Lecture Notes 2025

Optimisation, Control, and Data

Dario Prandi
dario.prandi@centralesupelec.fr

October 7, 2025

Contents

Part I

Optimisation

Chapter 1

Convex Analysis

This chapter closely follows chapter 5 of the lecture notes [3]. For most of the proof in this chapter we refer to [1] or [4].

1.1 Convex sets

Definition 1.1.1
A set $K \subset \mathbb{R}^d$ is *convex* if

$$tx + (1-t)y \in K \qquad \forall x, y \in K,\ t \in [0,1] \tag{1.1}$$

We have the following fact.

Proposition 1.1.1
The set $K \subset \mathbb{R}^d$ is convex if and only if for any $n \in \mathbb{N}$, and $t_1, \dots, t_n \geq 0$ such that $\sum_{i=1}^n t_i = 1$, it holds

$$x_1, \dots, x_n \in K \implies \sum_{i=1}^n t_i x_i \in K. \tag{1.2}$$

Proof: Property (1.2) with $n = 2$ is exactly the definition of K is convex. The statement then follows by induction on n. ■

Definition 1.1.2
The convex hull $\mathrm{conv}(\Omega)$ of $\Omega \subset \mathbb{R}^d$ is the smallest convex set K containing Ω.

By Proposition 1.1.1, it is immediate to observe that

$$\mathrm{conv}(\Omega) = \left\{ \sum_{i=1}^n t_i x_i \mid t_i \geq 0, \sum_{i=1}^n t_i = 1,\ x_i \in \Omega \right\}. \tag{1.3}$$

Example 1.1.1 (Convex sets)

- Unit ball w.r.t. any norm.
- Vector subspaces.
- Hyperplanes, i.e., for any $v \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$,

$$H_{v,\lambda} := \{x \in \mathbb{R}^d \mid \langle v, x \rangle \geq \lambda\}. \tag{1.4}$$

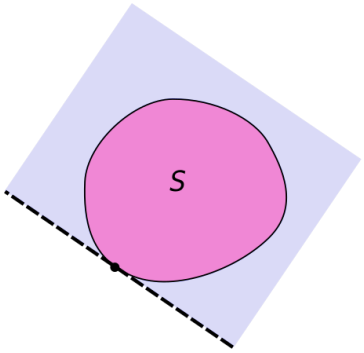

Figure 1.1: Supporting hyperplane for a set S.

An important result (that we will not prove) on convex sets is the following.

Theorem 1.1.1 Separation theorem

Let $K_1, K_2 \subset \mathbb{R}^d$ be two convex sets such with disjoint interior. Then there exists $v \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$ such that

$$K_1 \subset H_{v,\lambda}, \qquad K_2 \subset \mathbb{R}^d \setminus H_{v,\lambda}. \tag{1.5}$$

Here, $H_{v,\lambda}$ is defined in (1.4).

As a direct consequence, we have the following (see Figure 1.1).

Corollary 1.1.2 Supporting hyperplane theorem

Let $K \subset \mathbb{R}^d$ be a convex set and $x \in \partial K$. Then, there exists a supporting hyperplane of K containing x_0. That is, there exists $v \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$ such that $K \subset H_{v,\lambda}$ and $x_0 \in \partial H_{v,\lambda}$.

When K is a convex polygon, it is natural to expect it to be determined by its vertices. In order to formalize this intuition we need the following.

Definition 1.1.3

Let $K \subset \mathbb{R}^d$ be a convex set. A point $x \in K$ is an extremum of K if for any $y, z \in K$ and $t \in (0,1)$ we have that

$$x = ty + (1-t)z \implies x = y = z. \tag{1.6}$$

The set of exterma of K is denoted by $\operatorname{extr}(K)$.

In particular, for a convex polygon $\operatorname{extr}(K)$ is the set of its vertices.

Proposition 1.1.2

Let $K \subset \mathbb{R}^d$ be a convex set that is compact. Then,

$$\operatorname{conv}(K) = \operatorname{conv}\left(\operatorname{extr}(K)\right). \tag{1.7}$$

Figure 1.2: Two examples of polar cone.

1.2 Cones

Definition 1.2.1
A set $K \subset \mathbb{R}^d$ is a cone if

$$tx \in K \qquad \forall x \in K,\ t \geq 0. \tag{1.8}$$

Observe that every cone contains the origin.

Example 1.2.1 (Cones)

- The second order cone
$$C = \{x = (x', x_n) \in \mathbb{R}^d \times \mathbb{R} \mid \|x'\|_2 \leq x_n\}.$$
- Positive orthant $\mathbb{R}^d_+ = \{x \in \mathbb{R}^d \mid x_i \geq 0, \quad \forall i \in [\![1, d]\!]\}$.
- The set of positive semidefinite matrices $\mathrm{Sym}_+(\mathbb{R}^d)$.

Definition 1.2.2
The conic hull $\mathrm{cone}(\Omega)$ of a set $\Omega \subset \mathbb{R}^d$ is the smallest cone containing Ω. Namely,

$$\mathrm{cone}(\Omega) = \left\{ \sum_{i=1}^{n} t_i x_i \mid t_i \geq 0 \text{ and } x_i \in \Omega \text{ for any } i \in [\![1, n]\!] \right\}. \tag{1.9}$$

Definition 1.2.3
The dual cone K^* of a cone $K \subset \mathbb{R}^d$ is the set

$$K^* := \left\{ y \in \mathbb{R}^d \mid \langle x, y \rangle \geq 0 \quad \forall x \in K \right\}. \tag{1.10}$$

We have the following properties for the polar cone.

Proposition 1.2.1
The dual cone K^* is a closed, convex cone. If, moreover, the cone K is closed and convex, then $K^{**} = K$.

Proof: The fact that K^* is closed follows immediately by continuity of $y \mapsto \langle x, y \rangle$ for any $x \in K$. To show that K^* is convex, let $y, z \in K^*$, $x \in K$, and compute

$$\langle ty + (1-t)z, x \rangle = t\langle y, x \rangle + (1-t)\langle z, x \rangle \geq 0 \qquad \forall t \in [0, 1]. \tag{1.11}$$

Consider now $x \in K$. By definition of K^* we have that $\langle x, y \rangle \geq 0$ for all $y \in K^*$, which implies that $K \subset K^{**}$.

In order to show the opposite inclusion, we will show that $x \notin K$ implies that $x \notin K^{**}$. Let $x \notin K$ and observe that then $\{x\}$ is a convex set whose interior is disjoint from K. By the Separation Theorem 1.1.1, there exists $v \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$

such that

$$\langle y, v\rangle \geq \lambda \qquad \forall y \in K \qquad \text{and} \qquad \langle x, v\rangle < \lambda. \tag{1.12}$$

Since $0 \in K$, for the above to be true it has to hold $\lambda \leq 0$, and hence it holds

$$\langle y, v\rangle \geq 0 \qquad \forall y \in K \qquad \text{and} \qquad \langle x, v\rangle < 0. \tag{1.13}$$

In particular, the first part of the above yields $v \in K^*$. Hence, the second part of the above yields $x \notin K^*{}^*$, as desired. ■

Remark: The convexity assumption in the above is essential, a counterexample is easily constructed by considering K to be the union of two half-lines. In general, $K^{**} = \overline{\text{conv}(K)}^{**}$.

1.3 Convex functions

We will work with extended functions $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. The domain of an extended function is

$$\text{dom}(F) = \{x \in \mathbb{R}^d \mid F(x) < +\infty\}. \tag{1.14}$$

An extended function such that $\text{dom}(F) \neq \emptyset$ is called *proper*.

Given a standard function $F : \Omega \to \mathbb{R}$, we can identify it with the extended function $\bar{F} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\bar{F}(x) = \begin{cases} F(x) & \text{if } x \in \Omega, \\ +\infty & \text{if } x \in \mathbb{R}^d \setminus \Omega. \end{cases} \tag{1.15}$$

Definition 1.3.1 Convex functions

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be an extended function. Then,

- F is convex if
$$F(tx + (1-t)y) \leq tF(x) + (1-t)F(y) \qquad \forall x, y \in \mathbb{R}^d,\ t \in [0,1]. \tag{1.16}$$
- F is strictly convex if
$$F(tx + (1-t)y) < tF(x) + (1-t)F(y) \qquad \forall x, y \in \mathbb{R}^d,\ x \neq y,\ t \in [0,1]. \tag{1.17}$$
- F is strongly convex if there exists $\gamma > 0$ such that
$$F(tx + (1-t)y) \leq tF(x) + (1-t)F(y) - \frac{\gamma}{2} t(t-1)\|x - y\|_2^2 \qquad \forall x, y \in \mathbb{R}^d,\ t \in [0,1]. \tag{1.18}$$

We say that F is *concave* if $-F$ is convex.

Observe that it holds

$$\text{convex} \impliedby \text{strongly convex} \impliedby \text{strictly convex} \tag{1.19}$$

We say that a standard function $F : K \to \mathbb{R}$ is convex, strictly convex, strongly convex, or concave, if the same is true for its extension $\bar{F}$. Observe that this requires K to be convex.

Example 1.3.1

- The prototypical convex function, used in the definition of strongly convex, is the quadratic function
$$F(x) = \frac{\|x\|_2^2}{2} = \frac{1}{2}\sum_{i=1}^{d} |x_i|^2. \tag{1.20}$$

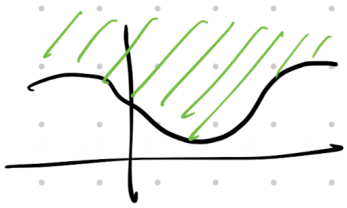

Figure 1.3: Epigraph of a function.

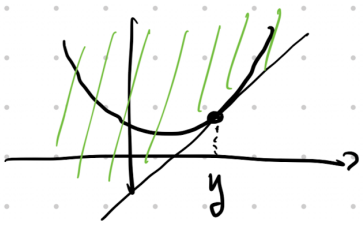

Figure 1.4: Graphical representation of Proposition 1.3.2.

- More generally, every norm is convex.
- The norm ℓ_p is strictly convex if and only if $p \in (1, +\infty)$.
- $F(x) = x^\top Ax$ is convex if A is positive semidefinite (i.e., $A \in \mathrm{Sym}_{\geq 0}(\mathbb{R}^d)$), and strongly convex if A is positive definite.

Proposition 1.3.1

A function $F : K \to \mathbb{R}$ is convex if and only if its epigraph $\mathrm{epi}(F) \subset \mathbb{R}^{d+1}$ is convex. Here, we let

$$\mathrm{epi}(F) = \{(x, r) \mid r \geq F(x)\}. \tag{1.21}$$

Proof: Assume F is convex and let $(x, r), (y, s) \in \mathrm{epi}(F)$. In particular, $r \geq F(x)$ and $s \geq F(y)$. Let $t \in [0, 1]$ and observe that

$$tr + (1-t)s \geq tF(x) + (1-t)F(y) \geq F(tx + (1-t)y). \tag{1.22}$$

Hence, $t(x, r) + (1-t)(y, s) \in \mathrm{epi}(F)$. A similar reasoning proves the opposite implication. ∎

Proposition 1.3.2 Differential characterisations of convexity

Let $F : \mathbb{R}^d \to \mathbb{R}$ be an everywhere differentiable function. Then,

- F is convex if and only if
$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle, \qquad \forall x, y \in \mathbb{R}^d. \tag{1.23}$$
- F is strongly convex with parameter $\gamma > 0$ if and only if
$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\gamma}{2}\|x - y\|_2^2, \qquad \forall x, y \in \mathbb{R}^d. \tag{1.24}$$
- If F is everywhere twice differentiable, then it is convex if and only if
$$\mathrm{Hess}\, F(x) \geq 0 \qquad \forall x \in \mathbb{R}^d, \tag{1.25}$$
and strongly convex if and only if there exists $\gamma > 0$ such that
$$\mathrm{Hess}\, F(x) \geq \frac{\gamma}{2} \qquad \forall x \in \mathbb{R}^d. \tag{1.26}$$
Here, we denoted by $\mathrm{Hess}\, F(x)$ the Hessian of F at x.

Proposition 1.3.3

Let $F : K \to \mathbb{R}$ be convex. Then, F is continuous on the interior of K.

Proof: Let $x_0 \in \mathrm{int}(K)$ and consider $r > 0$ such that $B(x_0, r) \subset K$. Without loss of generality, we assume $x_0 = 0$ (otherwise, replace the function F by its translation $G(x) = F(x) - F(x_0)$).

Convexity will allow to bound the difference $F(y) - F(0)$ with the values of F on the sphere $\partial B(0,r)$. However, without continuity, the function F need not be bounded on the compact set $\partial B(0,r)$, and hence we need some additional care.

Pick $d+1$ linearly independent points $v_0, \dots v_{d+1} \in \partial B(0,r)$, and consider the corresponding symplex

$$\Delta = \text{conv}\left(\{v_0, \dots, v_{d+1}\}\right) = \left\{ \sum_{i=1}^{d+1} t_i v_i \mid t_i \geq 0, \sum_i t_i = 1 \right\} \subset B(0,r). \tag{1.27}$$

Then, letting $M = \max_{i \in [\![1,d+1]\!]} F(v_i)$, the fact that F is convex yields that for any $x = \sum_{i=1}^{d+1} t_i v_i \in \Delta$ it holds

$$F(x) \leq \sum_{i=1}^{d+1} t_i F(v_i) \leq M. \tag{1.28}$$

In particular, we can fix a radius $r' < r$ such that $B(0,r') \subset \Delta$ where F is bounded.

We now proceed to bound the difference $F(x) - F(0)$. Let $x \in U \subset B(0,r')$ and set $t = \|x\|/r'$. In particular, $t \in [0,1]$ and the ray $\{sx \mid s \geq 0$ meets the sphere $\partial B(x_0, r')$ at the point

$$y = \frac{r'}{\|x\|}(x). \tag{1.29}$$

In particular, $x = (1-t)0 + ty$. By convexity and (1.28), we have

$$F(x) \leq (1-t)F(0) + tF(y) \leq .(1-t)F(0) + tM \implies F(x) - F(0) \leq t(M - F(0)). \tag{1.30}$$

To derive a bound from below, we proceed similarly, considering

$$z = \frac{r'}{\|x\| - r'}x. \tag{1.31}$$

Indeed, we then have $0 = (1-t)x + tx$, where $t = \|x\|/r'$ as above. Then, convexity and the fact that $z \in B(0,r')$ yield

$$F(0) \leq (1-t)F(x) + tM \implies F(x) - F(0) \geq -\frac{t}{1-t}(M - F(0)). \tag{1.32}$$

Combining (1.30) and (1.32), we obtain

$$-\frac{t}{1-t}(M - F(0)) \leq F(x) - F(0) \leq t(M - F(0)), \qquad t = \frac{\|x\|}{r'} \tag{1.33}$$

Since x was arbitrary in $B(0,r')$ we can take the limit as $x \to 0$, which implies $t \to 0$ and thus that

$$\lim_{x \to 0} |F(x) - F(0)| = 0, \tag{1.34}$$

concluding the proof. ∎

The following result is at the core of the relation between optimisation and convexity.

Theorem 1.3.1

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex extended function. Then,

- Any local minimum of F is global.
- The set of minima of F is convex.
- If F is strictly convex and admits a minimum, this minimum is unique.
- If F is real-valued and strongly convex, then it has a unique minimum.

Proof: Assume that $x^\star$ is a local minimum, i.e., there exists $r > 0$ such that $F(x^\star) \leq F(x)$ for any $x \in B(0,r)$. Let $y \in \mathbb{R}^d$ and consider a point on the ray starting at $x^\star$ and passing through y:

$$z = x^\star + s(y - x^\star) = (1-s)x^\star + sy \qquad s \geq 0. \tag{1.35}$$

Taking $s < \min\{1, r'/\|x^\star - y\|\}$ we have that $z \in B(0,r)$. Hence, by local minimality of $x^\star$ and convexity of F we have

$$F(x^\star) \leq F(z) \leq (1-s)F(x^\star) + sF(y) \implies F(x^\star) \leq F(y). \tag{1.36}$$

This concludes the proof of the first point.

Assume now that x_1, x_2 are minima for F. This clearly implies that $F(x_1) = F(x_2) =: m$, and thus, by convexity of F, for any $t \in [0,1]$ we have

$$m \leq F(tx_1 + (1-t)x_2) \leq tF(x_1) + (1-t)F(x_2) = m \implies F(tx_1 + (1-t)x_2). \tag{1.37}$$

This implies that $tx_1 + (1-t)x_2$ is a minimum for any $t \in [0,1]$, thus proving the second point.

The same argument as above in the case of a strictly convex function yields to

$$m \leq F(tx_1 + (1-t)x_2) < m \qquad \text{if } x_1 \neq x_2. \tag{1.38}$$

This implies immediately that the minimum is unique.

Assume, finally, that F is strongly convex. Since it is strictly convex, we just need to prove the existence of a minimum. By Proposition 1.3.3 we have that F is continuos, and thus it suffices to prove its coercivity: $F(x) \to +\infty$ if $\|x\| \to +\infty$. We provide a proof of this fact in the case where F is differentiable (the general case can be obtained similarly using Proposition 1.4.2, proven later on). In this case, by Proposition 1.3.2 we have that

$$F(y) \geq F(0) + \langle \nabla F(0), y\rangle + \frac{\gamma}{2}\|y\|_2^2 \qquad \forall y \in \mathbb{R}^d. \tag{1.39}$$

Since $\langle \nabla F(0), y\rangle \leq \|y\|_2$, the quadratic term on the right-hand side of the above equation, implies that the limit as $\|y\|_2 \to +\infty$ is $+\infty$. ∎

Remark: Strict convexity is not enough to ensure the existence of a minimum. Consider, for example, $F(x) = e^x$.

1.4 Convex conjugate and sub-differential

Definition 1.4.1

The convex conjugate (of Fenchel dual) of an extended function $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is the function $F^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined by

$$F^*(y) = \sup_{x \in \mathbb{R}^d} \left[\langle x, y\rangle - F(x)\right]. \tag{1.40}$$

Recall the following.

Definition 1.4.2

A function $F : \mathbb{R}^d \to \mathbb{R}$ is *lower semicontinuous* (l.s.c.) if

$$\liminf_{x \to x_0} F(x) \geq F(x_0), \qquad \forall x_0 \in \mathbb{R}^d. \tag{1.41}$$

Equivalently, F is l.s.c. if its epigraph is closed.

Example 1.4.1

- Every continuous function is lower semicontinuous.

- For any set $\Omega \subset \mathbb{R}^d$, the $0-\infty$ characteristic function

$$\chi_K = \begin{cases} 0 & \text{if } x \in \Omega, \\ +\infty & \text{otherwise,} \end{cases} \tag{1.42}$$

is lower semicontinuous, but not continuous.

Proposition 1.4.1

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Then,

1. The convex conjugate F^* is a lower semicontinuous convex function.
2. We have the Fenchel (or Young, or Fenchel-Young) inequality

$$\langle x, y \rangle \leq F(x) + F^*(y) \tag{1.43}$$

Proof: For any $y_1, y_2 \in \mathbb{R}^d$ and $t \in [0,1]$ we have

$$\langle x, ty_1 + (1-t)y_2 \rangle - F(x) = t\left(\langle x, y_1 \rangle - F(x)\right) + (1-t)\left(\langle x, y_2 \rangle - F(x)\right). \tag{1.44}$$

Taking the supremum for $x \in \mathbb{R}^d$ of the above, and recalling that $\sup(g(x)+h(x)) \leq \sup g(x) + \sup h(x)$ proves convexity of F^*.

Lower semicontinuity of F^* follows since it is the supremum for $x \in \mathbb{R}^d$ of $g_x(y) := \langle x, y \rangle - F(x)$, which is affine and in particular lower semicontinuous. Indeed, the supremum of a family of l.s.c. functions is l.s.c..

The second point (Fenchel inequality) is a direct consequence of the definition of F^*. ■

Example 1.4.2

- Let $F(x) = \frac{1}{2}\|x\|_2^2$. Then, $F^*(y) = \frac{1}{2}\|y\|_2^2 = F(y)$. This is the only function with this property.
- Let $F = \chi_K$ be the $0-\infty$ characteristic function of a convex set $K \subset \mathbb{R}^d$ defined in (1.42). Then,

$$F^*(y) = \sup_{x \in K} \langle x, y \rangle. \tag{1.45}$$

Definition 1.4.3

The *subdifferential* of a convex extended function $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ at $x \in \mathbb{R}^d$ is the set

$$\partial F(x) = \left\{ v \in \mathbb{R}^d \mid F(y) \geq F(x) + \langle v, y - x \rangle, \qquad \forall y \in \mathbb{R}^d \right\}. \tag{1.46}$$

A vector $v \in \partial F(x)$ is called a *subgradient* for F at x.

Example 1.4.3

Consider $F(x) = |x|$. Then,

$$\partial F(x) = \begin{cases} \{\operatorname{sgn}(x)\} & \text{if } x \neq 0, \\ [-1, 1] & \text{if } x = 0. \end{cases} \tag{1.47}$$

Here, $\operatorname{sgn}(x) = x/|x|$ is the sign function. See Figure 1.5.

Theorem 1.4.1

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Then, $x \in \mathbb{R}^d$ is a minimum for F if and only if $0 \in \partial F(x)$

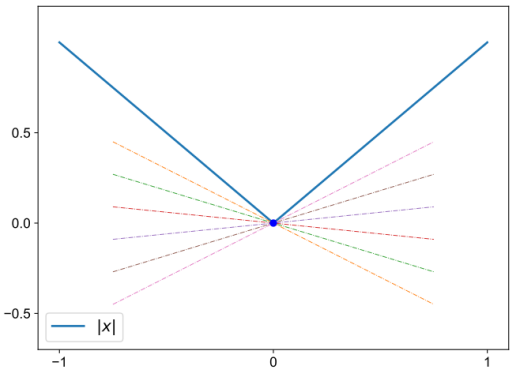

Figure 1.5: Visualization of the subgradients of $F(x) = |x|$ at $x = 0$. Image from this website.

Proof: The fact that x is a minimum means that $F(x) \leq F(y)$ for any $y \in \mathbb{R}^d$, which is the definition of $0 \in \partial F(x)$. ■

We have the following.

Proposition 1.4.2

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Then,

- For any $x \in \mathbb{R}^d$ the subdifferential $\partial F(x)$ is non-empty.
- It holds that
$$\partial F(x) = \left\{ v \in \mathbb{R}^d \mid F^*(v) + F(x) = \langle x, v \rangle \right\}. \tag{1.48}$$
- If F is differentiable at $x \in \mathbb{R}^d$, then $\partial F(x) = \{\nabla F(x)\}$.

Proof: The first part of the theorem is a consequence of the Supporting Hyperplane Theorem (see Corollary 1.1.2) and Proposition 1.3.1. Indeed, the latter implies that the epigraph $\operatorname{epi}(F)$ is convex and hence, by the former, any of its boundary point admits a supporting hyperplane. Using the fact that $\partial \operatorname{epi}(F) = \{(x, F(x)) \mid x \in \mathbb{R}^d\}$ allows to conclude.

To prove the second statement, observe that $v \in \partial F(x)$ is equivalent to

$$\langle y, v \rangle - F(y) \leq \langle x, v \rangle - F(x), \qquad \forall y \in \mathbb{R}^d. \tag{1.49}$$

Taking the sup for $y \in \mathbb{R}^d$ yields that $F^*(v) \leq \langle x, v \rangle - F(x)$. The opposite inequality follows from Fenchel inequality (see Proposition 1.4.1).

Concerning the proof of the last statement, the fact that $\nabla F(x) \in \partial F(x)$ follows from the characterisation of convexity for differentiable functions given in Proposition 1.3.2. To prove the opposite implication, let $v \in \partial F(x)$ and observe that by definition of subgradient the directional derivative $\partial_h F(x)$ of f in the direction $h \in \mathbb{R}^d$ at x satisfies

$$\partial_h F(x) = \lim_{t \to 0} \frac{F(x + th) - F(x)}{t} \geq \langle v, h \rangle. \tag{1.50}$$

Since we know that $\partial_h F(x) = \langle \nabla F(x), h \rangle$, we have that

$$\langle \nabla F(x) - v, h \rangle \geq 0, \qquad \forall h \in \mathbb{R}^d. \tag{1.51}$$

But this implies that $\nabla F(x) = v$, concluding the proof. ■

Thanks to the previous result, we are in a position to prove the following property of the convex biconjugate.

Theorem 1.4.2 Fenchel-Moreau Theorem

The biconjugate F^{**} is the largest convex lower semicontinuous function satisfying $F^{**}(x) \leq F(x)$ for any $x \in \mathbb{R}^d$.

In particular, $F^{**} = F$ if F is convex and proper.

Proof: We have that $-F^*(y) = \inf_{x\in\mathbb{R}^d}(F(x) - \langle x, y\rangle)$, which implies that for any $y, z \in \mathbb{R}^d$ it holds

$$\langle z, y\rangle - F^*(y) \leq \langle z - x, y\rangle + F(x), \qquad \forall x \in \mathbb{R}^d. \tag{1.52}$$

In particular, considering $z = x$ we have

$$F^{**}(x) = \sup_{y\in\mathbb{R}^d}(\langle x, y\rangle - F^*(y)) \leq F(x), \tag{1.53}$$

proving the first part of the statement.

Since F^{**} is convex and l.s.c. by Proposition 1.4.1, in order to complete the proof it suffices to show that if F is convex, then

$$F^{**}(x) \geq F(x), \qquad \forall x \in \mathbb{R}^d. \tag{1.54}$$

Let $v \in \partial F(x)$, which exists thanks to Proposition 1.4.2. For such a v, using the characterisation of the subdifferential in Proposition 1.4.2, we have

$$F^*(v) = \langle x, v\rangle - F(x), \tag{1.55}$$

so that $F^{**}(z) \geq \langle v, z - x\rangle + F(x)$ for any $z \in \mathbb{R}^d$. Picking $z = x$ allows to conclude. ■

Proposition 1.4.3 subdiff-conj

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $x, y \in \mathbb{R}^d$. Then, the following are equivalent:

i. $y \in \partial F(x)$.

ii. $F(x) + F^*(y) = \langle x, y\rangle$.

If, additionally, F is l.s.c. , then the above are also equivalent to

iii. $x \in \partial F^*(y)$.

Proof: To show that *i* is equivalent to *ii*, we just need to show that $y \in \partial F(x)$ is equivalent to

$$F(x) + F^*(y) \leq \langle x, y\rangle. \tag{1.56}$$

Indeed, the opposite inequality is always true due to Fenchel's inequality (see Proposition 1.4.1).

Observe that the fact that $y \in \partial F(x)$ means that

$$\langle x, y\rangle F(x) \geq \langle z, y\rangle F(z), \qquad \forall z \in \mathbb{R}^d. \tag{1.57}$$

That is, the function $z \mapsto \langle z, y\rangle F(z)$ attains its maximum at $z = x$. But, by definition of F^*, this is equivalent to (1.56), thus proving that *i* is equivalent to *ii*.

To complete the proof, observe that by Theorem 1.4.2 the lower semicontinuity of F yield that $F^{**} = F$, so that *ii* is equivalent to $F^{**}(x) + F^*(y) = \langle x, y\rangle$. Using the fact that $i \iff ii$ with F replaced by F^* completes the proof. ■

1.5 Proximal operator

Definition 1.5.1 Proximal operator

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The proximal operator associated with F is

$$P_F(y) = \arg\min_{x\in\mathbb{R}^d}\left\{F(x) + \frac{1}{2}\|x - y\|_2^2\right\}. \tag{1.58}$$

The above definition makes sense, since $x \mapsto F(x) + \frac{1}{2}\|x - y\|^2$ is a strongly convex function and hence has a unique minimum by Theorem 1.3.1.

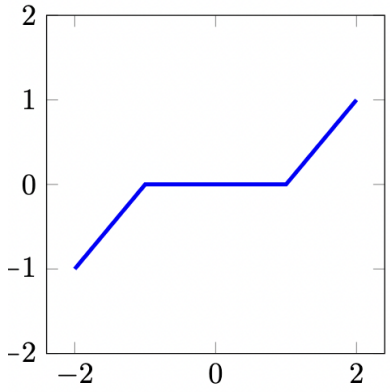

Figure 1.6: Soft-thresholding function.

Example 1.5.1 (Proximal operator of a convex set)

Let $F = \chi_K$ be the 0-∞ characteristic function of a convex set. Then, $P_K = P_{\chi_K}$ is the orthogonal projection onto K, that is

$$P_F(y) = \arg\min_{x \in K} \left\{ \|x - y\|^2 \right\}. \tag{1.59}$$

Example 1.5.2 (Soft-thresholding)

Let $F(x) = |x|$ for $x \in \mathbb{R}$. Then, for any $\lambda > 0$,

$$P_{\lambda F}(y) := S_\lambda(y) = \begin{cases} y + \lambda & \text{if } y \leq -\lambda, \\ 0 & \text{if } |y| < \lambda, \\ y - \lambda & \text{if } y \geq \lambda, \end{cases} \tag{1.60}$$

This function is known as soft-thresholding. See Figure 1.6.

The following proposition shows the relation between the proximal operator and the subdifferential, and justifies the notation

$$P_F = (\mathrm{Id} + \partial F)^{-1}. \tag{1.61}$$

Proposition 1.5.1

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function, then, for any $x, y \in \mathbb{R}^d$,

$$P_F(y) = x \iff y \in x + \partial F(x). \tag{1.62}$$

Proof: By Theorem 1.4.1, we have that $x = P_F(y)$ if and only if

$$0 \in \partial \left[\frac{1}{2} \| \cdot - y \|_2^2 + F(\cdot) \right](x) = x - y + \partial F(x). \tag{1.63}$$

Here, we used differentiability of $x \mapsto \|x - y\|_2^2$. This completes the proof. ∎

An important property of the proximal operator is the following.

Proposition 1.5.2 Non-expansiveness of the proximal operator

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function, then the proximal operator is non-expansive. Namely,

$$\|P_F(y_1) - P_F(y_2)\|_2^2 \leq \|y_1 - y_2\|_2^2, \qquad \forall y_1, y_2 \in \mathbb{R}^d. \tag{1.64}$$

Proof: Let $x_i = P_F(y_i)$. Then, by Proposition 1.5.1 we have $y_i \in x_i + \partial F(x_i)$. In particular, $y_i - x_i \in \partial F(x_i)$ and thus, by definition of subdifferential,

$$F(x_2) \geq F(x_1) + \langle y_1 - x_1, x_2 - x_1 \rangle \qquad \text{and} \qquad F(x_1) \geq F(x_2) + \langle y_2 - x_2, x_1 - x_2 \rangle. \tag{1.65}$$

Summing up, we get $0 \geq \langle x_1 - y_1 + y_2 - x_1, x_1 - x_2 \rangle$, which yields

$$\|x_1 - x_2\|_2^2 \leq \langle y_1 - y_2, x_1 - x_2 \rangle. \tag{1.66}$$

Applying Cauchy-Schwarz inequality allows to conclude. ■

The following relates the proximal operator of F and of its complex conjugate F^*, defined in Section 1.4.

Theorem 1.5.1 Moreau's Identity

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous convex function. Then,

$$P_F(y) + P_{F^*}(y) = y, \qquad \forall y \in \mathbb{R}^d. \tag{1.67}$$

Proof: Let $x = P_F(y)$ and set $z = y - x$. By Proposition 1.5.1 we thus have $z \in \partial F(x)$. Since F is lower semicontinuous, we have from Proposition **??** that $x \in \partial F^*(z)$ Since this is equivalent to $y \in z + \partial F^*(z)$, Proposition 1.5.1 implies that $z = P_{F^*}(y)$. Thus,

$$P_F(y) + P_{F^*}(y) = x + z = y. \tag{1.68}$$

■

Chapter 2

Optimization problems

2.1 Convex optimization problems

Definition 2.1.1

An optimization problem is a minimization problem of the form

$$\min_{x\in\mathbb{R}^d} F_0(x) \qquad \text{subject to} \qquad Ax = y \qquad \text{and} \qquad F_j(x) \le 0, \qquad j \in [\![1, M]\!]. \tag{OP}$$

Here,

1. $F_0 : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is the *objective function*;
2. $F_1, \dots, F_M : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ are the *constraing functions*;
3. $A \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^m$ provide the *equality constraints*;

The optimization problem is *convex* (resp. *linear*) if $F_0, \dots, F_M$ are convex (resp. linear) functions.

Definition 2.1.2

Consider an optimization problem (OP). Then,

- The set $\Phi \subset \mathbb{R}^d$ of points $x \in \mathbb{R}^d$ satisfying the constraints is the set of *feasible points*. That is,

$$\Phi = \left\{x \in \mathbb{R}^d \mid Ax = y, \qquad F_j(x) \le 0 \qquad \forall j \in [\![1, M]\!]\right\}. \tag{2.1}$$

  In particular, Φ is convex if (OP) is convex.
- Problem (OP) is *feasible* if it admits at least a feasible point (i.e., $\Phi \neq \varnothing$).
- The *optimal value* is $p^\star = \min_{x\in\Phi} F(x_0)$.
- A *minimizer* is a feasible point $x^\star$ such that $F_0(x^\star) \le F_0(x)$ for all feasible $x \in \Phi$. That is, $F_0(x^\star) = p^\star$.

Observe that the constrained optimization problem (OP) is equivalent to the uncostrained optimization problem

$$\min_{x\in\mathbb{R}^d} F_0(x) + \chi_\Phi, \tag{2.2}$$

where χ_Φ is the $0 - \infty$ characteristic function defined in (1.42).

Let us introduce the notation

$$\mathbb{R}^M = \{v \in \mathbb{R}^M \mid v_j \ge 0 \quad \forall j \in [\![1, M]\!]\}. \tag{2.3}$$

Definition 2.1.3 Lagrange and Lagrange dual functions

The *Lagrange function* of the optimization problem (OP) is the function $F : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^M_+ \to \mathbb{R} \cup \{+\infty\}$ defined by

$$L(x, \xi, \nu) = F_0(x) + \langle \xi, Ax - y \rangle + \sum_{j=1}^{m} \nu_j F_j(x). \tag{2.4}$$

The *Lagrange dual function* is the function $H : \mathbb{R}^m \times \mathbb{R}^M_+ \to \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$, defined by

$$H(\xi, \nu) = \inf_{x \in \mathbb{R}^d} L(x, \xi, \nu). \tag{2.5}$$

Proposition 2.1.1

The dual function is always concave. Moreover, if $x^\star$ is a minimizer of (OP), we have

$$H(\xi, \nu) \leq F(x^\star), \qquad \forall \xi \in \mathbb{R}^m,\ \nu \in \mathbb{R}^M_+. \tag{2.6}$$

Proof: Observe that $-H$ is the supremum w.r.t. $x \in \mathbb{R}^d$ of the functions $g_x(\xi, \nu) = -L(x, \xi, \nu)$. The function g_x is affine, and thus convex. Hence, $-H$ is the pointwise supremum of the family $\{g_x\}_{x \in \mathbb{R}^d}$ of convex function. It is immediate to check that it is convex, and thus that H is concave.

On the other hand, for any feasible point $x \in \Phi$, since $\nu_j \geq 0$ for any $j \in [\![1, M]\!]$, we have

$$\langle \xi, Ax - y \rangle + \sum_{j=1}^{m} \nu_j F_j(x) \leq 0. \tag{2.7}$$

Then, $L(x, \xi, \nu) \leq F_0(x) \leq F_0(x^\star)$ and, as a consequence,

$$H(\xi, \nu) \leq \inf_{x \in \Phi} L(x, \xi, \nu) \leq F_0(x^\star). \tag{2.8}$$

This completes the proof of the statement. ∎

The previous result suggests to introduce the following.

Definition 2.1.4 Primal and dual problem

The *dual problem* to (OP), which is called the *primal problem*, it the optimization problem

$$\max_{\xi \in \mathbb{R}^m,\ \nu \in \mathbb{R}^M} H(\xi, \nu) \qquad \text{subject to} \qquad \nu_j \geq 0 \quad \forall j \in [\![1, M]\!]. \tag{DP}$$

- A pair $(\xi, \nu) \in \mathbb{R}^m \times \mathbb{R}^M_+$ is called *dual feasible.*
- The *dual optimal value* is the solution $d^\star$ of (DP).
- A *dual optimal* or *optimal Lagrange multiplier* is a feasible maximizer $(\xi^\star \nu^\star) \in \mathbb{R}^m \times \mathbb{R}^M_+$.
- A *primal-dual optimal* is a triple $(x^\star, \xi^\star, \nu^\star)$ where $x^\star$ is a minimizer for (OP) and $(\xi^\star, \nu^\star)$ is a dual optimal.

Definition 2.1.5 Duality

The primal-dual problems always satisfy *weak duality*, that is $d^\star \leq p^\star$ where $d^\star$ is the dual optimal value and $p^\star$ is the primal optimal value.
We say that the problems enjoy *strong duality* if it holds

$$p^\star = d^\star. \tag{2.9}$$

The above shows the interest of the dual problem: when strong duality holds, in order to solve the minimization problem (OP) it suffices to solve the dual problem (DP).

The following is the most used criterion for strong duality.

Theorem 2.1.1 Slater's constraint quantification

Assume that $F_0, \dots, F_M$ are convex functions with domain $\mathrm{dom}(F_i) = \mathbb{R}^d$ for $i \in [\![1, M]\!]$. Then, strong duality holds if there exists $x \in \Phi \subset \mathbb{R}^d$ such that $F_j(x) < 0$ for any $j \in [\![1, M]\!]$. In particular, strong duality always holds for feasible optimization problems with no inequality constraints.
If, moreover, $F_0, \dots, F_M$ are lower semicontinuous, then the existence of a prima-dual optimal is guaranteed.

For a proof of the above result, we refer to [1, Section 5.3.2].

Example 2.1.1 (ℓ_1-minimization problem)

Consider the optimization problem

$$\min_{x \in \mathbb{R}^d} \|x\|_1 \qquad \text{subject to} \qquad Ax = y, \tag{2.10}$$

for some $A \in \mathbb{R}^{m \times d}$ and $y \in \mathbb{R}^m$.
The Lagrange function is independent of v, since there are no inequality constraints, and is

$$L(x, \xi) = \|x\|_1 + \langle \xi, Ax - y \rangle. \tag{2.11}$$

We have that the dual Lagrange function is

$$H(\xi) = \begin{cases} -\langle \xi, y \rangle & \text{if } \|A^\top \xi\|_\infty \leq 1, \\ -\infty & \text{otherwise.} \end{cases} \tag{2.12}$$

Indeed, it holds

$$H(\xi) = \inf_{x \in \mathbb{R}^d} \left[\|x\|_1 + \langle A^\top \xi, x \rangle - \langle \xi, y \rangle \right] \tag{2.13}$$

If $\|A^\top \xi\|_\infty \leq 1$ then $\langle A^\top \xi, x \rangle \geq -\|x\|_1$ and thus the infimum is attained for $x = 0$, yielding $H(\xi) = -\langle \xi, y \rangle$. On the other hand, for $\|A^\top \xi\|_\infty > 1$ let $i \in [\![1, m]\!]$ be the index such that $|(A^\top \xi)_i| = \|A^\top \xi\|$ and consider $x = -\operatorname{sgn}((A^\top \xi)_i) e_i$, so that $\langle A^\top \xi, x \rangle = -\|A^\top \xi\|_\infty$ and $\|x\|_1 = 1$. Thus, for any $\lambda > 0$,

$$H(\xi) \leq \lambda \left[1 - \|A^\top \xi\|_\infty \right] - \langle \xi, y \rangle \xrightarrow{\lambda \to +\infty} -\infty. \tag{2.14}$$

Hence, the dual program is given by

$$\max_{\xi \in \mathbb{R}^m} (-\langle \xi, y \rangle) \qquad \text{subject to} \qquad \|A^\top \xi\|_\infty \leq 1. \tag{2.15}$$

By Theorem 2.1.1 strong optimization holds for this primal-dual problems, provided the primal problem be feasible.

Geometric interpretation

Let us follow [1, Section 5.3] and present a geometric interpretation of the previous discussion. Assume that there are no equality constraints and a single inequality constraint, and define

$$\mathcal{G} = \{(F_1(x), F_0(x)) \mid x \in \mathbb{R}^d\}. \tag{2.16}$$

By construction, the problem is feasible if and only if $\mathcal{G}$ intersects the left-half plane. Furthermore, we have

$$p^\star = \min\{t \mid (u, t) \in \mathcal{G},\ u \leq 0\}. \tag{2.17}$$

Since $L(x, v) = (v, 1)^\top (F_1(x), F_0(x))$, we also have

$$H(v) = \inf\{(v, 1)^\top (u, t) \mid (u, t) \in \mathcal{G}\}. \tag{2.18}$$

Hence, if this infimum is finite, the inequality $(v, 1)^\top (u, t) \geq H(v)$ defines a supporting hyperplane for $\mathcal{G}$.

If the problem is convex, then $\mathcal{G}$ is convex and under Slater's condition its interior intersects the left-hand plane. This insures that strong duality holds.

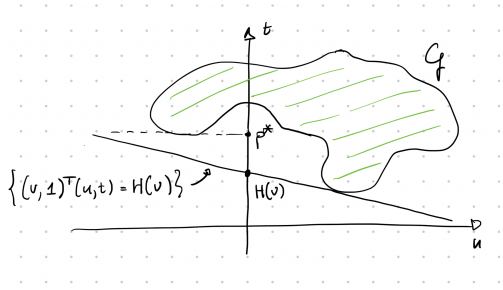

Figure 2.1: Geometric interpretation. The value of the dual function $H(v)$ identifies a supporting hyperplane for the set $\mathcal{G}$.

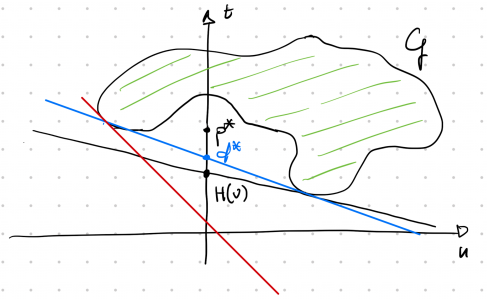

Figure 2.2: Geometric interpretation. Solving the dual problem yields the blue hyperplane. In this case $p^\star > d^\star$ and strong duality does not hold.

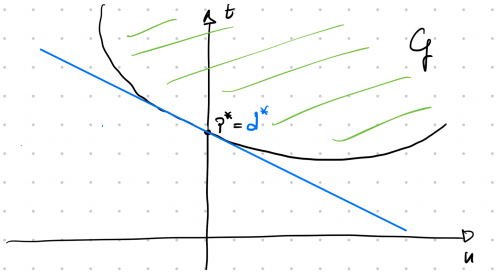

Figure 2.3: Geometric interpretation of Slater's condition. When the set $\mathcal{G}$ is convex and has interior that intersects the left-hand plane, the best supporting hyperplane yields the optimal value $p^\star$.

2.2 Conic optimization problems

Definition 2.2.1

A *conic optimization problem* is a minimization problem of the form

$$\min_{x\in\mathbb{R}^d} F_0(x) \qquad \text{subject to} \qquad x \in K \qquad \text{and} \qquad F_j(x) \le 0, \qquad j \in [\![1, M]\!]. \tag{COP}$$

Here, $F_0, \dots, F_M : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ are convex functions and $K \subset \mathbb{R}^d$ is a convex cone.

Also conic optimization problems have their duality theory.

Definition 2.2.2 Duality for conic optimization problems

The *Lagrange function* of the optimization problem (COP) is the function $F : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^M_+ \to \mathbb{R} \cup \{+\infty\}$ defined by

$$L(x, \xi, v) = F_0(x) - \langle \xi, x\rangle + \sum_{j=1}^{m} v_j F_j(x). \tag{2.19}$$

The *Lagrange dual function* is the function $H : \mathbb{R}^m \times \mathbb{R}^M_+ \to \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$, defined by

$$H(\xi, v) = \inf_{x\in\mathbb{R}^d} L(x, \xi, v). \tag{2.20}$$

The dual problem associated with (COP) is then

$$\max_{\xi\in\mathbb{R}^m, v\in\mathbb{R}^M} H(\xi, v) \qquad \text{subject to} \qquad \xi \in K^*,\ v \in \mathbb{R}^M_+. \tag{2.21}$$

Here, K^* is the dual cone of K (see Definition 1.2.3)

The duality theory is set up in order to have weak duality.

Proposition 2.2.1

The dual function is always concave. Moreover, if $x^\star$ is a minimizer of (COP), we have

$$H(\xi, \nu) \leq F(x^\star), \qquad \forall \xi \in \mathbb{R}^m,\ \nu \in \mathbb{R}^M_+. \tag{2.22}$$

The proof of this result can be done as for Proposition 2.1.1, taking into account the definition of dual cone.

Similar conditions as in Slater's constraint qualification (Theorem 2.1.1) ensure strong duality for conic problems; for instance, if there exists a point in the interior of K such that all inequality constraints hold strictly, see e.g., [1, Section 5.9].

Example 2.2.1

For a convex cone $K \subset \mathbb{R}^d$ and a vector $v \in \mathbb{R}^d$, consider the conic optimization problem

$$\min_{x \in \mathbb{R}^d} \langle x, v \rangle \qquad \text{subject to} \qquad x \in K,\ \|x\|_2^2 \leq 1. \tag{2.23}$$

The Lagrange function is given by

$$L(x, \xi, \nu) = \langle x, v \rangle - \langle \xi, x \rangle + \nu(\|x\|_2^2 - 1), \qquad \xi \in K^*,\ \nu \geq 0. \tag{2.24}$$

Minimizing the above w.r.t. x one immediately obtains

$$H(\xi, \nu) = -\nu - \frac{1}{4\nu}\|\xi - v\|_2^2, \qquad \xi \in K^*,\ \nu \geq 0. \tag{2.25}$$

For fixed ξ it is easy to maximize $H(\xi, \nu)$ w.r.t. $\nu \geq 0$, yielding $\nu = \|\xi - v\|_2^2/2$. Thus, the dual problem simplifies to

$$\max_{\xi \in \mathbb{R}^m} \left(-\frac{\|\xi - v\|_2}{2} \right) \qquad \text{subject to} \qquad \xi \in K^*. \tag{2.26}$$

That is, the optimal value of the dual problem is the optimal value of the above, and any dual optimal $(\xi^\star, \nu^\star)$ is such that $\nu^\star = \|\xi^\star - v\|_2^2/2$ and $\xi^\star$ is optimal for the above.
Observe, that minimizers for (2.26) are the orthogonal projections of v on the dual cone K^*.

2.3 Saddle-point interepretation and penalty method

Theorem 2.3.1 Saddle-point property

Consider an optimisation problem (convex of conical). Then, the primal-dual optimal values $p^\star$ and $d^\star$ satisfy

$$p^\star = \inf_{x \in \mathbb{R}^d} \sup_{\xi \in \mathbb{R}^m,\ \nu \in \mathbb{R}^M_+} L(x, \xi, \nu) \qquad \text{and} \qquad d^\star = \sup_{\xi \in \mathbb{R}^m,\ \nu \in \mathbb{R}^M_+} \inf_{x \in \mathbb{R}^d} L(x, \xi, \nu). \tag{2.27}$$

In particular,

- Strong duality is equivalent to the fact that

$$\inf_{x \in \mathbb{R}^d} \sup_{\xi \in \mathbb{R}^m,\ \nu \in \mathbb{R}^M_+} L(x, \xi, \nu) = \sup_{\xi \in \mathbb{R}^m,\ \nu \in \mathbb{R}^M_+} \inf_{x \in \mathbb{R}^d} L(x, \xi, \nu). \tag{2.28}$$

- Primal-dual optimizer $(x^\star, \xi^\star, \nu^\star)$ are exactly the saddle points of L. That is,

$$L(x^\star, \xi, \nu) \leq L(x^\star, \xi^\star, \nu^\star) \leq L(x, \xi^\star, \nu^\star), \qquad \forall x \in \mathbb{R}^d,\ (\xi, \nu) \in \mathbb{R}^m \times \mathbb{R}^M_+. \tag{2.29}$$

Proof: We consider a convex optimisation problem (the same considerations hold *mutas mutandis* for conical problems).

The fact that $d^\star$ satisfyies (2.27) is a direct consequence of the definition. On the other hand, we have

$$\sup_{\xi\in\mathbb{R}^m,\nu\in\mathbb{R}^M_+} L(x,\xi,\nu) = F_0(x) + \sup_{\xi\in\mathbb{R}^m}\langle\xi, Ax-y\rangle + \sup_{\nu_j\geq 0}\sum_{j=1}^M \nu_j F_j(x) = \begin{cases} F_0(x) & \text{if } Ax=y \text{ and } F_j(x)\leq 0\,\forall j\in[\![1,M]\!], \\ +\infty & \text{otherwise.}\end{cases} \tag{2.30}$$

In other words, the above supremum is $+\infty$ if x is not feasible (i.e., $x\notin\Phi$). This shows that minimizing the above w.r.t. $x\in\mathbb{R}^d$ yields $p^\star$. ∎

As a consequence of the saddle-point property, we recover a classical method (penalty method or Tychonoff regularization), that allows to transform constrained problems in different but equivalent unconstrained problems that are typically easier to solve.

Given two parameters $\eta>0$ and $\lambda\geq 0$, we consider the following two problems:

$$\min_{x\in\mathbb{R}^d} F_0(x) \qquad \text{subject to} \qquad F_1(x)\leq\eta, \tag{PO$_2(\eta)$}$$

and

$$\min_{x\in\mathbb{R}^d} F_0(x)+\lambda F_1(x). \tag{PO$_2(\lambda)$}$$

We have the following.

Theorem 2.3.2 Penalty method

Assume that F_0, F_1 are lower semicontinuous, satisfy the assumptions of Theorem 2.1.1, and that $F_1(x)\geq 0$ for all $x\in\mathbb{R}^d$. Then, for $x^\star\in\mathbb{R}^d$ the following statements are equivalent:

- There exists $\eta\geq 0$ such that $x^\star$ is a minimizer of (PO$_2(\eta)$).
- There exists $\lambda\geq 0$ such that $x^\star$ is a minimizer of (PO$_2(\lambda)$).

Proof: Assume that $x^\star$ is a minimizer for (PO$_2(\eta)$), and recall that for this problem

$$L(x,\nu) = F_0(x)+\nu(F_1(x)-\eta), \qquad x\in\mathbb{R}^d,\ \nu\geq 0. \tag{2.31}$$

By Theorem 2.1.1, we have that strong duality holds for (PO$_2(\eta)$) and that there exists a primal-dual optimal $(x^\star,\nu^\star)$. By the saddle-point propery (Theorem 2.3.1) it holds that $L(x^\star,\nu^\star)\leq L(x,\nu^\star)$ for any $x\in\mathbb{R}^d$. Since the constant term $-\nu^\star\eta$ does not affect the minimization, this proves that $x^\star$ is a minimizer of (PO$_2(\lambda)$) with $\lambda=\nu$.

For the converse statement, assume now that $x^\star$ is a minimizer of (PO$_2(\lambda)$). Choose $\eta=F_1(x^\star)\geq 0$, so that the dual function of problem (PO$_2(\eta)$) satisfies

$$H(\lambda) = \inf_{x\in\mathbb{R}^d} L(x,\lambda) = \inf_{x\in\mathbb{R}^d}[F_0(x)+\lambda F_1(x)] - \lambda F_1(x^\star) = F_0(x^\star). \tag{2.32}$$

Here, we used that $x^\star$ is a minimizer for (PO$_2(\lambda)$). Since weak duality implies that $H(\lambda)\leq F_0(x)$ for any feasible $x\in\mathbb{R}^d$, and $x^\star$ is feasible due to the choice of η, it follows that $x^\star$ is a minimizer of (PO$_2(\eta)$). ∎

Remark: The non-negativity assumption on F_1 can be removed by replacing $F_1(x)$ in (PO$_2(\lambda)$) by $\min\{0,F_1(x)\}$.

Chapter 3

Numerical methods for (non)convex optimization

3.1 Gradient descent

We are concerned with the following unconstrained minimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{3.1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is a sufficiently smooth (say C^1) function.

Since the gradient $\nabla f(x)$ encodes the direction of maximal growth of f at x, it is natural to expect that moving in the direction $-\nabla f(x)$ will lead to a minimum $x^\star$ (3.1).

Definition 3.1.1 Gradient flow

Assume that $f\in C^1(\mathbb{R}^d)$. The associated gradient flow from $x_0\in\mathbb{R}^d$ is ordinary differential equation:

$$\dot{x} = -\nabla f(x), \qquad x(0)=x_0. \tag{3.2}$$

By Cauchy-Lipschitz theorem, (3.2) admits local solutions. In order for them to be global, we assume the following.

Assumption 3.1.1

The function $f\in C^1(\mathbb{R}^d)$ and, moreover, $\nabla f:\mathbb{R}^d\to\mathbb{R}^d$ is Lipschitz continuous with constant $L>0$. That is,

$$\|\nabla f(x)-\nabla f(y)\|_2 \le L\|x-y\|_2, \qquad \forall x,y\in\mathbb{R}^d. \tag{3.3}$$

Observe that if $f\in C^2(\mathbb{R}^d)$, the above assumption is verified by assuming that $\operatorname{Hess} f(x)\le L\,\mathrm{Id}$ in the sense of symmetric matrices (i.e., all eigenvalues of $\operatorname{Hess} f(x)$ are bounded above by L).

Recall also that a function $f\in C^2(\mathbb{R}^d)$ that is strongly convex (i.e., there exists $\gamma>0$ such that $\operatorname{Hess} f(x)\ge\gamma/2$ for any $x\in\mathbb{R}^d$) always admits a unique minimizer (see Proposition 1.3.2 and Theorem 1.3.1).

Theorem 3.1.1

Assume that $f\in C^2(\mathbb{R}^d)$ is strongly convex function satisfying Assumption 3.1.1. Then, for any initial condition $x_0\in\mathbb{R}^d$, the gradient flow (3.2) converges to the unique minimizer $x^\star=\arg\min_{x\in\mathbb{R}^d} f(x)$.

Proof: Since $x^\star$ is unique, it is characterized by property $\nabla f(x^\star)=0$. In particular, the statement is trivial if $x_0=x^\star$. Otherwise, let $V(x)=\frac{1}{2}\|\nabla f(x)\|_2^2$ and observe that $\nabla V(x)=\operatorname{Hess} f(x)\nabla f(x)$. Computing V along a solution $t\mapsto x(t)$ of

(3.2) we have

$$\begin{aligned}\frac{d}{dt}V(x(t)) &= \nabla V(x(t))^\top \dot{x}(t)\\ &= \nabla f(x(t))^\top \operatorname{Hess} f(x(t))\dot{x}(t)\\ &= -\nabla f(x(t))^\top \operatorname{Hess} f(x(t))\nabla f(x(t))\\ &\leq -\frac{\gamma}{2}\|\nabla f(x(t))\|_2^2 = -\gamma V(x(t)).\end{aligned} \tag{3.4}$$

Here, we used the fact that $\operatorname{Hess} f(x) \geq \gamma/2$. By integration of the above, we then obtain that for any $t \geq 0$ it holds

$$V(x(t)) \leq V(x_0)e^{-\gamma t} \iff \|\nabla f(x(t))\|_2 \leq \|\nabla f(x_0)\|_2 e^{-\gamma t/2}. \tag{3.5}$$

Recall that $\nabla f(x^\star) = 0$, this implies that

$$\lim_{t\to+infty} \|\nabla f(x(t)) - \nabla f(x^\star)\|_2 \leq \|\nabla f(x_0)\|_2 \lim_{t\to+\infty} e^{-\gamma t/2} = 0. \tag{3.6}$$

Let us now use the above to show that $x(t) \to x^\star$ as $t \to +\infty$. Indeed, we have

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\|x(t) - x^\star\|_2^2 &= \langle x(t) - x^\star, \dot{x}(t)\rangle\\ &= -\langle x(t) - x^\star, \nabla f(x(t))\rangle\\ &= -\langle x(t) - x^\star, \nabla f(x(t)) - \nabla f(x^\star)\rangle.\end{aligned} \tag{3.7}$$

By Theorem 3.1.2, to be proven later, this implies

$$\frac{d}{dt}\frac{1}{2}\|x(t) - x^\star\|_2^2 \leq -\frac{1}{L}\|\nabla f(x(t)) - \nabla f(x_0)\|_2^2 \leq 0. \tag{3.8}$$

Integrating the above, we obtain $\|x(t) - x^\star\|_2 \leq \|x_0 - x^\star\|_2$, which implies that $x(t) \in \bar{B}(x^\star, \|x_0 - x^\star\|_2)$ (i.e., the evolution $t \mapsto x(t)$ is uniformly bounded in time). By compactness, there exists a sequence of times $(t_n)_n \subset \mathbb{R}_+$ and a point x_∞ such that $x(t_n) \to x_\infty$. However, the fact that $f \in C^2(\mathbb{R}^d)$ implies that $\nabla f(x(t_n)) \to \nabla f(x_\infty)$, and thus (3.6) yields that $\nabla f(x_\infty) = \nabla f(x^\star) = 0$. By existence and uniqueness of the minimizer of f, this shows that $x_\infty = x^\star$. ■

Theorem 3.1.2 Co-coercivity

Let $f \in C^1(\mathbb{R}^d)$ be convex and such that ∇f is Lipschitz continuous with constant $L > 0$. Then, we have that

1. Quadratic upper bound (compare with Proposition 1.3.2, this does not require convexity):

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^d. \tag{3.9}$$

2. If $x^\star$ is a minimizer for f,

$$\frac{1}{2L}\|\nabla f(y)\|_2^2 \leq f(y) - f(x^\star) \leq \frac{1}{2}\|y - x^\star\|_2^2, \qquad \forall y \in \mathbb{R}^d. \tag{3.10}$$

3. Co-coercivity:

$$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq L\langle x - y, \nabla f(x) - \nabla f(y)\rangle, \qquad \forall x, y \in \mathbb{R}^d. \tag{3.11}$$

Proof: The quadratic upper bound is a direct consequence of the fact that

$$f(y) = f(x) + \int_0^1 \frac{d}{dt} f(x + t(y - x))\, dt = f(x) + \int_0^1 \langle \nabla f(x + t(y - x)), y - x\rangle\, dt. \tag{3.12}$$

Indeed, the L-Lipschitz property of ∇f and the Cauchy-Schwarz inequality yields

$$\langle \nabla f(x + t(y - x)), y - x\rangle \leq \langle \nabla f(x), y - x\rangle + L\|y - x\|^2. \tag{3.13}$$

To prove the second part of the statement, since $\nabla f(x^\star) = 0$ the right-hand inequality follows from the quadratic lower bound (3.9) with $x = x^\star$. On the other hand, still by (3.9) with $x = y$ and $y = z$ we obtain

$$\begin{aligned} f(x^\star) &= \inf_z f(z) \\ &\leq \inf_z \left[f(y) + \langle \nabla f(y), z - y \rangle + \frac{L}{2}\|y - z\|^2 \right] \\ &\leq \inf_{\|v\|=1} \inf_{t \geq 0} \left[f(y) + t\langle \nabla f(y), v \rangle + \frac{Lt^2}{2} \right] \\ &= \inf_{\|v\|=1} \left[f(y) - \frac{1}{2L} \left(\langle \nabla f(y), x \rangle \right)^2 \right] \\ &= f(y) - \frac{1}{2L}\|\nabla f(y)\|_2^2. \end{aligned} \tag{3.14}$$

To conclude the proof, let us fix $x \in \mathbb{R}^d$ and consider the two functions

$$f_x(z) = f(z) - \langle \nabla f(x), z \rangle \qquad \text{and} \qquad f_y(z) = f(z) - \langle \nabla f(y), z \rangle. \tag{3.15}$$

It is immediate to check that $\nabla f_x(z) = \nabla f(z) - \nabla f(x)$ is Lipschitz continuous of constant L, and that f_x is convex. Since $z = x$ minimizes f_x by convexity, the left-hand inequality in (3.10) applied to f_x at $x^\star = x$ shows that

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = f_x(y) - f_x(x) \geq \frac{1}{2L}\|\nabla f_x(y)\|_2^2 = \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2. \tag{3.16}$$

The same reasoning applied to f_y, which is minimized by $z = y$, yields

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2. \tag{3.17}$$

Combining the two inequalities completes the proof. ∎

A natural time-discretization of the gradient flow (3.2) is via the Explicit Euler scheme, which yields the gradient descent algorithm.

Definition 3.1.2 Gradient descent algorithm

The gradient descent algorithm for $f \in C^1(\mathbb{R}^d)$ starting at $x_0 \in \mathbb{R}^d$ with parameters $(\alpha_n)_n \subset \mathbb{R}_+$, is

$$x_{n+1} = x_n - \alpha_n \nabla f(x_n). \tag{3.18}$$

Here the parameters α_n are tyipically chosen adaptively with a line search. However, also a constant choice $\alpha = \alpha_n$ sufficiently small will suffice. Let us show now a result of convergence of the algorithm above.

Theorem 3.1.3

Let $f \in C^1(\mathbb{R}^d)$ be coercive (i.e., $f(x) \to +\infty$ as $x \to +\infty$), and such that ∇f is Lipschitz continuous with constant $L > 0$, and admitting a unique minimizer $x^\star$. Assume that

$$\sum_{n=0}^{+\infty} \alpha_n = +\infty \qquad \text{and} \qquad \alpha_n \leq \frac{1}{L}, \quad \forall n \in \mathbb{N}. \tag{3.19}$$

Then, the iterations of the algorithm (3.18) converge to $x^\star$ (i.e., $x_n \to x^\star$ as $n \to +\infty$).

Proof: Fix $n \in \mathbb{N}$. Then, by the quadratic upper bound of Theorem 3.1.2, we have

$$f(x_{n+1}) \leq f(x_n) + \langle \nabla f(x_n), x_{n+1} - x_n \rangle + \frac{L}{2}\|x_{n+1} - x_n\|_2^2 = f(x_n) - \alpha_n \left(1 - \frac{\alpha_n L}{2}\right) \|\nabla f(x_n)\|_2^2. \tag{3.20}$$

The above implies that

$$\sum_{n=0}^{N-1} \alpha_n \left(1 - \frac{\alpha_n L}{2}\right) \|\nabla f(x_n)\|_2^2 \leq \sum_{n=0}^{N-1} (f(x_n) - f(x_{n+1})) = f(x_0) - f(x_N) \leq f(x_0) - f(x^\star). \tag{3.21}$$

Since the r.h.s. is independent of N and $(1 - L\alpha_n/2) \geq 1/2$, this shows that

$$\sum_{n=0}^{N-1} \alpha_n \|\nabla f(x_n)\|_2^2 \leq \frac{1}{2} \sum_{n=0}^{N-1} \alpha_n \left(1 - \frac{\alpha_n L}{2}\right) \|\nabla f(x_n)\|_2^2 < +\infty. \tag{3.22}$$

Since $\sum_n \alpha_n = +\infty$ by assumption, one easily shows by contradiction that $\|\nabla f(x_n)\|_2 \to 0$ as $n \to +\infty$.

Observe that, by (3.20) and the assumption $\alpha_n \leq 1/L$, we deduce that $f(x_{n+1}) < f(x_n)$ for any $n \in \mathbb{N}$. In particular, by coercivity, there exists $R > 0$ such that $x_n \in \bar{B}(0, R)$ for any $n \in \mathbb{N}$, and thus the same argument used at the end of the proof of Theorem 3.1.1 shows that $\|\nabla f(x_n)\|_2^2$ implies convergence to the unique minimum. ■

Remark: The coercivity assumption can be relaxed by simply asking that the level set $\{x \mid f(x) \leq f(x_0)\}$ be bounded. Is it also possible to remove the assumption of uniqueness of the minimum, in which case we have that every cluster point of the iterates is a minimum.

Under additional assumptions on the function f, we can quantify its the rate of convergence.

Theorem 3.1.4

Under the assumptions of Theorem 3.1.3, letting $\alpha_n = \alpha \leq 1/L$, we have

- If f is convex, the convergence is *sublinear*: there exists $c > 0$ such that

$$f(x_n) - f(x^\star) \leq c\|x_0 - x^\star\|_2^2 \tag{3.23}$$

- If f is γ-strongly convex, the convergence is *exponential*: there exists $c > 0$ and $\mu \in (0, 1)$ such that

$$\|x_n - x^\star\|_2 \leq \mu^n \|x_0 - x^\star\|_2, \qquad \forall n \in \mathbb{N}. \tag{3.24}$$

Proof: By definition of x_{n+1} we have

$$\|x_{n+1} - x^\star\|_2^2 = \|x_n - x^\star\|_2^2 - 2\alpha\langle \nabla f(x_n), x_n - x^\star\rangle + \alpha^2 \|\nabla f(x_n)\|_2^2. \tag{3.25}$$

Let us start by assuming that f is convex, so that

$$f(x_n) - f(x^\star) \leq \langle \nabla f(x_n), x_n - x^\star\rangle = \frac{1}{2\alpha}\left[\|x_n - x^\star\|_2^2 + \alpha^2 \|\nabla f(x_n)\|_2^2\right]. \tag{3.26}$$

Applying point 2 of Theorem 3.1.2 to bound the gradient term, we thus get

$$(1 - \alpha L) f(x_n) - f(x^\star) \leq \frac{1}{2\alpha}\|x_n - x^\star\|_2^2. \tag{3.27}$$

Since $1 - \alpha L > 0$, the statement follows by a recursion argument.

Let us now assume that f is γ-strongly convex. Using the fact that $\nabla f(x^\star) = 0$, from the differential characterization of convexity (Theorem 1.3.2) we obtain

$$\langle \nabla f(x_n), x_n - x^\star\rangle \geq \frac{\gamma}{2}\|x_n - x^\star\|_2^2. \tag{3.28}$$

Hence, using also (3.3), from (3.25) we obtain

$$\|x_{n+1} - x^\star\|_2^2 \leq \mu(\alpha)\|x_n - x^\star\|_2^2, \qquad \mu(\alpha) = 1 - \alpha\gamma + \alpha^2 L^2. \tag{3.29}$$

One easily checks that $\mu(\alpha) \in (0, 1)$ for $\alpha \in (1, L^{-1})$. The statement follows by a recursion argument. ■

3.2 Stochastic gradient descent

For this section, we mainly refer to [5].

The gradient descent algorithm of Definition 3.1.2 is simple to implement and widely used in machine learning. However, it requires to be able to compute $\nabla f(x)$ fairly efficiently at any point $x \in \mathbb{R}^d$.

A situation frequently encountered in machine learning is when the objective function is of the form

$$f(x) = \mathbb{E}\left[f(x,\cdot)\right] = \int_\Omega f(x,\omega)\, d\mathbb{P}(\omega), \tag{3.30}$$

where $f(x,\cdot) : \Omega \to \mathbb{R}$ are all random variables over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. This probability space tipically represents the space of training data: each realization $f(\cdot,\omega)$ is a loss function on the specific sample $\omega \in \Omega$: it tells "how bad" picking the parameter $x \in \mathbb{R}^d$ for our model performs on this sample. Since Ω can be extremely large, the computation of

$$\nabla f(x) = \int_\Omega \nabla f(x,\omega)\, d\mathbb{P}(\omega), \tag{3.31}$$

can become extremely expensive and the implementation of a gradient descent unreasonable. This is known as the *curse of dimensionality*.

Example 3.2.1

If we are trying to learn a regression model, $\omega = (\hat{z},\hat{y})$ would be the observations, and the model would be $y = mz + q$, with parameter $x = (m,q)$. Then, the loss function for the sample $(\hat{z},\hat{y})$ is

$$\ell((m,q),(\hat{z},\hat{y})) = (\hat{y} - m\hat{z} - q)^2. \tag{3.32}$$

The final loss function over N samples is then

$$\ell((m,q)) = \frac{1}{N}\sum_{i=1}^{N} \ell((m,q),(\hat{z}_i,\hat{y}_i)) = \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - m\hat{z}_i - q)^2. \tag{3.33}$$

In this case, $\Omega = \{(\hat{z}_i,\hat{y}_i) \mid i \in [\![1,N]\!]\}$ endowed with a uniform probability.

For this particular setting, there is a way out which stems from replacing the actual gradient $\nabla f(x)$ with a "stochastic approximation", obtained by sampling (3.31).

Definition 3.2.1 Stochastic gradient descent

Let f be given as in (3.30), with $f(\cdot,\omega) \in C^1(\mathbb{R}^d)$ for any $\omega \in \Omega$. The stochastic gradient descent algorithm starting at $x_0 \in \mathbb{R}^d$ with parameters $(\alpha_n)_n \subset \mathbb{R}_+$ is

$$x_{n+1} = x_n - \alpha_k \nabla_x f(x_n,\omega_n), \tag{3.34}$$

where $(\omega_n)_n \in \Omega$ is a sequence of indipendent and identically distributed random variables.

Due to the randomness in the choice of ω_n, this algorithm is probabilistic. We can, however, control its average behavior. For this we need to require certain assumptions on the stochastic gradient.

Assumption 3.2.1

At any iteration $n \in \mathbb{N}$ of (3.34), the random variable ω_n satisfies

1. The stochastic gradient $\nabla f(x,\omega_n)$ is an unbiased estimator of the real gradient, meaning that $\mathbb{E}_{\omega_n}\left[\nabla f(x,\omega_n)\right] = \nabla f(x)$ for any $x \in \mathbb{R}^d$.
2. There exists σ^2 such that the following variance estimate holds

$$\mathbb{E}_{\omega_n}\left[\|\nabla f(x,\omega_n)\|_2^2\right] \le \|\nabla f(x)\|_2^2 + \sigma^2, \qquad \forall x \in \mathbb{R}^d. \tag{3.35}$$

Remark: In the above, one has to pay attention to the fact that $\nabla f(x, \omega_n)$ is a random variable through ω_n. E.g., in the case where Ω is discrete

$$\mathbb{E}_{\omega_n}\left[\nabla f(x, \omega_n)\right] = \sum_{\omega \in \Omega} \nabla f(x, \omega) \mathbb{P}(\omega_n = \omega). \tag{3.36}$$

Example 3.2.2 (Uniform sampling)

In the case where $\Omega = [\![1, N]\!]$ and ω_n is just a uniform sampling over Ω (i.e., $\mathbb{P}(\omega_n = k) = 1/N$ for any $k \in \Omega$), we have

$$\mathbb{E}_{\omega_n}\left[\nabla f(x, \omega_n)\right] = \sum_{k=1}^{N} \nabla f(x, k) \mathbb{P}(\omega_n = k) = \frac{1}{N} \sum_{k=1}^{N} \nabla f(x, k) = \nabla f(x). \tag{3.37}$$

In particular, this shows that the first requirement above is always satisfied in this case.

Under these assumptions we have the following.

Proposition 3.2.1

The objective function f in (3.30) is $C^1(\mathbb{R}^d)$, strongly convex, and such that ∇f is Lipschitz continuous with constant $L > 0$. Moreover, assume that Assumption 3.2.1 holds.
Then, the n-th iterate of the stocastich gradient descent satisfies

$$\mathbb{E}_{\omega_n}[f(x_{n+1})] \leq f(x_n) - \alpha_n \left(1 - \frac{L\alpha_n}{2}\right) \|\nabla f(x_n)\|_2^2 + \frac{L\alpha_n}{2}\sigma^2. \tag{3.38}$$

Proof: Fix $n \in \mathbb{N}$. Then, by the quadratic upper bound of Theorem 3.1.2, we have

$$f(x_{n+1}) \leq f(x_n) + \langle \nabla f(x_n), x_{n+1} - x_n \rangle + \frac{L}{2}\|x_{n+1} - x_n\|_2^2 = f(x_n) - \alpha_k \langle \nabla f(x_n), \nabla f(x_n, \omega_n) \rangle + \frac{L\alpha_n^2}{2}\|\nabla f(x_{n+1}, \omega_n)\|_2^2 \tag{3.39}$$

Taking expectation w.r.t. ω_n (recall that here x_{n+1} depends on ω_n, while x_n does not) and applying Assumption 3.2.1 yields the statement. ∎

Remark: Due to the independence of the ω_n and the fact that x_n is independent of ω_k for $k \geq n$, it holds

$$\mathbb{E}(f(x_n)) = \mathbb{E}_{\omega_0}\left[\mathbb{E}_{\omega_1}\left[\dots \mathbb{E}_{\omega_{n-1}}\left[f(x_n)\right]\right]\right] \tag{3.40}$$

Note that this quantity will be deterministic (fixed) with respect to every ω_k with $k \geq n$.

We then have the following.

Theorem 3.2.1 Stochastic gradient with constant stepsize

Assume that the objective function f in (3.30) is $C^1(\mathbb{R}^d)$, γ-strongly convex, and such that ∇f is Lipschitz continuous with constant $L > 0$. Moreover, assume that Assumption 3.2.1 holds.
Then, if $\alpha_n = \alpha$ with $0 < \alpha \leq 1/L$, the n-th iterate of the stocastich gradient descent satisfies

$$\mathbb{E}\left[f(x_n) - f(x^\star)\right] \leq \frac{\alpha L \sigma^2}{2\gamma} + (1 - \alpha\gamma)^n \left[f(x_0) - f(x^\star) - \frac{\alpha L \sigma^2}{2\gamma}\right]. \tag{3.41}$$

Remark: It follows from the above that for any $\varepsilon > 0$ there exist α and n_0 such that

$$\mathbb{E}\left[f(x_n) - f(x^\star)\right] \leq \varepsilon \qquad \text{if } n \geq n_0. \tag{3.42}$$

However, in general

$$\lim_{n \to +\infty} \mathbb{E}\left[f(x_n) - f(x^\star)\right] \neq 0. \tag{3.43}$$

Proof: Fix $n \in \mathbb{N}$. By Proposition 3.2.1 and Theorem 3.1.2, we have

$$\mathbb{E}_{\omega_n}\left[f(x_{n+1})\right] - f(x_n) \leq \alpha\left(1 - \frac{L\alpha}{2}\right)\|\nabla f(x_n)\|_2^2 + \frac{L\alpha}{2}\sigma^2 \leq 2L\alpha\left(1 - \frac{L\alpha}{2}\right)\left(f(x_n) - f(x^\star)\right) + \frac{L\alpha}{2}\sigma^2 \tag{3.44}$$

Observe that $\mathbb{E}_{\omega_n}[f(x_n)] = f(x_n)$ and $\mathbb{E}_{\omega_n}[f(x^\star)] = f(x^\star)$. Hence,

$$\mathbb{E}_{\omega_n}\left[f(x_{n+1})\right] - f(x_n) = \mathbb{E}_{\omega_n}\left[f(x_{n+1}) - f(x^\star)\right] - \left[f(x_n) - f(x^\star)\right]. \tag{3.45}$$

Plugging this into the preceding equation, by simple manipulations, similar to those for the deterministic case (see Theorem 3.1.3), we obtain

$$\mathbb{E}_{\omega_n}\left[f(x_{n+1}) - f(x^\star)\right] - \frac{L\alpha}{2\gamma}\sigma^2 \leq (1 - \gamma\alpha)\left[f(x_n) - f(x^\star) - \frac{L\alpha}{2\gamma}\sigma^2\right] \tag{3.46}$$

Taking the expected value with respect to $\omega_{n-1}, \omega_{n-2}, \ldots, \omega_0$ of the above yield

$$\mathbb{E}\left[f(x_{n+1}) - f(x^\star)\right] - \frac{L\alpha}{2\gamma}\sigma^2 \leq (1 - \gamma\alpha)\left[\mathbb{E}\left[f(x_n) - f(x^\star)\right] - \frac{L\alpha}{2\gamma}\sigma^2\right] \tag{3.47}$$

Finally, recursively applying the above allows to prove the statement. ■

We present a simple result concerning variable stepsize, showing that a correct choice allows to have

$$\lim_{n \to +\infty} \mathbb{E}\left[f(x_n) - f(x^\star)\right] = 0. \tag{3.48}$$

Theorem 3.2.2 Stochastic gradient with variable stepsize

Assume that the objective function f in (3.30) is $C^1(\mathbb{R}^d)$, γ-strongly convex, and such that ∇f is Lipschitz continuous with constant $L > 0$. Moreover, assume that Assumption 3.2.1 holds.
Consider the variable stepsize

$$\alpha_n = \frac{\beta}{\delta + n}, \qquad \forall n \in \mathbb{N}, \tag{3.49}$$

where $\beta \geq 1/\gamma$ and $\delta > 0$ is such that $\alpha_0 = \beta/\delta < 1/L$. Then, the n-th iterate of the stocastich gradient descent satisfies

$$\mathbb{E}\left[f(x_n) - f(x^\star)\right] \leq \frac{\nu}{\delta + n}, \qquad \text{where} \qquad \nu = \max\left\{\delta\left[f(x_0) - f(x^\star)\right], \frac{\beta^2 L\sigma^2}{2(\beta\gamma - 1)}\right\}. \tag{3.50}$$

Proof: Proceeding as in the proof of the constant stepsize case (see Theorem 3.2.1), we arrive at (3.47) with α_n instead of α, that we rearrange as

$$\mathbb{E}\left[f(x_{n+1}) - f(x^\star)\right] \leq (1 - \gamma\alpha_n)\mathbb{E}\left[f(x_n) - f(x^\star)\right] - \frac{L\alpha_n^2}{2}\sigma^2. \tag{3.51}$$

To complete the proof of the statement it suffices to apply a recurrence over $n \in \mathbb{N}$. ■

3.3 Proximal methods

Often times the objective function is not smooth, or at least not entirely smooth. Indeed, a common problem encountered in optimization is the following

$$\min_{x \in \mathbb{R}^d} f(x) + g(x), \tag{3.52}$$

where f is sufficiently smooth (say, C^1 with Lipschitz gradient), while g is not smooth but "simple".

Example 3.3.1

- Constrained optimization: for a set K, letting χ_K be its 0-∞ characteristic function (1.42), consider

$$\min_{x\in K} f(x) = \min_{x\in\mathbb{R}^d} f(x) + \chi_K(x). \tag{3.53}$$

- Lasso regression:

$$\min_{x\in\mathbb{R}^d} \|Ax-b\|_2^2 + \lambda\|x\|_1. \tag{3.54}$$

Here, we will use the proximal operator introduced in Section 1.5, by considering the following algorithm.

Definition 3.3.1 Proximal gradient method

The proximal gradient method for (3.52), with $f \in C^1(\mathbb{R}^d)$ and g convex, starting at $x_0 \in \mathbb{R}^d$ with parameters $(\alpha_n)_n \subset \mathbb{R}_+$, is

$$x_{n+1} = P_{\alpha_n g}\left[x_n - \alpha_n \nabla f(x_n)\right]. \tag{3.55}$$

Remark: Observe that if g is smooth, the above reduces to

$$x_{n+1} = x_n - \alpha_n\left(\nabla f(x_n) + \nabla g(x_{n+1})\right). \tag{3.56}$$

In particular, this becomes *implicit* (the right-hand side depends on x_{n+1}), and does not reduce to the gradient descent.

On the other hand, if $g = \chi_K$ for some convex set $K \subset \mathbb{R}^d$, the above reduces to the *projected gradient descent* (see exercises).

The idea of the above algorithm is the following.

1. As a first guess, do a gradient descent step on the smooth part :

$$y_{n+1} = x_n - \alpha_n \nabla f(x_n). \tag{3.57}$$

2. Replace this first guess with

$$x_{n+1} = P_{\alpha_n g}(y_n) = \arg\min_{x\in\mathbb{R}^d}\left[g(x) + \frac{1}{2\alpha_n}\|x-y_{n+1}\|_2^2\right]. \tag{3.58}$$

Notice that x_{n+1} is a point near y_{n+1} (due to the $\|x-y_{n+1}\|_2^2$ in the above minimization) that makes g small.

Let us prove convergence in the constant step-size case.

Theorem 3.3.1

Let $g : \mathbb{R}^d \to \mathbb{R}$ be convex and lower semicontinuous, and $f \in C^1(\mathbb{R}^d)$ be such that ∇f is Lipschitz with constant $L > 0$. Assume, moreover that (3.52) admits a unique minimizer $x^\star$. Then, the iteration of (3.55) with $\alpha_n = \alpha < 1/L$ converges to $x^\star$.
Moreover, the same conclusions as in Theorem **??** holds for the rate of convergence.

Proof: To be added. We refer to [2], for the moment.

Let $n \in \mathbb{N}$, and denote $x^+ = x_{n+1}$ and $x = x_n$. We will bound the two parts of the objective function separately. By definition of proximal operator, we have that for any z it holds

$$g(x^+) + \frac{1}{2\alpha}\|x^+ - (x-\alpha\nabla f(x))\|_2^2 \le g(z) + \frac{1}{2\alpha}\|z-(x-\alpha\nabla f(x))\|_2^2. \tag{3.59}$$

Pick $z = x$ to obtain

$$g(x^+) + \frac{1}{2\alpha}\|x^+ - x - \alpha\nabla f(x)\|_2^2 \le g(x) + \frac{\alpha}{2}\|\nabla f(x)\|_2^2. \tag{3.60}$$

Rearranging the terms and expanding the squared norm allows to obtain

$$g(x^+) - g(x) \leq \frac{\alpha}{2}\|\nabla f(x)\|_2^2 - \frac{1}{2\alpha}\|x^+ - x - \alpha\nabla f(x)\|_2^2 = -\frac{1}{2\alpha}\|x^+ - x\|_2^2 - \langle \nabla f(x), x^+ - x\rangle. \tag{3.61}$$

Recall that, by the quadratic upper bound of Theorem 3.1.2, we have

$$f(x^+) \leq f(x) + \langle \nabla f(x), x^+ - x\rangle + \frac{L}{2}\|x^+ - x\|_2^2. \tag{3.62}$$

Summing up with (3.61), this yields

$$f(x^+) + g(x^+) \leq f(x) + g(x) + \frac{1}{2}\left(L - \frac{1}{\alpha}\right)\|x^+ - x\|_2^2. \tag{3.63}$$

Since $c = L - \alpha^{-1} > 0$, one can then sum over n to obtain

$$\sum_{n=1}^{+\infty} \|x_{n+1} - x_n\|_2^2 \leq \lim_{n\to+\infty} \frac{f(x_0) - f(x_n)}{c} \leq \frac{f(x_n) - f(x^\star)}{c} < +\infty. \tag{3.64}$$

In particular, $\|x_{n+1} - x_n\|_2 \to 0$ as $n \to +\infty$.

Let now x_∞ be a cluster point of $(x_n)_n$, i.e., there exists $(x_{n_k})_k \subset (x_n)_n$ such that $x_{n_k} \to x_\infty$. Since $x_{n+1} = P_{\alpha g}(x_n - \alpha\nabla f(x_n))$, reinterpreting the optimality condition for the definition of proximal operator in subgradient form we have

$$-\nabla f(x_n) - \frac{1}{\alpha}\|x_{n+1} - x_n\|_2^2 \in \partial h(x_{n+1}),$$

and passing to a convergent subsequence implies that

$$0 \in \nabla f(x_\infty) + \partial g(x_\infty). \tag{3.65}$$

Here, we use that the graph of ∂g is closed due to the lower semicontinuity of g. But by Theorem 1.4.1, this implies that $x^\star = x_\infty$, completing the proof.

Let us now assume that f is strongly convex and twice differentiable (we refer to [2] for a proof in the convex case). In particular, $\gamma\,\mathrm{Id} \leq \mathrm{Hess}\, f(x) \leq L\,\mathrm{Id}$ by Proposition 1.3.2 and the L-Lipschitz condition on ∇f.

We use the non-expansiveness of the proximal operator (Theorem 1.5.2), to obtain

$$\|x_{n+1} - x_n\|_2 = \|P_{\alpha g}(x_n - \alpha\nabla f(x_n)) - P_{\alpha g}(x_{n-1} - \alpha\nabla f(x_{n-1}))\|_2 \leq \|x_n - x_{n-1} - \alpha(\nabla f(x_n) - \nabla f(x_{n-1}))\|_2. \tag{3.66}$$

Observe that we have

$$\nabla f(x_n) - \nabla f(x_{n-1}) = \int_0^1 \frac{d}{dt}\nabla f(tx_n + (1-t)x_{n-1})\,dt = \left[\int_0^1 \mathrm{Hess}\, f(tx_n + (1-t)x_{n-1})\,dt\right](x_n - x_{n-1}) =: M(x_n - x_{n-1}), \tag{3.67}$$

where $\gamma\,\mathrm{Id} \leq M \leq L\,\mathrm{Id}$. Continuing from (3.66), we obtain

$$\|x_{n+1} - x_n\|_2 = \|(\mathrm{Id} - \alpha M)(x_n - x_{n-1})\|_2 \leq \|M\|\|x_n - x_{n-1}\|_2. \tag{3.68}$$

Since $\|M\| = \max\{|1 - \alpha\gamma|, |1 - \alpha L|\} < 1$, a recursion argument completes the proof. ∎

Example 3.3.2 (Regression with ℓ_1 regularization (Lasso))

Consider the problem

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2 + \lambda\|x\|_1, \tag{3.69}$$

where $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^d$. The $\|x\|_1$ term in the objective promotes sparsity in the solution $x^\star$.
This fits the template (3.52) with

$$f(x) = \|Ax - b\|_2^2 \quad \text{and} \quad g(x) = \lambda\|x\|_1. \tag{3.70}$$

We saw in Example 1.5.2 that the proximal operator of g is the soft-thresholding operator.
The proximal gradient method applied to this problem is called the *iterative shrinkage-thresholding algorithm (ISTA)* and takes the form

$$x_{k+1} = S_{\alpha\lambda}\left(x_k - 2tA^\top(Ax_k - b)\right),$$

where $S_{\alpha\lambda}$ is the soft-thresholding operator (1.60) with parameter $\alpha\lambda$.

3.4 Dual methods

We saw in Chapter 2.1 that to any convex optimization problem is possible to associate a dual problem. In some cases, the dual problem has a structure that is more amenable to algorithms than the original, primal, problem. We explore the possibility of applying various optimization methods to the dual problem.

Similarly to the previous section, we focus on optimization problems in the form

$$\min_{x\in\mathbb{R}^d} f(x) + g(Ax). \tag{3.71}$$

Here, $A \in \mathbb{R}^{m\times d}$, while f and g are convex function, with f typically smooth, and g non-smooth.

If $P_{g\circ A}$ is simple to compute, this problem can be solved by applying the proximal gradient method of the previous section. However, we can encounter situations where, although g has a simple proximal operatr, the proximal operator of $g \circ A$ is very hard to compute. In this section we see that, however, considering the dual problem allows to circumvent this issue.

Example 3.4.1 (Signal denoising using total variation)
Consider the problem of denoising a 1d signal $u \in \mathbb{R}^d$ with total-variation regularization:

$$\min_{x\in\mathbb{R}^d} \|x-u\|_2^2 + \lambda\sum_{i=1}^{d-1}|x_{i+1}-x_i|. \tag{3.72}$$

This problem can be put in the form (3.71) with

$$f(x) = \|x-u\|_2^2, \qquad \text{and} \qquad g(x) = \|x\|_1, \tag{3.73}$$

and A the discrete difference operator

$$Ax = \begin{bmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ x_n - x_{d-1} \end{bmatrix}.$$

Observe that (3.71) is easily rewritten as

$$\min_{(x,y)\in\mathbb{R}^{d\times m}} f(x) + g(y) \qquad \text{subject to} \qquad y = Ax. \tag{3.74}$$

We have the following.

Proposition 3.4.1
The dual problem to (3.74) is

$$\max_{\xi\in\mathbb{R}^m} \left[-f^*(-A^\top\xi) - g^*(\xi)\right]. \tag{3.75}$$

Here, f^* and g^* are the convex conjugate functions defined in Section 1.4.

Proof: The Lagrange function is $L : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m$

$$L(x,y,\xi) = f(x) + h(y) + \langle\xi, Ax - y\rangle. \tag{3.76}$$

The dual function then reads

$$\begin{aligned} H(z) &= \min_{x,y} L(x,y,z) \\ &= \min_{x,y} \{f(x) + \langle z, Ax\rangle + g(y) - \langle z, y\rangle\} \\ &= \min_{x} \{f(x) + \langle z, Ax\rangle\} + \min_{y} \{g(y) - \langle z, y\rangle\} \\ &= -f^*(-A^\top z) - h^*(z). \end{aligned} \tag{3.77}$$

This completes the proof. ■

We see that the dual problem concerns now the function $f^* \circ (-A^\top)$, which inherits the good properties of f, and g^*, whose proximal operator is easy to compute. Indeed, thanks to Moreau's identity (Theorem 1.5.1) we have

$$P_{g^*}(y) = y - P_g(y), \qquad \forall y \, in \mathbb{R}^d. \tag{3.78}$$

We thus have the following.

Theorem 3.4.1 Dual proximal gradient

Let g be convex, and $f \in C^1(\mathbb{R}^d)$ be γ-strongly convex. Then, the proximal gradient method (3.55) applied to the dual problem (3.75) reads

$$\begin{cases} x_{n+1} = \arg\min_{x\in\mathbb{R}^d} \left[f(x) + \langle z_n, Ax\rangle\right], \\ y_{n+1} = \arg\min_{y\in\mathbb{R}^m} \left[g(y) - \langle z_n, y\rangle + \frac{\alpha_n}{2}\|Ax_n - y\|_2^2\right], \\ z_{n+1} = z_n + \alpha_n(Ax_n - y_n). \end{cases} \tag{3.79}$$

In particular, if (3.71) admits a unique minimizer $x^\star$, and $\alpha_n = \alpha < \gamma/\|A\|$, the iterates $(z_n)_n$ converge to $x^\star$.

Remark: For the signal denoising example, it is possible to derive a closed form expression for x_n and y_n.

Proof: Since f is γ-strongly convex, it is straightforward to show that $y \mapsto -f^*(A^\top y)$ is C^1 and has Lipschitz constant $\gamma/\|A\|$ (here, we consider $\|A\| = \sup_{x\neq 0} \|Ax\|_2/\|x\|_2$).

The proximal method applied to the dual reads

$$z_{n+1} = P_{\alpha_n g}\left[z_n + \alpha_n A \nabla f^*(-A^\top z_n)\right]. \tag{3.80}$$

It is possible to show that the strong convexity of f implies

$$\nabla f^*(v) = \arg\max_{x\in\mathbb{R}^d} \left[\langle v, x\rangle - f(x)\right] = \arg\min_{x\in\mathbb{R}^d} \left[f(x) - \langle v, x\rangle\right] \tag{3.81}$$

Thus, (3.80) read

$$\begin{cases} x_{n+1} = \arg\min_{x\in\mathbb{R}^d} \left[f(x) + \langle z_n, Ax\rangle\right], \\ z_{n+1} = P_{\alpha_n g}\left[z_n + \alpha_n A x_n\right]. \end{cases} \tag{3.82}$$

The statement follows by using Moreau's identity (Theorem 1.5.1) and the following simple equalities for the conjugate and the proximal operator:

$$(\alpha g)^*(y) = \alpha g(y/\alpha) \qquad \text{and} \qquad P_{\alpha g(\cdot/\alpha)}(x) = \alpha P_{\alpha^{-1}g}(x/\alpha), \qquad \forall \alpha > 0. \tag{3.83}$$

■

Part II

Control

Chapter 4

Controllability

For this chapter we follow [6].

4.1 Control systems

Definition 4.1.1 Control system

A control system in $\mathbb{R}^n$ is the differential equation

$$\dot{x} = f(t, x, u), \tag{4.1}$$

where

- The *state* is a the function $x : \mathbb{R} \to \mathbb{R}^n$;
- $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is of class C^1 w.r.t. $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$, and locally integrable w.r.t. $t \in \mathbb{R}$;
- The *control* $u : \mathbb{R} \times U$ is a measurable and essentially bounded function of time, taking values in $U \subset \mathbb{R}^m$.

The control system is

- *Linear* is $f(t, x, u) = A(t)x + B(t)u + r(t)$ for $A : \mathbb{R} \to \mathbb{R}^{n\times n}$, $B : \mathbb{R} \to \mathbb{R}^{n\times m}$, $r : \mathbb{R} \to \mathbb{R}^n$. In this case, we assume these functions to be of class L^∞ on every compact interval.
- *Autonomous* if $f(t, x, u) = f(x, u)$ is independent of time. Otherwise, the system is *instationary* or *time-varying*.

Once a control u and an initial condition $x_0 \in \mathbb{R}^n$ are fixed, the existence and uniqueness of solutions to the non-autonomous equation (4.1) is guaranteed by the following.

Theorem 4.1.1 Carathéodory Existence Theorem

Consider the Cauchy problem

$$\begin{cases} \dot{x} = f(t, x), \\ x(0) = x_0 \in \mathbb{R}^n. \end{cases} \tag{4.2}$$

Assume that $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ satisfies the following conditions

- $f(t, \cdot)$ is Lipschitz continuous for any $t \in \mathbb{R}$ with Lispchitz constant $L(t)$ that is locally integrable;
- $f(\cdot, x)$ is measurable for any $x \in \mathbb{R}^n$;
- there exists $r, M > 0$ such that $\|f(t, x)\|_2 \leq M$ for any $(t, x) \in (-r, r) \times B(0, r)$.

Then, there exists a unique solution to (4.2), maximally defined on some open interval $I \subset \mathbb{R}$ such that $0 \in \mathbb{R}$.

When considering the control system on an interval $[0,T]$, we need its solutions to not blow up before the time T.

Definition 4.1.2 Admissible controls

Let $x_0 \in \mathbb{R}^n$ and $T > 0$. A control $u \in L^\infty([0,T],U)$ is *admissible* on $[0,T]$ at x_0 if the associated trajectory x_u of (4.1) such that $x_u(0) = x_0$ is well-defined on $[0,T]$. We let

$$\mathcal{U}_{x_0,T} = \{u \in L^\infty([0,T],U) \mid u \text{ is admissible}\}. \tag{4.3}$$

Definition 4.1.3 Controllability

The *end-point mapping* $\mathrm{End}_{x_0,T}$ is defined by

$$\mathrm{End}_{x_0,T} : \mathcal{U}_{x_0,T} \to \mathbb{R}^n, \qquad \mathrm{End}_{x_0,T}(u) = x_u(T). \tag{4.4}$$

The *reachable (or accessible) set* from x_0 in time $T > 0$ is

$$\mathrm{Reach}_{x_0,T} = \mathrm{End}_{x_0,T}(\mathcal{U}_{x_0,T}). \tag{4.5}$$

The system (4.1) is

- *Globally controllable* from x_0 in time $T > 0$ if $\mathrm{End}_{x_0,T}$ is surjective, that is,

$$\mathrm{Reach}_{x_0,T} = \mathbb{R}^n. \tag{4.6}$$

- *Locally controllable* from x_0 in time $T > 0$ around $x_1 \in \mathrm{Reach}_{x_0,T}$ if x_1 is in the interior of $\mathrm{Reach}_{x_0,T}$. That is, if $\mathrm{End}_{x_0,T}$ is locally surjective near x_1.

4.2 Controllability of linear autonomous systems

In this section we consider the linear autonomous control system

$$\dot{x} = Ax + Bu. \tag{4.7}$$

It is possible to show that there is no blow-up in finite time for linear systems, and thus $\mathcal{U}_{x_0,T} = L^\infty([0,T],U)$.

An essential tool for the study of these systems is the variation of constants formula (or Duhamel formula) for its solutions:

$$x_u(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu(s)\,ds, \qquad \forall t \in [0,T],\ u \in L^\infty([0,T],U). \tag{4.8}$$

We also need to recall the following celebrated result.

Theorem 4.2.1 Cayley-Hamilton Theorem

Let $A \in \mathbb{R}^{n\times n}$ be a matrix with characteristic polynomial

$$\chi_A(z) = \det(z\,\mathrm{Id} - A) = z^n + a_1 z^{n-1} + \ldots + a_{n-1}z + a_n. \tag{4.9}$$

Then, letting $p_A(A)$ be the number obtained by replacing the unknown z with the matrix A in this expression, we have $p_A(A) = 0$. That it,

$$A^n + a_1 A^{n-1} + \ldots + a_{n-1}A + a_n = 0. \tag{4.10}$$

In the case of linear systems, it turns out that controllability can be verified via a purely algebraic condition.

Definition 4.2.1 Kalman rank condition

We say that the pair $(A,B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ satisfies the *Kalman rank condition* if the Kalman matrix

$$K = [B, AB, \ldots, A^{n-1}B] \in \mathbb{R}^{n\times nm}, \tag{4.11}$$

is of maximal rank n.

Theorem 4.2.2 Kalman Theorem

Assume that $U = \mathbb{R}^n$. Then, (4.7) is controllable from $x_0 \in \mathbb{R}^n$ and in time $T > 0$ if and only if (A, B) satisfies the Kalman rank condition.
In particular, if a linear system is controllable from x_0 in time $T > 0$, then it is controllable from any initial point and in any time.

Proof: By the variation of constants formula (4.8), we see that controllability is equivalent to the surjectivity of the linear operator

$$L : L^\infty([0,T],\mathbb{R}^m) \to \mathbb{R}^n, \qquad Lu = \int_0^T e^{-As} B u(s)\, ds. \tag{4.12}$$

Here, we used that the exponential matrix e^{TA} is always invertible.

Let us prove that the fact that L invertible implies rank $K = n$. We proceed by contradiction and assume that rank $K < n$. That is, there exists $p \in \mathbb{R}^n$, $p \neq 0$, such that

$$p^\top K = 0 \iff p^\top A^i B = 0, \qquad \forall i \in [\![1, n]\!]. \tag{4.13}$$

Recall that by Cayley-Hamilton Theorem, we can write A^j as a linear combination of $\mathrm{Id}, A, \dots, A^{n-1}$. That is, for any $j \in \mathbb{N}$, there exists $a_0, \dots, a_{n-1}$ such that

$$A^j = \sum_{i=0}^{n-1} a_i A^i. \tag{4.14}$$

This implies that (4.13) actually holds for any $i \in \mathbb{N}$, which yields

$$p^\top e^{-As} B = \sum_{j=0}^{+\infty} p^\top \frac{(-As)^j}{j!} B = 0. \tag{4.15}$$

In particular, this shows that $p^\top Lu = 0$ for any $u \in L^\infty([0,T],\mathbb{R}^m)$ proving that L is not surjective.

To prove the opposite implication, assume that there exists $p \in \mathbb{R}^n$, $p \neq 0$, such that

$$p^\top Lu = 0 \qquad \forall u \in L^\infty([0,T],\mathbb{R}^m). \tag{4.16}$$

Consider, for $i \in [\![1, n]\!]$ and $\tau \in [0, T]$, the control

$$u(\tau) = \begin{cases} e_i, & \text{if } t \in [0,\tau], \\ 0 & \text{otherwise.} \end{cases} \tag{4.17}$$

Here, e_i is the i-th element of the canonical basis of $\mathbb{R}^n$. Thus, we have

$$Lu = \int_0^\tau e^{-As} B u\, ds = \left[\frac{\mathrm{Id} - e^{-\tau A}}{A}\right] Bu, \qquad \text{where} \qquad \frac{\mathrm{Id} - e^{-\tau A}}{A} = \sum_{j=1}^{+\infty} \frac{(-1)^{j-1}\tau^j}{j!} A^{j-1}. \tag{4.18}$$

Assumption (4.16) then yields

$$0 = p^\top \left[\frac{\mathrm{Id} - e^{-\tau A}}{A}\right] Bu = \sum_{j=1}^{+\infty} \frac{(-1)^{j-1}\tau^j}{j!} p^\top A^{j-1} Bu, \qquad \forall \tau \in [0,T]. \tag{4.19}$$

By analyticity[1] w.r.t. τ of the right-hand side, this implies that $p^\top A^{j-1} Bu = 0$, that is rank $K < n$. ∎

[1] Equivalently, one can observe that

$$0 = \frac{d^k}{d\tau^k}\left[\sum_{j=1}^{+\infty} \frac{(-1)^{j-1}\tau^j}{j!} p^\top A^{j-1} Bu\right]_{\tau=0} = p^\top A^{k-1} Bu, \qquad \forall k \in \mathbb{N}.$$

Corollary 4.2.3 Controllability with control constraints

Assume that $0 \in \operatorname{int}(U)$, and that the Kalman condition holds true. Then, the control system is locally controllable around $e^{TA}x_0$ for any $x_0 \in \mathbb{R}^n$ and any $T > 0$. Namely,

$$e^{TA}x_0 \in \operatorname{int} \operatorname{Reach}(x_0, T). \tag{4.20}$$

Proof: By the variation of constant formula (4.8), we have (for the unconstrained system)

$$\operatorname{End}_{x_0,T}(u) = e^{TA}x_0 + Lu, \tag{4.21}$$

where L is the operator defined in (4.12). Observe that L is a continuous linear map, since

$$\|Lu\|_2 \leq T \max_{s\in[0,T]} \|e^{-As}B\| \, \|u\|_\infty. \tag{4.22}$$

In particular, L is an open mapping and hence for any neighborhood $V \subset U$ of the origin, we have that $\operatorname{End}_{x_0,T}(V)$ is a neighborhood of $\operatorname{End}_{x_0,T}(0) = e^{TA}x_0 \in \operatorname{Reach}(x_0, T)$. ∎

Theorem 4.2.4 Hautus Test

The following assertions are equivalent

1. The pair (A, B) satisfies the Kalman rank condition.
2. $\operatorname{rank}[\lambda \operatorname{Id} - A, B] = n$ for any $\lambda \in \mathbb{C}$.
3. $\operatorname{rank}[\lambda \operatorname{Id} - A, B] = n$ for any $\lambda \in \operatorname{spec} A$.
4. For any eigenvector z of $A^\top$, we have $B^\top z \neq 0$.
5. There exists $c > 0$ such that

$$\|(\lambda \operatorname{Id} - A^\top)z\|_2^2 + \|B^\top z\|_2^2 \geq c\|z\|_2^2, \qquad \forall z \in \mathbb{R}^n, \forall \lambda \in \mathbb{C}. \tag{4.23}$$

Proof: We start by showing the equivalence of assertions 2 to 5.

2 $\iff$ 3 Since $\operatorname{spec} A \subset \mathbb{C}$, we have that assertion 2 implies assertion 3. On the other hand, if assertion 3 holds we obtain assertion 2 by recalling that $\lambda \operatorname{Id} - A$ is invertible for any $\lambda \in \mathbb{C} \setminus \operatorname{spec} A$.

3 $\iff$ 4 If assertion 4 does not hold for an eigenvector z associated to $\lambda \in \operatorname{spec} A$, we clearly have $z^\top(\lambda \operatorname{Id} - A) = z^\top B = 0$, which contradicts assertion 3. A similar reasoning shows the opposite implication.

2 $\iff$ 5 If assertion 2 does not hold, we contradict assertion 5 as above. To prove the other implication, let

$$M_\lambda = (\bar{\lambda} \operatorname{Id} - A)(\lambda \operatorname{Id} - A^\top) + BB^\top. \tag{4.24}$$

The matrix M_λ is symmetric and it holds

$$\|(\lambda \operatorname{Id} - A^\top)z\|_2^2 + \|B^\top z\|_2^2 \geq z^\top M_\lambda z \qquad \forall z \in \mathbb{R}^n. \tag{4.25}$$

Hence, letting $\mu(\lambda)$ be the smallest eigenvalue of M_λ, we have assertion 5 with $c = \inf_{\lambda\in\mathbb{C}} \mu(\lambda)$. We have that $c > 0$ since $\lambda \mapsto \mu(\lambda)$ is continuous and $\mu(\lambda) \to +\infty$ as $|\lambda| \to +\infty$.

1 $\iff$ 4 We are left to showing that assertion 1 is equivalent to the other equivalent assertions. It is immediate to observe that if assertion 4 does not hold, the same is true for assertion 1. To show the opposite implication, set

$$N = \{z \in \mathbb{R}^n \mid z^\top A^k B = 0 \, \forall k \in \mathbb{N}\}. \tag{4.26}$$

In particular, $N = \{0\}$ if and only if (A, B) satisfies the Kalman rank condition. Assume this is not the case. Then, since N is non-trivial $A^\top$ invariant (i.e., $A^\top N \subset N$) subspace, we have that $A^\top$ has at least one non-zero eigenvalue $z \in N \setminus \{0\}$. But then, $B^\top z = 0$ by definition of N, contradicting assertion 4. ∎

4.2.1 Similar systems and normal forms

In this section we look at what happens if we perform a change of basis $x_2 = Px_1$ for some $P \in \mathrm{GL}_2(\mathbb{R})$.

Definition 4.2.2 Similar systems

The linear control systems

$$\dot{x}_1 = A_1x_1 + B_1u_1 \qquad \text{and} \qquad \dot{x}_2 = A_2x_2 + B_2u_2, \tag{4.27}$$

are *similar* if there exists $P \in \mathrm{GL}_2(\mathbb{R})$ such that $A_2 = PA_1P^{-1}$ and $B_2 = PB_1$. In this case we say that the pairs (A_1, B_1) and (A_2, B_2) are similar.

Observe that the Kalman property is intrinsic, i.e., is invariant under the similarity transformation P. Indeed, letting K_1 and K_2 be the Kalman matrices associated to two similar systems, we have $K_2 = PK_1$.

An important application of similar systems is the existence of various normal forms, i.e., a change of coordinates (and sometimes a change of inputs) that transforms a nonlinear or linear control system into a simpler, standardized structure, where its controllability, observability, or stabilization properties become explicit.

The following result goes in that direction, and can be seen as an extension of Kalaman Theorem (Theorem 4.2.2) to non-controllable systems.

Proposition 4.2.1

Consider a linear system (4.7) whose Kalman matrix K satisfies $\operatorname{rank} K = r$. Then, letting $y = (y_1, y_2)^\top \in \mathbb{R}^{r\times(n-r)}$, the system is similar to

$$\dot{y}_1 = A_1'y_1 + B_1u + A_3'y_2 \tag{4.28}$$
$$\dot{y}_2 = A_2'y_2. \tag{4.29}$$

In particular, this splits the original system in a controllable part (the variable y_1) and an uncontrollable one (the variable y_2).

Proof: Let us assume that $\operatorname{rank} K < n$, otherwise there is nothing to prove. Consider the subspace $F = \operatorname{Ran} K$, and observe that it holds

$$F = \operatorname{Ran} B + \operatorname{Ran} AB + \ldots + \operatorname{Ran} A^{n-1}B. \tag{4.30}$$

Then, $\dim F = r$ and, using the Cayley-Hamilton Theorem is straightforward to verify that F is invariant under A (i.e., $AF \subset F$). Hence, $\mathbb{R}^n = F \oplus G$ for some subspace G such that $\dim G = n - r$. Pick a basis $(f_1, \ldots, f_r)$ of F, and a basis $(f_{r+1}, \ldots, f_n)$ of G, and let P be the matrix encoding the change of basis from $(f_1, \ldots, f_n)$ to the canonical basis of $\mathbb{R}^n$.

Using the invariance of F w.r.t. A, we obtain that

$$A' = PAP^{-1} = \begin{pmatrix} A_1' & A_3' \\ 0 & A_2' \end{pmatrix}, \tag{4.31}$$

where $A_1' \in \mathbb{R}^{r\times r}$. Moreover, since $\operatorname{Ran} B \subset F$, we have that

$$B' = PB = \begin{pmatrix} B_1' \\ 0 \end{pmatrix}. \tag{4.32}$$

■

Theorem 4.2.5 Brunovski normal form, single-input case

Consider the linear system (4.7) with scalar input (i.e., $m = 1$ so that $B \in \mathbb{R}^{n\times 1}$), and assume that (A, B) satisfies the Kalman rank condition. Then, letting the characteristic polynomial of A be

$$\chi_A(z) = \det(z\,\mathrm{Id} - A) = z^n + a_1z^{n-1} + \ldots + a_{n-1}z + a_n, \tag{4.33}$$

the control system is similar to the following chained form

$$\begin{cases} \dot{x}_1 = x_2 \\ \quad \vdots \\ \dot{x}_{n-1} = x_n \\ \dot{x}_n = -a_n x_1 - a_{n-1} x_2 - \cdots - a_1 x_n + u. \end{cases} \tag{4.34}$$

Proof: It suffices to show that the pair (A, B) is similar to $(\tilde{A}, \tilde{B})$ given by

$$\tilde{A} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{pmatrix} \quad \text{and} \quad \tilde{B} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{4.35}$$

Let us define the vectors

$$v_n = B,\ v_{n-1} = AB,\ \ldots,\ v_1 = A^{n-1}B. \tag{4.36}$$

These form a basis, since the Kalman matrix $K = [v_n, \ldots, v_1]$ is full rank. By definition of v_n it is immediate to observe that B transforms to $\tilde{B}$ under the change of basis defined by $\{v_1, \ldots, v_n\}$.

Let us check that this is true also for $\tilde{A}$ with respect to A. By definition, it trivially holds that

$$Av_j = A\left(A^{n-j}B\right) = A^{n-(j-1)}B = v_{j-1}, \qquad \forall j \in [\![2, n]\!]. \tag{4.37}$$

In other words, in the basis $\{v_1, \ldots, v_n\}$ the matrix A acts as $\tilde{A}$ on the last $n-1$ coordinates. To compute Av_1, we apply Cayley-Hamilton Theorem (Theorem 4.2.1) to obtain

$$Av_1 = A^n B = (-a_1 A^{n-1} - \ldots - a_{n-1}A - a_n)B = -a_1 v_1 - \ldots - a_n v_n. \tag{4.38}$$

This shows that indeed $\tilde{A}$ corresponds to the matrix A under the change of basis $\{v_1, \ldots, v_n\}$. ∎

In the general case $m > 1$, the system decomposes into m *controllability chains* (also called *Jordan chains of the couple* (A, B)). This, however, requires to perform also a linear change of input. More precisely, we have the following.

Theorem 4.2.6 Brunovski normal form, general case

Consider the linear system (4.7) and assume that (A, B) satisfies the Kalman rank condition. Then there exist invertible matrices $P \in \mathbb{R}^{n\times n}$ (change of state coordinates) and $R \in \mathbb{R}^{m\times m}$ (change of input coordinates) such that, under the transformations

$$x = P\tilde{x}, \qquad u = R\tilde{u},$$

the system becomes

$$\dot{\tilde{x}} = A_c\tilde{x} + B_c\tilde{u}, \qquad \text{with} \qquad A_c = T^{-1}AT, \qquad B_c = T^{-1}BR.$$

Here (A_c, B_c) has the block-diagonal *Brunovský form*

$$A_c = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_m \end{bmatrix}, \qquad B_c = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_m \end{bmatrix},$$

where each block (A_i, B_i) is a single-input Brunovský block of size r_i, as in Theorem 4.2.5. The integers $r_1, \ldots, r_m$ are the *controllability indices* of (A, B) and satisfy

$$r_1 + \cdots + r_m = n.$$

4.3 Controllability of time-varying linear systems

We now turn to time-varying control systems

$$\dot{x} = A(t)x + B(t)u, \tag{4.39}$$

where $A : \mathbb{R} \to \mathbb{R}^{n\times n}$ and $B : \mathbb{R} \to \mathbb{R}^{n\times m}$.

Definition 4.3.1 State-transition matrix

The *state-transition matrix* $R : \mathbb{R}\times\mathbb{R} \to \mathbb{R}^{n\times n}$ of sytem $\dot{x} = A(t)x$ is the unique solution of

$$\frac{\partial}{\partial t}R(t,s) = A(t)R(t,s), \qquad R(s,s) = \mathrm{Id}\,. \tag{4.40}$$

We have the following standard result.

Proposition 4.3.1

Let R be the state-transition matrix of $\dot{x} = A(t)x$. We have

- In the autonomous case (i.e., $A(t) \equiv A$), we have $R(t,s) = e^{(t-s)A}$.
- Semigroup property: It holds
$$R(t,s)R(s,\tau) = R(t,\tau) \qquad \forall t,s,\tau \in \mathbb{R}. \tag{4.41}$$
In particular, $R(t,s)^{-1} = R(s,t)$.
- Solutions to (4.39): For any $x_0 \in \mathbb{R}^n$, $T > 0$, and $u \in \mathcal{U}_{x_0,T}$, we have
$$x(t) = R(t,0)x_0 + \int_0^t R(t,s)Bu(s)\,ds. \tag{4.42}$$

Due to the time-varying nature of the system, a reasonable generalization of Kalman rank condition would be that the Kalman matrix $K(t)$ of $(A(t), B(t))$ be full rank at each time $t > 0$. This is however too strong, as the following shows.

Example 4.3.1

Consider the time-varying linear system with $n = 2$ and $m = 1$:

$$\dot{x} = B(t)u(t), \qquad B(t) = \begin{cases} (1,0)^\top & \text{if } t \in [0,1], \\ (0,1)^\top & \text{if } t > 1. \end{cases} \tag{4.43}$$

Since $A(t) = 0$ for all times, the instananeous Kalman matrix $K(t)$ is not full-rank. However, it is straightforward to explicitly show that $\mathrm{Reach}(x_0, T) = \mathbb{R}^2$ for all $T > 1$.

Indeed, in the time-varying case, the instantaneous lack of controllability for certain directions at a time t_0 is not an issue if at later times these directions are controllable. To formalize this idea, we introduce the following.

Definition 4.3.2 Controllability Gramian

The *Gramian matrix* of system (4.39) at time $T > 0$ is the matrix

$$G_T := \int_0^T R(T,t)B(t)B(t)^\top R(T,t)^\top\,dt \in \mathbb{R}^{n\times n}. \tag{4.44}$$

Before proving the controllability theorem, let us make the following observation.

Proposition 4.3.2 Observability inequality

Let $T > 0$. The Gramian matrix is a symmetric non-negative matrix, whose invertibility is equivalent to the

following *observability inequality*: There exists $C_T > 0$ such that

$$\int_0^T \|B(t)^\top R(T,t)^\top \psi\|_2^2 \, dt \geq C_T \|\psi\|_2^2 \qquad \forall \psi \in \mathbb{R}^n. \tag{4.45}$$

Proof: The symmetry of G_T is immediate from the definition. Moreover,

$$\psi^\top G_T \psi = \int_0^T \psi^\top R(T,t) B(t) B(t)^\top R(T,t)^\top \psi \, dt = \int_0^T \|B(t)^\top R(T,t)^\top \psi\|_2^2 \, dt \geq 0. \tag{4.46}$$

This proves both the non-negativity and the equivalence between the invertibility of G_T and (4.45). ∎

Remark: Inequality (4.45) is called an observability inequality for the following reason. Consider the *adjoint system* to (4.39), which is

$$\dot{z} = -A(t)^\top z, \qquad z(T) = \psi, \tag{4.47}$$

and assume that the output $y(t) = B(t)^\top z(t)$ is measured. In particular, the energy of this output over $[0, T]$ is the quantity

$$E(T) = \int_0^T \|y(t)\|_2^2 \, dt. \tag{4.48}$$

But, since $y(t) = B(t)^\top R(T,t)^\top \psi$, this coincide with the left-hand side of (4.45), which can then be recast as

$$\int_0^T \|y(t)\|_2^2 \, dt \geq C_T \|z(T)\|_2^2. \tag{4.49}$$

Namely, the output $y(t)$ controls the size of the final state $z(T)$. In particular, if $y(t) \neq 0$ for all $t \in [0, T]$ we are sure that $z(T) = 0$. More generally, one can show that this inequality allows to reconstruct the state $z(T)$ from the measurements $y : [0, T] \to \mathbb{R}$. This property is called *observability of the adjoint system.*

The following theorem shows that the observability property introduced above is actually equivalent to the controllability of the original system.

Theorem 4.3.1

Assume that $U = \mathbb{R}^m$. Then, the control system (4.39) is controllable from $x_0 \in \mathbb{R}^n$ in time $T > 0$ if and only if the Gramian matrix G_T is invertible. In particular, if a linear time-varying system is controllable from x_0 in times $T > 0$, then it is controllable for any time $T' > T$ and from any initial point.

Proof: By Proposition 4.3.1, given a control u we have that

$$\mathrm{End}_{x_0,T}(u) = x_u(T) = x^\star + Lu, \qquad \text{where} \qquad x^\star = R(T,0)x_0, \quad \text{and} \quad Lu = \int_0^T R(T,t)B(t)u(t) \, dt \tag{4.50}$$

Assume that G_T be invertible and let us prove that $\mathrm{End}_{x_0,T}$ is surjective (i.e., that the system is controllable). Fix $x_1 \in \mathbb{R}^n$ and let us construct a control u of the form $u(t) = B(t)^\top R(t,T)^\top \psi$ for some $\psi \in \mathbb{R}^n$ such that $\mathrm{End}_{x_0,T}(u) = x_1$. Since this choice of u implies that $Lu = G_T \psi$, we have

$$\mathrm{End}_{x_0,T}(u) = x^\star + G_T \psi. \tag{4.51}$$

By invertibility of G_T it then suffices to choose $\psi = G_T^{-1}(x_1 - x^\star)$.

Conversely, let us assume that G_T is not invertible. By Proposition 4.3.2 we then have that there exists $\psi \in \mathbb{R}^n$, $\psi \neq 0$, such that

$$\int_0^T \|B(t)^\top R(T,t)^\top \psi\|_2^2 \, dt = 0 \implies B(t)^\top R(T,t)^\top \psi = 0 \qquad \text{for a.e. } t \in [0,T]. \tag{4.52}$$

It follows that

$$\psi^\top Lu = 0, \qquad \forall u \in L^\infty([0,T], \mathbb{R}^m), \tag{4.53}$$

where L is defined in (4.50). Hence,

$$\psi^\top \operatorname{End}_{x_0,T}(u) = \psi^\top(x^\star + Lu) = \psi^\top x^\star, \qquad \forall u \in L^\infty([0,T],\mathbb{R}^m). \tag{4.54}$$

In particular, the range of $\operatorname{End}_{x_0,T}$ is contained in a proper affine subspace of $\mathbb{R}^n$:

$$\operatorname{ran} \operatorname{End}_{x_0,T} \subset \{z \in \mathbb{R}^n \mid \psi^\top(z - x^\star) = 0\}. \tag{4.55}$$

This contradicts the surjectivity of $\operatorname{End}_{x_0,T}$. ■

The same argument used to derive Corollary 4.2.3 can be used to derive the following.

Corollary 4.3.2 Controllability with control constraints

Assume that $0 \in \operatorname{int}(U)$ and that the Gramian matrix G_T is invertible for some $T > 0$. Then, the control system is locally controllable around $x_0 \in \mathbb{R}^n$ in time $T > 0$.

We conclude this section by stating a proper generalization of the Kalman rank condition to the time-varying case.

Theorem 4.3.3

Consider (4.39) with $U = \mathbb{R}^m$ and such that $t \mapsto A(t)$ and $t \mapsto B(t)$ are of class C^∞. Define the sequence of matrices

$$B_0(t) = B(t), \qquad B_{j+1} = A(t)B_j(t) - \frac{d}{dt}B_j(t), \qquad j \in \mathbb{N}. \tag{4.56}$$

Then, the system is controllable if there exists $t_0 \in [0,T]$ such that

$$\bigcup_{j\in\mathbb{N}} \operatorname{ran} B_j(t_0) = \mathbb{R}^n. \tag{4.57}$$

If, moreover, $t \mapsto A(t)$ and $t \mapsto B(t)$ are analytic, then the above property is equivalent to controllability and independent of $t_0 \in [0,T]$.

Bibliography

[1] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Version 29. Cambridge New York Melbourne New Delhi Singapore: Cambridge University Press, 2023. 716 pp. ISBN: 978-0-521-83378-3. URL: https://web.stanford.edu/boyd/cvxbook/bvcvxbook.pdf.

[2] Hamza Fawzi. *Lecture Notes for Topics in Convex Optimisation*. URL: https://www.damtp.cam.ac.uk/user/hf323/L22-III-OPT/.

[3] Massimo Fornasier. "Foundations of Data Analysis".

[4] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton: Princeton University Press, 2015. 470 pp. ISBN: 978-0-691-01586-6.

[5] Clement W Royer. "Lecture Notes on Stochastic Gradient Methods". In: (). URL: https://www.lamsade.dauphine.fr/croyer/ensdocs/SG/LectureNotesOML-SG.pdf.

[6] Emmanuel Trélat. *Control in Finite and Infinite Dimension*. SpringerBriefs on PDEs and Data Science. Singapore: Springer Nature Singapore, 2024. ISBN: 978-981-97-5947-7 978-981-97-5948-4. DOI: 10.1007/978-981-97-5948-4. URL: https://link.springer.com/10.1007/978-981-97-5948-4 (visited on 10/06/2025).