x) - \nabla g(y)\|^2.$$

Therefore,

$$\begin{aligned} (\nabla f(x) - \nabla f(y)) \cdot (x - y) &= (\nabla g(x) - \nabla g(y)) \cdot (x - y) + \mu\|y - x\|^2 \\ &\geq \frac{1}{L-\mu}\|\nabla g(x) - \nabla g(y)\|^2 + \mu\|y - x\|^2 \\ &= \frac{1}{L-\mu}\|\nabla f(x) - \nabla f(y) - \mu(x - y)\|^2 + \mu\|y - x\|^2 \\ &= \frac{1}{L-\mu}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{\mu L}{L-\mu}\|y - x\|^2 \\ &\quad -\frac{2\mu}{L-\mu}(\nabla f(x) - \nabla f(y)) \cdot (x - y). \end{aligned}$$

Multiplying by $L - \mu$ and rearranging the terms, we obtain

$$(L+\mu)(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \|\nabla f(x) - \nabla f(y)\|^2 + \mu L\|y - x\|^2,$$

which is equivalent to (15). □

4 Numerical approach to smooth unconstrained minimization

This section provides an overview of the main numerical methods that can be used to solve–exactly or approximately–unconstrained minimization problems when the objective function is smooth. We begin by introducing some notions and terminology around the concept of iterative algorithms in 4.1. The Gradient Method is a timeless representative of these procedures, not only because of its simplicity, but also because it is a crucial building block for more sophisticated methods used today. We shall analyze it in depth in 4.2. The Gradient Method comes with a natural question concerning the step-size selection. This is discussed in 4.3. Higher-order methods, both in time and in space, are considered in 2.7. Finally, we discuss the Conjugate-Gradient Method, particularly useful for minimizating quadratic functions, in 4.5.

4.1 Iterative algorithms

An *iterative algorithm* is a procedure that computes a sequence (x_k) of points in $\mathbb{R}^N$ that approximate a solution to a problem. It typically involves

- An *initial guess* x_0,
- Possibly a sequence (u_k) in $\mathbb{R}^M$ of *parameters*, either defined by the user or computed by a subroutine,
- An operator $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$ used to compute x_{k+1}, given x_k, as

$$x_{k+1} = F(x_k, u_k),$$

- A *stopping rule* that is activated and terminates the algorithm, when the approximation is considered *sufficiently good*, according to a criterion defined by the user.

Two important issues concerning the behavior of the algorithm, in connection with the problem it is meant to solve, are its *convergence*, which is whether or not the algorithm is able to produce a sufficiently good approximate solution, and its *complexity*, which has to do with how much of a given resource (number of iterations, computational time) it needs in order to produce such outcome.

Let $f : D \subset \mathbb{R}^N \to \mathbb{R}$. For the purpose of this course, we shall say that $\Phi : \mathbb{R}^N \to \mathbb{R}_+$ is a *merit function* for the problem of minimizing f if

$$\Phi(\hat{x}) = 0 \qquad \Longleftrightarrow \qquad \hat{x} \in \operatorname{argmin}(f),$$

and we consider $z \in \mathbb{R}^N$ to be a better approximate solution than $w \in \mathbb{R}^N$ if, and only if, $\Phi(z) < \Phi(w)$. This specifically implies that $\Phi(z) > 0$ for all $z \notin \operatorname{argmin}(f)$. We may interpret them as measures of optimality.

Example 4.1. Some commonly used merit functions are:

1. $\Phi(z) = f(z) - \inf(f)$.
2. $\Phi(z) = \operatorname{dist}\big(z, \operatorname{argmin}(f)\big)$ or $\Phi(z) = \operatorname{dist}\big(z, \operatorname{argmin}(f)\big)^2$.
3. $\Phi(z) = \|z - \hat{x}\|$ or $\Phi(z) = \|z - \hat{x}\|^2$, if $\operatorname{argmin}(f) = \{\hat{x}\}$.
4. $\Phi(z) = \|\nabla f(z)\|$ or $\Phi(z) = \|\nabla f(z)\|^2$, if f is convex and differentiable.

Merit functions serve multiple purposes. On the one hand, they are used in order to prove the convergence of an algorithm or establish theoretical performance guarantees. In other words, if an algorithm produces a sequence (x_k), we would like to determine whether $\lim\limits_{k\to\infty} \Phi(x_k) = 0$. On the other hand, when carrying out a practical implementation, we use merit function to decide when to accept the output of an algorithm as a valid approximate solution. More precisely, if the k-th iteration of the algorithm satisfies $\Phi(x_k) < \varepsilon$, the process terminates and x_k is reported as solution.

A *convergence rate* is a quantitative description of how fast a nonnegative sequence (ϕ_k) converges to zero. One standard way to express it is by comparison with another sequence that we take as a reference, such as a power or an exponential of k. Let (ρ_k) be positive, with $\lim\limits_{k\to\infty} \rho_k = 0$. We shall write $\phi_k = \mathcal{O}(\rho_k)$ if $\phi_k \to 0$ *at least as fast as* ρ_k, namely if

$$C := \sup_{k\geq 0} \left[\frac{\phi_k}{\rho_k}\right] < +\infty.$$

Example 4.2. It is easy to see that $\frac{1}{k^2+1} = \mathcal{O}\left(\frac{1}{k}\right)$, also $\frac{1}{k^2+1} = \mathcal{O}\left(\frac{1}{k^2}\right)$, but $\frac{1}{k^2+1} \neq \mathcal{O}\left(\frac{1}{k^3}\right)$.

On the other hand, the notation $\phi_k = o(\rho_k)$ means that $\phi_k \to 0$ *strictly faster than* ρ_k:

$$\lim_{k\to\infty} \left[\frac{\phi_k}{\rho_k}\right] = 0.$$

Example 4.3. We have $\frac{1}{k^2+1} = o\left(\frac{1}{k}\right)$, but $\frac{1}{k^2+1} \neq o\left(\frac{1}{k^2}\right)$.

Remark 4.4. Clearly, $\phi_k = o(\rho_k)$ implies $\phi_k = \mathcal{O}(\rho_k)$, but the constant C is important!

By extension, we define the *convergence rate* of a sequence (x_k) in $\mathbb{R}^N$ to a point $\hat{x} \in \mathbb{R}^N$ as the convergence rate of $\|x_k - \hat{x}\|$ to zero.

Comparison with a power sequence is particularly common. If $\phi_k = \mathcal{O}(q^k)$ for some $q \in (0,1)$, we say ϕ_k *converges linearly to zero with rate* q, or that it *converges* q*−linearly to zero.*

If $\phi_k \neq \mathcal{O}(q^k)$ for every $q \in (0,1)$, ϕ_k converges *sublinearly.* In many frequent cases, we still have $\phi_k = \mathcal{O}(k^{-p})$ for some $p > 0$. Finally, if $\phi_k = o(q^k)$ for all $q \in (0,1)$, then ϕ_k converges *superlinearly.* We shall see that this is the case for Newton's method, under favorable conditions.

Rates of convergence can also be expressed without recurring to a reference sequence, but by comparing successive iterates, instead. We say (ϕ_k) converges to zero with order $p > 0$ and rate $R \in (0,1)$, *in the sense of the ratio test*, if

$$\limsup_{k\to\infty} \left[\frac{\phi_{k+1}}{\phi_k^p}\right] = R.$$

Remark 4.5. If (ϕ_k) converges to zero with order 1 and rate $\rho \in (0,1)$, in the sense of the ratio test, it is easy to prove that (ϕ_k) converges ρ-linearly to zero. Some authors use this stronger notion as the definition of *linear convergence*, so we advise caution.

Exercise 4.6. If (ϕ_k) converges q−linearly to zero, can we say that it converges to zero with order 1 and rate q, in the sense of the ratio test?

If (ϕ_k) converges to zero with order 2, in the sense of the ratio test, we say it converges *quadratically* to zero.

Consider a problem with merit function Φ and a sequence (x_k) produced by an algorithm. Suppose that for every $\varepsilon > 0$, which represents a *tolerance* or *precision level*, the set

$$\Sigma_\varepsilon := \{k : \Phi(x_k) < \varepsilon\}$$

of *ε-approximate solutions* is nonempty. A *complexity bound* is description, estimation or approximation of the set Σ_ε. In more practical terms, we can express a complexity bound by stating that $\Phi(x_k) < \varepsilon$ for every $k \geq k_\varepsilon$, where k_ε is a number that can be computed as a function of ε.

Example 4.7. Complexity bounds can be obtained from convergence rates, as described in the following common cases:

1. Let $C, p > 0$, and suppose $\Phi(x_k) \leq C/k^p$ for all $k \geq 1$. Then, $\Phi(x_k) < \varepsilon$ for every $k > \sqrt[p]{C/\varepsilon}$.
2. Now, let $A, B > 0$. If $\Phi(x_k) \leq Ae^{-Bk}$ for all $k \geq 1$, then $\Phi(x_k) < \varepsilon$ for all $k > \ln\left(\sqrt[B]{A/\varepsilon}\right)$.

4.2 Descent directions and the Gradient Method

Let $D \subset \mathbb{R}^N$ be nonempty and open, and let $f : D \to \mathbb{R}$. A vector $d \neq 0$ is a *descent direction* for f at x if there is $\Gamma_d > 0$ such that

$$f(x + \gamma d) < f(x) \qquad \text{for all } \gamma \in (0, \Gamma_d).$$

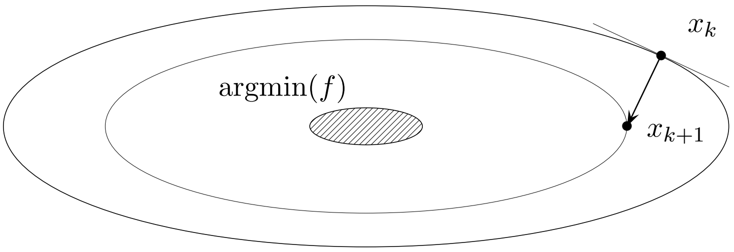
The numbers $\gamma \in (0, \Gamma_d)$ are *descent step sizes*. As shown in Proposition 2.36 (see also Remark 2.37), if f is differentiable at x and $\nabla f(x) \neq 0$, then $-\nabla f(x)$ is the *steepest descent direction* for f at x.

Exercise 4.8. Show that, if f is differentiable at x_0 and $\nabla f(x_0) \neq 0$, then every vector $d \neq 0$, satisfying $\nabla f(x_0) \cdot d < 0$, is a descent direction for f at x.

Descent methods consist in iterating a rule of the form

$$x_{k+1} = x_k + \gamma_k d_k, \quad k \geq 0, \tag{16}$$

where d_k is a descent direction and γ_k is a descent step size. Perhaps the most obvious choice is $d_k = -\nabla f(x_k)$. If f is L-smooth, then every $\gamma \in \left(0, \frac{2}{L}\right)$ is a descent step size.



4.2.1 *Vanilla* Gradient Method

Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth, let $x_0 \in \mathbb{R}^N$, and let $\gamma \in \left(0, \frac{2}{L}\right)$. The iterative algorithm defined by

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

is known as the *Gradient Method*. According to Remark 2.46, we have

$$f\big(x_{k+1}\big) \le f(x_k) + \gamma\left(\frac{\gamma L}{2} - 1\right)\|\nabla f(x_k)\|^2 \le f(x_k). \tag{17}$$

Theorem 4.9. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and bounded from below, let $\gamma \in \big(0, \frac{2}{L}\big)$, and let $x_0 \in \mathbb{R}^N$. For $k \ge 0$, iterate $x_{k+1} = x_k - \gamma\nabla f(x_k)$. Then, $\lim\limits_{k\to\infty} f(x_k)$ exists, $\lim\limits_{k\to\infty} \|\nabla f(x_k)\| = 0$ and, for every $k \ge 0$, we have*

$$\min\big\{\|\nabla f(x_j)\| : 0 \le j \le k\big\} \le \sqrt{\frac{2\big[f(x_0) - \inf(f)\big]}{\gamma(2-\gamma L)(k+1)}}.$$

Moreover, all subsequential limit points are critical, which means that if $x_{m_k} \to \hat{x}$, then $\nabla f(\hat{x}) = 0$. As a consequence, if f has no critical points, then $\lim\limits_{k\to\infty} \|x_k\| = +\infty$.

Proof. By (17), the sequence $\big(f(x_k)\big)$ is nonincreasing[1]. Since it is also bounded from below, $\lim\limits_{k\to\infty} f(x_k)$ exists. Moreover, for every $k \ge 0$, we have

$$\gamma\left(1 - \frac{\gamma L}{2}\right)\sum_{j=0}^{k}\|\nabla f(x_j)\|^2 \le f(x_0) - f(x_{k+1}) \le f(x_0) - \inf(f). \tag{18}$$

This shows the convergence of the series on the left-hand side, which implies that $\lim\limits_{k\to\infty} \|\nabla f(x_k)\| = 0$, and

$$(k+1)\min\big\{\|\nabla f(x_j)\|^2 : 0 \le j \le k\big\} \le \frac{2\big[f(x_0) - \inf(f)\big]}{\gamma(2-\gamma L)}.$$

The continuity of ∇f then gives that limit points are critical. The final consequence follows immediately, by contrapositive. □

Exercise 4.10. Suppose we use a *variable* step size in the gradient method, and iterate $x_{k+1} = x_k - \gamma_k\nabla f(x_k)$, for $k \ge 0$. Verify that the conclusions of Theorem 4.9 remain valid if $\inf_{k\ge 0}\gamma_k(2-\gamma_k L) > 0$.

4.2.2 The effect of geometry: (strong) convexity and the Polyak-Łojasiewicz inequality

When the objective function is convex, its critical points are minimizers. Also, it is possible to establish the convergence and minimization property of the sequences generated by the gradient method. More precisely, we have the following:

Theorem 4.11. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth, convex and bounded from below, let $\gamma \in \big(0, \frac{2}{L}\big)$, and let $x_0 \in \mathbb{R}^N$. For $k \ge 0$, iterate $x_{k+1} = x_k - \gamma\nabla f(x_k)$. Then,*

i) $\lim\limits_{k\to\infty} f(x_k) = \inf(f)$.

ii) If f has minimizers, then x_k converges to one of them, with

$$f(x_k) - \min(f) \le \frac{\operatorname{dist}\big(x_0, \operatorname{argmin}(f)\big)^2}{\gamma(2-\gamma L)(k+1)} \qquad \textit{and} \qquad \lim_{k\to\infty} k\,\big[f(x_k) - \min(f)\big] = 0.$$

Proof. By Proposition 3.37, for every $u \in \mathbb{R}^N$ and $k \ge 0$, we have

$$f(u) \ge f(x_k) + \nabla f(x_k)\cdot(u - x_k). \tag{19}$$

[1] Strictly decreasing as long as $\nabla f(x_k) \ne 0$!

Since $\gamma\nabla f(x_k) = x_k - x_{k+1}$, we can write

$$2\gamma\big(f(x_k) - f(u)\big) \le 2(x_k - x_{k+1})\cdot(x_k - u) = \|x_k - x_{k+1}\|^2 + \|x_k - u\|^2 - \|x_{k+1} - u\|^2,$$

where we have used the identity $2\,v\cdot w = \|v\|^2 + \|w\|^2 - \|v - w\|^2$, with $v = x_k - x_{k+1}$ and $w = x_k - u$. On the other hand, (17) shows that

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

It follows that

$$2\gamma\big(f(x_k) - f(u)\big) \le \frac{2\gamma}{2-\gamma L}\big(f(x_k) - f(x_{k+1})\big) + \|x_k - u\|^2 - \|x_{k+1} - u\|^2. \tag{20}$$

Since the left-hand side is nonincreasing, we deduce that

$$2\gamma(k+1)\big(f(x_k) - f(u)\big) \le 2\gamma\sum_{j=0}^{k}\big(f(x_j) - f(u)\big) \le \frac{2\gamma}{2-\gamma L}\big(f(x_0) - \inf(f)\big) + \|x_0 - u\|^2.$$

Dividing by $2\gamma(k+1)$, and letting $k\to\infty$, we conclude that $\lim\limits_{k\to\infty} f(x_k) \le f(u)$ for every $u\in\mathbb{R}^N$, and i) holds. For ii), take any $\hat{x}\in\operatorname{argmin}(f)$ and use (19) with $u = \hat{x}$, to write

$$\begin{aligned} 2\gamma\big(f(x_k) - f(\hat{x})\big) &\le 2\gamma\nabla f(x_k)\cdot(x_k - \hat{x}) \\ &= 2(x_k - x_{k+1})\cdot(x_k - \hat{x}) \\ &= \|x_k - x_{k+1}\|^2 + \|x_k - \hat{x}\|^2 - \|x_{k+1} - \hat{x}\|^2 \\ &= \gamma^2\|\nabla f(x_k)\|^2 + \|x_k - \hat{x}\|^2 - \|x_{k+1} - \hat{x}\|^2. \end{aligned} \tag{21}$$

In turn, the cocoercivity of ∇f (Theorem 3.38) gives

$$\frac{2\gamma}{L}\|\nabla f(x_k)\|^2 \le 2\gamma\nabla f(x_k)\cdot(x_k - \hat{x}) = \gamma^2\|\nabla f(x_k)\|^2 + \|x_k - \hat{x}\|^2 - \|x_{k+1} - \hat{x}\|^2,$$

which shows that

$$0 \le \frac{\gamma}{L}(2-\gamma L)\|\nabla f(x_k)\|^2 \le \|x_k - \hat{x}\|^2 - \|x_{k+1} - \hat{x}\|^2. \tag{22}$$

On the one hand, this implies that $k\mapsto\|x_k - \hat{x}\|$ is nonincreasing, whence (x_k) is bounded and $\lim\limits_{k\to\infty}\|x_k - \hat{x}\|$ exists. On the other, combining (21) and (22), we obtain

$$2\gamma\big(f(x_k) - \min(f)\big) \le \left[\frac{\gamma L}{2-\gamma L} + 1\right]\left(\|x_k - \hat{x}\|^2 - \|x_{k+1} - \hat{x}\|^2\right).$$

Since $k\mapsto f(x_k)$ is nonincreasing, we deduce that

$$(k+1)\big(f(x_k) - \min(f)\big) \le \sum_{j=0}^{k}\big(f(x_j) - \min(f)\big) \le \frac{\|x_0 - \hat{x}\|^2}{\gamma(2-\gamma L)}. \tag{23}$$

Since this holds for every $\hat{x}\in\operatorname{argmin}(f)$, we can take $\hat{x} = P_{\operatorname{argmin}(f)}x_0$ to conclude that

$$f(x_k) - \min(f) \le \frac{\operatorname{dist}\big(x_0, \operatorname{argmin}(f)\big)^2}{\gamma(2-\gamma L)(k+1)}.$$

Since (x_k) is bounded, it has a subsequence, say (x_{m_k}), that converges to some x_∞. By continuity,

$$f(x_\infty) = \lim_{k\to\infty} f(x_{m_k}) = \lim_{k\to\infty} f(x_k) = \min(f),$$

so $x_\infty \in \operatorname{argmin}(f)$. This also implies that

$$0 = \lim_{k\to\infty} \|x_{m_k} - x_\infty\| = \lim_{k\to\infty} \|x_k - x_\infty\|,$$

because the latter exists. We conclude that the whole sequence (x_k) converges to $x_\infty \in \operatorname{argmin}(f)$. It only remains to show that $\lim\limits_{k\to\infty} k\big(f(x_k) - \min(f)\big) = 0$. By (23), the nonincreasing sequence $f(x_k) - \min(f)$ is summable, so the result follows from Lemma 4.13 below. □

Lemma 4.12. *Let (a_k), (b_k) and (ε_k) be nonnegative sequences such that (ε_k) is summable and*

$$a_{k+1} - a_k + b_k \le \varepsilon_{k+1}$$

for every $k \ge 0$. Then, $\lim\limits_{k\to\infty} a_k$ exists and (b_k) is summable.

Proof. First, write

$$A_k = a_k + \sum_{j=0}^{k} \varepsilon_j,$$

so that

$$A_{k+1} \le A_{k+1} + b_k \le A_k$$

for every $k \ge 0$. Since (A_k) is then nonnegative and nonincreasing, it has a limit A. Since (ε_k) is summable,

$$\lim_{k\to\infty} a_k = A - \sum_{j=0}^{\infty} \varepsilon_j.$$

Finally, $\sum_{j=0}^{k} b_j \le A_0 - A_{k+1} \le A_0$, whence (b_k) is summable. □

Lemma 4.13. *Let (c_k) be nonnegative, nonincreasing and summable. Then $\lim\limits_{k\to\infty} kc_k = 0$.*

Proof. Since (c_k) is nonincreasing, we have

$$(k+1)c_{k+1} - kc_k = k(c_{k+1} - c_k) + c_{k+1} \le c_{k+1}.$$

Since (c_k) is summable, Lemma 4.12 (with $a_k = kc_k$, $b_k = 0$ and $\varepsilon_k = c_k$) shows that $\lim\limits_{k\to\infty} kc_k$ exists. But, the summability of (c_k) also implies that

$$\sum_{k=1}^{\infty} \frac{1}{k}(kc_k) = \sum_{k=1}^{\infty} c_k < \infty.$$

If $\lim\limits_{k\to\infty} kc_k > 0$, this would imply that the harmonic series $\sum_{k=1}^{\infty} \frac{1}{k}$ is convergent, which is not true. □

Remark 4.14. In Theorem 4.11, the best theoretical convergence rate is achieved when $\gamma = \frac{1}{L}$.

Exercise 4.15. In the context of Theorem 4.11, show that, if $\gamma \leq \frac{1}{L}$, then

$$f(x_k) - \min(f) \leq \frac{\operatorname{dist}\left(x_0, \operatorname{argmin}(f)\right)^2}{2\gamma k},$$

for all $k \geq 1$. How does this convergence rate compare to the one established in Theorem 4.11?

Under strong convexity, the sequence of values converges linearly to the optimum, as shown in the following:

Theorem 4.16. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and μ-strongly convex, let $\gamma \in \left(0, \frac{2}{L+\mu}\right]$, and let $x_0 \in \mathbb{R}^N$. For $k \geq 0$, iterate $x_{k+1} = x_k - \gamma \nabla f(x_k)$. Then,*

$$\|x_k - \hat{x}\| \leq \left(1 - \frac{2\gamma\mu L}{L+\mu}\right)^{k/2} \|x_0 - \hat{x}\|,$$

where $\hat{x}$ is the unique minimizer of f. In particular, taking $\gamma = \frac{2}{L+\mu}$, we obtain

$$\|x_k - \hat{x}\| \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|x_0 - \hat{x}\|.$$

Proof. Use Proposition 3.51, with $x = x_k$ and $y = \hat{x}$ to obtain

$$\nabla f(x_k) \cdot (x_k - \hat{x}) \geq \frac{\mu L}{L+\mu}\|x_k - \hat{x}\|^2 + \frac{1}{L+\mu}\|\nabla f(x_k)\|^2.$$

Multiply this by 2γ and rewrite the dot product as we did in (21), to get

$$\frac{2\gamma\mu L}{L+\mu}\|x_k - \hat{x}\|^2 + \frac{2\gamma}{L+\mu}\|\nabla f(x_k)\|^2 \leq \gamma^2\|\nabla f(x_k)\|^2 + \|x_k - \hat{x}\|^2 - \|x_{k+1} - \hat{x}\|^2,$$

which is equivalent to

$$\|x_{k+1} - \hat{x}\|^2 \leq \gamma\left(\gamma - \frac{2}{L+\mu}\right)\|\nabla f(x_k)\|^2 + \left(1 - \frac{2\gamma\mu L}{L+\mu}\right)\|x_k - \hat{x}\|^2 \leq \left(1 - \frac{2\gamma\mu L}{L+\mu}\right)\|x_k - \hat{x}\|^2,$$

as claimed. □

Exercise 4.17. In the context of Theorem 4.16, can you prove linear convergence of x_k to x^* if $\gamma \in \left(\frac{2}{L+\mu}, \frac{2}{L}\right)$?

Exercise 4.18. In the context of Theorem 4.16 and Exercise 4.17, can you establish convergence rates for $f(x_k) - \min(f)$?

Strong convexity is a restrictive assumption that can only be fulfilled by functions with a unique minimizer. We now present a weaker condition on the objective function that still ensures linear convergence of the gradient method.

A function $f : \mathbb{R}^N \to \mathbb{R}$ satisfies the *Polyak-Łojasiewicz* (PŁ) *Inequality* with constant $\mu > 0$ if

$$2\mu\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2 \tag{24}$$

for all $x \in \mathbb{R}^N$. Strongly convex functions satisfy the PŁ Inequality, as shown in the following:

Proposition 4.19. *If f is μ-strongly convex, it satisfies the PŁ Inequality with constant μ.*

Proof. Let $\hat{x}$ be the unique minimizer of f. From Proposition 3.45, we know that

$$f(x) - \min(f) \leq \nabla f(x) \cdot (x - \hat{x}) - \frac{\mu}{2}\|x - \hat{x}\|^2 \leq \frac{1}{2t}\|\nabla f(x)\|^2 + \frac{t}{2}\|x - \hat{x}\|^2 - \frac{\mu}{2}\|x - \hat{x}\|^2,$$

where we have used (3). It suffices to set $t = \mu$ to obtain (24). □

However, unlike strongly convex functions, a function satisfying the PŁ inequality may have multiple minimizers.

Exercise 4.20. Show that $f(x) = \frac{1}{2}\|Ax - b\|^2$ satisfies a PŁ Inequality, even if $\ker(A)$ is nontrivial. The constant μ is the smallest positive eigenvalue of the matrix $A^T A$.

Proposition 4.21. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be $L-$smooth and satisfy the Polyak-Łojasiewicz inequality with constant $\mu > 0$. Pick $\gamma \in \left(0, \frac{2}{L}\right)$, and iterate $x_{k+1} = x_k - \gamma \nabla f(x_k)$. Then, $f(x_k)$ converges linearly to* $\min(f)$*, with*

$$f(x_k) - \min(f) \leq \big(1 - \mu\gamma(2 - \gamma L)\big)^k \big(f(x_0) - \min(f)\big),$$

for all $k \geq 1$. In particular, for $\gamma = \frac{1}{L}$, we obtain

$$f(x_k) - \min(f) \leq \left(1 - \frac{\mu}{L}\right)^k \big(f(x_0) - \min(f)\big).$$

Proof. Using (17) and (24), we get

$$f(x_{k+1}) - f(x_k) \leq \gamma\left(\frac{\gamma L}{2} - 1\right)\|\nabla f(x_k)\|^2 \leq 2\mu\gamma\left(\frac{\gamma L}{2} - 1\right)\big(f(x_k) - \min(f)\big).$$

Adding and subtracting $\min(f)$ on the left-hand side, and then rearranging the terms, we are left with

$$f(x_{k+1}) - \min(f) \leq \big(1 - \mu\gamma(2 - \gamma L)\big)\big(f(x_k) - \min(f)\big).$$

It suffices to iterate this inequality from 0 to $k - 1$ to conclude. □

Notice the difference and similarity with Theorem 4.16.

4.3 Step size selection

Choosing descent step sizes is a relevant and subtle issue to consider when implementing descent methods. When chosen too large, the corresponding sequence can wander around, sometimes in a chaotic manner. If they are taken too short, convergence will be slow. In what follows, we discuss several options.

Fixed **or** *constant* **step sizes.** As explained in 4.2, if the objective function f is L-smooth, one can fix any $\gamma \in (0, 2/L)$ (the best theoretical choice being $\gamma = \frac{1}{L}$), and define $\gamma_k \equiv \gamma$. This strategy does however require the knowledge, or at least a good estimation, of the constant L, which is not always easily accessible. If one is running a large number of runs on similar problems, one can get a feeling about the order of magnitude of L by trial and error.

Vanishing **step sizes.** Alternatively, one can start with a large value of γ, and gradually decrease it: no matter how large L is, if $\lim_{k\to\infty} \gamma_k = 0$, the step sizes will eventually fall in the correct interval. Nevertheless, if γ_k tends to 0 too fast, the algorithm may converge slowly, or even fail to converge.

Example 4.22. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = \frac{1}{2}x^2$. The function f is both 1-smooth and 1-strongly convex. Given a sequence (γ_k) of step sizes converging to 0, the Gradient Method gives

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k) = (1-\gamma_k)x_k = \cdots = \left[\prod_{j=0}^{k}(1-\gamma_j)\right] x_0.$$

To simplify the ideas, assume $x_0 > 0$ and $\gamma_k < 1$ for all k. Then,

$$\ln\left(\frac{x_{k+1}}{x_0}\right) = \sum_{j=0}^{k} \ln(1-\gamma_j) = \sum_{j=0}^{k} \frac{\ln(1-\gamma_j)}{\gamma_j}\gamma_j. \tag{25}$$

Since $\lim\limits_{\tau\to 0} \frac{\ln(1-\tau)}{\tau} = -1$, the series on the right-hand side converges to a negative number if $\sum_{j=0}^{k} \gamma_j < \infty$, and diverges to $-\infty$ otherwise. If the series in (25) converges to S, then $\lim\limits_{k\to\infty} x_k = x_0 e^S > 0$, and so the limit *is not* a minimizer of f. On the other hand, if the said series is divergent, then $\lim\limits_{k\to\infty} x_k = 0$, which is the unique minimizer of f.

The preceding example shows that, if one wishes to establish a general convergence result for the Gradient Method with vanishing step sizes, which is applicable to functions that are either smooth or (strongly) convex, the non-summability of the step sizes should be amog the hypotheses. Indeed, we can prove the following:

Theorem 4.23. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and bounded from below, let (γ_k) be positive sequence such that $\lim\limits_{k\to\infty} \gamma_k = 0$ and $\sum_{k\geq 0} \gamma_k = \infty$. Starting from $x_0 \in \mathbb{R}^N$, iterate $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$ for $k \geq 0$. Then, $\lim\limits_{k\to\infty} f(x_k)$ exists, $\lim\limits_{k\to\infty} \|\nabla f(x_k)\| = 0$ and every subsequential limit point of (x_k) is critical.*

4.3.1 Line search

Line search strategies select a descent step size once a descent direction has been chosen. At each iteration, a decision is made, and the current step size is chosen based on the available information. In this sense, they are *feedback* strategies.

A *greedy* strategy. A natural aim is to use the step size that gives the smallest function value along the ray defined by the descent direction, namely:

$$\gamma_k := \operatorname{argmin}_{\gamma>0} f(x_k + \gamma d_k), \tag{26}$$

which is known as the *greedy* or *exact* line search strategy. Although this choice of γ_k is not necessarily in $(0, 2/L)$, it achieves at least as much reduction as any choice in that interval, in terms of Inequality (17), and hence the conclusions of Theorem 4.9 still hold.

Exercise 4.24. What about Theorem 4.11?

The subproblem given by (26) is, in principle, easier than the original minimization problem, as we are optimizing over $(0, \infty)$ instead of all of $\mathbb{R}^N$. Now, depending on the nature of the function f, the subproblem (26) might have an analytic solution, but this is rare. One must therefore consider the associated cost added to each iteration. A variant of this strategy is to compute the minimum over a bounded interval defined beforehand:

$$\gamma_k = \operatorname{argmin}_{\gamma\in(0,\bar{\gamma}]} f(x_k + \gamma d_k). \tag{27}$$

Implementable line search strategies. Rather than solving the subproblems (26) or (27) analytically, it is common to apply algorithms to find approximate solutions, which will serve as sufficiently good step sizes. First, one fixes some initial guess $\bar{\gamma} > 0$ that defines the part of the ray that will be tested, along with a shrinking factor $\beta \in (0,1)$. At the k-th iteration, one has computed the current iterate x_k, and selected a descent direction d_k. Then, one defines the step size as

$$\gamma_k = \beta^{m_k}\,\bar{\gamma},$$

where m_k is the smallest nonnegative integer m such that $\gamma = \beta^m\,\bar{\gamma}$ satisfies a given set of conditions. The process of finding m_k is usually referred to as *backtracking*, since it involves testing a larger step size $\bar{\gamma}$ first, and eventually shrinking it back to smaller values $\beta\,\bar{\gamma}$, $\beta^2\,\bar{\gamma}$, and so on. The most popular of such conditions is *Armijo's Rule*

$$f(x_k + \gamma d_k) \leq f(x_k) + \sigma\,\gamma\,\nabla f(x_k) \cdot d_k, \tag{28}$$

where $\sigma \in (0,1)$ is a number selected beforehand, typically of the order of 10^{-4}. Armijo's Rule can be used on its own, but some user combine it with the following reinforcements, which add a supplementary requirement to (28):

- The oldest one is *Goldstein's Rule*, which adds

  $$f(x_k + \gamma d_k) \geq f(x_k) + (1-\sigma)\gamma\nabla f(x_k) \cdot d_k. \tag{29}$$

  Inequalities (28) and (29) can be expressed together in a more succinct way as

  $$\sigma \leq \frac{f(x_k + \gamma d_k) - f(x_k)}{\gamma\nabla f(x_k) \cdot d_k} \leq 1 - \sigma$$

  (recall that the denominator is negative, in view of Exercise 4.8), whence the preference for smaller values of σ.

- The *Weak Wolfe Rule* adds

  $$\nabla f(x_k + \gamma d_k) \cdot d_k \geq \tau\nabla f(x_k) \cdot d_k, \tag{30}$$

  where $\tau \in (\sigma, 1)$, typically chosen between 0.1 and 0.9, while the *Strong Wolfe Rule* adds

  $$|\nabla f(x_k + \gamma d_k) \cdot d_k| \leq \tau|\nabla f(x_k) \cdot d_k|. \tag{31}$$

Exercise 4.25. Show that, in all the above cases, γ_k is well defined: there always exists an integer $m \geq 0$ such that $\gamma_k = \beta^m\,\bar{\gamma}$ satisfies the corresponding conditions.

4.4 Acceleration methods

Gradient descent, although the most well-known optimization algorithm, is sometimes inconveniently slow. In this section, we will see how one can speed up gradient descent through acceleration methods. An alternative would be to make use of second-order methods, as explored in Section 4.6.

Recall that for a function $f\colon \mathbb{R}^N \to \mathbb{R}$, a step size $\gamma > 0$ (considered constant) and an initial point $x_0 \in \mathbb{R}^N$, gradient descent is described by

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

One way to build intuition of this iterative procedure is to consider it as a discretization of a dynamical system. Rewriting the gradient descent as

$$\frac{x_{k+1} - x_k}{\gamma} = -\nabla f(x_k),$$

and taking $\gamma \to 0$, we notice that gradient descent is nothing but the forward discretization of the dynamical system described by

$$\dot{x}(t) = -\nabla f(x(t)), \tag{32}$$

where the time discretization is done with a discretization step γ. A simple analysis shows that

$$\frac{d}{dt}(f(x(t) - \min f)) = \nabla f(x(t)) \cdot \dot{x}(t) = -\|\nabla f(x(t))\|^2.$$

As such, $f(x(t)) - \min f$ decreases along the trajectory, as long as $\nabla f(x(t)) \neq 0$. In specific, if f is convex, the trajectory will end up in a minimizer of f.
System (32) is not the only dynamical system whose trajectories end up in minimizers of f. Consider for instance the second-order differential equation that governs a particle with mass moving in a potential defined by the gradient of f. The system is given by

$$\mu\ddot{x}(t) + \nabla f(x) + b\dot{x}(t) = 0, \tag{33}$$

where $\mu > 0$ corresponds to the mass of the particle and $b \geq 0$ corresponds to the friction of the particle. This system is obviously different from (32), but recovers it (up to a constant factor) as $\mu \to 0$. Moreover, it has the same property that its trajectories converge towards a minimizer of f.
One could thus be interested in algorithms arising from (33). A simple forward-discretization leads to

$$\mu\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + \nabla f(x_k) - b\frac{x_{k+1} - x_k}{h},$$

where $h > 0$ is the discretization step. One can rearrange the terms to obtain

$$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1}),$$

where α and β are defined appropriately. This algorithm is known as the Heavy-Ball algorithm, and can equivalently be presented as the following two-step algorithm

$$\begin{cases} y_k = x_k + \beta(x_k - x_{k-1}), \\ x_{k+1} = y_k - \alpha\nabla f(x_k). \end{cases} \tag{34}$$

The first step, defining y_k, may be viewed as an *extrapolation step*, or a *momentum step*, where rather than directly using the point x_k, we make use of the idea of physical momentum to keep moving a little in the direction we moved when shiting from x_{k-1} to x_k.
Although we leave the proof out of these notes, the following convergence result can be proven. Note that the convergence guarantee requires the function to be quadratic, which is rather restrictive.

Theorem 4.26. *Let $0 < \mu < L$ and let $f\colon \mathbb{R}^N \to \mathbb{R}$ be a μ-strongly convex L-smooth quadratic function. By setting*

$$\alpha = \frac{4}{(\sqrt{\mu} + \sqrt{L})^2} \quad \text{and} \quad \beta = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2,$$

there exists a constant $C > 0$, independent on μ and L, such that

$$f(x_k) - \min f \leq Ck^2 \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^{2k} \|x_0 - x_*\|^2,$$

where (x_k) is given by (34) and $x_* = \operatorname{argmin} f$.

By modifying (34) slightly, namely evaluating the gradient at y_k as well, we obtain the well-known *Nesterov accelerated gradient descent*, given by

$$\begin{cases} y_k = x_k + \beta(x_k - x_{k-1}), \\ x_{k+1} = y_k - \alpha \nabla f(y_k). \end{cases} \tag{35}$$

We again present the convergence results without proofs. We note that Nesterov's accelerated method converges for any strongly convex smooth function, whereas we required the function to be quadratic for Heavy-Ball.

Theorem 4.27. *Let $0 < \mu < L$ and let $f \colon \mathbb{R}^N \to \mathbb{R}$ be a μ-strongly convex L-smooth function. By setting*

$$\alpha = \frac{1}{L} \quad \textit{and} \quad \beta = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}},$$

then

$$f(x_k) - \min f \leq 2 \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}} \right)^k (f(x_0) - \min f),$$

where (x_k) is given by (35).

Exercise 4.28. Compare the convergence results of Heavy-Ball and Nesterov accelerated gradient descent to the standard gradient descent results. Are the momentum methods faster?

4.5 Minimization of quadratic functions: The Conjugate Gradient Method

Let us for a moment revisit the minimzation of accelerated gradient methods in the simple case of strongly convex quadratic functions. Such functions take the form

$$f(x) = \frac{1}{2} x^T Q x - b^T x + c,$$

where $Q \in \mathbb{R}^{N \times N}$ satisfies $Q \succ \mu I$ for some $\mu > 0$, and $b \in \mathbb{R}^N$. Recall that

$$\nabla f(x) = Qx - b,$$

as Q is symmetric.

Consider the Heavy Ball Gradient Descent from Section 4.4. We allow the acceleration parameter and the step-size to vary with the iterations, namely for a sequence of parameters (α_k) and (β_k), the method is given by

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1}).$$

We then introduce a vector $p_k \in \mathbb{R}^N$, which captures the search direction of the iterations. Specifically, we write $x_{k+1} = x_k + \alpha_k p_k$, where

$$p_k = -\nabla f(x_k) + \gamma_{k-1} p_{k-1}, \tag{36}$$

where we defined $\gamma_k = \frac{\beta_{k+1}\alpha_k}{\alpha_{k+1}}$. For consistency, we set $\alpha_{-1} = 0$, and hence $\gamma_0 = 0$ and $p_0 = -\nabla f(x_0)$. We moreover introduce the residual quantity $r_k = \nabla f(x_k) = Qx_k - b$, which is updated according to

$$r_{k+1} = Qx_{k+1} - b = Qx_k - b + \alpha_k Q p_k = r_k + \alpha_k Q p_k,$$

where we used the definition of p_k to rewrite x_{k+1}. The conjugate gradient method then is defined through

$$\begin{cases} x_{k+1} &= x_k + \alpha_k p_k, \\ r_{k+1} &= r_k + \alpha_k Q p_k, \\ p_{k+1} &= -r_{k+1} + \gamma_k p_k. \end{cases}$$

The choice of α_k is made such that $f(x_{k+1}) = f(x_k + \alpha_k p_k)$ is minimized. In specific, the step-size is chosen according to *greedy line search*, as presented in Section 4.3.1. This leads to

$$\alpha_k = -\frac{p_k^T r_k}{p_k^T Q p_k}. \tag{37}$$

The choice of γ_k is made such that the search directions p_k and p_{k+1} are conjugate with respect to Q, namely that $p_{k+1}^T Q p_k = 0$. Using the algorithm, this is solved through

$$\gamma_k = \frac{r_{k+1}^T Q p_k}{p_k^T Q p_k} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}. \tag{38}$$

Not only does this step-size choice ensure that p_{k+1} and p_k are conjugate with respect to Q, it also guarantees that $p_0, \ldots, p_{k+1}$ are all conjugate with respect to Q. As such, the search directions form a linearly independent set, and x_k is the minimizer of f over the affine subspace $x_0 + \operatorname{span}(p_0, \ldots, p_k)$. In specific, if $k \geq N$, $\operatorname{span}(p_0, \ldots, p_k) = \mathbb{R}^N$, and hence x_k must be the minimizer of f over $\mathbb{R}^N$. The conjugate gradient method is thus guaranteed to converge in finite time (at most N iterations) for strongly convex quadratic functions.

Exercise 4.29. Show that the function $f(x) = \frac{1}{2}\|Ax - b\|^2$ is a strongly convex quadratic function, provided that $A \in \mathbb{R}^{N \times N}$ has full-rank and $N < M$.

Exercise 4.30. Verify Equations (36), (37) and (38).

Exercise 4.31. Verify that $\{p_0, \ldots, p_k\}$ are all conjugate with respect to Q, and that x_k is the minimizer of f over the affine subspace $x_0 + \operatorname{span}(p_0, \ldots, p_k)$.

Exercise 4.32. Let $f \colon \mathbb{R}^N \to \mathbb{R}$ be as in Exercise 4.29.

1. Compute ∇f and prox_f. Which one is easier to implement numerically?

2. Use to above to write out explicitly convergence rates for the gradient method and for the proximal-point method, and compare the convergence rates.

4.6 Second-order methods

Section 4.4 introduced the notion of momentum to achieve faster convergence rates. Another way to achieve accelerated convergence rates so through second-order methods in space. Specifically, recall that for convex functions, we are looking for a point x that minimizes a given function f, such that $\nabla f(x) = 0$.

Using Taylor's approximation, we know that, for any point y close to x,

$$\nabla f(x) \approx \nabla f(y) + \nabla^2 f(y)(x - y).$$

We note that although we cannot solve $\nabla f(x) = 0$ analytically, we can find the roots of the approximation analytically, namely they are given by

$$x = y - [\nabla^2 f(y)]^{-1} \nabla f(y).$$

Iterating this with an initial point $x_0 \in \mathbb{R}^N$, specifically defining

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k),$$

is known as *Newton's method for optimization.* The convergence behaviour is slightly more complicated than the one of gradient descent, and shall be left out of these notes. We do however note that Newton's method converges much faster than the standard gradient descent.

Although Newton's method exhibits faster converge rates than gradient descent, each iteration is more costly. In specific, we require the evaluation of the Hessian matrix, and the computation of its inverse. As the problem dimension N becomes large, computing such an inverse is very expensive. To avoid this, we introduce a matrix D_k that shall approximate $[\nabla^2 f(x_k)]^{-1}$, and define

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

Such methods are called *quasi-Newton* methods, and are mainly defined by the definition of D_k.

A simple chain rule shows that

$$\frac{d}{dt} \nabla f\big(x(t)\big) = \nabla^2 f\big(x(t)\big) \dot{x}(t),$$

which, under a forward discretization, yields

$$\nabla f(x_{k+1}) - \nabla f(x_k) \approx \nabla^2 f(x_k)(x_{k+1} - x_k),$$

which is known as the secant condition. We construct D_{k+1} to be symmetric, like the Hessian, and to satisfy the secant condition:

$$D_{k+1} g_k = s_k,$$

where $g_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ and $s_k = x_{k+1} - x_k$. Multiple schemes to compute such a D_{k+1} exist. We mention the two most popular instances:

- DFP (Davidon (1959), Fletcher and Powell (1987)):

$$D_{k+1} = D_k - \frac{(D_k g_k)(D_k g_k)^T}{g_k \cdot D_k g_k} + \rho_k (s_k s_k^T), \qquad \rho_k = \frac{1}{g_k \cdot s_k}.$$

- BFGS (Broyden, Fletcher, Goldfarb and Shanno (1970)):

$$D_{k+1} = (I - \rho_k s_k g_k^T) D_k (I - \rho_k s_k g_k^T)^T + \rho_k (s_k s_k^T), \qquad \rho_k = \frac{1}{g_k \cdot s_k}.$$

In both instances, the Hessian is no longer required, and no matrices need to be inverted. The convergence rate naturally suffers from the approximations, but remains faster than gradient descent.

4.7 Exercise: Projected Gradient Method

Consider the problem

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

where $f : \mathbb{R}^N \to \mathbb{R}$ is L-smooth, and $C \subset \mathbb{R}^N$ is nonempty, closed, convex and easy to project onto. The set of solutions, which we assume to be nonempty, is denoted by S, and the optimal value by f^*. This problem can be solved by means of the *Projected Gradient Method*, which starts from some $x_0 \in \mathbb{R}^N$, and then iterates

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

for $k \geq 0$. In other words, at each iteration, we perform a gradient step and then project onto C. The purpose of this assignment is to prove that the Projected Gradient Methods converges.

Part I. If f is convex and $0 < \gamma L \leq 1$, then x_k converges to a point in S, with

$$f(x_k) - f^* \leq \frac{\operatorname{dist}(x_0, S)^2}{2\gamma k}. \tag{39}$$

To prove this, follow these steps:

1. Let $w \in \mathbb{R}^N$ and let $\overline{w} = P_C(w)$. Show that for all $z \in C$, we have

$$\|z - w\|^2 \geq \|\overline{w} - w\|^2 + \|z - \overline{w}\|^2.$$

2. Verify that, for every $\gamma \in (0, 1/L]$ and $k \geq 0$,

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

3. Deduce that, if f is convex, $\hat{x} \in S$, $\gamma \in (0, 1/L]$ and $k \geq 0$, then

$$2\gamma\big(f(x_{k+1}) - f^*\big) \leq \|x_k - \hat{x}\|^2 - \|x_{k+1} - \hat{x}\|^2.$$

4. Conclude that (39) holds for every $\gamma \in (0, 1/L]$ and $k \geq 1$.

5. Prove that the sequence x_k converges to a point in S.

Part II. If f is μ-strongly convex, with $\operatorname{argmin}(f) = \{x^*\}$, and $0 < \gamma \leq \frac{2}{L+\mu}$, then

$$\|x_k - x^*\| \leq \left(1 - \frac{2\gamma\mu L}{L+\mu}\right)^{\frac{k}{2}}\|x_0 - x^*\|. \tag{40}$$

To prove this, follow these steps:

1. Using the characterization of the projection, show that

$$P_C(x^* - \gamma\nabla f(x^*)) = x^*.$$

2. Prove the projection operator is nonexpansive, namely that

$$\|P_C(x) - P_C(y)\| \leq \|x - y\| \quad \forall x, y \in \mathbb{R}^N.$$

3. Use the strong convexity of f to conclude that (40) holds for every $\gamma \in (0, \frac{2}{L+\mu}]$ and $k \geq 0$.

4.8 Computational exercise: Linear Least Squares Problem

The *Linear Least Squares* problem is defined as follows: for $A \in \mathbb{R}^{M \times N}$ and $b \in \mathbb{R}^M$ we seek

$$\min_{x \in \mathbb{R}^N} \frac{1}{2}\|Ax - b\|^2. \tag{41}$$

In this computational exercise you will solve the least squares problem using different methods. The aim of the exercise is to compare the computed solutions, the complexity of the methods, their convergence properties. Implement and test on various matrix sizes the following methods:

1. Gradient Descent with constant step size,

2. Gradient Descent with exact (greedy) line search,

3. Gradient Descent with Armijo rule,

4. Conjugate Gradient Method,

5. Polyak's Heavy Ball accelerated Gradient Descent, and

6. Nesterov's momentum applied to Gradient Descent.

5 Minimization of nonsmooth convex functions

In the previous sections, we have always assumed our objective functions to be smooth. In this section, we consider the nonsmooth case, by first extending the notion of gradients to non-differentiable functions in Section 5.1. This allows us to naturally define the Subgradient Method, analogous to the Gradient Method, in Section 5.2. In Section 5.3 we consider functions defined on the extended reals, and show various properties of such functions. A key object in nonsmooth optimization is the proximity operator, which we discuss in Section 5.4, and which gives rise to the famous Proximal-Gradient Method, presented in Section 5.6.

5.1 Subgradients and subdifferential of a convex function

Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex. If f is differentiable at $x \in D$, Proposition 3.37 shows that

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$$

for all $y \in \mathbb{R}^N$. Motivated by this fact, we say that a vector $v \in \mathbb{R}^N$ is a *subgradient* of f at x if

$$f(y) \geq f(x) + v \cdot (y - x)$$

for all $y \in \mathbb{R}^N$. The *subdifferential* of f at x, denoted by $\partial f(x)$, is the set of all the subgradients of f at x. As in Proposition 3.37, we also have the following:

Proposition 5.1 (Monotonicity of ∂f)**.** *If $f : \mathbb{R}^N \to \mathbb{R}$ is convex, then ∂f is a monotone operator. Specifically, for all $x, y \in \mathbb{R}^N$ and $u \in \partial f(x)$, $v \in \partial f(y)$, it holds that $(x - y) \cdot (u - v) \geq 0$.*

Another immediate consequence of the definition is:

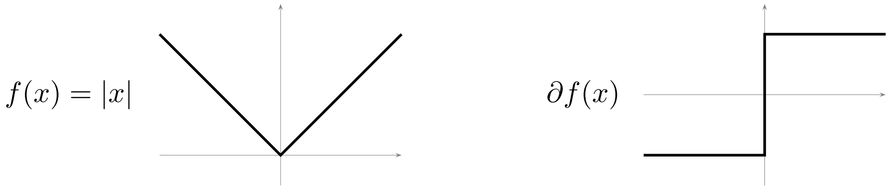
Theorem 5.2 (Fermat's Rule III)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}$. Then, $\hat{x} \in \operatorname{argmin}(f)$ if, and only if $0 \in \partial f(\hat{x})$.*

Exercise 5.3. Show that, for every $x \in \mathbb{R}^N$, $\partial f(x)$ is closed and convex.

Example 5.4 (A big ice cream cone)**.** The function $f : \mathbb{R}^N \to \mathbb{R}$, defined by $f(x) = \|x\|$, is differentiable in $\mathbb{R}^N \setminus \{0\}$. Determining $\partial f(0)$, is reduced to finding all vectors $v \in \mathbb{R}^N$ such that

$$\|y\| \geq v \cdot y$$

for every $y \in \mathbb{R}^N$. By the Cauchy-Schwarz Inequality, $\bar{B}(0,1) \subset \partial f(0)$. Conversely, if $\|v\| > 1$, the choice $y = v$ gives $\|v\| \geq \|v\|^2$, which implies $\|v\| \leq 1$, and is a contradiction. We conclude that $\partial f(0) = \bar{B}(0,1)$.



Exercise 5.5 (The ℓ^1-norm)**.** Let $\|\cdot\|_1 : \mathbb{R}^N \to \mathbb{R}$ be defined by $\|x\|_1 = |x_1| + \cdots + |x_N|$. Compute $\partial f(x)$ for each $x \in \mathbb{R}^N$.

The subdifferential is a generalization of the concept of gradient, as shown in the following:

Proposition 5.6. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex. If f is differentiable at $x \in \mathbb{R}^N$, then $\partial f(x) = \{\nabla f(x)\}$.*

Proof. Let $v \in \partial f(x)$, let $t > 0$ and let $d \in \mathbb{R}^N$. By the definition of subdifferential, we have

$$f(x + td) \geq f(x) + t\, v \cdot d,$$

which is equivalent to

$$\frac{f(x + td) - f(x)}{t} \geq v \cdot d.$$

Since f is differentiable at x, letting $t \to 0$, we deduce that

$$\nabla f(x) \cdot d \geq v \cdot d.$$

Since this holds for every $d \in \mathbb{R}^N$, we may choose $d = v - \nabla f(x)$ to conclude that $v = \nabla f(x)$. □

Proposition 5.6 shows that Theorem 2.39 is a corollary of Theorem 5.2.

Exercise 5.7. Write out the optimality conditions for the classical and robust SVM from Section 1.3.2. You may assume without proof that the subdifferential separates over the sum in this case.

By Proposition 3.36, convex functions are (Lipschitz-)continuous in the interior of their domain. Moreover, we have the following:

Proposition 5.8. *If $f : \mathbb{R}^N \to \mathbb{R}$ is convex, then $\partial f(x)$ is nonempty and bounded for every $x \in \mathbb{R}^N$.*

Proof. First, by Proposition 3.36, f is continuous in all of $\mathbb{R}^N$. By Proposition 3.34, epi(f) is has nonempty interior. Take $x \in \mathbb{R}^N$, and set $C = \operatorname{int}\big(\operatorname{epi}(f)\big) - \big\{\big(x, f(x)\big)\big\}$, so that C is a nonempty and convex (by Exercise 3.10) subset of $\mathbb{R}^N \times \mathbb{R}$, with $(0, 0) \notin C$. Proposition 3.15 then shows that there exist $w \in \mathbb{R}^N$ and $s \in \mathbb{R}$ such that $(w, s) \neq (0, 0)$ and

$$w \cdot (z - x) + s\big(t - f(x)\big) \leq 0$$

for every $(z, t) \in \operatorname{int}\big(\operatorname{epi}(f)\big)$. If $s > 0$, we can let $t \to \infty$ and break the inequality. If $s = 0$, then $w \cdot (z - x) \leq 0$ for all z in an open ball centered at x, and so w would have to be 0, contradicting that $(w, s) \neq (0, 0)$. As a consequence, $s < 0$. Dividing by s and writing $v = -\frac{1}{s}w$, we deduce that

$$-v \cdot (z - x) + \big(t - f(x)\big) \geq 0$$

for every $(z, t) \in \operatorname{int}\big(\operatorname{epi}(f)\big)$. Since $\operatorname{int}\big(\operatorname{epi}(f)\big)$ is dense in epi(f), we may let $(z, t) \to \big(y, f(y)\big)$ to deduce that

$$f(y) \geq f(x) + v \cdot (y - x)$$

for every $y \in \mathbb{R}^N$. This means that $v \in \partial f(x)$, and so $\partial f(x) \neq \emptyset$. To see it is bounded, recall, from Proposition 3.36, that there exist $L_x, r_x > 0$ such that $|f(y) - f(z)| \leq l_x \|y - z\|$ for every $y, z \in B(x, r_x)$. If $u \in \partial f(x)$ and we take $\delta \in \big(0, \frac{\|u\|}{r_x}\big)$, then

$$f(x + \delta u) \geq f(x) + u \cdot (x + \delta u - x) = f(x) + \delta \|u\|^2.$$

Therefore,

$$\delta \|u\|^2 \leq |f(x + \delta u) - f(x)| \leq L_x \|x + \delta u - x\| = \delta L_x \|u\|,$$

whence $\|u\| \leq L_x$, which shows that $\partial f(x) \subset \bar{B}(0, L_x)$. □

5.2 Subgradient Method

The *Subgradient Method* starts from $x_0 \in \mathbb{R}^N$, and then, for $k \geq 0$, iterates

$$x_{k+1} = x_k - \gamma d_k, \qquad \text{with} \qquad d_k \in \partial f(x_k).$$

We have the following convergence result:

Proposition 5.9. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and Lipschitz-continuous with constant M. Suppose $S := \operatorname{argmin}(f) \neq \emptyset$, and let (x_k) be defined by the subgradient method. Set $\bar{x}_k = \frac{1}{k+1}\sum_{j=0}^{k} x_j$. Then,*

$$f(\bar{x}_k) - \min(f) \leq \frac{\gamma M^2}{2} + \frac{\operatorname{dist}(x_0, S)^2}{2\gamma(k+1)}.$$

Proof. Take $p = P_S x_0$, so that $\|x_0 - p\| = \operatorname{dist}(x_0, S)$. We have

$$\begin{aligned} \|x_{k+1} - p\|^2 &= \|x_k - \gamma d_k - p\|^2 \\ &= \|x_k - p\|^2 + \gamma^2 \|d_k\|^2 - 2\gamma\, d_k \cdot (x_k - p) \\ &\leq \|x_k - p\|^2 + \gamma^2 M^2 + 2\gamma\big(f(p) - f(x_k)\big), \end{aligned} \tag{42}$$

in view of Proposition 5.8 and the definition of subgradient. Using the convexity of f, we see that

$$f(\bar{x}_k) - \min(f) = f\left(\tfrac{1}{k+1}\sum_{j=0}^{k} x_j\right) - \min(f) \leq \left[\frac{1}{k+1}\sum_{j=0}^{k} f(x_j)\right] - \min(f) = \frac{1}{k+1}\sum_{j=0}^{k}\big(f(x_j) - \min(f)\big).$$

Reorganizing the terms in (42), summing, and using the telescopic property, we deduce that

$$f(\bar{x}_k) - \min(f) \leq \frac{1}{k+1}\sum_{j=0}^{k}\big(f(x_j) - \min(f)\big) \leq \frac{1}{2\gamma(k+1)}\left[\|x_0 - p\|^2 - \|x_{k+1} - p\|^2 + \gamma^2 M^2\right],$$

from which we easily conclude. □

Exercise 5.10. Given $\varepsilon > 0$, after how many iterations and with which step size can we be sure to have found a point $\hat{x}$ such that $f(\hat{x}) - \min(f) \leq \varepsilon$?

5.3 Extended real-valued functions

We define the set of *extended real numbers* as $\mathbb{R} \cup \{+\infty\}$, with the conventions that $+\infty > \lambda$ for all $\lambda \in \mathbb{R}$, and some algebraic operations are allowed, namely:

- $+\infty + \lambda = +\infty$ for all $\lambda \in \mathbb{R}$,
- $\lambda(+\infty) = +\infty$ for all $\lambda > 0$, and
- $0(+\infty) = 0$.

The *(effective) domain* of an *extended real-valued function* $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is

$$\operatorname{dom}(f) = \{x \in \mathbb{R}^N \ : \ f(x) < +\infty\},$$

and we will *always* assume that $\operatorname{dom}(f) \neq \emptyset$.

Remark 5.11. If $f, g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ and $\lambda > 0$, then $\mathrm{dom}(\lambda f + g) = \mathrm{dom}(f) \cap \mathrm{dom}(g)$.

For the *epigraph* of $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, we use the same definition as for real-valued functions, namely:

$$\mathrm{epi}(f) = \{(x,t) \in \mathbb{R}^N \times \mathbb{R} \colon t \geq f(x)\}.$$

For $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, the definition of *convexity* can be adapted in a straightforward manner, as well as those of strict and strong convexity. The following proposition establishes relationships between the convexity of a function and that of its epigraph, domain and sublevel sets:

Proposition 5.12 (Properties of extended real-valued convex functions)**.** *For $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, we have the following:*

i) f is convex if, and only if, $\mathrm{epi}(f)$ is convex.

ii) If f is convex, then $\mathrm{dom}(f)$ and $\mathrm{argmin}(f)$ are convex. Also, for every $\gamma \in \mathbb{R}$, the sublevel set $[f \leq \gamma]$ is convex.

iii) If f is strictly convex, then $\mathrm{argmin}(f)$ is either empty or a singleton.

Given $f, g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, we can define the sum $f + g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ as $(f+g)(x) = f(x) + g(x)$, whenever $\mathrm{dom}(f) \cap \mathrm{dom}(g) \neq \emptyset$ (otherwise, $f + g \equiv +\infty$). In what follows, when we refer to the sum of two extended real-valued functions, we shall implicitly assume that it is well defined.

Exercise 5.13. Show that the sum of two convex functions is convex. If, moreover, one of them is strictly or strongly convex, so is the sum.

The *pointwise supremum* of a (finite or infinite) family (f_i) of extended real-valued functions is the function $\sup_i f_i : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, defined by $\left[\sup_i f_i\right](x) = \sup_i \{f_i(x)\}$.

Exercise 5.14. Show that the pointwise supremum of convex functions is convex.

A function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is *closed* if $\mathrm{epi}(f)$ is closed. This property can be characterized as follows:

Proposition 5.15. *Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$. The following statements are equivalent:*

i) f is closed;

ii) for every $\gamma > \inf(f)$, the set $[f \leq \gamma]$ is closed;

iii) the function $f\big|_{\mathrm{dom}(f)} : \mathrm{dom}(f) \subset \mathbb{R}^N \to \mathbb{R}$ is lower-semicontinuous (see Section 2.2); and

iv) for every $x \in \mathrm{dom}(f)$, and every sequence (x_k) that converges to x, we have $f(x) \leq \liminf_{k\to\infty} f(x_k)$.

Exercise 5.16. Show that, if $f, g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ are closed and $\lambda \geq 0$, then $\lambda f + g$ is closed.

Exercise 5.17. Show that the pointwise supremum of closed functions is closed.

Proposition 5.18 (Existence of an affine minorant)**.** *If $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is closed and convex, there exist $c \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$ such that*

$$f(x) \geq c \cdot x + \alpha$$

for every $x \in \mathbb{R}^N$.

Proof. Recall that the set $\text{epi}(f)$ is nonempty, closed and convex. Take any $x_0 \in \text{dom}(f)$ and $\lambda_0 < f(x_0)$, so that $(x_0, \lambda_0) \notin \text{epi}(f)$. Let $(\hat{x}, \hat{\lambda}) = P_{\text{epi}(f)}(x_0, \lambda_0)$. By Corollary 3.20 (applied in $\mathbb{R}^{N+1}$), we have

$$(x_0 - \hat{x}) \cdot (x - \hat{x}) + (\lambda_0 - \hat{\lambda})(\lambda - \hat{\lambda}) \leq 0$$

for every $(x, \lambda) \in \text{epi}(f)$. In particular, for every $x \in \text{dom}(f)$ and $\lambda = f(x)$. It follows that

$$(\hat{\lambda} - \lambda_0) f(x) \geq (x_0 - \hat{x}) \cdot x + \left[\hat{x} \cdot (\hat{x} - x_0) + \hat{\lambda} \cdot (\hat{\lambda} - \lambda_0)\right]$$

for every $x \in \text{dom}(f)$. Since $\hat{\lambda} > \lambda_0$ (why?), we deduce that

$$f(x) \geq c \cdot x + \alpha, \quad \text{where} \quad c = \frac{1}{\hat{\lambda} - \lambda_0}(x_0 - \hat{x}) \quad \text{and} \quad \alpha = \frac{\hat{x} \cdot (\hat{x} - x_0) + \hat{\lambda} \cdot (\hat{\lambda} - \lambda_0)}{\hat{\lambda} - \lambda_0}.$$

If $x \notin \text{dom}(f)$, then $f(x) = +\infty$ and the inequality holds trivially. □

Example 5.19. The *indicator function* of a nonempty set $C \subset \mathbb{R}^N$ is the function $\iota_C : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, defined as

$$\iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{otherwise.} \end{cases}$$

Here, $\text{dom}(\iota_C) = C$ and $\text{epi}(\iota_C) = C \times [0, +\infty)$. The indicator function ι_C is closed if, and only if, C is closed. It is convex if, and only if, C is convex. Also, for every $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, we have

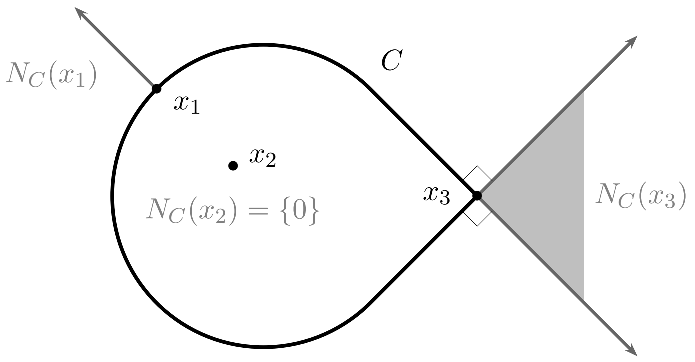
$$\min\{f(x) : x \in C\} = \min\left\{f(x) + \iota_C(x) : x \in \mathbb{R}^N\right\}.$$

For an extended real-valued function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, the *subgradients* and *subdifferential* are defined in the same way as in the finite case. Observe that, if $x \notin \text{dom}(f)$, then necessarily $\partial f(x) = \emptyset$.

Example 5.20. Let $C \subset \mathbb{R}^N$ be nonempty, closed and convex, and let us compute $\partial \iota_C(x)$ for each $x \in \mathbb{R}^N$. We already know that, if $x \notin C$, then $\partial \iota_C(x) = \emptyset$. If $x \in C$, then

$$\begin{aligned} \partial \iota_C(x) &= \left\{v \in \mathbb{R}^N : \iota_C(y) \geq \iota_C(x) + v \cdot (y - x) \quad \forall y \in \mathbb{R}^N\right\} \\ &= \left\{v \in \mathbb{R}^N : \iota_C(y) \geq 0 + v \cdot (y - x) \quad \forall y \in \mathbb{R}^N\right\} \\ &= \left\{v \in \mathbb{R}^N : 0 \geq v \cdot (y - x) \quad \forall y \in C\right\}. \end{aligned}$$

In other words, $\partial \iota_C(x)$ contains the vectors that form an obtuse or straight angle with all the vectors starting at x and pointing towards other elements of C. In particular, if $x \in \text{int}(C)$, then $\partial \iota_C(x) = \{0\}$. The fact that $\partial \iota_C(x) = \lambda \partial \iota_C(x)$ for every $\lambda > 0$ makes this set a *cone* (recall that, since it is a subdifferential, it is also closed and convex). The set $\partial \iota_C(x)$ is the *normal cone* to C at x, and is usually denoted by $N_C(x)$.



As earlier, we have the following:

Theorem 5.21 (Fermat's Rule IV). *Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$. Then, $\hat{x} \in \operatorname{argmin}(f)$ if, and only if $0 \in \partial f(\hat{x})$.*

Remark 5.22. For $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, the conclusion of Propositions 5.1 and 5.6 remain the same, while that of Proposition 5.8 holds for every $x \in \operatorname{int}\big(\operatorname{dom}(f)\big)$.

Proposition 5.23. *If $f, g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, then $\partial f(x) + \partial g(x) \subset \partial(f+g)(x)$.*

Proof. Let $x \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$ (otherwise there is nothing to prove), and suppose $v_f \in \partial f(x)$ and $v_g \in \partial g(x)$. For every $y \in \mathbb{R}^N$, we have

$$\begin{array}{rcl} f(y) & \geq & f(x) + v_f \cdot (y - x) \\ g(y) & \geq & g(x) + v_g \cdot (y - x). \end{array}$$

Adding the two inequalities, we obtain

$$(f+g)(y) \geq (f+g)(x) + (v_f + v_g) \cdot (y - x).$$

Since this holds for each $y \in \mathbb{R}^N$, $v_f + v_g \in \partial(f+g)(x)$. □

The converse is not true in general, as shown in the following:

Example 5.24. Let $f, g : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be given by

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ +\infty & \text{if } x > 0 \end{cases} \qquad \text{and} \qquad g(x) = \begin{cases} +\infty & \text{if } x < 0 \\ -\sqrt{x} & \text{if } x \geq 0. \end{cases}$$

We have

$$\partial f(x) = \begin{cases} \{0\} & \text{if } x < 0 \\ [0, +\infty) & \text{if } x = 0 \\ \emptyset & \text{if } x > 0 \end{cases} \quad \text{and} \quad \partial g(x) = \begin{cases} \emptyset & \text{if } x \leq 0 \\ \left\{-\frac{1}{2\sqrt{x}}\right\} & \text{if } x > 0. \end{cases}$$

Therefore, $\partial f(x) + \partial g(x) = \emptyset$ for every $x \in \mathbb{R}$. On the other hand, $f + g = \delta_{\{0\}}$, which implies $\partial(f+g)(x) = \emptyset$ if $x \neq 0$, but $\partial(f+g)(0) = \mathbb{R}$. We see that $\partial(f+g)(x)$ may differ from $\partial f(x) + \partial g(x)$.

Exercise 5.25. Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be closed and convex, let A be a positive semidefinite matrix of size $N \times N$, and let $b \in \mathbb{R}^N$. Compute the subdifferential of the function $g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, defined by $g(x) = f(x) + x^T A x + b \cdot x$.

We end this section with the following result concerning the subdifferentials of strongly convex functions:

Proposition 5.26. *Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be closed and strongly convex. For every $y^* \in \mathbb{R}^N$ there is $x^* \in \mathbb{R}^N$ such that $y^* \in \partial f(x^*)$.*

Proof. The function $g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, defined by $g(x) = f(x) - y^* \cdot x$, is closed and strongly convex. Therefore, it has a unique minimizer x^*, which must satisfy

$$0 \in \partial g(x^*) = \partial f(x^*) - y^*.$$

In other words, $y^* \in \partial f(x^*)$. □

5.4 The proximity operator

In this section, we present a key object in nonsmooth optimization, for which we require a few preliminaries. We begin by establishing the following:

Proposition 5.27. *If $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is closed and μ-strongly convex, it has exactly one minimizer. Moreover, if $v \in \partial f(x)$, then*

$$f(y) \geq f(x) + v \cdot (y - x) + \frac{\mu}{2}\|y - x\|^2 \tag{43}$$

for every $y \in \mathbb{R}^N$.

Proof. The existence and uniqueness of the minimizer is proved as in Exercise 2.23, so it is omitted. Now, suppose $v \in \partial f(x)$, and fix $y \in \mathbb{R}^N$. Take $\lambda \in (0, 1)$, and set $w = \lambda y + (1 - \lambda)x$. The definition of v and the strong convexity respectively imply

$$\begin{array}{rcl} f(w) & \geq & f(x) + v \cdot (w - x) \\ f(w) & \leq & \lambda f(y) + (1 - \lambda) f(x) - \frac{\mu}{2}\lambda(1 - \lambda)\|y - x\|^2. \end{array}$$

Combining these two inequalities, and observing that $w - x = \lambda(y - x)$, we obtain

$$\lambda\, v \cdot (y - x) \leq \lambda\big(f(y) - f(x)\big) - \frac{\mu}{2}\lambda(1 - \lambda)\|y - x\|^2,$$

and so

$$f(y) \geq f(x) + v \cdot (y - x) + \frac{\mu}{2}(1 - \lambda)\|y - x\|^2.$$

Letting $\lambda \to 0$, we obtain (43). □

Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be closed and convex, and let $z \in \mathbb{R}^N$. The function $f_z : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, defined by

$$f_z(x) = f(x) + \frac{1}{2}\|x - z\|^2,$$

is closed—because it is the sum of two closed functions—and strongly convex—because it is the sum of a convex function and a strongly convex one. By Proposition 5.27, it has a unique minimizer, which we denote by $\hat{x}_z$. The *proximity operator* associated to f is the function $\operatorname{prox}_f : \mathbb{R}^N \to \mathbb{R}^N$ defined by $\operatorname{prox}_f(z) = \hat{x}_z$. In other words,

$$\operatorname{prox}_f(z) = \operatorname{argmin}\left\{ f(x) + \frac{1}{2}\|x - z\|^2 \right\}.$$

According to Theorem 5.21, we have

$$0 \in \partial f_z(\hat{x}_z). \tag{44}$$

It turns out that we can express the subdifferential of f_z in terms of that of f.

Proposition 5.28. *With the notation introduced above, $\partial f_z(x) = \partial f(x) + x - z$ for every $x, z \in \mathbb{R}^N$.*

Proof. Fix $z \in \mathbb{R}^N$ and $x \in \operatorname{dom}(f) = \operatorname{dom}(f_z)$ (otherwise, there is nothing to prove). Since the quadratic term in f_z is differentiable, Proposition 5.23 shows that $\partial f(x) + x - z \subset \partial f_z(x)$. Let us verify that $\partial f_z(x) + z - x \subset \partial f(x)$. To this end, take $v \in \partial f_z(x)$. Since f_z is strongly convex with parameter $\mu = 1$, Proposition 5.27 guarantees that

$$f_z(y) \geq f_z(x) + v \cdot (y - x) + \frac{1}{2}\|y - x\|^2$$

for every $y \in \mathbb{R}^N$. This inequality is equivalent to

$$\begin{aligned} f(y) &\geq f(x) + v \cdot (y - x) + \frac{1}{2}\|x - z\|^2 - \frac{1}{2}\|y - z\|^2 + \frac{1}{2}\|y - x\|^2 \\ &= f(x) + (v + z - x) \cdot (y - x), \end{aligned}$$

because $\|y - z\|^2 = \|y - x\|^2 + \|x - z\|^2 + 2(y - x) \cdot (x - z)$. Therefore, $v + z - x \in \partial f(x)$ and this completes the proof. □

Combining this with (44), we conclude that $\operatorname{prox}_f(z) = \hat{x}_z$ is characterized by

$$z \in \hat{x}_z + \partial f(\hat{x}_z). \tag{45}$$

As a consequence, $\hat{x} \in \operatorname{argmin}(f)$ if, and only if, $\hat{x}$ is a *fixed point* of prox_f, which means that $\hat{x} = \operatorname{prox}_f(\hat{x})$.

Exercise 5.29. Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be closed and convex, let $\ell : \mathbb{R}^N \to \mathbb{R}$ be affine, and set $g = f + \ell$. Given $\lambda > 0$, find a formula for $\operatorname{prox}_{\lambda g}$ in terms of $\operatorname{prox}_{\lambda f}$.

Remark 5.30. Since $\operatorname{prox}_f(z)$ is uniquely determined by z, we often write $\operatorname{prox}_f(z) = \big(I + \partial f\big)^{-1}(z)$. In view of its similarity with the concept from Functional Analysis, the function $\operatorname{prox}_f = \big(I + \partial f\big)^{-1}$ is known as the *resolvent* of the (set-valued) operator ∂f.

Example 5.31. If $C \subset \mathbb{R}^N$ is nonempty, closed and convex, then $\operatorname{prox}_{\iota_C}(x) = P_C(x)$.

Proposition 5.32. *For every closed convex function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, and every $x, y \in \mathbb{R}^N$, we have*

$$\|\operatorname{prox}_f(x) - \operatorname{prox}_f(y)\| \leq \|x - y\|.$$

In particular, prox_f is a continuous function.

Proof. Fix x, y, and set $u = \operatorname{prox}_f(x)$ and $v = \operatorname{prox}_f(y)$, so that

$$u - x \in \partial f(x) \quad \text{and} \quad v - y \in \partial f(y),$$

in view of (45). The monotonicity of ∂f (see Proposition 5.1 and Remark 5.22) then shows that

$$(u - x - (v - y)) \cdot (x - y) \geq 0 \quad \Longrightarrow \quad (x - y) \cdot (u - v) \geq \|x - y\|^2.$$

Using the Cauchy-Schwarz Inequality, we conclude that $\|u - v\| \geq \|x - y\|$, as wanted. □

5.5 Exercise: Proximal Point Method

In this assignment, we shall prove the convergence of the Proximal-Point Method (PPM). We consider a function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ which is proper, convex, lower semi-continuous, and which has minimizers. Note that we do not assume differentiability of the function. As usually, the aim is to find a minimizer of f.

Recall the *proximity operator* is defined by

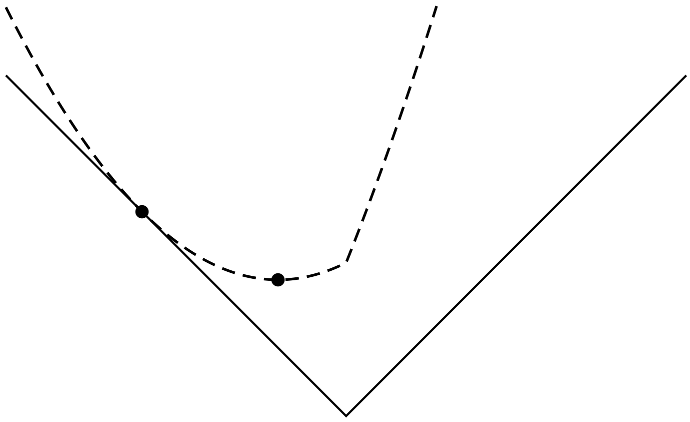
$$\mathrm{prox}_f(x) = \mathrm{argmin}_{y\in\mathbb{R}^d}\left[f(y) + \frac{1}{2}\|x-y\|^2\right].$$

The *Proximal-Point Method* (PPM) consists of iteratively applying the above operator to an initial guess $x_0 \in \mathbb{R}^d$, as follows

$$x_{k+1} = \mathrm{prox}_f(x_k).$$

1. Justify why $\mathrm{prox}_f\colon \mathbb{R}^d \to \mathbb{R}^d$ is well-defined, that is that it is defined and single-valued at each point.

2. Explain how the following pictures relates to the PPM. That is, identify the objective function, the points x_k and x_{k+1}, and any other objects present on the figure.



3. Show that, for each $k \geq 1$, $x_k - x_{k+1} \in \partial f(x_{k+1})$, where ∂f represents the subdifferential of f.

4. Deduce that $f(x_{k+1}) \leq \min f - (x_k - x_{k+1}) \cdot (x^* - x_{k+1})$, where $x^* \in \mathbb{R}^d$ is a minimiser of f.

5. Use the previous to deduce that $f(x_{k+1}) - \min f \leq \frac{1}{2}(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2)$.

6. Prove that
$$\sum_{i=0}^{k-1} f(x_i) - \min f \leq \frac{1}{2}\|x_0 - x^*\|^2.$$

7. Prove that $(f(x_k))$ is non-decreasing and conclude that $f(x_k) - \min f \leq \frac{C}{k}$, for some $C > 0$ not dependent on k.

5.6 Structured problems and the Proximal-Gradient Method

It is common to encounter problems of the form

$$\min\left\{f(x) + g(x) : x \in \mathbb{R}^N\right\},$$

where $f : \mathbb{R}^N \to \mathbb{R}$ is convex and smooth, but $g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is just closed and convex. These problems can be tackled using the Subgradient Method. However, it is not a descent method, and its complexity is of the order of ε^{-2}, rather than ε^{-1} as for the Gradient Method. Using the procedure in Section 5.5 may not be easy to implement, since there is no general formula for the proximity operator associated to a sum of functions, except, for instance, if one is affine.

An alternative is to *split* the problem in order to combine the gradient iterations−applied to the smooth part− with the proximity operator described above. The result is the *proximal-gradient method*, which we now present. A first interpretation of this numerical scheme is that we first linearize the smooth function f, and then apply the proximity operator to the sum of g plus the said linearization of f. More precisely, given $x_0 \in \mathbb{R}^N$, we define a sequence (x_k) by iterating

$$x_{k+1} = \operatorname{argmin}\left\{\gamma\big(g(z) + f(x_k) + \nabla f(x_k)\cdot(z - x_k)\big) + \frac{1}{2}\|z - x_k\|^2 \ :\ z \in \mathbb{R}^N\right\}.$$

As in Exercise 5.29, we can complete the square and discard the constant terms (which do not change the optimization subproblem), to obtain

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}\left\{\gamma g(z) + \frac{1}{2}\|\gamma\nabla f(x_k)\|^2 + \gamma\nabla f(x_k)\cdot(z - x_k) + \frac{1}{2}\|z - x_k\|^2 \ :\ z \in \mathbb{R}^N\right\} \\ &= \operatorname{argmin}\left\{\gamma g(z) + \frac{1}{2}\big\|z - \big(x_k - \gamma\nabla f(x_k)\big)\big\|^2 \ :\ z \in \mathbb{R}^N\right\}. \end{aligned} \tag{46}$$

Therefore, the proximal-gradient iteration can be interpreted as a two-step proces, in which we first compute the auxiliary point

$$x_{k+\frac{1}{2}} = x_k - \gamma\nabla f(x_k),$$

and then apply the proximity operator on $x_{k+\frac{1}{2}}$, to finally obtain

$$x_{k+1} = \operatorname{prox}_{\gamma g}\left(x_{k+\frac{1}{2}}\right) = \operatorname{prox}_{\gamma g}\left(x_k - \gamma\nabla f(x_k)\right).$$

The point x_{k+1} is also characterized by the optimality condition associated to the problem in (46), as

$$-\frac{x_{k+1} - x_k}{\gamma} \in \nabla f(x_k) + \partial g(x_{k+1}). \tag{47}$$

Theorem 5.33. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and L-smooth, let $g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be closed and convex, and set $\gamma \leq 1/L$. Define (x_k) by the proximal-gradient method (46). If $\operatorname{argmin}(f+g) \neq \emptyset$, then x_k converges, as $k \to \infty$, to a minimizer of $f + g$. Moreover, for every $k \geq 1$, we have*

$$(f+g)(x_k) - \min(f+g) \leq \frac{\operatorname{dist}(x_0, \operatorname{argmin}(f+g))^2}{2\gamma k}. \tag{48}$$

Proof. In view of (47), the definition of subgradient gives

$$\begin{aligned} g(x_k) &\geq g(x_{k+1}) + \left(-\frac{x_{k+1} - x_k}{\gamma} - \nabla f(x_k)\right)\cdot(x_k - x_{k+1}) \\ &= g(x_{k+1}) + \frac{1}{\gamma}\|x_{k+1} - x_k\|^2 + \nabla f(x_k)\cdot(x_{k+1} - x_k). \end{aligned}$$

For f, we use the Baillon-Haddad Theorem 3.38 to obtain

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

Combining these two inequalities, we deduce that

$$(f+g)(x_{k+1}) + \left(\frac{1}{\gamma} - \frac{L}{2}\right)\|x_{k+1} - x_k\|^2 \leq (f+g)(x_k), \tag{49}$$

where $\frac{1}{\gamma} - \frac{L}{2} > 0$. Therefore, $(f+g)(x_k)$ is nonincreasing. On the other hand, from (47), we also know that

$$w_{k+1} := -\frac{x_{k+1} - x_k}{\gamma} + \nabla f(x_{k+1}) - \nabla f(x_k) \in \nabla f(x_{k+1}) + \partial g(x_{k+1}).$$

Since $\gamma < 2/L$, using the Baillon-Haddad Theorem 3.38 once again, we get

$$\begin{aligned} \|w_{k+1}\|^2 &= \frac{1}{\gamma^2}\|x_{k+1} - x_k\|^2 + \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 - \frac{2}{\gamma}\big(\nabla f(x_{k+1}) - \nabla f(x_k)\big) \cdot (x_{k+1} - x_k) \\ &\leq \frac{1}{\gamma^2}\|x_{k+1} - x_k\|^2 + \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 - \frac{2}{\gamma L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\ &\leq \frac{1}{\gamma^2}\|x_{k+1} - x_k\|^2. \end{aligned} \tag{50}$$

Combining this inequality with (49), we obtain

$$(f+g)(x_{k+1}) + \gamma^2\left(\frac{1}{\gamma} - \frac{L}{2}\right)\|w_{k+1}\|^2 \leq (f+g)(x_k), \tag{51}$$

with $w_{k+1} \in \partial(f+g)(x_{k+1})$.

Now, take $\hat{x} \in \operatorname{argmin}(f+g)$. The Baillon-Haddad Theorem 3.38 and the convexity of f give, respectively,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k) \cdot (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ f(x_k) &\leq f(\hat{x}) + \nabla f(x_k) \cdot (x_k - \hat{x}). \end{aligned}$$

Summing the two inequalities, and multiplying the result by 2γ, we get

$$2\gamma f(x_{k+1}) \leq 2\gamma f(p) + 2\gamma \nabla f(x_k) \cdot (x_{k+1} - \hat{x}) + \gamma L\|x_{k+1} - x_k\|^2.$$

Since $x_k - x_{k+1} - \gamma\nabla f(x_k) \in \gamma\partial g(x_{k+1})$, the convexity of$g$, implies that

$$2\gamma g(x_{k+1}) \leq 2\gamma g(\hat{x}) + 2\big(x_k - x_{k+1} - \gamma\nabla f(x_k)\big) \cdot (x_{k+1} - \hat{x}).$$

Upon combining both inequalities, it follows that

$$2\gamma\big((f+g)(x_{k+1}) - \min(f+g)\big) \leq 2(x_k - x_{k+1}) \cdot (x_{k+1} - \hat{x}) + \gamma L\|x_{k+1} - x_k\|^2.$$

We may then rewrite the dot product, to deduce that

$$\begin{aligned} 2\gamma\big((f+g)(x_{k+1}) - \min(f+g)\big) &\leq \|x_k - p\|^2 - \|x_{k+1} - p\|^2 + (\gamma L - 1)\|x_{k+1} - x_k\|^2 \\ &\leq \|x_k - p\|^2 - \|x_{k+1} - p\|^2 \end{aligned}$$

Summing for $k = 0, \dots, n-1$, using the telescopic property, and keeping in mind that $(f+g)(x_{k+1}) \leq (f+g)(x_k)$ for $k \leq n$, we finally see that

$$2\gamma n\big((f+g)(x_k) - \min(f+g)\big) \leq \|x_0 - \hat{x}\|^2 - \|x_k - \hat{x}\|^2 \leq \|x_0 - \hat{x}\|^2.$$

It suffices to take $\hat{x} = P_{\operatorname{argmin}(f+g)} x_0$ and divide by $2\gamma n$ to obtain (48). The convergence is left as an exercise to the reader. □

A function $\Phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ satisfies the *Polyak-Łojasiewicz* (PŁ) *Inequality* with constant $\mu > 0$ if

$$2\mu\big(f(x) - \min(f)\big) \leq \min\left\{\|v\|^2 : v \in \partial f(x)\right\} \tag{52}$$

for all $x \in \operatorname{dom}(\partial f)$. As in the finite case, strongly convex functions have this property. For these kinds of functions, the following holds:

Theorem 5.34. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and L-smooth, let $g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be closed and convex, and set $\gamma < 2/L$. Define (x_k) by the proximal-gradient method (46). If $\Phi := f+g$ satisfies the PŁ Inequality with constant $\mu > 0$, and $\operatorname{argmin}(\Phi) \neq \emptyset$, then*

$$\Phi(x_k) - \min(\Phi) \leq \frac{\Phi(x_0) - \min(\Phi)}{(1+\eta)^k},$$

where $\eta = \mu(2 - \gamma L)$.

Proof. As in the proof of Theorem 5.33, since $w_{k+1} \in \partial(f+g)(x_{k+1})$, (51) holds. On the other hand, the Polyak-Łojasiewicz Inequality gives

$$\|w_{k+1}\|^2 \geq \|\partial(f+g)^0(x_{k+1})\|^2 \geq \mu^2\big((f+g)(x_{k+1}) - \min(f+g)\big).$$

Combining this with (51), we conclude that

$$(1+\eta)\big((f+g)(x_{k+1}) - \min(f+g)\big) \leq \big((f+g)(x_k) - \min(f+g)\big),$$

which immediately gives the result. □

6 Duality

Duality in optimization refers to the principle of viewing a problem from two distinct perspectives, namely the one of the primal or the one of the dual. There are two major types of duality. The first relies on the concept of conjugate functions (Subsection 6.1), namely the Fenchel-Rockafeller duality (Subsection 6.2), and gives rise to the so-called Primal-Dual Method (Subsection 6.3). The second is based on Lagrange multipliers (Subsection 6.4), namely the Lagrangian duality (Subsection 6.5).

6.1 Conjugate functions

The *Fenchel conjugate* of $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is the extended real-valued function $f^* : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, defined by

$$f^*(y) = \sup_{x\in\mathbb{R}^N} \{y \cdot x - f(x)\}. \tag{53}$$

According to (53), f^* is the supremum of continuous affine functions. Therefore, Exercises 5.14 and 5.17 give:

Proposition 6.1. *If $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is not identically $+\infty$, then f^* is closed and convex.*

The following proposition shows that the Fenchel conjugate is well defined if f is closed and convex. Its proof is technical, so it can be omitted in a first reading.

Proposition 6.2. *If $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is closed and convex, then $f^*(y) > -\infty$ for every $y \in \mathbb{R}^N$ and f^* is not identically $+\infty$.*

Proof. First, take any $\bar{x} \in \mathrm{dom}(f)$, so that

$$f^*(y) \sup_{x\in\mathbb{R}^N} \{y \cdot x - f(x)\} \geq y \cdot \bar{x} - f(\bar{x}) > -\infty.$$

Next, by Proposition 5.18, there exist $c \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$ such that $f(x) \geq c \cdot x + \alpha$ for all $x \in \mathbb{R}^N$. As a consequence,

$$f^*(c) = \sup_{x\in\mathbb{R}^N} \{c \cdot x - f(x)\} \leq \sup_{x\in\mathbb{R}^N} \{c \cdot x - c \cdot x - \alpha\} = -\alpha < +\infty,$$

so f^* is not identically $+\infty$. □

Combining the last two results, we obtain:

Corollary 6.3. *If $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is closed, convex and not identically $+\infty$, so is f^*.*

An immediate, yet useful consequence of the definition given by (53) is the following:

Proposition 6.4 (Fenchel-Young Inequality)**.** *If $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is closed, convex and not identically $+\infty$, then*

$$f(x) + f^*(y) \geq y \cdot x$$

for every $x, y \in \mathbb{R}^N$. Equality holds if, and only if, $y \in \partial f(x)$.

Another straightforward consequence of the definition is the subject of the following:

Exercise 6.5. Show that if $f \leq g$, then $g^* \leq f^*$.

Let us now examine some examples.

Example 6.6. Let $a > 0$, and define $f_a : \mathbb{R} \to \mathbb{R}$ by $f_a(x) = ax^2$. Then,

$$f_a^*(y) = \sup\{yx - ax^2 : x \in \mathbb{R}\} = -\inf\{ax^2 - yx : x \in \mathbb{R}\}.$$

The unique minimizer of the function $x \mapsto ax^2 - yx$ is $\hat{x} = \frac{y}{2a}$. Therefore,

$$f_a^*(y) = y\hat{x} - a\hat{x}^2 = \frac{y^2}{2a} - \frac{y^2}{4a} = \frac{y^2}{4a}.$$

In particular, $f_{\frac{1}{2}}^* = f_{\frac{1}{2}}$. Actually, this is the only function with this property. Indeed, suppose φ satisfies $\varphi^* = \varphi$. On the one hand, The Fenchel-Young Inequality gives

$$2\varphi(x) = \varphi(x) + \varphi^*(x) \geq x^2,$$

whence $\varphi(x) \geq \frac{1}{2}x^2 = f_{\frac{1}{2}}(x)$. By Exercise 6.5,

$$\varphi(x) = \varphi^*(x) \leq f_{\frac{1}{2}}^*(x) = f_{\frac{1}{2}}(x) \leq \varphi(x),$$

from which we conclude that $\varphi = f_{\frac{1}{2}}$.

Example 6.7. Let us now compute the conjugate of the exponential function $f(x) = e^x$. We have

$$f^*(y) = \sup\left\{yx - e^x : x \in \mathbb{R}\right\}.$$

If $y < 0$, we can let $x \to -\infty$, so that $yx - e^x \to +\infty$. Therefore, $f^*(y) = +\infty$ for every $y < 0$. For $y = 0$, we clearly have $f^*(0) = 0$. Finally, if $y > 0$, the function $x \mapsto e^x - yx$ is coercive, and its unique minimizer $\hat{x}$ is characterized by $0 = e^{\hat{x}} - y$, which we can write as $y = e^{\hat{x}}$ or $\hat{x} = \ln(y)$. It follows that $f^*(y) = y\ln(y) - y$ for $y > 0$. Summarizing, we have

$$f^*(y) = \begin{cases} y\ln(y) - y & \text{if } y > 0 \\ 0 & \text{if } y = 0 \\ +\infty & \text{if } y < 0. \end{cases}$$

This function is known as the *Boltzmann–Shannon Entropy.*

Exercise 6.8. Compute the conjugate of the Boltzmann–Shannon Entropy.

Example 6.9 (Affine functions)**.** Given $c \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$, define $f : \mathbb{R}^N \to \mathbb{R}$ by $f(x) = c \cdot x + \alpha$. From (53), have

$$f^*(y) = \sup\left\{y \cdot x - c \cdot x - \alpha : x \in \mathbb{R}^N\right\} = \sup\left\{(y - c) \cdot x : x \in \mathbb{R}^N\right\} - \alpha.$$

The supremum is $+\infty$ unless $y = c$, in which case it is $-\alpha$. Therefore, $f^*(y) = \iota_c(y) - \alpha$.

Example 6.10 (The support functional)**.** Given a nonempty, closed convex set $C \subset \mathbb{R}^N$, the conjugate of the indicator function ι_C, namely

$$\sigma_C(y) := \iota_C^*(y) = \sup\{y \cdot x : x \in C\},$$

is called the *support functional* of C.

The *biconjugate* of $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is $f^{**} := (f^*)^*$. According to the Fenchel-Young Inequality (Proposition 6.4) we have

$$x \cdot y - f^*(y) \leq f(x)$$

for every $x, y \in \mathbb{R}^N$. Taking the supremum over $y \in \mathbb{R}^N$, we deduce that $f^{**} \leq f$. Actually, we have the following:

Proposition 6.11. *If $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is closed, convex and not identically $+\infty$, then $f^{**} = f$.*

A consequence of this is:

Corollary 6.12 (Legendre-Fenchel Reciprocity Formula)**.** *Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be closed, convex and not identically $+\infty$. Then $x^* \in \partial f(x)$ if, and only if, $x \in \partial f^*(x^*)$.*

By combining the Reciprocity Formula with Theorem 3.38 and Proposition 3.45, we obtain the following:

Corollary 6.13. *Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be closed, convex and not identically $+\infty$, and let $\mu L = 1$. Then, f is μ-strongly convex if, and only if, f^* is L-smooth.*

6.2 Fenchel-Rockafellar duality

Consider a matrix $A \in \mathbb{R}^{M \times N}$, and let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^M \to \mathbb{R} \cup \{+\infty\}$ be closed and convex. We are concerned with solving

$$\min_{x \in \mathbb{R}^N} \big\{ f(x) + g(Ax) \big\},$$

which we call the *primal problem*. We shall denote its optimal value by $v \in \mathbb{R}$, and collect its solutions in a set $S \subset \mathbb{R}^N$. We shall refer to v as the *primal value* and the elements of S as the *primal solutions*.

We also define a *dual problem* as

$$\min_{y \in \mathbb{R}^M} \big\{ f^*(-A^T y) + g^*(y) \big\},$$

with *dual optimal value* $v^* \in \mathbb{R}$, and set of *dual solutions* $S^* \subset \mathbb{R}^M$.

The quantity $v + v^*$, known as the *duality gap*, is nonnegative, a fact known as *weak duality*. Indeed, by the Fenchel-Young Inequality (Proposition 6.4), we have

$$\begin{aligned} f(x) + f^*(-A^T y) &\geq -x \cdot A^T y \\ g(Ax) + g^*(y) &\geq Ax \cdot y. \end{aligned}$$

Summing the two inequalities, and recalling that $Ax \cdot y = x \cdot A^T y$, we obtain

$$\big[f(x) + g(Ax) \big] + \big[f^*(-A^T y) + g^*(y) \big] \geq 0. \tag{54}$$

We conclude by minimizing on x and y. If the duality gap is zero, we speak of *strong duality*. The primal and dual solutions can be characterized as follows:

Theorem 6.14 (Strong Duality)**.** *The following statements concerning $\hat{x} \in \mathbb{R}^N$ and $\hat{y} \in \mathbb{R}^M$ are equivalent:*

i) $\hat{x} \in S$, $\hat{y} \in S^*$;

ii) $-A^T \hat{y} \in \partial f(\hat{x})$ *and* $\hat{y} \in \partial g(A\hat{x})$;

iii) $f(\hat{x}) + f^*(-A^T \hat{y}) = -A^T \hat{y} \cdot \hat{x}$ *and* $g(A\hat{x}) + g^*(\hat{y}) = \hat{y} \cdot A\hat{x}$; *and*

iv) $f(\hat{x}) + g(A\hat{x}) + f^*(-A^T \hat{y}) + g^*(\hat{y}) = 0$.

If the statements above hold, then $v + v^* = 0$. Suppose, moreover, that either f is continuous and $g \circ A$ is not identically $+\infty$, or that g is continuous. If $\hat{x} \in S$, there is $\hat{y} \in \mathbb{R}^M$ such that all four statements hold.

Let us first analyze the case of linear equality constraints.

Example 6.15 (Dualizing linear equality constraints)**.** Consider a matrix $A \in \mathbb{R}^{M\times N}$, a vector $b \in \mathbb{R}^M$, and let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be closed and convex. Observe that

$$\min\big\{f(x) : Ax = b\big\} = \min\big\{f(x) + \iota_{\{b\}}(Ax)\big\}, \tag{55}$$

so this is a primal problem in the Fenchel-Rockafellar duality with $g = \iota_{\{b\}}$. Since $g^*(y) = b \cdot y$, the dual is

$$\min\big\{f^*(-A^T y) + b \cdot y\big\}. \tag{56}$$

According to Theorem 6.14, $\hat{x}$ is a solution to (55) if, and only if, $A\hat{x} = b$ and there is $\hat{y} \in \mathbb{R}^M$ such that $-A^T\hat{y} \in \partial f(\hat{x})$. In turn, this $\hat{y}$ is a solution to (56).

Exercise 6.16. Contrast Examples 6.15 and 3.19. Does there seem to be an error in the signs?

6.3 The Primal-Dual Method

In Example 6.15, suppose that f is continuous and strongly convex. Then, the primal problem always has a solution, so that, by Theorem 6.14, the dual also has a solution, and there is no duality gap. According to Corollary 6.13, the function f^* is smooth. Therefore, the gradient method can be applied to solve the dual problem.

If f is not strongly convex, then f^* is not smooth. In principle, one could apply the proximal algorithm. However, from a practical perspective, the computation of the proximity operator corresponding to the composition $f^* \circ (-A)$ may be challenging. This difficulty can be overcome by simultaneously solving the primal and the dual problems.

Consider a *structured* optimization problem, of the form

$$\min\{f(x) + g(Ax) + h(x)\}, \tag{57}$$

where $A \in \mathbb{R}^{M\times N}$, $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^M \to \mathbb{R} \cup \{+\infty\}$ are closed and convex, and $h : \mathbb{R}^N \to \mathbb{R}$ is ℓ-smooth and convex.

The *Primal-Dual Method* (Chambolle-Pock, 2011; Condat-Vũ, 2013) starts from an initial guess (x_0, y_0), and iterates:

$$\left\{\begin{array}{rcl} x_{k+1} & = & \operatorname{prox}_{\tau f}\big(x_k - \tau\nabla h(x_k) - \tau A^T y_k\big) \\ y_{k+1} & = & \operatorname{prox}_{\sigma g^*}\big(y_k + \sigma A(2x_{k+1} - x_k)\big), \end{array}\right.$$

with $\tau, \sigma > 0$. We have the following:

Proposition 6.17. *Every subsequential limit point of (x_k, y_k) is a primal-dual solution.*

Proof. By continuity, if $(x_{m_k}, y_{m_k}) \to (\bar{x}, \bar{y})$, we can pass to the limit in the equalities defining the Primal-dual Algorithm to obtain

$$\left\{\begin{array}{rcl} \bar{x} & = & \operatorname{prox}_{\tau f}\big(\bar{x} - \tau\nabla h(\bar{x}) - \tau A^T \bar{y}\big) \\ \bar{y} & = & \operatorname{prox}_{\sigma g^*}\big(\bar{y} + \sigma A\bar{x}\big). \end{array}\right.$$

The first equality is equivalent to

$$\bar{x} + \tau\partial f(\bar{x}) \ni \bar{x} - \tau\nabla h(\bar{x}) - \tau A^T\bar{y},$$

which we can simplify to

$$-A^T\bar{y} \in \partial f(\bar{x}) + \nabla h(\bar{x}) = \partial(f+h)(\bar{x}).$$

The second one is

$$\bar{y} + \sigma\partial g^*(\bar{y}) \ni \bar{y} + \sigma A\bar{x},$$

or, equivalently,

$$A\bar{x} \in \partial g^*(\bar{y}).$$

By the Reciprocity Formula, this is precisely

$$\bar{y} \in \partial g(A\bar{x}).$$

Theorem 6.14 then shows that $(\bar{x}, \bar{y})$ is a primal-dual solution. □

Exercise 6.18. Let (x_k, y_k) be generated by the Primal-Dual Method. Show that the pair (x_k, y_k) is a primal-dual solution if, and only if, $x_{k+1} = x_k$ and $y_{k+1} = y_k$.

Theorem 6.19. *If $\tau\sigma\|A\|^2 + \frac{\tau\ell}{2} \leq 1$, the sequence (x_k, y_k) converges to a primal-dual solution.*

The following result allows us to circumvent the computation of g^*, and is useful for the purpose of implementation.

Proposition 6.20 (Moreau's Identity)**.** *Let $g : \mathbb{R}^M \to \mathbb{R} \cup \{+\infty\}$ be closed and convex, and let $\sigma > 0$. For every $y \in \mathbb{R}^M$, we have*

$$\operatorname{prox}_{\sigma g^*}(y) = y - \sigma \operatorname{prox}_{\sigma^{-1}g}\left(\sigma^{-1}y\right).$$

Proof. Write $u = \operatorname{prox}_{\sigma g^*}(y)$, so that $u + \sigma\partial g^*(u) \ni y$. By the Reciprocity Formula (Corollary 6.12), this is equivalent to

$$u \in \partial g\left(\frac{y-u}{\sigma}\right),$$

which we rewrite as

$$\frac{y}{\sigma} \in \left(\frac{y-u}{\sigma}\right) + \frac{1}{\sigma}\partial g\left(\frac{y-u}{\sigma}\right).$$

But this is the same as

$$\left(\frac{y-u}{\sigma}\right) = \operatorname{prox}_{\frac{1}{\sigma}g}\left(\frac{y}{\sigma}\right).$$

Multiplying by σ and recalling the meaning of u, we recover precisely Moreau's Identity. □

Example 6.21 (Linear equality constraints)**.** Consider the problem

$$\min\left\{f(x) + h(x) : Ax = b\right\},$$

which is a particular case of (57) with $g = \iota_{\{b\}}$, as in Example 6.15. Using Moreau's Identity, we have $\operatorname{prox}_{\sigma g^*}(y) = y - \sigma b$, for each $y \in \mathbb{R}^M$, and the Primal-dual Algorithm is reduced to

$$\begin{cases} x_{k+1} &= \operatorname{prox}_{\tau f}\left(x_k - \tau\nabla h(x_k) - \tau A^T y_k\right) \\ y_{k+1} &= y_k + \sigma(Ax_{k+1} - b) + \sigma A(x_{k+1} - x_k), \end{cases}$$

with $\tau, \sigma > 0$.

Example 6.22 (The ROF Model)**.** The *Total Variation Regularization Problem* (Rudin-Osher-Fatemi, 1992) is

$$\min_{x \in \mathbb{R}^{N_1 \times N_2}} \left\{ \frac{1}{2} \|Fx - b\|^2 + \rho \|Dx\|_1 \right\},$$

where F models or approximates the process by which an image x has been modified (usually deteriorated) to produce b, and D is the *discrete gradient*. This model is discussed further in Section 6.8.

Exercise 6.23. Can we apply the Primal-Dual Method to this problem?

6.4 A complementary approach to linear equality constraints

Given a differentiable function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, a matrix A of size $M \times N$ and a vector $b \in \mathbb{R}^M$, consider the problem

$$(P) \qquad \min \{ f(x) : Ax = b \}.$$

As in Examples 3.19 and 6.15, if $\hat{x}$ is a solution of (P), then $A\hat{x} = b$ and there is $\hat{z} \in \mathbb{R}^M$ such that

$$-A^T \hat{z} = \nabla f(\hat{x}),$$

which can be rewritten as

$$0 = \nabla f(\hat{x}) + A^T \hat{z} = \nabla f(\hat{x}) + \nabla G_{\hat{z}}(\hat{x}), \qquad \text{where} \qquad G_z(x) = z^T(Ax - b).$$

The *Lagrangian* associated to (P) is the function $L : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$, defined by

$$L(x, z) = f(x) + z^T(Ax - b).$$

Here, x is the *primal variable* and z is the *dual variable*, more commonly referred to as *(Lagrange) multiplier*.

From the discussion above, if $\hat{x}$ is a solution to (P) and $\hat{z}$ is the corresponding multiplier, then

$$0 = \nabla L(\hat{x}, \hat{z}) = \begin{pmatrix} \nabla f(\hat{x}) + A^T \hat{z} \\ A\hat{x} - b \end{pmatrix}.$$

In this context, the equalities

$$\begin{cases} A\hat{x} &= b \\ -A^T \hat{z} &= \nabla f(\hat{x}) \end{cases}$$

are usually referred to as *primal* and *dual feasibility*, respectively.

6.4.1 Multiplier methods

In what follows, we discuss a family of algorithms, based on Lagrangian duality, that allow us to solve different versions of this problem.

The *Method of Multipliers* uses the *augmented Lagrangian*

$$\mathcal{L}_\alpha(x, z) = f(x) + z^T(Ax - b) + \frac{\alpha}{2} \|Ax - b\|^2,$$

and computes $(x_k, z_k) \mapsto (x_{k+1}, z_{k+1})$ as follows:

$$\left\{ \begin{array}{lcl} x_{k+1} & \in & \operatorname{argmin} \left\{ \mathcal{L}_\alpha(x, z_k) : x \in \mathbb{R}^N \right\}, \\ z_{k+1} & = & z_k + \alpha(Ax_{k+1} - b). \end{array} \right. \tag{58}$$

If f is differentiable, the optimality condition for the definition of x_{k+1} is

$$0 = \nabla f(x_{k+1}) + A^T z_k + \alpha A^T(Ax_{k+1} - b).$$

Multiplying the second equality from the left by A^T, we obtain

$$\nabla f(x_{k+1}) + A^T z_{k+1} = \nabla f(x_{k+1}) + A^T z_k + \alpha A^T(Ax_{k+1} - b) = 0.$$

Therefore, dual feasibility is automatically enforced.

The Method of Multipliers has its limitations:

1. The iterate x_{k+1} may not be well defined. Indeed, the function

$$\mathcal{L}_\alpha(x, z_k) = f(x) + z_k^T(Ax - b) + \frac{\alpha}{2}\|Ax - b\|^2$$

   may not have a minimizer if f is not strongly convex and A has a nontrivial kernel. Also, even if there is a minimizer, it may not be unique.

2. If f is not differentiable, the computation of x_{k+1} may be more involved.

If the function f is closed and convex, we can implement a *proximal* version of the Method of Multipliers

$$\left\{ \begin{array}{lcl} x_{k+1} & \in & \operatorname{argmin} \left\{ \mathcal{L}_\alpha(x, z_k) + \frac{\alpha}{2}\|x - x_k\|^2 : x \in \mathbb{R}^N \right\}, \\ z_{k+1} & = & z_k + \alpha(Ax_{k+1} - b). \end{array} \right. \tag{59}$$

The function $\mathcal{L}_\alpha(x, z_k) + \frac{\alpha}{2}\|x - x_k\|^2$ is closed and strongly convex, so it always has a unique minimizer. This solves the first of the issues mentioned above. It does not solve the second one, though, even if the proximity operator for f is easy to compute. Indeed, we have

$$\mathcal{L}_\alpha(x, z_k) + \frac{\alpha}{2}\|x - x_k\|^2 = f(x) + z_k^T(Ax - b) + \frac{\alpha}{2}\|Ax - b\|^2 + \frac{\alpha}{2}\|x - x_k\|^2.$$

The second term is affine, and can be absorbed into either the third or the last one. The difficulty comes from the quadratic term, since there is no explicit formula for the proximity operator of a sum of functions.

Exercise 6.24. Show that, if $A^T A$ is a multiple of the identity, then x_{k+1} in (59) can be explicitly written in terms of simple algebraic operations and the proximity operator of f.

One way to overcome this problem is to use the *Predictor-corrector Method*, either in its standard form:

$$\left\{ \begin{array}{lcl} p_{k+1} & = & z_k + \alpha(Ax_k - b) \\ x_{k+1} & = & \operatorname{argmin} \left\{ L(x, p_{k+1}) + \frac{1}{2\alpha}\|x - x_k\|^2 \ : \ x \in \mathbb{R}^N \right\} \\ z_{k+1} & = & z_k + \alpha(Ax_{k+1} - b), \end{array} \right.$$

or in its *proximal* form, which the reader can easily guess.

Finally, it is frequent to encounter problems involving two sets of variables, structured as

$$\min\left\{f(x)+g(y):(x,y)\in\mathbb{R}^{N_1\times N_2},\ Ax+By=c\right\}.$$

In principle, we could use the multiplier method as it is, but this would imply solving a minimization problem in the product space, where the different sets of variables are linked by the quadratic term, instead of exploiting the structure of the objective function, where each summand depends only on one set of variables. With this in mind, we can minimize *in series*: first only with respect to the x variable, and then with respect to y. The result of this process is the *Alternating Direction Method of Multipliers (ADMM)*:

$$\begin{cases} x_{k+1} & = & \operatorname{argmin}\{\mathcal{L}_\alpha(x,y_k,z_k):x\in\mathbb{R}^{N_1}\}, \\ y_{k+1} & = & \operatorname{argmin}\{\mathcal{L}_\alpha(x_{k+1},y,z_k):y\in\mathbb{R}^{N_2}\}, \\ z_{k+1} & = & z_k+\alpha(Ax_{k+1}+By_{k+1}-c). \end{cases}$$

Parallel, proximal and predictor-corrector variants can also be implemented in this case.

6.5 Optimality conditions: mathematical programming and Lagrangian duality

Let $f,g_j,h_m\in\mathcal{C}^1(\mathbb{R}^N;\mathbb{R})$, and consider the *mathematical programming* problem, namely:

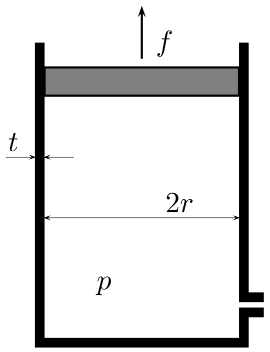
$$\min_{x\in C} f(x), \tag{$\mathcal{P}$}$$

where the feasible set C is given by

$$C=\{\,x\in\mathbb{R}^N\ :\ h_m(x)=0,\ g_j(x)\le 0,\ \forall m,j\,\}.$$

Example 6.25 (Hydraulic Cylinder Design)**.** We wish to design a hydraulic cylinder (as shown in the figure) of minimal width $r+t$, which is to provide a force $f=\pi r^2p$, where p is the pressure, of at least f_{min}, with a hoop stress $s=\frac{pr}{t}$ of at most s_{max}, and a wall thickness t no less than t_{min}. This means:

$$\begin{aligned} \text{minimize}\quad & r+t \\ \text{subject to}\quad & pr-ts_{max}\le 0 \\ & f_{min}-\pi r^2p\le 0 \\ & t_{min}-t\le 0. \end{aligned}$$



In what follows, we shall establish optimality conditions to characterize the solutions to the mathematical programming problem.

6.5.1 Step 1: no inequality constraints

We begin by establishing *necessary conditions* for optimality, in case no inequality constraints are present.

A feasible point $\hat{x} \in C$ is *regular* if the set

$$V = \{\nabla h_1(\hat{x}), \nabla h_2(\hat{x}), \dots, \nabla h_M(\hat{x})\}$$

is linearly independent.

Theorem 6.26 (Lagrange Multiplier Theorem)**.** *Let $\hat{x}$ be a regular local minimizer of ($\mathcal{P}$). There is a unique $\hat{\lambda} \in \mathbb{R}^M$, called Lagrange multiplier vector, such that*

$$\nabla f(\hat{x}) + \sum_{m=1}^{M} \hat{\lambda}_m \nabla h_m(\hat{x}) = 0. \tag{60}$$

If, in addition, $f, h_1, h_2, \dots, h_M \in \mathcal{C}^2(\mathbb{R}^N; \mathbb{R})$, then

$$y \cdot \left(\nabla^2 f(\hat{x}) + \sum_{m=1}^{M} \hat{\lambda}_m \nabla^2 h_m(\hat{x}) \right) y \geq 0 \tag{61}$$

for all $y \in V^\perp$.

Proof. Take $r > 0$, such that $\hat{x}$ minimizes f over $C \cap B$, for the compact set $B = \bar{B}(\hat{x}, r)$, and let c be a lower bound for $f(x)$ over B. For each $k \in \mathbb{N}$, the function $f_k : \mathbb{R}^N \to \mathbb{R}$, defined by

$$f_k(x) = f(x) + \frac{\varepsilon}{2}\|x - \hat{x}\|^2 + \frac{k}{2}\sum_{m=1}^{M} |h_m(x)|^2,$$

attains its minimum over B at some $x_k \in B$. In particular, we have

$$f(x_k) + \frac{\varepsilon}{2}\|x_k - \hat{x}\|^2 + \frac{k}{2}\sum_{m=1}^{M} |h_m(x_k)|^2 = f_k(x_k) \leq f_k(\hat{x}) = f(\hat{x}).$$

This implies that

$$\sum_{m=1}^{M} |h_m(x_k)|^2 \quad \leq \quad \frac{2\big(f(\hat{x}) - c\big)}{k} \tag{62}$$

$$f(x_k) + \frac{\varepsilon}{2}\|x_k - \hat{x}\|^2 \quad \leq \quad f(\hat{x}). \tag{63}$$

We shall prove that (x_k) converges to $\hat{x}$ as $k \to \infty$. Since (x_k) is bounded, it suffices to show that $\hat{x}$ is the only possible subsequential limit of (x_k). Suppose, then, that $x_{m_k} \to x_\infty$. Passing to the limit in (62), we deduce that $h_m(x_\infty) = 0$ for every m, which says that $x_\infty \in C$. Next, passing to the limit in (63), we get

$$f(x_\infty) + \frac{\varepsilon}{2}\|x_\infty - \hat{x}\|^2 \leq f(\hat{x}) \leq f(x_\infty),$$

by the optimality of $\hat{x}$ on C, and this shows that $x_\infty = \hat{x}$. As a consequence, there is $k_0 \in \mathbb{N}$ such that x_k minimizes f_k in $\operatorname{int}(B)$, whence

$$0 = \nabla f_k(x_k) = \nabla f(x_k) + \varepsilon(x_k - \hat{x}) + \sum_{m=1}^{M} \big[k h_m(x_k)\big] \nabla h_m(x_k). \tag{64}$$

For each m, let V_m be the subspace generated by $\{\nabla h_m(\hat{x}) : i \neq m\}$, and write $v_m = P_{V_m^\perp}\big(\nabla h_m(\hat{x})\big)$, which is not zero by linear independence. Multiplying (64) by v_m, passing to the limit as $k \to \infty$, and using the perpendicularity, we deduce that

$$\lim_{k\to\infty} kh_m(x_k)\big[v_m \cdot \nabla h_m(\hat{x})\big] = -v_m \cdot \nabla f(\hat{x}).$$

Since $v_m \cdot \nabla h_m(\hat{x}) \neq 0$, $\lim\limits_{k\to\infty} kh_m(x_k)$ exists, for each m. Writing $\hat{\lambda}_m = \lim\limits_{k\to\infty} kh_m(x_k)$, and passing to the limit in (64), we obtain precisely (60). The second order condition (61) is left as an exercise to the reader. □

The conclusion of the Lagrange Multiplier Theorem may not hold if the point $\hat{x}$ is not regular, as seen in the following:

Example 6.27. Let $f, h_1, h_2 : \mathbb{R}^2 \to \mathbb{R}$ be defined as follows:

$$\left\{\begin{array}{rcl} f(x,y) & = & x \\ h_1(x,y) & = & x^2 - y \\ h_2(x,y) & = & y. \end{array}\right.$$

We have $C = \{(0,0)\}$, so $\min_C(f) = f(0,0) = 0$. On the other hand,

$$\left\{\begin{array}{rcl} \nabla f(0,0) & = & (1,0)^T \\ \nabla h_1(0,0) & = & (0,-1)^T \\ \nabla h_2(0,0) & = & (0,1)^T. \end{array}\right.$$

Clearly, $\nabla f(0,0)$ cannot be written as a linear combination of $\nabla h_1(0,0)$ and $\nabla h_2(0,0)$, and the conclusion of Theorem 6.26 is not valid. Of course, since $\nabla h_1(0,0) = -\nabla h_2(0,0)$, the solution $(0,0)$ *is not* regular.

Sufficient conditions for optimality are given by the following:

Theorem 6.28. *Let $f, h_1, h_2, \ldots, h_M : \mathbb{R}^N \to \mathbb{R}$ be twice continuously differentiable, and let $\hat{x} \in C$ and $\hat{\lambda} \in \mathbb{R}^M$ satisfy*

$$\nabla f(\hat{x}) + \sum_{m=1}^{M} \hat{\lambda}_m \nabla h_m(\hat{x}) = 0$$

and

$$y \cdot \left(\nabla^2 f(\hat{x}) + \sum_{m=1}^{M} \hat{\lambda}_m \nabla^2 h_m(\hat{x}) \right) y > 0$$

for all $y \in V^\perp \setminus \{0\}$. Then, $\hat{x}$ is a strict local minimizer of $(\mathcal{P})$.

6.5.2 The general case

The set of *active inequality constraints* at a feasible point $\hat{x}$ is

$$A(\hat{x}) = \{\, j \ : \ g_j(\hat{x}) = 0 \,\}.$$

We redefine the notion of regularity as follows: a feasible point $\hat{x} \in C$ is *regular* if the set

$$V = \big\{ \nabla g_j(\hat{x}), \nabla h_m(\hat{x}) \ : \ j \in A(\hat{x}),\ m = 1, \ldots, M \big\}$$

is linearly independent. In the absence of inequality constraints, this is consistent with the terminology introduced above.

Theorem 6.29 (Karush-Kuhn-Tucker conditions)**.** *Let $\hat{x}$ be a regular local minimizer of $(\mathcal{P})$. There exist unique Lagrange multiplier vectors $\hat{\lambda} \in \mathbb{R}^M$ and $\hat{\mu} \in \mathbb{R}^J_+$, such that*

$$\nabla f(\hat{x}) + \sum_{j=1}^{J} \hat{\mu}_j \nabla g_j(\hat{x}) + \sum_{m=1}^{M} \hat{\lambda}_m \nabla h_m(\hat{x}) = 0,$$

and $\hat{\mu}_j g_j(\hat{x}) = 0$ for all j. If, in addition, $f, g_j, h_m \in \mathcal{C}^2(\mathbb{R}^N; \mathbb{R})$, then

$$y \cdot \left(\nabla^2 f(\hat{x}) + \sum_{m=1}^{M} \hat{\lambda}_m \nabla^2 h_m(\hat{x}) \right) y \geq 0$$

for all $y \in \mathcal{V}^{\perp}$, where $\mathcal{V} = \{\nabla g_j(\hat{x}), \nabla h_m(\hat{x}),\ j \in A(\hat{x}),\ m = 1, \ldots M\}$.

6.6 Exercise: Linear programming

The *Linear Programming* (LP) problem is given by

$$\min_{x \in \mathbb{R}^N} \{c \cdot x \colon Ax \leq b\},$$

where $c \in \mathbb{R}^N$, A is a matrix of size $M \times N$, and $b \in \mathbb{R}^M$.

1. Write the LP as a primal problem in the sense of Fenchel-Rockafeller.
2. Show the dual problem (in the sense of Fenchel-Rockafeller) of the LP problem is given by

$$\min_{y \in \mathbb{R}^M} \{b \cdot y \colon A^T y + c = 0, \text{ and } y \geq 0\}.$$

3. Show the dual problem (in the sense of Lagrange) coincides with the Fenchel-Rockafeller dual in the case of linear programming.
4. Compute the dual of the dual, both in the sense of Fenchel-Rockafeller and Lagrange, and show that it coincides with the primal problem.

6.7 Exercise: Comparison between primal and dual

This assignment aims to show that there may be an advantage of solving either the primal or the dual problem over solving the other. This this end, follow the subsequent steps:

1. Compute the Fenchel conjugates of the functions
   (a) $g : \mathbb{R}^N \to \mathbb{R}$, defined by $g(x) = \|x\|_1$.
   (b) $h : \mathbb{R}^M \to \mathbb{R}$, given by $h(y) = \frac{1}{2}\|y - b\|^2$, where $b \in \mathbb{R}^M$.
2. Determine the Fenchel-Rockafellar dual of the primal problem:

$$\min_{x \in \mathbb{R}^N} \left\{ g(x) + h(Ax) \right\},$$

where A is a real matrix of size $M \times N$. Observe that the proximal-gradient method can be applied to solve the dual problem. Specifically, discuss whether it is computationally feasible to apply the proximal-gradient method.

3. What is the convergence rate of the proximal-gradient method applied to the dual? Use only information available this far, specifically do not use the information provided in question 4.

4. Is the function $y \mapsto g^*(-A^*y) + h^*(y)$ strongly convex? Would you expect this to influence your answer in 3? Compare the primal and the dual problem, and elaborate which one you think is easier to solve.

6.8 Computational exercise: Image deblurring

This computational exercise focuses on the task of image deblurring. You will begin with a known non-blurry image, apply a blurring effect to it, and subsequently implement optimization methods learned in class to reverse the blur. Figure 1 illustrates an example of the expected outcome.

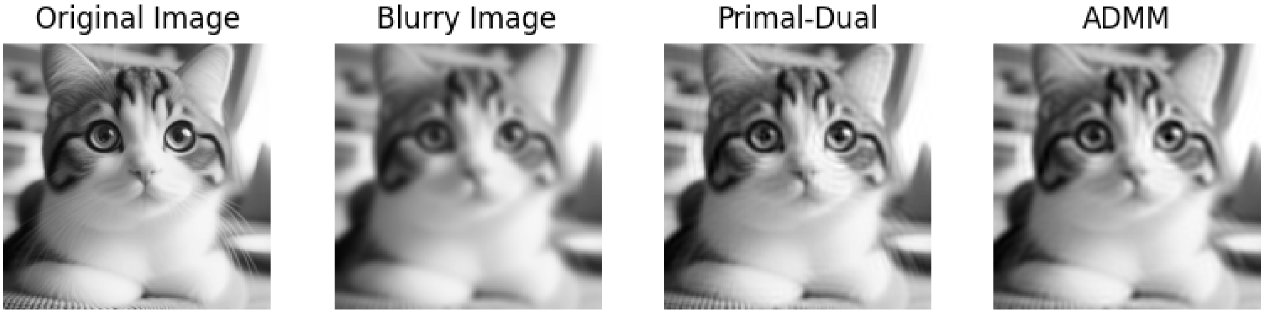


Figure 1: Expected final result.

The core challenge in image deblurring involves balancing two competing objectives:

1. Sharpness: The resulting image X should be as little blurry as possible. This is encouraged by minimizing the total variation, measured by the 1-norm of its discrete gradient, $\|\mathrm{grad}(X)\|_1$.

2. Data Fidelity: The result X, if re-blurred by the blurring operator R, should closely match the observed blurred image X_{blur}. This is measured by the squared 2-norm of the difference, $\|R(X) - X_{\mathrm{blur}}\|^2$.

The *Rudin-Osher-Fatemi (ROF) model* allows to handle both objectives simultaneously, by combining them linearly as

$$\min_{X \in \mathbb{R}^{256\times 256}} \left\{ \frac{\lambda}{2} \|R(X) - X_{\mathrm{blur}}\|_2^2 + \|\mathrm{grad}(X)\|_1 \right\}, \tag{65}$$

where R is the blurring operator, X_{blur} is the input blurred image, and $\mathrm{grad}(X)$ is the discrete gradient of X. Note both R and grad are linear operators. You may assume without proof that $\|\mathrm{grad}\| \leq 2\sqrt{2}$.
By introducing an auxiliary variable $Y \in \mathbb{R}^{2\times 256\times 256}$, the problem may equivalently be written as

$$\min_{X \in \mathbb{R}^{256\times 256}, Y \in \mathbb{R}^{2\times 256\times 256}} \left\{ \frac{\lambda}{2} \|R(X) - X_{\mathrm{blur}}\|_2^2 + \|Y\|_1 \right\} \quad \text{subject to} \quad \mathrm{grad}(X) = Y. \tag{66}$$

The norms used in these models are defined for the underlying image matrices (treated as vectors when calculating the norm). Specifically, for $X, Y_1, Y_2 \in \mathbb{R}^{256\times 256}$,

$$\|X\|_2 \coloneqq \sqrt{\sum_{i,j}(X_{i,j})^2} \quad \text{and} \quad \|(Y_1, Y_2)\|_1 \coloneqq \sum_{i,j}\sqrt{((Y_1)_{i,j})^2 + ((Y_2)_{i,j})^2}.$$

Solve the image deblurring problem defined by the ROF model using the following two optimization methods;

1. the Primal-Dual Method using formulation (65),

2. the Alternating Direction Method of Multipliers using formulation (66).

To facilitate this exercise, functions for blurring images, implementing the discrete gradient operator and its adjoint, and plotting images are provided within the `RuG_IntroToOptimization` Python library. You must install this library. The documentation[2] contains installation instructions and examples of its usage. Therefore, your implementation efforts should concentrate solely on coding the algorithms themselves, as the necessary data handling and visualization tools are supplied.

[2] See https://pypi.org/project/RuG-IntroToOptimization/.

7 Self-Assessment

This section serves as a self-assessment, allowing the reader to verify their knowledge of all of the previous sections.

7.1 Exercise: Mock Exam

Consider the problem $(\mathcal{P})$ of minimizing a continuous convex function $f\colon \mathbb{R}^N \to \mathbb{R}$ over the affine subspace $V = \{x \in \mathbb{R}^N \colon Ax = b\}$, for given $A \in \mathbb{R}^{M\times N}$ and $b \in \mathbb{R}^M$.

1. If ι_V denotes the indicator function of V, prove that $\partial\iota_V(x) = \operatorname{ran}(A^*)$ for each $x \in \mathbb{R}^N$. Suggestion: How is V related to $\ker(A)(= \operatorname{ran}(A^*)^\perp)$?

2. Use the first order optimality condition for $(\mathcal{P})$, obtained from Fermat's Rule, to show that $\hat{x}$ is a solution of $(\mathcal{P})$ if, and only if, $A\hat{x} = b$ and there exists $\hat{y} \in \mathbb{R}^M$ such that $-A^*\hat{y} \in \partial f(\hat{x})$.[3] We say $(\hat{x}, \hat{y})$ is an *optimal pair*. Is this related to Lagrange multipliers?

3. Define the *Lagrangian* of the problem by $\mathcal{L}(x, y) = f(x) + y \cdot (Ax - b)$, for $(x, y) \in \mathbb{R}^N \times \mathbb{R}^M$. Show that if $(\hat{x}, \hat{y})$ is an optimal pair, then

$$\mathcal{L}(\hat{x}, y) \le \mathcal{L}(\hat{x}, \hat{y}) \le \mathcal{L}(x, \hat{y})$$

for all $(x, y) \in \mathbb{R}^N \times \mathbb{R}^M$.

In what follows, we establish the convergence of the algorithm given by

$$\begin{cases} p_{k+1} &= \operatorname{argmax}\left\{\mathcal{L}(x_k, y) - \frac{1}{2\gamma}\|y - y_k\|^2 \colon y \in \mathbb{R}^M\right\} \\ x_{k+1} &= \operatorname{argmin}\left\{\mathcal{L}(x, p_{k+1}) + \frac{1}{2\gamma}\|x - x_k\|^2 \colon x \in \mathbb{R}^N\right\} \\ y_{k+1} &= \operatorname{argmax}\left\{\mathcal{L}(x_{k+1}, y) - \frac{1}{2\gamma}\|y - y_k\|^2 \colon y \in \mathbb{R}^M\right\}, \end{cases}$$

with $\gamma > 0$, and starting from an initial point $(x_0, y_0) \in \mathbb{R}^N \times \mathbb{R}^M$.

4. Write the optimality conditions corresponding to the three subiterations, in order to find closed formulates for p_{k+1} and y_{k+1}, and to express x_{k+1} in terms of a proximal step.

In parts 5, 6 and 7, $(\hat{x}, \hat{y})$ is any optimal pair.

5. Prove that

$$\begin{aligned} 2\gamma(\mathcal{L}(x_{k+1}, p_{k+1}) - \mathcal{L}(\hat{x}, p_{k+1})) &\le \|x_k - \hat{x}\|^2 - \|x_{k+1} - \hat{x}\|^2 - \|x_{k+1} - x_k\|^2 \\ 2\gamma(\mathcal{L}(x_{k+1}, \hat{y}) - \mathcal{L}(x_{k+1}, y_{k+1})) &\le \|y_k - \hat{x}\|^2 - \|y_{k+1} - \hat{x}\|^2 - \|y_{k+1} - y_k\|^2 \\ 2\gamma(\mathcal{L}(x_{k+1}, y_{k+1}) - \mathcal{L}(x_{k+1}, p_{k+1})) &\le \delta\|y_{k+1} - p_{k+1}\|^2 + \frac{1}{\delta}\|y_{k+1} - y_k\|^2 \end{aligned}$$

for every $k \ge 0$. Suggestion: Remember the definition of the subgradient, and that $2ab \le \delta a^2 + \frac{1}{\delta}b^2$ for $a, b, \delta > 0$.

6. Show that if $\gamma\|A\| < 1$, there is $\varepsilon > 0$ such that

$$\|x_{k+1} - \hat{x}\|^2 + \|y_{k+1} - \hat{y}\|^2 + 2\gamma(\mathcal{L}(x_{k+1}, \hat{y}) - \mathcal{L}(\hat{x}, p_{k+1})) + \varepsilon\|Ax_{k+1} - b\|^2 \le \|x_k - \hat{x}\|^2 + \|y_k - \hat{y}\|^2$$

[3] Since f is continuous, we have $\partial(f + \iota_V) = \partial f + \partial\iota_V$. You do not need to prove this.

for every $k \geq 0$.

7. Deduce that $\lim_{k\to\infty} f(x_k) = f(\hat{x})$ and $\lim_{k\to\infty} Ax_k = b$.

8. Prove that (x_k, y_k) converges to an optimal pair. Suggestion: Verify that for every optimal pair $(\hat{x}, \hat{y})$, $\lim_{k\to\infty}[\|x_k - \hat{x}\|^2 + \|y_k - \hat{y}\|^2]$ exists.

7.2 Exercise: Exam 2022-2023

Let $C \subset \mathbb{R}^N$ be nonempty, convex and compact (closed and bounded). This part concerns

$$(\mathcal{P}_1) \qquad \min\{\, a \cdot z \,:\, z \in C \,\},$$

where $a \in \mathbb{R}^N$.

1. Show that, for every $a \in \mathbb{R}^N$, $(\mathcal{P}_1)$ has at least one solution.

2. Set $f(z) = \iota_C(z)$ and $g(z) = a \cdot z$.

   (a) Compute f^* and g^*, and determine the (Fenchel-Rockafellar) dual of $(\mathcal{P}_1)$. Observe that there is no duality gap.

   (b) Show that the primal-dual method is reduced to the projected gradient method.

3. Write the first-order optimality condition for $(\mathcal{P}_1)$.

Now let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and L-smooth, and let us analyze the convergence of the following algorithm to approximate

$$(\mathcal{P}_2) \qquad \hat{f} = \min\{\, f(z) \,:\, z \in C \,\}.$$

Starting with $x_0 \in C$, define a sequence (x_k, y_k) inductively by, for $k \geq 1$,

$$\begin{cases} y_k & \in & \operatorname{argmin}\{\, \nabla f(x_k) \cdot z \,:\, z \in C \,\} \\ \gamma_k & = & \operatorname{argmin}\big\{\, f\big(x_k + \gamma(y_k - x_k)\big) \,:\, \gamma \in [0,1] \,\big\} \\ x_{k+1} & = & x_k + \gamma_k(y_k - x_k). \end{cases}$$

4. Verify that $x_k \in C$ for all $k \geq 0$.

5. For each $k \geq 1$ and $\gamma \in [0,1]$, set $z_k(\gamma) = x_k + \gamma(y_k - x_k)$. Show that $f(x_{k+1}) \leq f\big(z_k(\gamma)\big)$. Deduce that $f(x_{k+1}) \leq f(x_k)$.

6. Prove that there is a constant $D \geq 0$ such that

$$f(x_{k+1}) - \hat{f} \leq (1-\gamma)\big(f(x_k) - \hat{f}\big) + \gamma^2 D$$

for all $\gamma \in [0,1]$ and $k \geq 1$.

Suggestion: (1) Use question 5 and the Descent Lemma; (2) Which sets have finite diameter?

7. For each k, substitute $\gamma = \frac{2}{k+2}$ in the preceding inequality to deduce that

$$(k+2)^2 \cdot \big(f(x_{k+1}) - \hat{f}\big) \leq (k+1)^2 \cdot \big(f(x_k) - \hat{f}\big) + 4D.$$

Do this for each k to show that

$$f(x_k) - \hat{f} \leq \frac{4D}{k+1}$$

for all $k \geq 1$.

8. Given $\varepsilon > 0$, how many iterations of this method give you the certainty that you have found a point $\hat{x} \in C$ such that $f(\hat{x}) \leq \hat{f} + \varepsilon$?

7.3 Exercise: Exam 2023-2024

Let $\phi : \mathbb{R}^K \to \mathbb{R} \cup \{+\infty\}$ be closed and $\mu-$strongly convex.

1. Use the Reciprocity formula to show why ϕ^* (the Fenchel conjugate of ϕ) must be $\frac{1}{\mu}-$smooth.

In all that follows, we consider the problem $(\mathcal{P})$ of minimizing a continuous and $\mu-$strongly convex function $f : \mathbb{R}^N \to \mathbb{R}$ on $V = \{x \in \mathbb{R}^N : Ax = b\}$, where $A \in \mathbb{R}^{M\times N}$ and $b \in \mathbb{R}^M$.

2. Why can we assure that problem $(\mathcal{P})$ has a unique solution?

3. Use the first order optimality condition to show that $\hat{x}$ is a solution of $(\mathcal{P})$ if, and only if, $A\hat{x} = b$ and there exists $\hat{y} \in \mathbb{R}^M$ such that $-A^T\hat{y} \in \partial f(\hat{x})$.[4] We say $(\hat{x}, \hat{y})$ is an *optimal pair.*

[4] Since f is continuous, we have $\partial(f + \iota_V) = \partial f + \partial\iota_V = \partial f + \mathrm{ran}(A^T)$. You do not need to prove this.

In the rest of the exam, we shall establish the convergence of a sequence (x_k, y_k), constructed from an initial point $(x_0, y_0) \in \mathbb{R}^N \times \mathbb{R}^M$, by iterating

$$\begin{cases} x_{k+1} & = & \operatorname{argmin}\left\{L(x, y_k) \ : \ x \in \mathbb{R}^N\right\} \\ y_{k+1} & = & y_k - \alpha(Ax_{k+1} - b), \end{cases}$$

with $\alpha > 0$ and $L(x, y) = f(x) + y \cdot (Ax - b) = f(x) + (A^T y) \cdot x - y \cdot b$, for each $(x, y) \in \mathbb{R}^N \times \mathbb{R}^M$.

4. Why is x_{k+1} well defined?

5. Write the optimality condition satisfied by x_{k+1} (this comes from the first subiteration).

6. Show that the dual $(\mathcal{D})$ of problem $(\mathcal{P})$ is $\min\{h(y) : y \in \mathbb{R}^M\}$, where $h(y) = f^*(-A^T y) + b \cdot y$.

7. Compute ∇h, and verify that h is ℓ−smooth, with $\ell = \frac{\|A\|^2}{\mu}$.

8. Show that the sequence (y_k) satisfies $y_{k+1} = y_k - \alpha \nabla h(y_k)$.

9. For which values of α can we guarantee that the sequence (x_k, y_k) converges to an optimal pair $(\hat{x}, \hat{y})$ as $k \to \infty$ (express the result in terms of μ and $\|A\|$).

10. (Bonus) What can you say about the convergence rate of this algorithm?