Introduction to Convex Optimisation

MBL 2023 - Daniel Cortild

Abstract

Optimisation is omnipresent in our lives nowadays, such as within data science, finance, medicine or machine learning. These notes present introductory notions of convex optimisation, which solves specific types of optimisation problems. We aim to cover, with full proofs of convergence and motivation, the algorithms of gradient descent and of projected gradient descent. The notes assume preliminary knowledge of one-dimensional calculus, such as limits and derivatives.

1. Introduction

Optimisation is a term widely used to refer to the act of improving results. When we talk about mathematical optimisation, we mainly refer to minimisation or maximisation problems. Of course, such problems include the aforementionned, as long as the quality of the result is measurable. For instance, a shop figuring out which product to sell could be formulated as a cost-minimisation or a profit-maximisation problem.

Minimisation or maximisation problems, what we refer to as mathematical optimisation problems, occur in every scientific domain. In machine learning, the objective is to find a model that minimises the error of a certain measuring function on a certain training set. In image processing, the objective is to find an image ressembling the blurry image whilst not being blurry. If one measures blurriness quantitatively, this may again be formulated as a minimisation problem.

In all generality, we will consider the following minimisation problem: For a function $f\colon \mathbb{R}^d \to \mathbb{R}$ and a set $C \subset \mathbb{R}^d$, find $\hat{\mathbf{x}} \in C$ such that

$$f(\hat{\mathbf{x}}) = \min_{\mathbf{x}\in C} f(\mathbf{x}).$$

The function f, which is being minimised, is called the *objective function*, the set C over which f is being minimised is called the *constraint set*, and the solution $\hat{\mathbf{x}}$ is called a *minimiser*.

A special case of the above is when $C = \mathbb{R}^d$. In that case, we minimise f over the entire space, and we say the problem is *unconstrained*. Otherwise, we say the problem is *constrained*.

Solving the above problem in all generality is an absurdly complicated task. As such, we shall add more restrictions to make the task possible. For starters, we shall consider f to be convex and have a smooth gradient. Moreover, we shall consider either $C = \mathbb{R}^d$, or C to be closed and convex. These assumptions will be further explained through the notes, and are the reason why these methods are referred to as *convex optimisation*.

These notes are structured in several sections. Section **??** studies a specific example of optimisation applied to data analysis. In Section 3, we list the used notation, assumptions and preliminary knowledge required. In Section 4, we outline some basic properties about convex sets and functions. In Sections 5, 6 and 7, the gradient descent and projected gradient descent algorithms are explained, and their convergence proved. Finally, in Section 8, a series of exercises relying on the same ideas as the rest of the notes, which may be solved independently to enhance the understanding of the concepts exposed, are listed.

2. Optimisation and Data Analysis

??

One area in which optimisation is widely used is in the field of data science, and more specifically in data analysis. The typical problem is, given a set of collected data, to find a model that describes this data, whilst reflecting a set of common beliefs we set up to describe a "good" model.

We can describe the set of collected that $\mathcal{D}$ by m objects, each constiting of a pair of a *feature vector* (data known a priori) and a *observation vector* (data measured a posteriori). More mathematically speaking, we can write

$$\mathcal{D} = \{(a_i, y_i), j = 1, 2, \ldots, m\}.$$

On example of this could be in the medical world. We have a file of each patient, constiting of various data such as their age, height, weight, hair colour, etc (All summarised in a_i for patient i), and for each patient we measure whether they dispose of a certain disease (Collected in y_i for patient i). With enough data, the goal of the data analyst is to create a model that, only provided with a new feature vector a_{new}, to predict whether this patient has the given disease, without testing the patient.

Mathematically speaking, the objective is to discover a function ϕ such that $\phi(a_i) \approx y_i$. The process of discovering such a function ϕ is often called "learning", since we are only basing our work on known data. Once the function ϕ found, we can approximate the observation of a new feature vector by $\phi(a_{\text{new}})$, without measuring it. Realistically, we cannot aim to find the best function due to the large amount of distinct functions. As such, we often restrict ourselves to some predefined functions, depending on some parameter, say a vector $\mathbf{x}$. For each vector $\mathbf{x}$, there is thus a function $\phi_{\mathbf{x}}$, and we want to find the $\hat{\mathbf{x}}$ such that $\phi_{\hat{\mathbf{x}}}$ is the best of all of them. This might still lead to infinitely many distinct functions, but a less scary infinite.

In order to write this as a standard optimisation problem, we introduce a *loss* function, which given a feature vector a_i, an observation vector y_i and a parameter $\mathbf{x}$, computes how good $\phi_{\mathbf{x}}(a_i)$ approximates y_i. Such a loss function is denoted by ℓ, and takes three parameters, namely the feature vector, the observation vector and the parameter. Many loss functions exist, but a simple example could be given by

$$\ell(a_i, y_i, \mathbf{x}) = \|\phi_{\mathbf{x}}(a_i) - y_i\|.$$

Now we can formulate the average error as

$$\mathcal{L}_{\mathcal{D}}(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \ell(a_i, y_i, \mathbf{x}).$$

The function $\mathcal{L}_{\mathcal{D}}$ is a function of $\mathbf{x}$, which we wish to minimise, since it is equivalent to minimising the average error over all data points. Finding the best function can thus be done by solving

$$\min_{\mathbf{x}} \mathcal{L}_{\mathcal{D}}(\mathbf{x}),$$

which is of the same form as previously.

3. Preliminaries

Throughout these notes we shall be working with functions in several variables. We first introduce some notations and explanations for them, followed by notions of "size" of vectors in $\mathbb{R}^d$, obtained

through norms, and conclude this section by a generalisation of the notion of differentiability in one dimension.

3.1. Notation and Assumptions

Recall that $\mathbb{R}$ denotes the set of real numbers. By $\mathbb{R}^d$ we denote the set of vectors with d real entries, namely

$$\mathbb{R}^d = \{\mathbf{x} = (x_1, \ldots, x_d) \colon x_1, \ldots, x_d \in \mathbb{R}\}.$$

We denote the set of $m \times n$ matrices with real entries by $\mathbb{R}^{m \times n}$.

Throughout these notes, we shall mainly work with functions $f \colon \mathbb{R}^d \to \mathbb{R}$. Unless otherwise specified, we assume such functions to be differentiable. The exact notion of differentiability is not introduced, but Lemma 3.3 provides the required consequences, and is sufficient for our purpose.

3.2. Norms

We recall the *dot product* of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ with components $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_d)$, defined as

$$\mathbf{x} \cdot \mathbf{y} = x_1 \cdot y_1 + x_2 \cdot y_2 + \cdots + x_d \cdot y_d.$$

We also recall the *standard norm* on $\mathbb{R}^d$, also referred to as the *Euclidean norm* or the 2-*norm*, defined as

$$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}.$$

For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we can write

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= (x_1 + y_1)^2 + \cdots + (x_d + y_d)^2 \\ &= x_1^2 + \cdots + x_d^2 + y_1^2 + \cdots + y_d^2 + 2(x_1 \cdot y_1 + \cdots + x_d \cdot y_d) \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\mathbf{x} \cdot \mathbf{y}. \end{aligned}$$

This identity will be used without reference in the future.

Another important link between the norm and the dot product resides in the Cauchy-Schwarz inequality.

Theorem 3.1 (Cauchy-Schwarz)**.** For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it holds that

$$|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|.$$

Proof. If $\mathbf{x} = \mathbf{0}$, the inequality is trivial. Thus suppose that $\mathbf{x} \neq \mathbf{0}$, and thus that $\|\mathbf{x}\| \neq 0$. Define

$$P(t) = \|\mathbf{x}\|^2 t^2 + 2(\mathbf{x} \cdot \mathbf{y})t + \|\mathbf{y}\|^2 = \|\mathbf{x} \cdot t + \mathbf{y}\|^2,$$

which is a second degree polynomial in t. Since $P(t) \geq 0$ for all $t \in \mathbb{R}$, the discriminant must be nonpositive, namely

$$\Delta = 4(\mathbf{x} \cdot \mathbf{y})^2 - 4\|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \leq 0,$$

which may be rewritten as the wanted inequality. □

Although we called $\|\cdot\|$ a *norm*, this does not guarantee that it satisfies the properties of a norm. Luckily for us, it does.

Proposition 3.2. The Euclidean norm $\|\cdot\|$ is indeed a norm, i.e. it satisfies the following properties:

- Positive-definiteness: For all $\mathbf{x}\in\mathbb{R}^d$, it holds that $\|\mathbf{x}\|\geq 0$. Moreover, $\|x\|=0$ if, and only if, $\mathbf{x}=\mathbf{0}$.
- Homogeneity: For all $\mathbf{x}\in\mathbb{R}^d$ and $\lambda\in\mathbb{R}$, it holds that $\|\lambda\cdot\mathbf{x}\|=|\lambda|\cdot\|\mathbf{x}\|$.
- Triangle Inequality: For all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, it holds that $\|\mathbf{x}+\mathbf{y}\|\leq\|\mathbf{x}\|+\|\mathbf{y}\|$.

Proof. The two first properties are immediate from the definition of $\|\cdot\|$, and are left as an exercise to the reader. The last property follows by expanding the square and by the Cauchy-Schwarz Inequality 3.1

$$\|\mathbf{x}+\mathbf{y}\|^2=\|\mathbf{x}\|^2+\|y\|^2+2\mathbf{x}\cdot\mathbf{y}\leq\|\mathbf{x}\|^2+\|y\|^2+2\|\mathbf{x}\|\cdot\|\mathbf{y}\|=(\|\mathbf{x}\|+\|\mathbf{y}\|)^2,$$

from which the conclusion follows since both $\|\mathbf{x}+\mathbf{y}\|$ and $\|\mathbf{x}\|+\|\mathbf{y}\|$ are positive. □

It is worth noting that a lot of different norms exist, but for convience we shall restrict ourselves to the standard norm above.

3.3. Multivariable Calculus

We present certain well-known facts from multivariable calculus which will be helpful later on. Note that, as stated earlier, we always assume f to be differentiable, and do not need to worry about problems related to that. Moreover, we will not look into the exact definition of differentiability, as it requires more technicalities than for the one-dimensional case.

For a function $f\colon\mathbb{R}^m\to\mathbb{R}^n$, we define its *partial derivative* of the ith component with respect to x_j at a point $\mathbf{a}\in\mathbb{R}^m$ to be

$$\frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}}=\lim_{h\to 0}\frac{f_i(\mathbf{a}+h\mathbf{e}_j)-f_i(\mathbf{a})}{h}\in\mathbb{R}.$$

When f is assumed differentiable, these limits always exist.

We define the *gradient* of f at $\mathbf{a}$ to be

$$\nabla f(\mathbf{a})=\begin{pmatrix}\frac{\partial f_1}{\partial x_1}\big|_{\mathbf{a}} & \cdots & \frac{\partial f_1}{\partial x_m}\big|_{\mathbf{a}}\\ \vdots & \ddots & \vdots\\ \frac{\partial f_n}{\partial x_1}\big|_{\mathbf{a}} & \cdots & \frac{\partial f_n}{\partial x_m}\big|_{\mathbf{a}}.\end{pmatrix}$$

As for the one-dimensional derivative, the gradient tells us information about the tangent hyperplane to the curve at a point. Namely, the gradient is a vector tangent to the curve that points in the direction of **largest increase** at that point.

The following result shall be useful later, and is presented without proof.

Lemma 3.3. If the function $f\colon \mathbb{R}^m \to \mathbb{R}^n$ is differentiable at $\mathbf{a} \in \mathbb{R}^m$, then, for all $\mathbf{u} \in \mathbb{R}^n$,

$$\lim_{h \to 0} \frac{f(\mathbf{a}+h\mathbf{u}) - f(\mathbf{a})}{h}$$

is well-defined and equal to $\nabla f(\mathbf{a}) \cdot \mathbf{u}$.

Most properties known for the one-dimensional derivative have an analogue for functions in several variables. For starters, the addition rule holds.

Theorem 3.4 (Addition Rule)**.** For two maps $f, g\colon \mathbb{R}^m \to \mathbb{R}^n$, it holds that $f+g$ is differentiable at $\mathbf{a} \in \mathbb{R}^m$ when both f and g are differentiable at $\mathbf{a}$. In that case, it holds that

$$\nabla(f+g)(\mathbf{a}) = \nabla f(\mathbf{a}) + \nabla g(\mathbf{a})$$

Theorem 3.5 (Chain Rule)**.** For two maps $f\colon \mathbb{R}^m \to \mathbb{R}^n$ and $g\colon \mathbb{R}^n \to \mathbb{R}^k$, it holds that $g \circ f$ is differentiable at $\mathbf{a} \in \mathbb{R}^m$ when f is differentiable at $\mathbf{a}$ and g is differentiable at $f(\mathbf{a})$. In that case, it holds that

$$\nabla(g \circ f)(\mathbf{a}) = \nabla g(f(\mathbf{a})) \cdot \nabla f(\mathbf{a}).$$

We call a map $F\colon \mathbb{R}^m \to \mathbb{R}^n$ is called *Lipschitz continuous* with parameter $L > 0$ if, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\|F(\mathbf{x}) - F(\mathbf{y})\| \leq L \cdot \|\mathbf{x} - \mathbf{y}\|.$$

Moreover, a function $f\colon \mathbb{R}^m \to \mathbb{R}$ is called *L-smooth* for $L > 0$ if f is differentiable and ∇f is Lipschitz continuous with parameter L.

4. Notion of Convexity

We call a set $C \subset \mathbb{R}^d$ *convex* if, for all $\mathbf{x}, \mathbf{y} \in C$ and all $\lambda \in [0,1]$, it holds that

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

We call $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ a *convex combination* of $\mathbf{x}$ and $\mathbf{y}$. As such, a set C is convex if it is closed under convex combinations.

Intuitively, a set C is convex if the line segment between any two points in C is fully contained in C. This is illustrated in Figure 4.1.

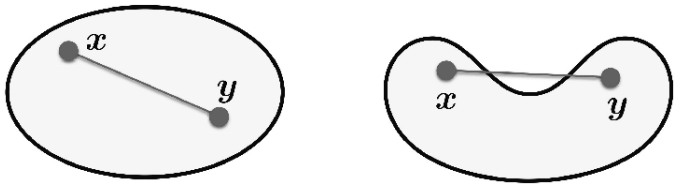

Figure 4.1: Example of a convex set (left) and a non-convex set (right).

Analogously to convex sets, a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called *convex* if, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\lambda \in [0,1]$, it holds that

$$f((1-\lambda)\mathbf{x} + \lambda\mathbf{y}) \leq f(1-\lambda)f(\mathbf{x}) + \lambda f(\mathbf{y}).$$

As such, a function is convex is the image of a convex combination lies below the convex combination of the images. Intuitively, this means that for any pair of points $(\mathbf{x}, f(\mathbf{x})), (\mathbf{y}, f(\mathbf{y}))$, any point on the line connecting these points is above the graph of the function.

Geometrically, a convex function is a function with a certain upward rounded curvature, as shown in Figure 4.2.

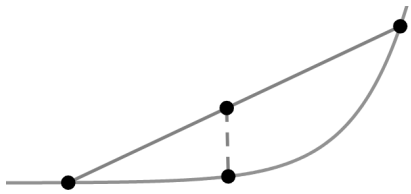

Figure 4.2: Example of a convex function.

There are many tight links between convex functions and sets. One such link is presented, also pointing out why we are interested in convex problems.

Proposition 4.1. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then the set of minimisers of f forms a convex set.

Proof. Let S be the set of minimisers of f. If $S = \emptyset$, the statement is trivial. Assume $S \neq \emptyset$ and take two points $\mathbf{x}, \mathbf{y} \in S$. By definition of S, it holds that

$$f(\mathbf{x}) = f(\mathbf{y}) = \min_{\mathbf{z}\in\mathbb{R}^d} f(\mathbf{z}).$$

As such, any convex combination of $\mathbf{x}$ and $\mathbf{y}$, say $\mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$, must satisfy

$$f(\mathbf{z}) = f((1-\lambda)\mathbf{x} + \lambda\mathbf{y}) \leq (1-\lambda)\mathbf{x} + \lambda f(\mathbf{y}) = (1-\lambda)\min f + \lambda \min f = \min f,$$

where the first inequality following by the convexity of f. Since $f(\mathbf{z}) \leq \min f$, it must hold that $f(\mathbf{z}) = \min f$, such that $\mathbf{z} \in S$. As such, since any convex combination of any two points in S is in S, we conclude that S is a convex set. □

If we further impose the function f to be differentiable, we may get an alternative characterisation for convex functions.

Lemma 4.2. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then f is convex if, and only if, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it holds that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x}) \cdot \nabla f(\mathbf{x}).$$

Proof. Suppose f is convex, and fix $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and let $0 < \lambda \leq 1$. Then, by convexity,

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

which rewrites as

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

Taking $\lambda \to 0$ and using Lemma 3.3, we obtain

$$\nabla f(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}),$$

which is equivalent to the wanted inequality.

Conversly, suppose the inequality holds and fix $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ and $\lambda\in[0,1]$, and set $\mathbf{z}=(1-\lambda)\mathbf{x}+\lambda\mathbf{y}$. Then it holds that

$$f(\mathbf{x})\geq f(\mathbf{z})+(\mathbf{x}-\mathbf{z})\cdot\nabla f(\mathbf{z})=f((1-\lambda)\mathbf{x}+\lambda\mathbf{y})+\lambda\nabla f(\mathbf{z})\cdot(x-y),$$

and

$$f(\mathbf{y})\geq f(\mathbf{z})+(\mathbf{y}-\mathbf{z})\cdot\nabla f(\mathbf{z})=f((1-\lambda)\mathbf{x}+\lambda\mathbf{y})+(1-\lambda)\nabla f(\mathbf{z})\cdot(y-x).$$

Multiplying the first inequality by $(1-\lambda)$, the second by λ, and adding the results yields

$$(1-\lambda)f(\mathbf{x})+\lambda f(\mathbf{y})\geq f((1-\lambda)\mathbf{x}+\lambda\mathbf{y}),$$

which implies the convexity of f since $\mathbf{x},\mathbf{y}$ and λ were arbitrary. □

Theorem 4.3 (Fermat's Rule)**.** Let $f\colon\mathbb{R}^d\to\mathbb{R}$ be convex and differentiable. Then $\hat{\mathbf{x}}\in\mathbb{R}^d$ is a global minimiser of f if, and only if, $\nabla f(\hat{\mathbf{x}})=0$.

Proof. If $\nabla f(\hat{\mathbf{x}})=0$, then Lemma 4.2 with $\mathbf{x}=\hat{\mathbf{x}}$ yields $f(\mathbf{y})\geq f(\hat{\mathbf{x}})$ for all $\mathbf{y}\in\mathbb{R}^d$, implying $\hat{\mathbf{x}}$ is a global minimiser of f.

Conversly, if $\hat{\mathbf{x}}$ is a global minimiser, it must hold that $f(\mathbf{y})\geq f(\hat{\mathbf{x}})$ for all $y\in\mathbb{R}^d$, such that Lemma 4.2 implies $\nabla f(\hat{\mathbf{x}})\cdot(\mathbf{y}-\hat{\mathbf{x}})\leq 0$. Considering $\mathbf{y}=\mathbf{z}$ and $\mathbf{y}=2\hat{\mathbf{x}}-\mathbf{z}$ for some $\mathbf{z}\in\mathbb{R}^d$, we get

$$(\mathbf{z}-\hat{\mathbf{x}})\cdot\nabla f(\hat{\mathbf{x}})\leq 0\quad\text{and}\quad(\hat{\mathbf{x}}-\mathbf{z})\cdot\nabla f(\hat{\mathbf{x}})\leq 0,$$

such that $(\mathbf{z}-\hat{\mathbf{x}})\cdot\nabla f(\hat{\mathbf{x}})=0$ for all $\mathbf{z}\in\mathbb{R}^d$. Setting $\mathbf{z}=\nabla f(\hat{\mathbf{x}})+\hat{\mathbf{x}}$ then yields $\|\nabla f(\hat{\mathbf{x}})\|=0$, showing that $\nabla f(\hat{\mathbf{x}})=0$, as wanted. □

Lemma 4.4 (Descent Lemma)**.** Let $f\colon\mathbb{R}^d\to\mathbb{R}$ be a convex L-smooth function. For all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, it holds that

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

Proof. Fix $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ and define $g\colon[0,1]\to\mathbb{R}$ by $g(t)=f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$, such that $g(0)=f(\mathbf{x})$ and $g(1)=f(\mathbf{y})$. Then, by the fundamental theorem of calculus and the Cauchy-Schwarz Inequality 3.1,

$$\begin{aligned}
f(\mathbf{y})-f(\mathbf{x})+(\mathbf{x}-\mathbf{y})\cdot\nabla f(\mathbf{x}) &= g(1)-g(0)+(\mathbf{x}-\mathbf{y})\cdot\nabla f(\mathbf{x})\\
&=\int_0^1 g'(t)dt+(\mathbf{x}-\mathbf{y})\cdot\nabla f(\mathbf{x})\\
&=\int_0^1(\mathbf{y}-\mathbf{x})\cdot(\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\nabla f(\mathbf{x}))dt\\
&\leq\|\mathbf{y}-\mathbf{x}\|\cdot\int_0^1\|\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\nabla f(\mathbf{x})\|dt\\
&\leq\|\mathbf{y}-\mathbf{x}\|\cdot\int_0^1 L\|t(\mathbf{y}-\mathbf{x})\|dt\\
&\leq L\|\mathbf{y}-\mathbf{x}\|^2\cdot\int_0^1 tdt\\
&=\frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2,
\end{aligned}$$

which, rewritten, yields the wanted inequality. □

5. Gradient Descent

We are now interested in solving the following problem: For a convex L-smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$, find $\hat{\mathbf{x}} \in \mathbb{R}^d$ such that

$$f(\hat{\mathbf{x}}) = \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}).$$

In other words, we are looking for a minimiser of f.

To simplify the subsequent analysis, we shall suppose f is such that it does admit a minimiser.

In order to find such a minimiser, we shall follow a naive yet working approach. We start at a certain *initial point*, and take steps. These steps shall always be aimed at the direction of steepest descent, such that we always go downwards, or at least can hope for it. The only parameter of the process to be choosen is the *step-size*, which determines how large steps we take.

As such, we shall study the following algorithm, known as the Gradient Descent Algorithm, for which an initial guess $\mathbf{x}_0 \in \mathbb{R}^d$ and a step-size $\alpha > 0$ are given, and $\mathbf{x}_k$ is defined iteratively as

$$\mathbf{x}_{k+1} \coloneqq \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k). \tag{5.1}$$

5.1. Convergence Proof

Theorem 5.1. Suppose the function $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth with $L > 0$, and that f admits a minimiser. Defining $\mathbf{x}_k$ iteratively with initial guess $\mathbf{x}_0$ through Algorithm (5.1) where $\alpha \leq \frac{1}{L}$, will result in a sequence satisfying

$$f(\mathbf{x}_k) - \min f \leq \frac{C}{k},$$

for some constant $C > 0$. In specific, $f(\mathbf{x}_k) \to \min f$.

Proof. Let $\hat{\mathbf{x}} \in \mathbb{R}^d$ be a minimiser of f, which exists by assumption.

Set $\mathbf{x} = \mathbf{x}_k$ and $\mathbf{y} = \mathbf{x}_{k+1}$ in the Descent Lemma 4.4 and use that $\alpha \leq \frac{1}{L}$ to obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq \alpha\left(\frac{\alpha L}{2} - 1\right)\|\nabla f(\mathbf{x}_k)\|^2 \leq -\frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|^2.$$

As such, $f(\mathbf{x}_k)$ is a nonincreasing sequence.

Set $\mathbf{x} = \mathbf{x}_k$ and $\mathbf{y} = \hat{\mathbf{x}}$ in the alternate characterisation of convexity from Lemma 4.2 to observe

$$f(\mathbf{x}_k) - f(\hat{\mathbf{x}}) \leq (\mathbf{x}_k - \hat{\mathbf{x}}) \cdot \nabla f(\mathbf{x}_k).$$

Summing the previous two yields

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\hat{\mathbf{x}}) &\leq (\mathbf{x}_k - \hat{\mathbf{x}}) \cdot \nabla f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|^2 \\ &= \frac{1}{2\alpha}\left(-\|\alpha \nabla f(\mathbf{x}_k) - (\mathbf{x}_k - \hat{\mathbf{x}})\|^2 + \|\mathbf{x}_k - \hat{\mathbf{x}}\|^2\right) \\ &= \frac{1}{2\alpha}\left(-\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\|^2 + \|\mathbf{x}_k - \hat{\mathbf{x}}\|^2\right). \end{aligned}$$

Since $f(\mathbf{x}_k)$ is nonincreasing, we get, combining with the previous,

$$\begin{aligned} f(\mathbf{x}_k) - f(\hat{\mathbf{x}}) &\leq \frac{1}{k}\sum_{i=0}^{k-1}(f(\mathbf{x}_i) - f(\hat{\mathbf{x}})) \\ &\leq \frac{1}{2\alpha k}\sum_{i=0}^{k-1}\left(-\|\mathbf{x}_{i+1} - \hat{\mathbf{x}}\|^2 + \|\mathbf{x}_i - \hat{\mathbf{x}}\|^2\right) \\ &= \frac{-\|\mathbf{x}_k - \hat{\mathbf{x}}\|^2 + \|\mathbf{x}_0 - \hat{\mathbf{x}}\|^2}{2\alpha k} \\ &\leq \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|^2}{2\alpha k}. \end{aligned}$$

As such, setting $C = \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|^2}{2\alpha}$ yields the wanted conclusion. □

5.2. Numerical Example

Let us consider the example $f\colon \mathbb{R} \to \mathbb{R}$, defined by $f(x) = x^2$. It is easy to compute that $f'(x) = 2x$ has Lipschitz constant $L = 2$. As such, we know convergence is guaranteed for $\alpha \leq \frac{1}{2}$. We can also easily compute that the minimum is located at $\hat{x} = 0$, with objective value of 0.

By runnning a simulation with the above algorithm, we get the different sequences for different values of α, depicted in Figure 5.1. It is worth noticing that the sequence still convergences for $\alpha = 0.8$. Indeed, although not done in these notes, convergence is assured for $\alpha \leq \frac{2}{L}$. We also notice that $\alpha = 0.4$ converges faster than $\alpha = 0.1$, which is to be expected, since $\alpha = 0.4$ is closer to the limiting value of $\frac{1}{L} = 0.5$.

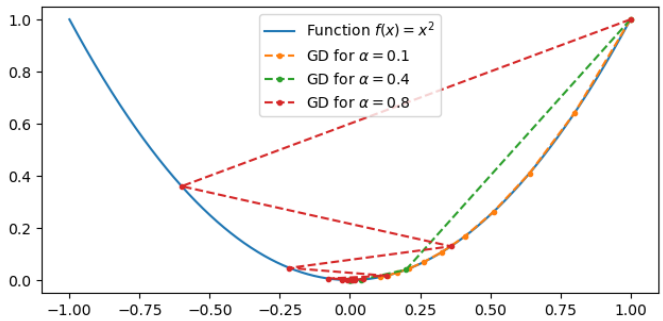

Figure 5.1: Example of gradient descent algorithm.

6. Strongly Convex Functions

7. Projected Gradient Descent

The gradient descent algorithm solves minimisation problems over an entire space. In this section, we consider a constrained minimisation problem: For a convex L-smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$ and

a convex set $C \subset \mathbb{R}^d$, find $\hat{\mathbf{x}} \in C$ such that

$$f(\hat{\mathbf{x}}) = \min_{\mathbf{x} \in C} f(\mathbf{x}).$$

As such, we are looking for a minimiser of f within the set C, which is not necessarily a global minimiser of f.

To solve the above problem, we shall slightly adapt the gradient descent algorithm to ensure that the generated sequence remains within the set C. As such, after each gradient descent iteration, we shall project the obtained point onto the set C. This projection shall be further defined in the following subsection, and is denoted by P_C. The studied algorithm is given by

$$\mathbf{x}_{k+1} \coloneqq P_C(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)), \tag{7.1}$$

where $\mathbf{x}_0 \in \mathbb{R}^d$ is an initial guess and $\alpha > 0$ is the step-size.

7.1. Projections

Note: For technical reasons, we require the set C to be closed. In order to simply the notes, we omit the details about closedness, and its implications. Note that the projection map might not be defined without this additional assumption.

For a given convex set $C \subset \mathbb{R}^d$, we define a projection map $P_C : \mathbb{R}^d \to C$ which maps a point $\mathbf{x} \in \mathbb{R}^d$ to the point in C closest to $\mathbf{x}$. Mathematically, we define

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

namely $P_C(\mathbf{x})$ is the point $\mathbf{y} \in C$ that minimises the distance between $\mathbf{x}$ and $\mathbf{y}$.

Importantly, notice that $\|\mathbf{y} - \mathbf{x}\| \geq 0$ for all $\mathbf{y} \in C$. As such, if $\mathbf{x} \in C$, it holds that $P_C(\mathbf{x}) = \mathbf{x}$. This makes sense, as the closest point to a point already in C must be itself. If $\mathbf{x} \notin C$, then $\|\mathbf{y} - \mathbf{x}\| > 0$ for all $\mathbf{y} \in C$ since $\mathbf{y} \neq \mathbf{x}$, and as such $P_C(\mathbf{x}) \neq \mathbf{x}$.

Example 7.1. Let us consider the convex set $C = [a, b] \subset \mathbb{R}$. If $x \in [a, b]$, then $P_C(x) = x$, as described above. If $x \leq a$, then $P_C(x) = a$, since $\|y - x\| = |y - x| = y - x$ is minimised for $y = a$. If $x \geq b$, then $P_C(x) = b$, since $\|y - x\| = |y - x| = x - y$ is minimised for $y = b$. As such,

$$P_C(x) = \begin{cases} a & \text{if } x \leq a \\ x & \text{if } a \leq x \leq b \\ b & \text{if } b \leq x \end{cases}$$

7.2. Convergence proof

The following inequality involving projections holds and will be useful in the convergence proof.

Lemma 7.2. Let $C \subset \mathbb{R}^d$ be convex and closed, $\mathbf{w} \in \mathbb{R}^d$ and $\overline{\mathbf{w}} = P_C(\mathbf{w})$. Then it holds that, for all $\mathbf{z} \in C$,

$$(\mathbf{z} - \overline{\mathbf{w}}) \cdot (\mathbf{w} - \overline{\mathbf{w}}) \leq 0.$$

Moreover, it holds that

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

PROOF. Notice that by definition of the projection map, for all $\mathbf{z} \in C$

$$\|\mathbf{w} - \overline{\mathbf{w}}\| \leq \|\mathbf{w} - \mathbf{z}\|.$$

Moreover, notice that since $\overline{\mathbf{w}}$ and $\mathbf{z}$ are in C, and C is convex, $\overline{\mathbf{z}} = \lambda \mathbf{z} + (1-\lambda)\overline{\mathbf{w}}$ also is in C for all $\lambda \in [0,1]$. Plugging in $\overline{\mathbf{z}}$ as $\mathbf{z}$ in the previous yields that

$$\|\mathbf{w} - \overline{\mathbf{w}}\| \leq \|\mathbf{w} - \overline{\mathbf{z}}\| = \|\mathbf{w} - \lambda\mathbf{z} - (1-\lambda)\overline{\mathbf{w}}\| = \|\mathbf{w} - \overline{\mathbf{w}} + \lambda(\overline{\mathbf{w}} - \mathbf{z})\|.$$

By squaring and expanding the square in the right-most expression, we observe that

$$\|\mathbf{w} - \overline{\mathbf{w}}\|^2 \leq \|\mathbf{w} - \overline{\mathbf{w}}\|^2 + \lambda^2 \|\mathbf{z} - \overline{\mathbf{w}}\|^2 + 2\lambda(\mathbf{w} - \overline{\mathbf{w}}) \cdot (\overline{\mathbf{w}} - \mathbf{z}),$$

which may be rewritten as

$$(\mathbf{w} - \overline{\mathbf{w}}) \cdot (\mathbf{z} - \overline{\mathbf{w}}) \leq \frac{\lambda}{2} \|\mathbf{z} - \overline{\mathbf{w}}\|^2.$$

Taking $\lambda \to 0$ then yields the first result. The second result follows from noticing that

$$\|\mathbf{z} - \mathbf{w}\|^2 = \|\mathbf{z} - \overline{\mathbf{w}} + \overline{\mathbf{w}} - \mathbf{w}\|^2 = \|\mathbf{z} - \overline{\mathbf{w}}\|^2 + \|\overline{\mathbf{w}} - \mathbf{w}\|^2 + 2(\mathbf{z} - \overline{\mathbf{w}}) \cdot (\overline{\mathbf{w}} - \mathbf{w}) \geq \|\mathbf{z} - \overline{\mathbf{w}}\|^2 + \|\overline{\mathbf{w}} - \mathbf{w}\|^2,$$

where the inequality follows from the first result. □

We are now ready to tackle the proof of convergence for the sequence generated by the projected gradient descent algorithm.

THEOREM 7.3. Suppose the function $f \colon \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth with $L > 0$, and that f admits a minimiser. Defining $\mathbf{x}_k$ iteratively with initial guess $\mathbf{x}_0 \in \mathbb{R}^d$ through Algorithm (7.1), where $\alpha < \frac{1}{L}$, will result in a sequence satisfying

$$f(\mathbf{x}_k) - \min f \leq \frac{C}{k},$$

for some constant $C > 0$ not depending on k. In specific, $f(\mathbf{x}_k) \to \min f$.

PROOF. Let $\hat{\mathbf{x}} \in \mathbb{R}^d$ be a minimiser of f, which exists by assumption.

Set $\mathbf{w} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$ and $\overline{\mathbf{w}} = \mathbf{x}_{k+1}$ in Lemma 7.2 to obtain that, for all $\mathbf{z} \in C$,

$$\|\mathbf{z} - \mathbf{x}_k + \alpha \nabla f(\mathbf{x}_k)\|^2 \geq \|\mathbf{x}_{k+1} - \mathbf{x}_k + \alpha \nabla f(\mathbf{x}_k)\|^2 + \|\mathbf{z} - \mathbf{x}_{k+1}\|^2.$$

This may be rewritten as, by expanding the squares and dividing by 2α,

$$(\mathbf{x}_{k+1} - \mathbf{x}_k) \cdot \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq (\mathbf{z} - \mathbf{x}_k) \cdot \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha} \|\mathbf{z} - \mathbf{x}_k\|^2 - \frac{1}{2\alpha} \|\mathbf{z} - \mathbf{x}_{k+1}\|^2.$$

Now set $\mathbf{x} = \mathbf{x}_k$ and $\mathbf{y} = \mathbf{x}_{k+1}$ in the Descent Lemma 4.4 and use that $\alpha \leq \frac{1}{L}$ to observe that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + (\mathbf{x}_{k+1} - \mathbf{x}_k) \cdot \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.$$

Combining the previous two inequalities yields that, for all $\mathbf{z} \in C$,

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + (\mathbf{z} - \mathbf{x}_k) \cdot \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha} \|\mathbf{z} - \mathbf{x}_k\|^2 - \frac{1}{2\alpha} \|\mathbf{z} - \mathbf{x}_{k+1}\|^2.$$

By the alternate characterisation of convexity in Lemma 4.2, we get that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{z}) + \frac{1}{2\alpha}\|\mathbf{z}-\mathbf{x}_k\|^2 - \frac{1}{2\alpha}\|\mathbf{z}-\mathbf{x}_{k+1}\|^2.$$

If we set $\mathbf{z} = \mathbf{x}_{k+1}$, we observe that $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$, such that the sequence $(f(\mathbf{x}_k))$ is nonincreasing.

If we instead set $\mathbf{z} = \hat{\mathbf{x}}$, where $\hat{x}$ is the minimiser of f, we obtain that

$$f(\mathbf{x}_{k+1}) - \min f \leq \frac{1}{2\alpha}\left(\|\hat{\mathbf{x}}-\mathbf{x}_k\|^2 - \|\hat{\mathbf{x}}-\mathbf{x}_{k+1}\|^2\right).$$

As such, using that $(f(\mathbf{x}_k))$ is nonincreasing, we may write

$$\begin{aligned} f(\mathbf{x}_k) - \min f &\leq \frac{1}{k}\sum_{i=0}^{k-1}(f(\mathbf{x}_i) - \min f) \\ &\leq \frac{1}{2\alpha k}\sum_{i=0}^{k-1}\left(\|\hat{\mathbf{x}}-\mathbf{x}_i\|^2 - \|\hat{\mathbf{x}}-\mathbf{x}_{i+1}\|^2\right) \\ &= \frac{\|\hat{\mathbf{x}}-\mathbf{x}_0\|^2 - \|\hat{\mathbf{x}}-\mathbf{x}_k\|^2}{2\alpha k} \\ &\leq \frac{\|\hat{\mathbf{x}}-\mathbf{x}_0\|^2}{2\alpha k}. \end{aligned}$$

As such, setting $C = \frac{\|\hat{\mathbf{x}}-\mathbf{x}_0\|^2}{2\alpha}$ yields the wanted conclusion. □

Notice that if $C = \mathbb{R}$, the above proof boils down to the proof of Theorem 5.1.

8. Exercises

This section is aimed at complementing the lecture notes, and to help the readers achieve a deeper level of understanding of the material. They are structured in such a way that they build on top of each other, and of the theory in the notes. As such, most exercises will require some previous ones to be solved.

Problem 8.1. Elaborate on the details of positive-definiteness and homogeneity of the Euclidean norm in Proposition 3.2.

Problem 8.2. As mentionned previously, the standard norm is not the only norm in $\mathbb{R}^d$. Prove that the 1-norm, defined by $\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_d|$ for $\mathbf{x} \in \mathbb{R}^d$, also verifies the three properties of Proposition 3.2.

Problem 8.3. Compute the partial derivatives and the gradient of the following functions at the given point:

- $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x$, at $\mathbf{a} = 1$,
- $f\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = (x+y)^2$, at $\mathbf{a} = (0,0)$,
- $f\colon \mathbb{R}^4 \to \mathbb{R}$ defined by $f(x,y,z,w) = xyzw$, at $\mathbf{a} = (1,2,3,4)$, and

- $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ defined by $f(x,y) = (y,x)$, at $\mathbf{a} = (0,0)$.

Problem 8.4. Given a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$, give the vector of largest decrease at a point.

Problem 8.5. Fix some $\mathbf{a} \in \mathbb{R}^d$. Let $f, g\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}$ and $g(\mathbf{x}) = \|\mathbf{x}\|^2$ for $\mathbf{x} \in \mathbb{R}^d$. You may assume f and g are differentiable everywhere. Determine ∇f and ∇g.

Hint: Determine each partial derivative individually.

Problem 8.6. Give an example of a convex set and of a non-convex set in $\mathbb{R}$. Repeat the exercise for $\mathbb{R}^2$.

Problem 8.7. Let $a, b \in \mathbb{R}$ be such that $a < b$. Show that the set $C = [a,b] \subset \mathbb{R}$ is convex. Also show that the set $C' = [a,b] \times [a,b] \subset \mathbb{R}^2$ is convex.

Problem 8.8. Give an example of a convex function $f\colon \mathbb{R} \to \mathbb{R}$. Also give an example of a non-convex function.

Problem 8.9. Note that not all convex functions are differentiable. Give an example of a convex function $f\colon \mathbb{R} \to \mathbb{R}$ that is not differentiable.

Problem 8.10. Fix $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Prove that the following functions are convex:

- $f\colon \mathbb{R}^d \to \mathbb{R}$ given by $f(\mathbf{x}) = \mathbf{a}$,
- $f\colon \mathbb{R}^d \to \mathbb{R}$ given by $f(\mathbf{x}) = \mathbf{a} \cdot (\mathbf{x} - \mathbf{b})$,
- $f\colon \mathbb{R}^d \to \mathbb{R}$ given by $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{a}\|$,
- $f\colon \mathbb{R}^d \to \mathbb{R}$ given by $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{a}\|^2$.

Problem 8.11. Let $f, g\colon \mathbb{R}^d \to \mathbb{R}$ be convex, and let $\mu \in \mathbb{R}_{\geq 0}$ be a positive constant. Show that $\mu \cdot f$ and $f + g$ are convex as well.

Problem 8.12. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Prove that f is convex if, and only if,

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})) \cdot (\mathbf{x} - \mathbf{y}) \geq 0$$

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

Hint: Use the alternative characterisation of convexity from Lemma 4.2 twice.

Problem 8.13. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex L-smooth function. For a given $\mathbf{x} \in \mathbb{R}^d$, define $f_{\mathbf{x}}(\mathbf{z}) = f(\mathbf{z}) - \nabla f(\mathbf{x}) \cdot \mathbf{z}$. Show that $f_{\mathbf{x}}$ is convex and L-smooth, compute $\nabla f_{\mathbf{x}}(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{R}^d$, and determine a minimiser of $f_{\mathbf{x}}$.

Problem 8.14. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex L-smooth function with minimiser $\hat{\mathbf{x}}$. Prove that, for all $\mathbf{y} \in \mathbb{R}^d$, it holds that

$$\frac{1}{2L} \|\nabla f(\mathbf{y})\|^2 \leq f(\mathbf{y}) - f(\hat{\mathbf{x}}).$$

Hint: Argue that both the RHS and the LHS of the Descent Lemma 4.4 are convex in $\mathbf{y}$, and minimise both sides using Fermat's Rule 4.3.

Problem 8.15. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex L-smooth function. Prove that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it holds that

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

Hint: Define $f_{\mathbf{x}}$ as in Problem 8.13, and define $f_{\mathbf{y}}$ analogously. Use Problem 8.14 to determine a lower bound on $f_{\mathbf{x}}(\mathbf{y}) - f_{\mathbf{x}}(\mathbf{x})$ and on $f_{\mathbf{y}}(\mathbf{x}) - f_{\mathbf{y}}(\mathbf{y})$, and sum the obtained inequalities.

Note: The above is known as the *cocoercivity* property of the gradient. It plays an important role in more advanced results in convex optimisation.

Problem 8.16. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function. Show that if the inequality of Problem 8.15 holds for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, then it holds that f is L-smooth.

Problem 8.17. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex L-smooth function with minimiser $\hat{\mathbf{x}}$. Verify that if $\mathbf{x}_0 = \hat{\mathbf{x}}$, then $\mathbf{x}_k = \hat{\mathbf{x}}$ for all $k \geq 0$, where $(\mathbf{x}_k)$ is generated by the gradient descent algorithm with any step-size $\alpha > 0$.

Problem 8.18. Give an example of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ and a convex set $C \subset \mathbb{R}^d$ such that $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ and $\min_{\mathbf{x} \in C} f(\mathbf{x})$ coincide. Give another example in which they differ.

Problem 8.19. Let $C = [0,1] \times [0,1] \subset \mathbb{R}^2$. Compute $P_C(\mathbf{x})$.

Hint: Draw out the region C, and compute $P_C(\mathbf{x})$ for different regions in $\mathbb{R}^2$.

Problem 8.20. Let $C \subset \mathbb{R}^d$ be a convex set and let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex L-smooth function with minimiser $\hat{\mathbf{x}} \in C$. Verify that if $\mathbf{x}_0 = \hat{\mathbf{x}}$, then $\mathbf{x}_k = \hat{\mathbf{x}}$ for all $k \geq 0$, where $(\mathbf{x}_k)$ is generated by the projected gradient descent algorithm with any step-size $\alpha > 0$.

Problem 8.21. Compare the proofs of Theorems 5.1 and 7.3 in the case where $C = \mathbb{R}$. Observe the proofs are based on the same arguments, and that Theorem 5.1 is a special case of 7.3.

References

[1] Boyd, Stephen P., and Lieven Vandenberghe. Convex optimization. Cambridge University Press, 2004.

[2] Garrigos, Guillaume, and Robert M. Gower. "Handbook of convergence theorems for (stochastic) gradient methods." arXiv preprint arXiv:2301.11235 (2023).

[3] Nesterov, Yurii. Introductory lectures on convex optimization: A basic course. Vol. 87. Springer Science & Business Media, 2003.

[4] Peypouquet, Juan. Convex optimization in normed spaces: theory, methods and examples. Springer, 2015.

[5] Polyak, Boris T. "Introduction to optimization. optimization software." Inc., Publications Division, New York 1 (1987): 32.

[6] Wright, Stephen J., and Benjamin Recht. Optimization for data analysis. Cambridge University Press, 2022.