Lecture Notes 2025

Optimisation, Control, and Data

Dario Prandi
dario.prandi@centralesupelec.fr

November 17, 2025

Contents

Part I

Optimisation

Chapter 1

Convex Analysis

This chapter closely follows chapter 5 of the lecture notes [5]. For most of the proof in this chapter we refer to [2] or [12].

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 $\lambda \in \mathbb{R}$,

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

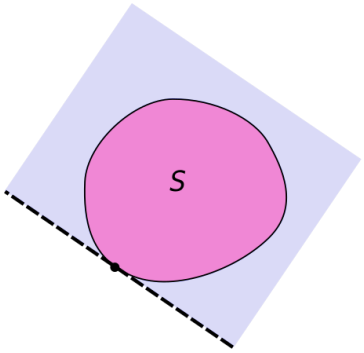

Figure 1.1: Supporting hyperplane for a set S.

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
$$F(x) = \frac{\|x\|_2^2}{2} = \frac{1}{2}\sum_{i=1}^{d} |x_i|^2. \tag{1.20}$$

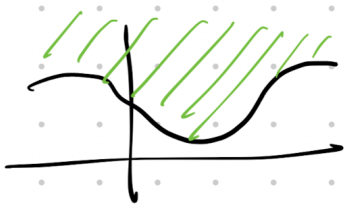

Figure 1.3: Epigraph of a function.

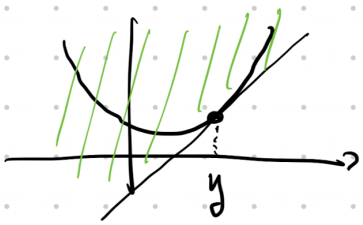

Figure 1.4: Graphical representation of Proposition 1.3.2.

- 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

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

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 $0 \in \partial F(x)$

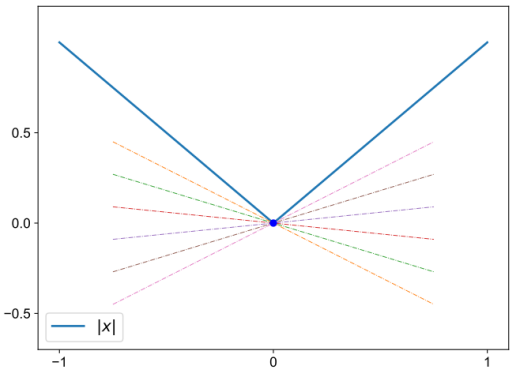

Figure 1.5: Visualization of the subgradients of $F(x) = |x|$ at $x = 0$. Image from this website.

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.

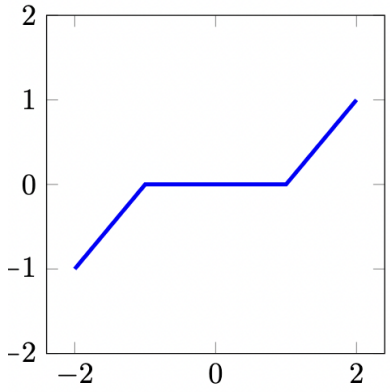

Figure 1.6: Soft-thresholding function.

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

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

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

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

Proposition 2.1.1

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

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

Proof: Observe that $-H$ is the supremum w.r.t. $x \in \mathbb{R}^d$ of the functions $g_x(\xi, v) = -L(x, \xi, v)$. 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 $v_j \geq 0$ for any $j \in [\![1, M]\!]$, we have

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

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

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

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

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

- A pair $(\xi, v) \in \mathbb{R}^m \times \mathbb{R}^M_+$ 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 $(\xi^\star v^\star) \in \mathbb{R}^m \times \mathbb{R}^M_+$.
- A *primal-dual optimal* is a triple $(x^\star, \xi^\star, v^\star)$ where $x^\star$ is a minimizer for (OP) and $(\xi^\star, v^\star)$ 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 [2, 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.

2.1.1 Geometric interpretation

Let us follow [2, 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.

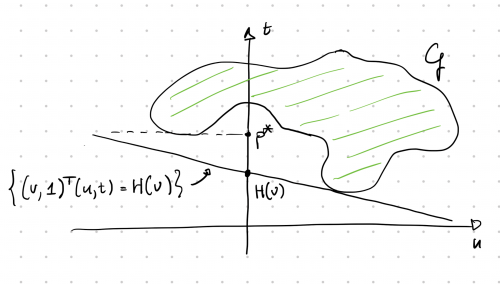

Figure 2.1: Geometric interpretation. The value of the dual function $H(v)$ identifies a supporting hyperplane for the set $\mathcal{G}$.

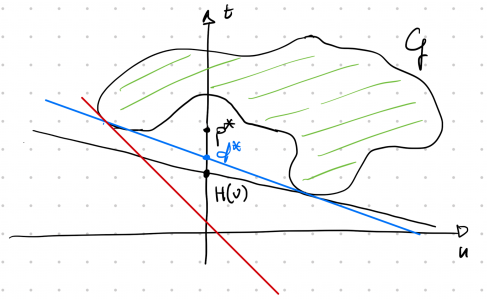

Figure 2.2: Geometric interpretation. Solving the dual problem yields the blue hyperplane. In this case $p^\star > d^\star$ and strong duality does not hold.

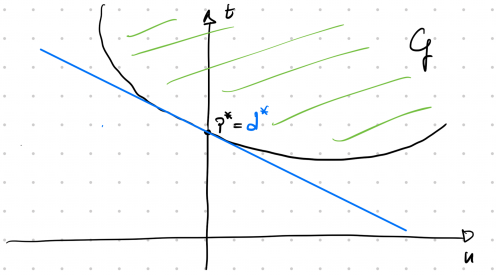

Figure 2.3: Geometric interpretation of Slater's condition. When the set $\mathcal{G}$ is convex and has interior that intersects the left-hand plane, the best supporting hyperplane yields the optimal value $p^\star$.

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., [2, 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 [13].

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 [4], 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 [4] 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 [16].

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.

Example 4.1.1 (Trivial control system)

Consider the control system, for $x \in \mathbb{R}^d$,

$$\dot{x} = u, \qquad u \in L^\infty([0,T],\mathbb{R}^d). \tag{4.7}$$

Then,

$$\mathrm{End}_{x_0,T}(u) = x_0 + \int_0^T u(s)\,ds. \tag{4.8}$$

In particular, the system is globally controllable from any $x_0 \in \mathbb{R}^d$ and in any time $T > 0$. Indeed, for any $x_1 \in \mathbb{R}^d$, the constant control

$$u(t) = \frac{x_1 - x_0}{T}, \qquad t \in [0,T], \tag{4.9}$$

steers the system from x_0 to x_1 in time T.

Example 4.1.2 (Not globally controllable system)

Consider the control system, for $x \in \mathbb{R}$,

$$\dot{x} = u, \qquad u \in L^\infty([0,T],[0,+\infty)). \tag{4.10}$$

Since $\dot{x} \geq 0$, the system cannot be globally controllable from $x_0 \in \mathbb{R}$ in any time $T > 0$. Indeed, proceeding as in Example 4.1.1 one shows that

$$\mathrm{Reach}_{x_0,T} = [x_0,+\infty), \qquad \forall x_0 \in \mathbb{R},\ T > 0. \tag{4.11}$$

In particular, the system is locally controllable around any $x_1 > x_0$.

4.2 Controllability of linear autonomous systems

In this section we consider the linear autonomous control system

$$\dot{x} = Ax + Bu. \tag{4.12}$$

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:

$$x_u(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu(s)\,ds, \qquad \forall t \in [0,T],\ u \in L^\infty([0,T],U). \tag{4.13}$$

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

$$\chi_A(z) = \det(z\,\mathrm{Id} - A) = z^n + a_1 z^{n-1} + \ldots + a_{n-1}z + a_n. \tag{4.14}$$

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,

$$A^n + a_1 A^{n-1} + \ldots + a_{n-1}A + a_n = 0. \tag{4.15}$$

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

$$K = [B, AB, \ldots, A^{n-1}B] \in \mathbb{R}^{n\times nm}, \tag{4.16}$$

is of maximal rank n.

Theorem 4.2.2 Kalman Theorem

Assume that $U = \mathbb{R}^n$. Then, (4.12) 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 (4.13), we see that controllability is equivalent to the surjectivity of the linear operator

$$L : L^\infty([0,T],\mathbb{R}^m) \to \mathbb{R}^n, \qquad Lu = \int_0^T e^{-As}Bu(s)\,ds. \tag{4.17}$$

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

$$p^\top K = 0 \iff p^\top A^i B = 0, \qquad \forall i \in [\![1,n]\!]. \tag{4.18}$$

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

$$A^j = \sum_{i=0}^{n-1} a_i A^i. \tag{4.19}$$

This implies that (4.18) actually holds for any $i \in \mathbb{N}$, which yields

$$p^\top e^{-As}B = \sum_{j=0}^{+\infty} p^\top \frac{(-As)^j}{j!}B = 0. \tag{4.20}$$

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

$$p^\top Lu = 0 \qquad \forall u \in L^\infty([0,T],\mathbb{R}^m). \tag{4.21}$$

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

$$u(\tau) = \begin{cases} e_i, & \text{if } t \in [0,\tau], \\ 0 & \text{otherwise.} \end{cases} \tag{4.22}$$

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

$$Lu = \int_0^\tau e^{-As} Bu\, ds = \left[\frac{\mathrm{Id} - e^{-\tau A}}{A}\right] Bu, \qquad \text{where} \qquad \frac{\mathrm{Id} - e^{-\tau A}}{A} = \sum_{j=1}^{+\infty} \frac{(-1)^{j-1}\tau^j}{j!} A^{j-1}. \tag{4.23}$$

Assumption (4.21) then yields

$$0 = p^\top \left[\frac{\mathrm{Id} - e^{-\tau A}}{A}\right] Bu = \sum_{j=1}^{+\infty} \frac{(-1)^{j-1}\tau^j}{j!} p^\top A^{j-1} Bu, \qquad \forall \tau \in [0,T]. \tag{4.24}$$

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

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,

$$e^{TA}x_0 \in \operatorname{int} \operatorname{Reach}(x_0, T). \tag{4.25}$$

Proof: By the variation of constant formula (4.13), we have (for the unconstrained system)

$$\operatorname{End}_{x_0,T}(u) = e^{TA}x_0 + Lu, \tag{4.26}$$

where L is the operator defined in (4.17). Observe that L is a continuous linear map, since

$$\|Lu\|_2 \leq T \max_{s\in[0,T]} \|e^{-As}B\|\, \|u\|_\infty. \tag{4.27}$$

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. $\operatorname{rank}[\lambda \mathrm{Id} - A, B] = n$ for any $\lambda \in \mathbb{C}$.
3. $\operatorname{rank}[\lambda \mathrm{Id} - A, B] = n$ for any $\lambda \in \operatorname{spec} A$.
4. For any eigenvector z of $A^\top$, we have $B^\top z \neq 0$.

[1] Equivalently, one can observe that

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

5. There exists $c > 0$ such that

$$\|(\lambda \operatorname{Id} - A^\top)z\|_2^2 + \|B^\top z\|_2^2 \geq c\|z\|_2^2, \qquad \forall z \in \mathbb{R}^n, \forall \lambda \in \mathbb{C}. \tag{4.28}$$

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

$$M_\lambda = (\bar{\lambda} \operatorname{Id} - A)(\lambda \operatorname{Id} - A^\top) + BB^\top. \tag{4.29}$$

The matrix M_λ is symmetric and it holds

$$\|(\lambda \operatorname{Id} - A^\top)z\|_2^2 + \|B^\top z\|_2^2 \geq z^\top M_\lambda z \qquad \forall z \in \mathbb{R}^n. \tag{4.30}$$

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

$$N = \{z \in \mathbb{R}^n \mid z^\top A^k B = 0 \, \forall k \in \mathbb{N}\}. \tag{4.31}$$

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

$$\dot{x}_1 = A_1 x_1 + B_1 u_1 \qquad \text{and} \qquad \dot{x}_2 = A_2 x_2 + B_2 u_2, \tag{4.32}$$

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 (4.12) 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

$$\dot{y}_1 = A_1' y_1 + B_1 u + A_3' y_2 \tag{4.33}$$

$$\dot{y}_2 = A_2' y_2. \tag{4.34}$$

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

$$F = \operatorname{Ran} B + \operatorname{Ran} AB + \ldots + \operatorname{Ran} A^{n-1}B. \tag{4.35}$$

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

$$A' = PAP^{-1} = \begin{pmatrix} A'_1 & A'_3 \\ 0 & A'_2 \end{pmatrix}, \tag{4.36}$$

where $A'_1 \in \mathbb{R}^{r\times r}$. Moreover, since $\operatorname{Ran} B \subset F$, we have that

$$B' = PB = \begin{pmatrix} B'_1 \\ 0 \end{pmatrix}. \tag{4.37}$$

■

Theorem 4.2.5 Brunovski normal form, single-input case

Consider the linear system (4.12) 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

$$\chi_A(z) = \det(z\,\mathrm{Id} - A) = z^n + a_1 z^{n-1} + \ldots + a_{n-1}z + a_n, \tag{4.38}$$

the control system is similar to the following chained form

$$\begin{cases} \dot{x}_1 = x_2 \\ \quad \vdots \\ \dot{x}_{n-1} = x_n \\ \dot{x}_n = -a_n x_1 - a_{n-1}x_2 - \cdots - a_1 x_n + u. \end{cases} \tag{4.39}$$

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

$$\tilde{A} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{pmatrix} \quad \text{and} \quad \tilde{B} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{4.40}$$

Let us define the vectors

$$v_n = B,\ v_{n-1} = AB,\ \ldots,\ v_1 = A^{n-1}B. \tag{4.41}$$

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

$$Av_j = A\left(A^{n-j}B\right) = A^{n-(j-1)}B = v_{j-1}, \qquad \forall j \in [\![2, n]\!]. \tag{4.42}$$

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

$$Av_1 = A^n B = (-a_1 A^{n-1} - \ldots - a_{n-1}A - a_n)B = -a_1 v_1 - \ldots - a_n v_n. \tag{4.43}$$

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 (4.12) 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 $r_1, \dots, r_m$ 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

$$\dot{x} = A(t)x + B(t)u, \tag{4.44}$$

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

$$\frac{\partial}{\partial t}R(t, s) = A(t)R(t, s), \qquad R(s, s) = \mathrm{Id}\,. \tag{4.45}$$

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
$$R(t, s)R(s, \tau) = R(t, \tau) \qquad \forall t, s, \tau \in \mathbb{R}. \tag{4.46}$$
In particular, $R(t, s)^{-1} = R(s, t)$.
- Solutions to (4.44): For any $x_0 \in \mathbb{R}^n$, $T > 0$, and $u \in \mathcal{U}_{x_0,T}$, we have
$$x(t) = R(t, 0)x_0 + \int_0^t R(t, s)Bu(s)\, ds. \tag{4.47}$$

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 $m = 1$:

$$\dot{x} = B(t)u(t), \qquad B(t) = \begin{cases} (1,0)^\top & \text{if } t \in [0,1], \\ (0,1)^\top & \text{if } t > 1. \end{cases} \tag{4.48}$$

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 (4.44) at time $T > 0$ is the matrix

$$G_T := \int_0^T R(T,t)B(t)B(t)^\top R(T,t)^\top \, dt \in \mathbb{R}^{n\times n}. \tag{4.49}$$

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

$$\int_0^T \|B(t)^\top R(T,t)^\top \psi\|_2^2 \, dt \geq C_T \|\psi\|_2^2 \qquad \forall \psi \in \mathbb{R}^n. \tag{4.50}$$

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

$$\psi^\top G_T \psi = \int_0^T \psi^\top R(T,t)B(t)B(t)^\top R(T,t)^\top \psi \, dt = \int_0^T \|B(t)^\top R(T,t)^\top \psi\|_2^2 \, dt \geq 0. \tag{4.51}$$

This proves both the non-negativity and the equivalence between the invertibility of G_T and (4.50). ∎

Remark: Inequality (4.50) is called an observability inequality for the following reason. Consider the *adjoint system* to (4.44), which is

$$\dot{z} = -A(t)^\top z, \qquad z(T) = \psi, \tag{4.52}$$

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

$$E(T) = \int_0^T \|y(t)\|_2^2 \, dt. \tag{4.53}$$

But, since $y(t) = B(t)^\top R(T,t)^\top \psi$, this coincide with the left-hand side of (4.50), which can then be recast as

$$\int_0^T \|y(t)\|_2^2 \, dt \geq C_T \|z(T)\|_2^2. \tag{4.54}$$

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 (4.44) 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

$$\mathrm{End}_{x_0,T}(u) = x_u(T) = x^\star + Lu, \qquad \text{where} \qquad x^\star = R(T,0)x_0, \quad \text{and} \quad Lu = \int_0^T R(T,t)B(t)u(t)\,dt \tag{4.55}$$

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

$$\mathrm{End}_{x_0,T}(u) = x^\star + G_T\psi. \tag{4.56}$$

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

$$\int_0^T \|B(t)^\top R(T,t)^\top \psi\|_2^2\,dt = 0 \implies B(t)^\top R(T,t)^\top \psi = 0 \qquad \text{for a.e. } t \in [0,T]. \tag{4.57}$$

It follows that

$$\psi^\top Lu = 0, \qquad \forall u \in L^\infty([0,T],\mathbb{R}^m), \tag{4.58}$$

where L is defined in (4.55). Hence,

$$\psi^\top \mathrm{End}_{x_0,T}(u) = \psi^\top(x^\star + Lu) = \psi^\top x^\star, \qquad \forall u \in L^\infty([0,T],\mathbb{R}^m). \tag{4.59}$$

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

$$\mathrm{ran}\,\mathrm{End}_{x_0,T} \subset \{z \in \mathbb{R}^n \mid \psi^\top(z - x^\star) = 0\}. \tag{4.60}$$

This contradicts the surjectivity of $\mathrm{End}_{x_0,T}$. ∎

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 $0 \in \mathrm{int}(U)$ 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 (4.44) 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

$$B_0(t) = B(t), \qquad B_{j+1} = A(t)B_j(t) - \frac{d}{dt}B_j(t), \qquad j \in \mathbb{N}. \tag{4.61}$$

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

$$\bigcup_{j\in\mathbb{N}} \mathrm{ran}\,B_j(t_0) = \mathbb{R}^n. \tag{4.62}$$

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

Chapter 5

Optimal control theory for linear systems

Given a control system

$$\dot{x} = f(t, x, u), \qquad x(0) = x_0, \tag{5.1}$$

that is controllable at x_1 in a certain time $T > 0$, it is natural to wonder which control $u \in \mathcal{U}_{x_0,T}$ is the "best" to reach x_1. To answer this question, we need to introduce a cost functional $J : \mathcal{U}_{x_0,T} \to \mathbb{R}$, and then look for a control $u^\star$ minimizing J among all controls that steer x_0 to x_1 in time T. This is the goal of optimal control theory.

Definition 5.0.1 Optimal control problem

Given $x_0 \in \mathbb{R}^d$, $T > 0$, and $x_1 \in \mathrm{Reach}_{x_0,T}$, the *optimal control problem* associated with a cost functional $J : \mathcal{U}_{x_0,T} \to \mathbb{R}$ is

$$\begin{cases} \dot{x} = f(t, x, u) \\ x(0) = x_0 \\ x(T) = x_1 \\ J(u) \to \min. \end{cases} \tag{5.2}$$

Many variants of this problem exist, e.g., where the initial or final positions are not fixed but belong to a certain subset of the state space, or where the final time is free.

Definition 5.0.2 Admissible controls for (5.2)

The set of *admissible controls* to (5.2) is

$$\Omega_{x_0,x_1,T} = \{u \in \mathcal{U}_{x_0,T} \mid x_u(T) = x_1\}. \tag{5.3}$$

Then, (5.2) can be recast as an (infinite-dimensional) optimization problem under constraints. Namely,

$$\min_{u \in \Omega_{x_0,x_1,T}} J(u). \tag{5.4}$$

5.1 Linear-quadratic optimal control problems

In this chapter, we follow [9], and we only consider linear autonomous control systems of the form (4.12). For this kind of systems, we consider the following quadratic cost functional

$$J(u) = \int_0^T c(x_u(t), u(t))\, dt, \tag{5.5}$$

where $c : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a quadratic form. That is,

$$c(x, u) = \frac{1}{2} u^\top P u + u^\top Q x + \frac{1}{2} x^\top R x, \qquad P \in \mathrm{Sym}(\mathbb{R}^m),\ Q \in \mathbb{R}^{m \times d},\ R \in \mathrm{Sym}(\mathbb{R}^d). \tag{5.6}$$

The first question to be answered is the existence of solutions to the corresponding linear-quadratic (LQ) optimal control problem:

$$\begin{cases} \dot{x} = Ax + Bu \\ x(0) = x_0 \\ x(T) = x_1 \\ J(u) = \int_0^T c(x_u(t), u(t))\, dt \to \min. \end{cases} \tag{5.7}$$

Theorem 5.1.1 Existence of solution to (5.7)

An optimal solution for (5.7) exists if and only if

$$J(u) \geq 0 \qquad \forall u \in \Omega_{0,0,T}. \tag{5.8}$$

If, moreover, there exists $\alpha > 0$ such that $J(u) > \alpha \|u\|_2^2$ for all $u \in \Omega_{0,0,T}$ then the optimal solution is unique.

Proof: The idea is to use controls in $\tilde{u} \in \Omega_{0,0,T}$ as variations to test the optimality of $u \in \Omega_{x_0,x_1,T}$. Indeed, by linearity of the control system, the control $u_\lambda = u + \lambda\tilde{u}$ belongs to $\Omega_{x_0,x_1,T}$ for any $\lambda \in \mathbb{R}$, i.e., it steers the system from x_0 to x_1 in time T. Indeed, by the variation of constants formula (4.13) we have

$$\mathrm{End}_{x_0,T}(u_\lambda) = e^{TA}x_0 + \int_0^T e^{(T-s)A}B(u + \lambda\tilde{u}(s))\, ds = x_1 + \lambda \underbrace{\int_0^T e^{(T-s)A}B\tilde{u}(s)\, ds}_{\mathrm{End}_{0,T}(\tilde{u})=0} = x_1. \tag{5.9}$$

In particular, letting x (resp. $\tilde{x}$, and x_λ) be the trajectory corresponding to u (resp. $\tilde{u}$, and u_λ), it then follows that $x_\lambda = x + \lambda\tilde{x}$. Then, by the expression of $c(x,u)$, we have

$$c(x_\lambda, u_\lambda) = \lambda^2 c(\tilde{x}, \tilde{u}) + \lambda\left[\frac{1}{2}\tilde{u}^\top P u + \tilde{u}^\top Q x + u^\top Q\tilde{x} + \tilde{x}^\top R x\right] + c(x,u), \qquad \forall t \in [0,T] \tag{5.10}$$

Assume that u is optimal. In particular, $J(u_\lambda) \geq J(u)$ for all $\lambda \in \mathbb{R}$ and $\tilde{u} \in \Omega(0,0,T)$. Hence, integrating (5.10) over $[0,T]$, we then have

$$\lambda^2 \underbrace{\int_0^T c(x_0, \tilde{u})\, ds}_{=J(\tilde{u})} + \lambda \int_0^T \left[\frac{1}{2}\tilde{u}^\top P u^\star + \tilde{u}^\top Q x^\star + (u^\star)^\top Q x_0 + x_0^\top R x^\star\right] ds \geq 0, \qquad \forall \lambda \in \mathbb{R},\ \tilde{u} \in \Omega(0,0,T). \tag{5.11}$$

This immediately yields $J(\tilde{u}) \geq 0$ for all $\tilde{u} \in \Omega(0,0,T)$.

Let us now prove the converse implication. Assume that $J(\tilde{u}) \geq 0$ for all $\tilde{u} \in \Omega_{0,0,T}$. Up to fixing any $u_0 \in \Omega_{x_0,x_1,T}$, we have that for any $u \in \Omega_{x_0,x_1,T}$ it holds $u - u_0 \in \Omega_{0,0,T}$, i.e., $u = u_0 + \tilde{u}$ for some $\tilde{u} \in \Omega_{0,0,T}$. Hence, minimizing J over $\Omega_{x_0,x_1,T}$ is equivalent to minimizing the functional $J(u_0 + \tilde{u})$ over $\tilde{u} \in \Omega_{0,0,T}$. From (5.10) with $\lambda = 1$, this is equivalent to minimizing the functional $\phi : \Omega_{0,0,T} \to \mathbb{R}$ defined as

$$\phi(\tilde{u}) = J(\tilde{u}) + L\tilde{u}, \qquad \text{where } L\tilde{u} := \int_0^T \left[\frac{1}{2}\tilde{u}^\top P u_0 + \tilde{u}^\top Q x + u_0^\top Q\tilde{x} + \tilde{x}^\top R x\right] dt, \tag{5.12}$$

Here, J is a quadratic form and L is a linear functional. Even though the problem is posed on $\Omega_{0,0,T} \subset L^\infty([0,T], \mathbb{R}^m)$, we can equivalently consider it on the Hilbert space $X \subset L^2([0,T], \mathbb{R}^m)$ of L^2 controls steering the system from 0 to itself in time T, since L^∞ is dense in L^2 and both J and L extends continuously to the latter. Let $N = \ker(J) \subset L^2([0,T], \mathbb{R}^m)$ and observe that $\ker(L) \supset N$. Indeed, if this was not the case $J(u_0 + \lambda\tilde{u})$ would be unbounded from below for $\tilde{u} \in N$ such that $L\tilde{u} < 0$ and $\lambda \to +\infty$, contradicting the non-negativity of J on $\Omega_{0,0,T}$.

Let now $N^\perp$ be the orthogonal complement of N in X. Since J is non-negative by assumption, it is strictly positive on $N^\perp$. Thus, ϕ is continous, coercive and strictly convex on $N^\perp$. In particular, there exists a unique $\tilde{u}^\star \in N^\perp$ such that

$$\phi(\tilde{u}^\star) = \min_{\tilde{u} \in N^\perp} \phi(\tilde{u}). \tag{5.13}$$

Since ϕ is zero on N, it follows that $\tilde{u}^\star$ is a minimizer of ϕ on the whole X. To conclude the proof of the converse implication it then suffices to observe that $\tilde{u}^\star \in \Omega_{0,0,T}$. We omit a precise argument for this fact, but this can be done exploiting the necessary conditions for optimality of the Pontryiagin Maximum Principle, to be presented in the next section.

Finally, if $J(u) \geq \alpha \|u\|_2^2$ for all $u \in \Omega_{0,0,T}$, then J is strictly convex on $\Omega_{0,0,T}$, thus on X, and thus the minimizer is unique. ■

We henceforth consider the following assumptions.

Assumption 5.1.1

- The pair (A, B) satisfies the Kalman rank condition.
- There exists $T_{\max} > 0$ such that the functional J is non-negative on $\Omega_{0,0,T_{\max}}$.
- The system is *dissipative*, i.e., for any $T \in (0, T_{\max}]$ there exists a control $u \in \Omega_{0,0,T}$ such that $J(u) > 0$.

The first assumption ensures that the system is controllable, and thus that the optimal control problem (5.7) can have a solution. By Theorem 5.1.1, the second assumption guarantees the existence of a solution to (5.7) for any $T < T_{\max}$ and any $x_0, x_1 \in \mathbb{R}^d$. The third assumption is a technical condition, and it guarantees the uniqueness of the solution.

An important consequence of the third assumption is the following.

Proposition 5.1.1

Let $u \in \Omega_{x_0,x_1,T}$. Then, for any $\varepsilon > 0$ there exists a control $u_\varepsilon \in \Omega_{x_0,x_1,T}$ such that

$$J(u_\varepsilon) = J(u) + \varepsilon. \tag{5.14}$$

Proof: Let $\bar{u} \in \Omega_{0,0,T}$ be such that $J(\bar{u}) > 0$, which exists by the dissipativity assumption. Then, for any $\lambda \in \mathbb{R}$ the control $u_\lambda = u + \lambda \bar{u}$ belongs to $\Omega_{x_0,x_1,T}$. Then, $\lambda \mapsto J(u_\lambda)$ is continuous and strictly increasing, with $J(u_0) = J(u)$ and $\lim_{\lambda \to +\infty} J(u_\lambda) = +\infty$. The conclusion follows by the intermediate value theorem. ■

Observe that the second assumption is always satisfied if P is non-negative definite (i.e., $c(x, u)$ is strictly convex in u for any fixed x). It turns out that this is actually necessary.

Proposition 5.1.2

We have that $P \in \mathrm{Sym}_{\geq 0}(\mathbb{R}^m)$.

Proof: Assume by contradiction that there exists a splitting $E_1 \oplus E_2 = \mathbb{R}^m$ such that P is negative definite on E_2. Then, there exists $K > 0$ such that $u^\top P u \leq -K\|u\|_2^2$ for all $u \in E_2$. This implies that

$$c(x, u) \leq -\frac{K}{2}\|u\|_2^2 + \|u\|_2 \|Q\| \|x\|_2 + \frac{1}{2}\|R\| \|x\|_2^2 \qquad \forall (x, u) \in \mathbb{R}^d \times E_2. \tag{5.15}$$

By controllability, for any $T > 0$ there exists a control $u \in \Omega_{0,0,T}$ with values in E_2 and arbitrarily large L^2 norm. But then, the above inequality implies that $J(u)$ can be made arbitrarily negative, contradicting the non-negativity of J on $\Omega_{0,0,T}$. ■

5.1.1 Pontryagin Maximum Principle for LQ problems

Let us consider the extended system on $(x, \bar{x}) \in \mathbb{R}^d \times \mathbb{R}$, given by

$$\begin{cases} \dot{x} = Ax + Bu, \\ \dot{\bar{x}} = c(x, u). \end{cases} \qquad (x(0), \bar{x}(0)) = (x_0, 0). \tag{5.16}$$

Denote the corresponding reachable set at time $T > 0$ by $\mathrm{Reach}^{\mathrm{ext}}_{x_0,T}$. We then have the following, to be compared with the geometric interpretation for optimality, presented in Section 2.1.1.

Theorem 5.1.2 Optimality and extended system

Assume that the control $u^\star \in \Omega_{x_0,x_1,T}$ is an optimal solution to (5.7). Then,

$$(x_1, J(u^\star)) \in \partial \mathrm{Reach}^{\mathrm{ext}}_{x_0,T}. \tag{5.17}$$

Conversely, every point $(x_1, \bar{J}) \in \partial \mathrm{Reach}^{\mathrm{ext}}_{x_0,T}$ corresponds to an optimal solution $u^\star \in \Omega_{x_0,x_1,T}$ with $J(u^\star) = \bar{J}$.

Proof: The key observation is that any point $(x, \bar{J}) \in \mathrm{Reach}^{\mathrm{ext}}_{x_0,T}$ is of the form $(x_u(T), J(u))$ for some $u \in \mathcal{U}_{x_0,T}$.

Then, if $u^\star$ is optimal, the extended terminal point $(x_1, J(u^\star))$ must be the lowest point of the line $\{x_1\} \times \mathbb{R}$ to belong to $\mathrm{Reach}^{\mathrm{ext}}_{x_0,T}$. In particular, this implies (5.17).

Conversely, assume that $(x_1, \bar{J}(\bar{u})) \in \partial \mathrm{Reach}^{\mathrm{ext}}_{x_0,T}$ for some control $\bar{u} \in \Omega_{x_0,x_1,T}$, but that v is not optimal for (5.7). Then, there exists a control $v \in \Omega_{x_0,x_1,T}$ such that $J(v) = J(\bar{u}) - \varepsilon$ for a certain $\varepsilon > 0$. Given $\delta > 0$, let $V_\delta = \{u \in L^\infty([0,T], \mathbb{R}^m) \mid \|u - v\|_\infty < \delta\}$. Since the system is controllable (see Corollary 4.2.3), $\mathrm{End}_{x_0,T}$ is an open mapping, and thus $\mathrm{End}_{x_0,T}(V_\delta)$ covers a neighborhood of x_1 in $\mathbb{R}^d$. This implies that there exists $\delta > 0$ sufficiently small such that

$$J(u) \le J(v) + \frac{\varepsilon}{2} = J(\bar{u}) - \frac{\varepsilon}{2}, \qquad \forall u \in V_\delta. \tag{5.18}$$

In particular, by Proposition 5.1.1, for any $y \in \mathrm{End}_{x_0,T}(V_\delta)$ and any $\alpha \in [J(\bar{u}) - \frac{\varepsilon}{2}, J(\bar{u}) + \frac{\varepsilon}{2}]$ there exists $u_\alpha \in \Omega_{x_0,x_1,T}$ such that $J(u_\eta) = J(\bar{u}) - \frac{\varepsilon}{2} + \eta$. Namely,

$$W := \left[J(\bar{u}) - \frac{\varepsilon}{2}, J(\bar{u}) + \frac{\varepsilon}{2}\right] \times \mathrm{End}_{x_0,T}(V_\delta) \subset \mathrm{Reach}^{\mathrm{ext}}_{x_0,T}. \tag{5.19}$$

Since W is a neighborhood of $(x_1, J(\bar{u}))$, this contradicts the fact that $(x_1, J(\bar{u})) \in \partial \mathrm{Reach}^{\mathrm{ext}}_{x_0,T}$. ∎

Example 5.1.1 (Convex LQ systems)

The LQ optimal control problem (5.7) is *convex* if $c : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is a convex function. This is the case, e.g., if $P \in \mathrm{Sym}_{>0}(\mathbb{R}^m)$, $R \in \mathrm{Sym}_{\ge 0}(\mathbb{R}^d)$, and $Q = 0$.

An LQ optimal control problem always satisfies $c(x,u) \ge 0$. Moreover, in this case the reachable set $\mathrm{Reach}^{\mathrm{ext}}_{x_0,T}$ is convex.

Theorem 5.1.3 Pontryagin Maximum Principle for LQ problems

Let $u^\star \in \Omega_{x_0,x_1,T}$ be an optimal solution to (5.7), and let $x^\star$ be the corresponding trajectory. Then, there exists a nontrivial Lipschitz function $p : [0,T] \to \mathbb{R}^d$ such that

$$\begin{cases} \dot{x}^\star = Ax^\star + Bu^\star, \\ \dot{p} = -A^\top p + Rx^\star + Q^\top u^\star, \end{cases} \qquad \text{and} \qquad u^\star(t) = \arg\max_{v \in \mathbb{R}^m} H(x^\star(t), p(t), v), \tag{5.20}$$

where the Hamiltonian $H : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is given by

$$H(x,p,u) = p^\top (Ax + Bu) - c(x,u). \tag{5.21}$$

Proof: We have that $(x_1, J(u^\star)) \in \partial \mathrm{Reach}^{\mathrm{ext}}_{x_0,T}$. Let us consider variations of the form $u_\lambda = u^\star + \lambda \tilde{u}$, where $\tilde{u} \in L^\infty([0,T], \mathbb{R}^m)$. Then, by the same argument used in the proof of Theorem 5.1.1, we have that $x_\lambda = x^\star + \lambda \tilde{x}$, where $\tilde{x}$ is the trajectory corresponding to $\tilde{u}$ with initial condition $\tilde{x}(0) = 0$. Observe that $x_\lambda(T) \ne x_1$ in general and that, as in (5.10),

$$J(u_\lambda) = J(u^\star) + \lambda \int_0^T \left[\frac{\partial}{\partial u} c(x^\star, u^\star)\, \tilde{u} + \frac{\partial}{\partial x} c(x^\star, u^\star)\, \tilde{x}\right] dt + \lambda^2 J(\tilde{u}). \tag{5.22}$$

Consider the map $F : L^\infty([0,T], \mathbb{R}^m) \to \mathbb{R}^{d+1}$ given by

$$F(u) = (x_u(T), J(u)). \tag{5.23}$$

Then, $F(u_\lambda) = (x_\lambda(T), J(u_\lambda))$ belongs to $\mathrm{Reach}^{\mathrm{ext}}_{x_0,T}$ for all $\lambda \in \mathbb{R}$. We compute its directional derivative $L(\tilde{u}) := DF_{u^\star}(\tilde{u}) = \lim_{\lambda\to 0} \frac{F(u_\lambda)-F(u^\star)}{\lambda}$ at $u^\star$ in the direction $\tilde{u}$, obtaining

$$L(\tilde{u}) = \left(\tilde{x}(T), \int_0^T \left[\frac{\partial}{\partial u}c(x^\star,u^\star)\,\tilde{u} + \frac{\partial}{\partial x}c(x^\star,u^\star)\,\tilde{x}\right] dt\right). \tag{5.24}$$

The key observation is then the following: there exists a linear form $p_T : \mathbb{R}^d \to \mathbb{R}$ such that

$$L(\tilde{u}) = (\tilde{x}(T), p_T(\tilde{x}(T))), \qquad \forall \tilde{u} \in L^\infty([0,T],\mathbb{R}^m). \tag{5.25}$$

Indeed, L is a linear map from the infinite-dimensional space $L^\infty([0,T],\mathbb{R}^m)$ to the finite-dimensional space $\mathbb{R}^{d+1}$, and thus $\operatorname{ran} L$ is a linear subspace of $\mathbb{R}^{d+1}$ of dimension at most $d+1$. But, on one hand, controllability implies that $\operatorname{ran} L$ contains the subspace $\mathbb{R}^d \times \{0\}$, which has dimension d. On the other hand, if $\operatorname{ran} L = \mathbb{R}^{d+1}$ then for any $\alpha > 0$ there would exists v such that $L(v) = (0,-\alpha)$. In particular, the corresponding trajectory y satisfies $y(T) = 0$ and hence $x_\lambda(T) = x^\star(T) + \lambda y(T) = x_1$. As a consequence, (5.22) yields

$$J(u_\lambda) = J(u^\star) - \lambda\alpha + \lambda^2 J(v), \qquad \forall \lambda \in \mathbb{R} \implies J(u_\lambda) < J(u^\star), \qquad \lambda \ll 1. \tag{5.26}$$

This contradicts the optimality of $u^\star$. Hence, $\operatorname{ran} L$ has dimension exactly d, and thus it is the graph of a linear form $p_T : \mathbb{R}^d \to \mathbb{R}$.

To complete the proof, we need to show that p_T is of the form $p_T(\tilde{x}(T)) = p(T)^\top \tilde{x}(T)$ for some absolutely continuous function $p : [0,T] \to \mathbb{R}^d$ satisfying the adjoint equation

$$\dot{p} = -A^\top p + Rx^\star + Q^\top u^\star = -A^\top p + \frac{\partial}{\partial x}c(x^\star,u^\star). \tag{5.27}$$

Let us consider a solution to the above with terminal condition $p(T) = p_T$. Then, recalling that $\dot{\tilde{x}} = A\tilde{x} + B\tilde{u}$ and that $\tilde{x}(0) = 0$, by (5.25) we obtain

$$\int_0^T \left[\frac{\partial}{\partial u}c(x^\star,u^\star)\,\tilde{u} + \frac{\partial}{\partial x}c(x^\star,u^\star)\,\tilde{x}\right] dt = p(T)^\top(x(T)) = \int_0^T \frac{d}{dt}\left[p^\top x\right] dt = \int_0^T \left[\dot{p}^\top x + p^\top(Ax+Bu)\right] dt. \tag{5.28}$$

Replacing (5.27) in the above, we obtain

$$\int_0^T \left[\frac{\partial}{\partial u}c(x^\star,u^\star) - B^\top p\right] \tilde{u} = 0, \qquad \forall \tilde{u} \in L^\infty([0,T],\mathbb{R}^m). \tag{5.29}$$

That is,

$$\frac{\partial}{\partial u}c(x^\star,u^\star) - B^\top p = 0 \qquad \text{for a.e. } t \in [0,T], \tag{5.30}$$

which is equivalent to the maximization condition $u^\star(t) = \arg\max_{v\in\mathbb{R}^m} H(x^\star(t),p(t),v)$, since

$$\frac{\partial}{\partial u}H(x,p,u) = \frac{\partial}{\partial u}c(x,u) - B^\top p, \tag{5.31}$$

and H is concave in u for any fixed (x,p) due to the fact that $P \in \mathrm{Sym}_{\geq 0}(\mathbb{R}^m)$ by Proposition 5.1.2. ∎

5.1.2 Lagrangian interpretation of PMP for LQ problems

For simplicity, in this section we assume $Q = 0$, so that the running cost takes the form

$$c(x,u) = \frac{1}{2}\left(x^\top Rx + u^\top Pu\right).$$

Let $\mathrm{Lip}([0,T],\mathbb{R}^d)$ denote the space of Lipschitz functions $x : [0,T] \to \mathbb{R}^d$, endowed with the norm

$$\|x\|_{\mathrm{Lip}} := \|x\|_\infty + L_x,$$

where L_x is the Lipschitz constant of x. Consider the extended cost functional

$$J : \mathrm{Lip}([0,T],\mathbb{R}^d) \times L^\infty([0,T],\mathbb{R}^m) \to \mathbb{R}, \qquad J(x,u) = \int_0^T c(x(t),u(t))\,dt. \tag{5.32}$$

Then, the linear–quadratic optimal control problem (5.7) can be rewritten as an infinite-dimensional optimization problem with dynamical constraints:

$$\min_{(x,u)\in \mathrm{Lip}([0,T],\mathbb{R}^d)\times L^\infty([0,T],\mathbb{R}^m)} J(x,u) \quad \text{subject to} \quad \dot{x} = Ax + Bu, \quad x(0) = x_0, \quad x(T) = x_1. \tag{5.33}$$

Following the ideas of Chapter 2.1, we introduce the Lagrangian

$$\mathcal{L}(x,u,p) := \frac{1}{2}\int_0^T \left(x^\top R x + u^\top P u\right) dt + \int_0^T p^\top (\dot{x} - Ax - Bu)\,dt, \tag{5.34}$$

where $p \in \mathrm{Lip}([0,T],\mathbb{R}^d)$ is the Lagrange multiplier associated with the dynamical constraint $\dot{x} = Ax + Bu$.

Integrating the term $\int_0^T p^\top \dot{x}\,dt$ by parts and using the boundary conditions $x(0) = x_0$, $x(T) = x_1$ gives

$$\mathcal{L}(x,u,p) = \frac{1}{2}\int_0^T \left(x^\top R x + u^\top P u\right) dt - \int_0^T p^\top (Ax + Bu)\,dt + p(T)^\top x_1 - p(0)^\top x_0 - \int_0^T \dot{p}^\top x\,dt. \tag{5.35}$$

An optimal solution $(x^\star, u^\star)$ should satisfy the stationarity condition

$$\frac{\partial}{\partial(x,u)}\mathcal{L}(x^\star,u^\star,p) = 0 \qquad \text{for some } p \in \mathrm{Lip}([0,T],\mathbb{R}^d). \tag{5.36}$$

By considering appropriate variations of $\mathcal{L}$ with respect to x and u, concentrated on small time intervals, it is possible to show that the above condition is equivalent to the corresponding pointwise conditions (i.e., without integration in time).

This yields the following conditions:

- Variation with respect to u:
$$0 = Pu^\star + B^\top p, \tag{5.37}$$
which gives the optimal feedback law
$$u^\star = -P^{-1}B^\top p.$$
- Variation with respect to x:
$$0 = Rx^\star - \dot{p} - A^\top p, \tag{5.38}$$
which yields the adjoint (costate) equation
$$\dot{p} = -A^\top p + Rx^\star.$$

Thus, the Pontryagin Maximum Principle (PMP) for the linear–quadratic problem can be interpreted as a first-order stationarity condition for the Lagrangian $\mathcal{L}$, with the adjoint p serving as the Lagrange multiplier for the dynamics.

5.2 Linear-quadratic problems with unconstrained terminal state

The PMP given in Theorem 5.1.3 for problem (5.7) presents an important drawback: even though the control $u^\star$ is completely determined by the pair $(x^\star, p^\star)$, the ODE determining the latter cannot be easily solved due to the fact that the boundary conditions are on $x^\star(0) = x_0$ and $x^\star(T) = x_1$. Namely, $p^\star$ is determined implicitly and is typically found by "shooting" methods (i.e., one chooses initial conditions $x^\star(0) = x_0$ and $p^\star(0) = \lambda$ and then looks for λ such that $x^\star(T) = x_1$).

In this section, to circumvent this problem, we consider the following relaxed version of (5.7):

$$\begin{cases} \dot{x} = Ax + Bu \\ x(0) = x_0 \\ J(u) = \int_0^T c(x_u(t),u(t))\,dt \to \min. \end{cases} \tag{5.39}$$

Since no terminal condition is required, to avoid trivial optimal solutions $u \equiv 0$ one needs to adjust the cost J by adding a "final cost" as

$$J(u) = \int_0^T c(x,u)\,dt + \frac{1}{2}x(T)^\top W x(T). \tag{5.40}$$

Here, $W \in \mathrm{Sym}_{\geq 0}(\mathbb{R}^d)$ is a matrix. (Compare with the penalty methods discussed in Section **??**.)

For simplicity we consider the following.

Assumption 5.2.1

- The pair (A, B) satisfies the Kalman rank condition.
- The running cost is given by
$$c(x,u) = \frac{1}{2}u^\top P u + \frac{1}{2}x^\top R x, \tag{5.41}$$
with $P \in \mathrm{Sym}_{>0}(\mathbb{R}^d)$ and $R \in \mathrm{Sym}_{\geq 0}(\mathbb{R}^d)$.

The results previously discussed still hold in this case (observe that $J(u) > 0$ for any non-zero control). We resume them in the following and refer to [17] for the proofs.

Theorem 5.2.1 Pontryiagin Maximum Principle for LQ systems with terminal cost

There exists a unique optimal solution $u^\star$ to (5.39), with corresponding trajectory $x^\star$. Moreover, there exists a nontrivial Lipschitz function $p^\star : [0,T] \to \mathbb{R}^d$ such that

$$\begin{cases} \dot{x}^\star = Ax^\star + Bu^\star, & x^\star(0) = x_0 \\ \dot{p} = -A^\top p + Rx^\star, & p^\star(T) = -Wx(T), \end{cases} \tag{5.42}$$

and yielding

$$u^\star = P^{-1}B^\top p, \qquad \text{on } [0,T]. \tag{5.43}$$

We also need the following concept.

Definition 5.2.1 Value function

The *value function* associated with the optimal control problem (5.7) with final time $T > 0$ is the map $V_T : \mathbb{R}^d \to \mathbb{R}$ defined as

$$V_T(x_0) = \inf\{J(u) \mid x_u(0) = x_0\}. \tag{5.44}$$

Proposition 5.2.1

For any $T > 0$ there exists $c > 0$ such that

$$V_s(x_0) \leq c\|x_0\|_2^2, \qquad \forall x_0 \in \mathbb{R}^d,\ s \in [0,T]. \tag{5.45}$$

Proof: Fix $x_0 \in \mathbb{R}^d$ and let us consider the control $u \equiv 0$ on $[0,s]$. Then, the corresponding trajectory satisfies $x(t) = e^{tA}x_0$. Observing that $\|x(t)^\top R x(t)\|_2^2 \leq \|R\|\|x(t)\|_2^2$ and that $\|x(t)\| \leq \|e^{tA}\|\|x_0\|_2^2$, we get

$$|J(0)| = \frac{1}{2}\left|\int_0^s x(t)^\top R x(t)\,dt + x(s)^\top W x(s)\right| \leq \frac{1}{2}\left[\int_0^s \|e^{tA}\|^2\|R\|\,dt + \|W\|\|e^{sA}\|\right]\|x_0\|_2^2. \tag{5.46}$$

The statement follows since $V_s(0) \leq J(0)$. ∎

5.2.1 The Riccati equation and the optimal feedback law

Although the PMP provides necessary conditions for optimality, it does not directly yield an explicit expression for the optimal control $u^\star$ in terms of the state $x^\star$. However, in the case of linear–quadratic problems, it is actually possible to

express the adjoint state $p^\star$ as a function of the state $x^\star$ via the solution of a matrix differential equation known as the Riccati equation. In particular, this allows us to derive an explicit feedback law for the optimal control, i.e., an expression of the form $u^\star(t) = K(t)x^\star(t)$ for some time-varying gain matrix $K(t)$.

We have the following.

Theorem 5.2.2 Riccati Equation

The unique optimal solution $u^\star$ to (5.7) is given by

$$u^\star(t) = -P^{-1}B^\top E(t)x^\star(t). \tag{5.47}$$

Here, $x^\star$ is the corresponding optimal trajectory, and $E : [0,T] \to \mathrm{Sym}(\mathbb{R}^d)$ is the unique solution to the matrix Riccati differential equation

$$\begin{cases} \dot{E} = R - A^\top E - EA - EBP^{-1}B^\top E, \\ E(T) = -W, \end{cases} \tag{5.48}$$

In particular, the value function is given by

$$V_T(x_0) = -\frac{1}{2}x_0^\top E(0)x_0. \tag{5.49}$$

Proof: Let $x^\star$ be the optimal trajectory corresponding to the unique optimal control $u^\star$, and let $p^\star$ be the associated adjoint state, satisfying the PMP conditions of Theorem 5.2.1. In particular, we have

$$u^\star(t) = -P^{-1}B^\top p^\star(t). \tag{5.50}$$

This implies that the optimal pair $(x^\star, p^\star)$ satisfies the closed-loop system

$$\begin{cases} \dot{x}^\star = Ax^\star - BP^{-1}B^\top p^\star, \\ \dot{p}^\star = -A^\top p^\star + Rx^\star, \end{cases} \qquad \text{with } x^\star(0) = x_0, \quad x^\star(T) = x_1, \quad p^\star(T) = -Wx(T). \tag{5.51}$$

Let E be a solution to (5.48) (which we will show to exist on $[0,T]$ in Lemma 5.2.3), and observe that $E(t) \in \mathrm{Sym}(\mathbb{R}^d)$ for all $t \in [0,T]$. Indeed, $E(0) \in \mathrm{Sym}(\mathbb{R}^d)$, and the right-hand side of (5.48) is symmetric whenever $E(t)$ is symmetric.

Let now $p = Ex$, where x is the solution to (4.12), i.e., $\dot{x} = Ax + Bu$, w.r.t. the feedback control $u = P^{-1}B^\top Ex$. Then, by (5.50), it follows that

$$\dot{p} = \dot{x}^\top E + x^\top \dot{E} = -A^\top p + Rx^\star. \tag{5.52}$$

That is, (u, x, p) satisfies the conditions of the PMP. By unicity of the optimal control we then get $u = u^\star$, which in turn implies that $x = x^\star$, and finally that $p = p^\star$.

We are left to show (5.49). We have,

$$\frac{1}{2}\frac{d}{dt}x^\top Ex = \frac{1}{2}\frac{d}{dt}p^\top x = \frac{1}{2}\left(\dot{p}^\top x + p^\top \dot{x}\right) = \frac{1}{2}\left(x^\top Rx + p^\top Bu\right) = \frac{1}{2}\left(x^\top Rx + u^\top Pu\right) = c(x,u). \tag{5.53}$$

Here, we used the fact that a simple computation, relying on the feedback expression of u and the fact that $p = Ex$, yields $p^\top Bu = u^\top Pu$. Finally, recalling that $E(T) = -W$,

$$V_T(x) = \int_0^T c(x^\star, u^\star)\,dt + \frac{1}{2}x(T)^\top Wx(T) = \frac{1}{2}\int_0^T \frac{d}{dt}x^\top Ex\,dt + \frac{1}{2}x(T)^\top Wx(T) = -\frac{1}{2}x(0)^\top E(0)x(0). \tag{5.54}$$

■

Lemma 5.2.3 Existence of solutions to the Riccati equation

The solution to (5.48) is defined on $[0,T]$.

Proof: By contradiction, let us assume that there exists $t_* \in (0,T)$ such that $\|E(t)\| \to +\infty$ as $t \downarrow t_*$. In particular, this implies that

$$\forall M > 0\ \exists x_0 \in \mathbb{R}^d \qquad \text{such that} \qquad \frac{|x_0^\top E(t_0)x_0|}{\|x_0\|^2} \geq M. \tag{5.55}$$

We now claim that for any $t_0 \in [0,T)$ and $x_0 \in \mathbb{R}^d$, there exists $c > 0$ such that

$$|x_0^\top E(t_0) x_0| \leq c\|x_0\|_2^2. \tag{5.56}$$

This contradicts (5.55), and thus concludes the proof.

Let $t_0 \in [0,T)$ and $x_0 \in \mathbb{R}^d$ and consider $x^\star$ to be the unique optimal trajectory for (5.7) on the interval $[t_0,T]$ such that $x(t_0) = x_0$ (this exists since $P \in \mathrm{Sym}_{>0}(\mathbb{R}^d)$). Then, the same computations carried out at the end of the proof of Theorem 5.2.2, show that it must hold

$$V_{T-t_0}(x_0) = -x_0^\top E(t_0) x_0. \tag{5.57}$$

The statement then follows from Proposition 5.2.1 ∎

5.2.2 The tracking problem

Consider the problem of finding a trajectory of a linear control system $\dot{x} = Ax + Bu$ in such a way that it "best" follows a given trajectory $\gamma : [0,T] \to \mathbb{R}^d$. In general the trajectory γ is not a solution of the control system, and thus we set up an optimal control problem in order to minimize the error

$$z(t) := x(t) - \gamma(t). \tag{5.58}$$

We will also try to minimize the L^2 norm of the control u.

Observe that z is a solution of the following control system

$$\dot{z} = Az + Bu + r, \qquad z(0) = z_0, \tag{5.59}$$

where $z_0 = x_0 - \gamma(0)$ and $r(t) = A\gamma(T) - \dot{\gamma}(t)$. It is then reasonable to look for controls u minimising a cost in the form (5.40).

We then have the following.

Proposition 5.2.2 Tracking problem

Let $\gamma : [0,T] \to \mathbb{R}^d$ and consider the tracking problem for the linear control system

$$\dot{x} = Ax + Bu, \qquad x(0) = x_0, \tag{5.60}$$

with cost

$$J(u) = (x(T) - \gamma(T))^\top W (x(T) - \gamma(T)) + \int_0^T \left[(x-\gamma)^\top R (x-\gamma) + u^\top P u\right] dt. \tag{5.61}$$

Then, there exists a unique optimal control u which is given in feedback form by

$$u(t) = P^{-1}B^\top E(t)(x(t) - \gamma(t)) + P^{-1}B^\top h(t), \tag{5.62}$$

where $E : [0,T] \to \mathrm{Sym}(\mathbb{R}^d)$ and $h : [0,T] \to \mathbb{R}^d$ are the solutions of

$$\dot{E} = R - A^\top E - EA - EBP^{-1}B^\top E, \qquad E(T) = -W, \tag{5.63}$$

$$\dot{h} = -A^\top h - E(A\gamma - \dot{\gamma}) - EBP^{-1}B^\top h, \qquad h(T) = 0. \tag{5.64}$$

Proof: We consider the system (5.59). In order to get rid of the perturbation term r, we augment the dimension of the system, considering the system

$$\dot{z}_1 = A_1 z_1 + B_1 u, \qquad z_1(0) = \begin{pmatrix} z(0) \\ 1 \end{pmatrix}, \qquad A_1 = \begin{pmatrix} A & r \\ 0 & 0 \end{pmatrix}, \qquad B_1 = \begin{pmatrix} B \\ 0 \end{pmatrix} \tag{5.65}$$

where $z_1 = (z,1)^\top \in \mathbb{R}^{d+1}$ and R_1, W_1 are chosen so that the cost $J(u)$ reads

$$J(u) = \frac{1}{2}\int_0^T \left[u^\top P u + z_1^\top R_1 z_1\right] dt + z_1(T)^\top W_1 z_1(T). \tag{5.66}$$

Then, by the preceding section, we have that the optimal control is unique and reads

$$u(t) = P^{-1}B_1^\top E_1(t)z_1(t), \tag{5.67}$$

where $E_1 : [0,T] \to \mathrm{Sym}(\mathbb{R}^{d+1})$ satisfies the corresponding Riccati equation. Let

$$E_1 = \begin{pmatrix} E & h \\ h^\top & \alpha \end{pmatrix}, \tag{5.68}$$

for $E \in \mathrm{Sym}(\mathbb{R}^d)$, $h \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$. Then, the Riccati equation yields (5.63), (5.64), and

$$\dot{\alpha} = -2r^\top h - h^\top BP^{-1}B^\top h, \qquad \alpha(T) = 0. \tag{5.69}$$

The statement follows by the formula (5.47) in Theorem 5.2.2 for the optimal control. ■

Part III

Data

Chapter 6

Machine learning

The fundamental problem of machine learning can be formalized as follows.

Problem 6.1: Machine learning problem

Let X, Y be two sets, and $A(X, Y)$ be a set of functions from X to Y. Assume moreover that we are given a subset $\mathcal{F} \subset A(X, Y)$ of these functions parametrized over a set Θ, that is

$$\mathcal{F} = \{\phi_\theta \in A(X, Y) \mid \theta \in \Theta\}. \tag{6.1}$$

Then, given a function $\psi \in A(X, Y)$ the machine learning (ML) problem consists in finding $\hat{\theta} \in \Theta$ such that

$$\psi \approx \phi_{\hat{\theta}}. \tag{6.2}$$

This has to be done using the *available* information about ψ.

Example 6.0.1 (Linear approximation and pointwise observation)

Let $\psi : [0, 2\pi] \to \mathbb{R}$ be a continuous function (in particular, $\psi \in L^2([0, 2\pi])$).
Consider $\{\phi_i\}_{i=1}^{+\infty} \subset L^2([0, 2\pi]) \cap C([0, 2\pi])$ to be an orthonormal system for $L^2([0, 2\pi])$, and for any $N \in \mathbb{N}$ let $\Theta = \mathbb{R}^N$ and consider

$$\mathcal{F} = \left\{\phi_\theta = \sum_{i=1}^{N} \theta_i \phi_i \mid \theta_i \in \mathbb{R}^N\right\} \tag{6.3}$$

Assume to be given M observations $\{(x_i, y_i)\}_{i=1}^{M}$ such that

$$y_i = \psi(x_i), \qquad i \in [\![1, M]\!]. \tag{6.4}$$

The task here is to find $\hat{\theta} \in \Theta$ such that

$$\sum_{i=1}^{M} \left|\phi_{\hat{\theta}}(x_i) - y_i\right|^2 = \min_{\theta} \sum_{i=1}^{M} |\phi_\theta(x_i) - y_i|^2. \tag{6.5}$$

Example 6.0.2 (Finite elements)

Let $B_1(0) \subset \mathbb{R}^d$ and $\psi : B_1(0) \to \mathbb{R}$ be a real-valued function such that $\psi \in H_0^2(B_1(0))$. Then, given independent functions $\{\phi_i\}_{i=1}^{N}$ we can consider $\Theta = \mathbb{R}^N$ and let

$$\mathcal{F} = \left\{\phi_\theta = \sum_{i=1}^{N} \theta_i \phi_i \mid \theta_i \in \mathbb{R}^N\right\} \tag{6.6}$$

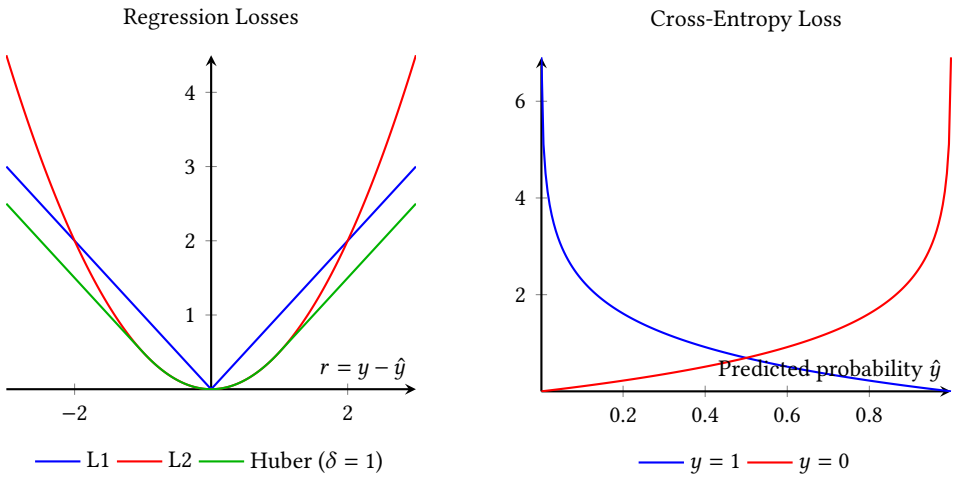

Figure 6.1: Typical loss functions for $Y = \mathbb{R}$ (right) and cross-entropy loss.

A typical situation, and the basis for the *finite element methods*, is to know $f = \Delta\psi$, in which case one looks for $\hat{\theta} \in \Theta$ such that

$$\|\phi_{\hat{\theta}} - \psi\|_{L^2} = \min_{\theta} \|\phi_\theta - \psi\|_{L^2}. \tag{6.7}$$

In modern machine learning one is interested in "going beyond" these examples, by considering a nonlinear parametrization of the approximating family $\mathcal{F}$. That is,

$$\text{the map } \theta \mapsto \phi_\theta \text{ is nonlinear.}$$

In the following we will focus on the problem of data-fitting exoked in Example 6.0.1, see also Example **??**. This is the following.

Definition 6.0.1 Data-fitting problem

Consider the following data:

- A function $\psi : X \to Y$ to approximate, of which only a set of samples $\mathfrak{X} = \{(x_i, \psi(x_i)) \mid i \in [\![1, M]\!]\}$ is known;
- A set $\mathcal{F} = \{\psi_\theta : X \to Y \mid \theta \in \Theta\}$ of possible approximations;
- A loss function $\mathcal{L} : Y \times Y \to \mathbb{R}$.

Then, the data-fitting problem (or minimization of the empirical risk) is the following:

$$\text{Find } \hat{\theta} \in \Theta \text{ such that } \hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{(x_i, \hat{y}_i) \in \mathfrak{X}} \mathcal{L}(\phi_\theta(x_i), \hat{y}_i). \tag{6.8}$$

When Y has a vector space structure, it is natural to consider $\mathcal{L}(y, \hat{y}) = \ell(r)$ where $r = y - \hat{y}$ is called the residual. In this case, typical choices for the loss function are appropriate norms on the space Y. For example, if $Y \in \mathbb{R}^n$ one tyically considers $\mathcal{L}$ to be the ℓ_1 or the ℓ_2. Another common choice is the *Huber error loss*, which "mixes" the two. It is defined for $\delta > 0$ as

$$\mathfrak{h}_\delta(r) = \begin{cases} \frac{1}{2}\|r\|_2^2, & \text{if } \|r\|_1 \le \delta \\ \delta\left(\|r\|_1 - \frac{\delta}{2}\right) & \text{otherwise.} \end{cases} \tag{6.9}$$

See Figure 6.1, left.

A special mention need to be done in the case for *classification problems*, i.e., problems where the function ψ assigns a discrete label (e.g., cat or dog) to the inputs. In this the relevant values for ψ are $\{0, 1\}$, say, but one approximate it with

functions taking value in $Y = \mathbb{R}$ and interprets values different from 0 or 1 as uncertain. In this case, the most used loss function is the *cross-entropy* function, which is defined by

$$\mathcal{L}(y, \hat{y}) = -\left[y \log \hat{y} + (1 - y) \log(1 - \hat{y})\right]. \tag{6.10}$$

This function has a probabilistic interpretation and encodes, roughly speaking, the "surprise" of seeing $\hat{y}$ after the model predicted y. See Figure 6.1, right.

Chapter 7

Feed-forward neural networks

In this chapter we introduce and discuss the simplest example of neural network[1]. These are obtained by concatenating the simplest possible nonlinear operations:

$$\phi_\theta(x) = \sigma(Wv + b), \qquad \text{where } \theta = (W, b) \in \mathbb{R}^{n\times m} \times \mathbb{R}^n. \tag{7.1}$$

Here, the function $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinear *activation* function that is assumed to be applied element-wise to the vector $Wx + b$. Typically activation functions are depicted in Figure 7.1.

Definition 7.0.1 One-hidden layer feed-forward ANN

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function and let $d_1 \in \mathbb{N}$. A scalar-valued one-hidden layer feed-forward ANN $\psi_\theta : \mathbb{R}^d \to \mathbb{R}$ is identified by the set of parameters

$$\Theta = \left\{(W_2, b_2, W_1, b_1) \in \mathbb{R}^{1\times d_1} \times \mathbb{R} \times \mathbb{R}^{d_1\times d} \times \mathbb{R}^{d_1}\right\}, \tag{7.2}$$

and it reads

$$\psi_\theta(x) = W_2\,\sigma(W_1 + b_1) + b_2 = \sum_{j=1}^{d_1} w_{2j}\sigma\left(\sum_{i=1}^{d} w_{ji}x_i + b_{1j}\right). \tag{7.3}$$

Here, the *architecture parameters* of the network are

- $d_1 \in \mathbb{N}$ is the *width* of the hidden layer;
- W_i it the *weight* matrix for the i-th layer;
- b_1 is the *bias* of the 1^{st} layer.

In Figure 7.2 we present an example of a one-hidden layer feed-forward ANN
The general expression for a feed-forward ANN can then be inferred, see Figure 7.3 for a graphical representation.

[1] We mostly follow [11], to which we refer for further clarifications.

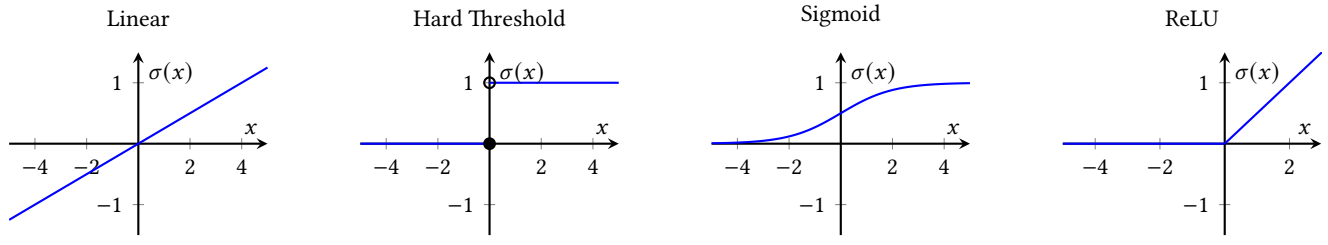

Figure 7.1: Activation functions

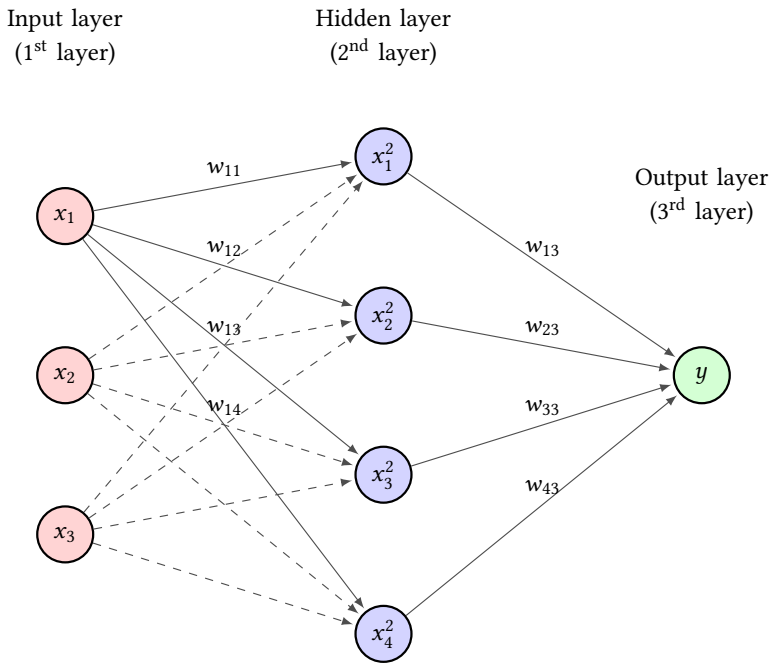

Figure 7.2: Graphical illustration of an single hidden layer ANN, yielding a function from $\mathbb{R}^3$ to $\mathbb{R}$. We only explicited the weights relative to the first variable.

Definition 7.0.2 Feed-forward fully-connected ANN

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function and let $d_1, \dots, d_L \in \mathbb{N}$. An N-hidden layer feed-forward ANN $\phi\theta : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$ is identified by the set of parameters

$$\Theta = \left\{(W_{L+1}, b_{L+1}, \dots, W_1, b_1) \mid W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}},\ b_\ell \in \mathbb{R}^{d_\ell}\right\}. \tag{7.4}$$

The corresponding function is then

$$\phi_\theta(x) = \phi^{L+1} \circ \phi^L \circ \cdots \phi^1(x), \tag{7.5}$$

where

$$\phi^\ell : x^{(\ell-1)} \in \mathbb{R}^{d_{\ell-1}} \mapsto x^{(\ell)} := \sigma(W_\ell x^{(\ell-1)} + b_\ell) \in \mathbb{R}^{d_\ell}, \qquad \forall \ell \in [\![1, L]\!] \tag{7.6}$$

$$\phi^{L+1} : x^{(L)} \in \mathbb{R}^{d_L} \mapsto x^{(L+1)} := W_L x^{(L)} \in \mathbb{R}^{d_{L+1}}. \tag{7.7}$$

Here, the *architecture parameters* of the network are

- $d_\ell \in \mathbb{N}$, $\ell \in [\![1, L]\!]$, is the *width* of the ℓ^{th} hidden layer;
- W_ℓ it the *weight* matrix for the ℓ^{th} layer;
- b_ℓ is the *bias* of the ℓ^{th} layer;
- x^ℓ collects the output of the *neurons* in the ℓ^{th} layer;

Neural networks of depth one are called *shallow*, if the depth is larger than one they are called *deep*.

Remark: In this chapter we focus on this type of neural networks, although many adjustments are possible (as we will see in Chapter 8). Some notable ones are:

- The activation functions σ can be different layer to layer, or even neuron to neuron.
- The output of layer ℓ could not only depend on layer $\ell - 1$, but also on any combination of the preceding layers from 0 to $\ell - 1$ (i.e., ϕ^ℓ can depend on $(x^{(0)}, \dots, x^{(\ell-1)})$). These are called *skip connections* and the corresponding networks are *residual*.

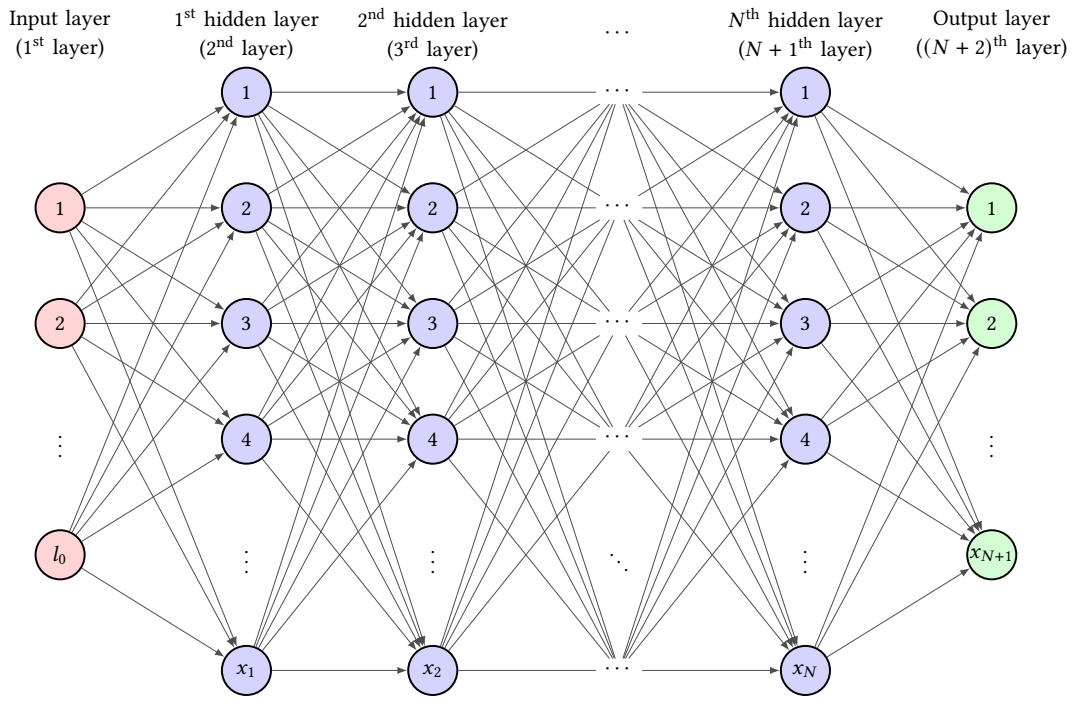

Figure 7.3: Graphical illustration of a fully-connected feedforward ANN consisting of $N + 2 \in \mathbb{N}$ affine transformations (i.e., consisting of $N + 1$ layers: one input layer, N hidden layers, and one output layer). Image from [8].

- Conversely, information could flow backwards, in the sense that layers $\ell - 1$ to $L + 1$ could serve as input for layer ℓ. This creates loops in the flow of information, and one has to introduce a time index $t \in \mathbb{N}$, as the output of a node at time step t might be different from the output at time step $t + 1$. These networks are called *recurrent*.

7.1 Learning and backpropagation

Once a loss function has been chosen, the data-fitting problem of Definition 6.0.1 can be solved using the algorithms presented in Chapter 3, along with their various extensions. All of these algorithms, however, require the computation of the gradient of the loss function, which in turn involves differentiating the function ϕ_θ from (8.2). This task is particularly challenging because ϕ_θ is defined as a composition of many functions. Even in the simplified case where σ is the identity (i.e., $\sigma(x) = x$), expanding (8.2) explicitly leads to a sum with an order of $d_1 \times \cdots \times d_N$ terms. In contemporary neural networks, the layer dimensions $d_1, \ldots, d_N$ are so large that explicitly computing such a sum is completely infeasible, which is why efficient methods like backpropagation are essential.

Assumption 7.1.1

Assume that the loss function $\mathcal{L} : \mathbb{R}^{d_{L+1}} \times \mathbb{R}^{d_{L+1}} \to \mathbb{R}$ is differentiable, and that a set $\mathfrak{X} = \{(x_i, y_i) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_{L+1}} \mid i \in [\![1, M]\!]\}$ is given.

Recall that our goal is to minimize the empirical risk

$$f(\theta) := \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(\phi_\theta(x_i), y_i), \tag{7.8}$$

over all possible neural network parameters $\theta = (W_{L+1}, b_{L+1}, \ldots, W_1, b_1)$. We thus need to find an efficient way to compute

$$\frac{\partial}{\partial b_\ell} \mathcal{L}(\phi_\theta(x), y) \in \mathbb{R}^{d_\ell} \qquad \text{and} \qquad \frac{\partial}{\partial W_\ell} \mathcal{L}(\phi_\theta(x), y) \in \mathbb{R}^{d_\ell \times d_{\ell-1}}. \tag{7.9}$$

We have the following.

Theorem 7.1.1 Backpropagation

Let $x, y \in \mathbb{R}^{d_0}$ be fixed and define $\bar{x}^{(\ell)}, \alpha^{(\ell)} \in \mathbb{R}^{d_\ell}$ by

$$\begin{cases} \bar{x}^{(1)} = W_1 x + b_1, \\ \bar{x}^{(\ell)} = W_\ell \sigma(\bar{x}^{\ell-1}) + b_\ell, \quad \ell \in [\![2, L+1]\!]. \end{cases} \qquad \begin{cases} \alpha^{(L+1)} = \frac{\partial}{\partial \bar{x}^{L+1}} \mathcal{L}(\bar{x}^{L+1}, y), \\ \alpha^{(\ell)} = \sigma'(\bar{x}^{\ell}) \odot (W_{\ell+1})^\top \alpha^{(\ell+1)}, \quad \ell \in [\![L, 1]\!]. \end{cases} \tag{7.10}$$

Here, we denoted by $\odot$ the elementwise product of vectors (i.e., $(v \odot w)_i = v_i w_i$).
Then, it holds

$$\frac{\partial}{\partial b_\ell} \mathcal{L}(\phi_\theta(x), y) = \alpha^{(\ell)}, \qquad \ell \in [\![1, L+1]\!], \tag{7.11}$$

$$\frac{\partial}{\partial W_1} \mathcal{L}(\phi_\theta(x), y) = \alpha^{(1)} x^\top, \qquad \text{and} \qquad \frac{\partial}{\partial W_\ell} \mathcal{L}(\phi_\theta(x), y) = \alpha^{(\ell)} \sigma(\bar{x}^{(\ell-1)})^\top, \qquad \ell \in [\![2, L+1]\!]. \tag{7.12}$$

Remark: The $(\bar{x}^\ell)$ defined in (7.10) satisfy $x^{(\ell)} = \sigma(\bar{x}^\ell)$ for $\ell \in [\![1, L]\!]$ and $x^{(L+1)} = \bar{x}^{(L+1)}$, where $x^{(\ell)} = \phi^\ell(x^{(\ell-1)})$ are the outputs of the neural network. For this reason, $\bar{x}^{(\ell)}$ are called the *preactivations*.

Remark: The above algorithm is called *backpropagation* since it allows to compute the gradients of the loss function via the quantities $\alpha^{(\ell)}$ that are "backpropagating" through the network: from the last layer $L+1$ to the 1^{st}.

We start by proving the following.

Lemma 7.1.2

It holds that

$$\alpha^{(\ell)} = \frac{\partial}{\partial \bar{x}^{(\ell)}} \mathcal{L}(\phi_\theta(x), y), \qquad \ell \in [\![1, L+1]\!] \tag{7.13}$$

Proof: The statement holds for $L+1$ by definition. Assume by induction that the statement holds for $\ell + 1 \leq L+1$ and compute by the chain rule

$$\frac{\partial \mathcal{L}}{\partial \bar{x}^{(\ell)}} = \left[\frac{\partial \bar{x}^{(\ell+1)}}{\partial \bar{x}^{(\ell)}}\right]^\top \frac{\partial \mathcal{L}}{\partial \bar{x}^{(\ell+1)}} = \left[\frac{\partial \bar{x}^{(\ell+1)}}{\partial \bar{x}^{(\ell)}}\right]^\top \alpha^{(\ell+1)}. \tag{7.14}$$

The statement follows by (7.10) and the direct computation

$$\left(\frac{\partial \bar{x}^{(\ell)}}{\partial \bar{x}^{(\ell+1)}}\right)_{ij} = (W_\ell)_{ij} \sigma'(\bar{x}^{(\ell-1)}). \tag{7.15}$$

■

Proof: We focus on proving the part concerning the derivative w.r.t. W_ℓ, the one w.r.t. b_ℓ being similar. Note that $\bar{x}^k$ depends only on (W_ℓ, b_ℓ) for $k \leq \ell$. Hence, the chain rule and Lemma 7.1.2 yield

$$\frac{\partial \mathcal{L}}{\partial W_\ell} = \frac{\partial \mathcal{L}}{\partial \bar{x}^{(\ell)}} \frac{\partial \bar{x}^\ell}{\partial W_\ell} = \alpha^{(\ell)} \frac{\partial \bar{x}^\ell}{\partial W_\ell}. \tag{7.16}$$

The proof follows by definition of the preactivations $\bar{x}^{(\ell)}$. Indeed, we have

$$\frac{\partial \bar{x}_k^{(1)}}{\partial (W_1)_{ij}} = \delta_{ki} x_j \quad \Longrightarrow \quad \frac{\partial \bar{x}^{(1)}}{\partial W_1} = x^\top, \tag{7.17}$$

$$\frac{\partial \bar{x}_k^{(\ell)}}{\partial (W_\ell)_{ij}} = \delta_{ki} \sigma(\bar{x}_j^{(\ell-1)}) \quad \Longrightarrow \quad \frac{\partial \bar{x}^{(\ell)}}{\partial W_\ell} = \sigma(\bar{x}^{(\ell-1)})^\top, \qquad \ell \in [\![2, L+1]\!]. \tag{7.18}$$

■

7.1.1 Issues with learning

We conclude this section by discussing two issues that might arise when training deep neural networks using backpropagation. These are the *vanishing gradient problem* and the *dead neuron problem*.

Vanishing gradient problem The vanishing gradient problem arises when the gradients $\alpha^{(\ell)}$ defined in (7.10) become very small as they are backpropagated through the layers of the network. This can happen when the activation function σ has a derivative σ' that is small in magnitude for a large portion of its domain (e.g., the *sigmoid function*). As a result, the product of these small derivatives can lead to gradients that are effectively zero, making it difficult for the optimization algorithm to update the weights in the earlier layers of the network. This can slow down or even halt the learning process, as the network struggles to adjust its parameters based on the loss function.

More precisely, if we assume that there exists $c \in (0,1)$ such that $|\sigma'(x)| \leq c$ for all $x \in \mathbb{R}$, then from (7.10) we have

$$\|\alpha^{(\ell)}\| \leq c^{L+1-\ell}\|W_{\ell+1}\| \cdots \|W_{L+1}\|\|\alpha^{(L+1)}\|. \tag{7.19}$$

Thus, if L is large and $c < 1$, then $\alpha^{(\ell)}$ can become very small, leading to the vanishing gradient problem: the gradients used to update the weights in the earlier layers of the network become so small that they effectively stop learning.

Dead neuron problem The dead neuron problem occurs when certain neurons in the network become inactive and stop contributing to the learning process. This can happen when the activation function σ outputs zero for a wide range of inputs, causing the gradients to be zero during backpropagation. As a result, the weights associated with these neurons do not get updated, leading to a situation where the neuron is effectively "dead" and does not participate in the network's computations. This issue is particularly common with activation functions like the *ReLU (Rectified Linear Unit)*, which outputs zero for all negative inputs. If a neuron consistently receives inputs that lead to zero outputs, it may never recover and continue to contribute to the network's learning.

More precisely, if there exists $a \in \mathbb{R}$ such that $\sigma'(x) = 0$ for all $x \leq a$, then any neuron receiving inputs that are always less than or equal to a will have zero gradients during backpropagation. Consequently, the weights associated with this neuron will not be updated, leading to a dead neuron that does not contribute to the network's output or learning process.

7.2 Universal approximation theorem

In this section we are interested in providing some theoretical results justifying the use of neural networks. More precisely, we show that (under certain assumptions) we can always find a neural network approximating a given continuous function on compact sets. To this aim we introduce the following definitions.

Definition 7.2.1 Universal approximator

A set of functions $\mathcal{H}$ form $\mathbb{R}^d$ to $\mathbb{R}$ is a *universal approximator* (of $C^0(\mathbb{R}^d)$) if for any $\varepsilon > 0$, $K \subset \mathbb{R}^d$ compact, and $f \in C^0(\mathbb{R}^d)$, there exists $g \in \mathcal{H}$ such that

$$\|f - g\|_K := \sup_{x \in K} |f(x) - g(x)| \leq \varepsilon. \tag{7.20}$$

Equivalently, $C^0(\mathbb{R}^d)$ is contained in the closure of $\mathcal{H}$ with respect to the convergence on compact sets.

Definition 7.2.2 Neural network spaces

Let $d, m, L, n \in \mathbb{N}$ and $\sigma : \mathbb{R} \to \mathbb{R}$. The set of all functions realized by ANNs $\phi_\theta : \mathbb{R}^d \to \mathbb{R}^m$, with at most depth L, width n, and activation function σ is

$$\mathcal{N}_d^m(\sigma; L, n). \tag{7.21}$$

Furthermore,

$$\mathcal{N}_d^m(\sigma; L) = \bigcup_{n \in \mathbb{N}} \mathcal{N}_d^m(\sigma; L, n). \tag{7.22}$$

7.2.1 Shallow neural networks

In this section, we establish the following result, which stresses why the nonlinear nature of the activation function is essential.

Theorem 7.2.1 Universal approximation for shallow neural networks

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an bounded piecewise continuous function[a]. Then, $\mathcal{N}_d^1(\sigma; 1)$ is a universal approximator of $C^0(\mathbb{R}^d)$ if and only if σ is not a polynomial.

[a]Here, we implicitly require that the discontinuity point be locally finite.

Remark: This theorem can be extended to different functional spaces and different norms.

The difficult part of this theorem (which holds for more general functions [10]) is to establish the universal approximation property for non-polynomial activation fucntions. The reverse implication is left as an exercice.

The result follows from the following three facts:

1. If $\mathcal{N}_1^1(\sigma; 1)$ is a universal approximator of $C^0(\mathbb{R})$, then $\mathcal{N}_d^1(\sigma; 1)$ is a universal approximator of $C^0(\mathbb{R}^d)$;
2. If $\sigma \in C^\infty(\mathbb{R})$ is not a polynomial, then $\mathcal{N}_1^1(\sigma; 1)$ is a universal approximator of $C^0(\mathbb{R})$;
3. If σ is bounded and piecewise continuous, functions in $\mathcal{N}_1^1(\sigma; 1)$ can approximate a smooth activation function $\tilde{\sigma}$ which is not a polynomial.

In particular, since $\mathcal{N}_1^1(\sigma; 1)$ is a universal approximator by the second point, it follows that the same is true for σ.

Let us start by showing the first point. An essential tool to establish universal approximation results is the Stone-Weirstrass Theorem, that we hereby recall.

Theorem 7.2.2 Stone-Weierstrass

Let $K \subset \mathbb{R}^d$ be compact, and $\mathcal{H} \subset C^0(K, \mathbb{R})$ be such that

- For all $x \in K$ there exists $f \in \mathcal{H}$ such that $f(x) \neq 0$;
- For all $x, y \in K$ there exists $f \in \mathcal{H}$ such that $f(x) \neq f(y)$;
- $\mathcal{H}$ is an algebra of functions, i.e., $\mathcal{H}$ is closed under addition, multiplication and scalar multiplication.

Then, $\mathcal{H}$ is dense in $C^0(K, \mathbb{R})$.
In particular, this holds when $\mathcal{H}$ is the space of real-valued polynomias and K is an arbitrary compact set.

Then, we have the following.

Lemma 7.2.3 Point 1 of the proof of Theorem 7.2.1

Assume that $\mathcal{H}$ is a universal approximator of $C^0(\mathbb{R})$. Let $d \in \mathbb{N}$, and define

$$\mathcal{X} = \text{span}\left\{x \mapsto g(w \cdot x) \mid w \in \mathbb{R}^d,\, g \in \mathcal{H}\right\}. \tag{7.23}$$

Then, $\mathcal{X}$ is a universal approximator of $C^0(\mathbb{R}^d)$.

Proof: Let us consider the space of k-homogeneous polynomials

$$\mathbb{H}_k = \{P : \mathbb{R}^d \to \mathbb{R} \mid P(x) = x^\alpha,\, \alpha \in \mathbb{N}_0^d,\, |\alpha| = k\}. \tag{7.24}$$

Then, by Stone-Weierstrass Theorem, it suffices to show that $\mathcal{X}$ is a universal approximator of $\mathbb{H}_k$ for all k.

Assume that this is not the case, namely that the closure $\bar{\mathcal{X}}$ of $\mathcal{X}$ w.r.t. compact convergence is a proper subset of $\mathbb{H}_k$ for some $k \in \mathbb{N}$. Then, by the Hahn-Banach Theorem, there exists a linear functional $\mathfrak{q} \in \mathbb{H}_k'$ such that $\mathfrak{q} \neq 0$ and $\ker \mathfrak{q} \supset \bar{\mathcal{X}}$. Let us show that this contradicts the universal approximation property of $\mathcal{H}$.

We start by claiming that the set $\{D^\alpha \mid |\alpha| = k\}$ is a basis for the topological dual $\mathbb{H}_k'$. Here, D^α is the derivative with multi-index α, and it is immediate to check that[2] $D^\alpha x^\beta = \delta_{\alpha,\beta}\alpha!$ for any $|\alpha| = |\beta| = k$. Since $p(x) = x^\beta$ for $|\beta| = k$ is a basis of $\mathbb{H}_k$, this proves the claim.

Consider now the k-homogeneous polynomial $p_w(x) = (w \cdot x)^k$ with $w \in \mathbb{R}^d$, and observe that it holds

$$p_w(x) = \left[\sum_{i=1}^d w_i x_i\right]^k = \sum_{|\alpha|=k} \frac{k!}{\alpha!} w^\alpha x^\alpha. \tag{7.25}$$

By the universal approximation property of $\mathcal{H}$ we have that $p_w \in \bar{\mathcal{X}}$ for any $w \in \mathbb{R}^d$. On the other hand, by the previous claim, we have that $\mathfrak{q} = \sum_{|\alpha|=k} q^\alpha D^\alpha$ for some $q^\alpha \in \mathbb{R}$, and thus by (7.25) we get

$$0 = \mathfrak{q}(p_w) = k! \sum_{|\alpha|=k} q^\alpha w^\alpha, \qquad \forall w \in \mathbb{R}^d \implies \mathfrak{q} = 0. \tag{7.26}$$

We have thus reached the desired contradiction. ∎

Lemma 7.2.4 Point 2 of the proof of Theorem 7.2.1

If $\sigma \in C^\infty(\mathbb{R})$ is not a polynomial, then $\mathcal{N}_1^1(\sigma; 1)$ is a universal approximator for $C^0(\mathbb{R})$.

Proof: Let $\mathcal{X} = \mathcal{N}_1^1(\sigma; 1)$. We proceed as in the proof of Lemma 7.2.3 and show that $\mathcal{X}$ is a universal approximator for the space of polynomials. This completes the proof by the Stone-Weierstrass Theorem.

Let $b \in \mathbb{R}$, and consider the function

$$f_b(x, w) := \sigma(wx + b), \qquad x, w \in \mathbb{R}. \tag{7.27}$$

Denoting ∂_w^k the k^{th} derivative w.r.t. the variable w, we have that

$$\partial_w^k f_b(x, w) = x^k \sigma^{(k)}(wx + b) \implies \partial_w^k f_b(x, w) = x^k \sigma^{(k)}(b). \tag{7.28}$$

Since $\sigma \in C^\infty(\mathbb{R})$ is not a polynomial, for any k there exsists $b_k \in \mathbb{R}$ such that $\sigma^{(k)}(b_k) \neq 0$. Thus, to prove the statement it suffices to show that $\partial_w^k f_b(\cdot, 0) \in \mathcal{X}$ for any b.

To this aim, consider the function $\phi_\theta \in \mathcal{X}$ given by

$$\phi_\theta(x) = \frac{1}{h}\sigma((w + h)x + b) - \frac{1}{h}\sigma(wx + b). \tag{7.29}$$

This is a one-hidde layer feed-forward ANN as per Definition 7.0.1 for the choice $d = 1$, $d_1 = 2$ and

$$W_1 = \begin{pmatrix} w \\ w \end{pmatrix}, \qquad b_1 = \begin{pmatrix} b \\ b \end{pmatrix}, \qquad W_2 = \begin{pmatrix} \frac{1}{h} & \frac{1}{h} \end{pmatrix}, \qquad b_2 = 0. \tag{7.30}$$

A simple Taylor development in w shows that for any $h > 0$ there exists $\xi \in [w, w + h]$ such that

$$\phi_\theta(x) = \partial_w f_b(x, w) + \frac{h}{2}\partial_w^2 f_b(x, \xi) = \partial_w f_b(x, w) + \frac{h}{2}x^2\sigma''(\xi x + b). \tag{7.31}$$

But σ'' is continous, and thus for any compact set $K \subset \mathbb{R}$ we have

$$\sup_{x\in K}\sup_{|h|<1} |x^2\sigma''(\xi x + b)| \leq \sup_{x\in K}\sup_{|\eta - w|<1} |x^2\sigma''(\eta x + b)| < +\infty. \tag{7.32}$$

In particular, letting $h \to 0$ we see that ϕ_θ converges uniformly on K to $\partial_w f_b(\cdot, w)$. Thus, $\partial_w f_b(\cdot, w) \in \mathcal{X}$. Applying the argument inductively, shows that $\partial_w^k f_b(\cdot, w) \in \mathcal{X}$, completing the proof. ∎

To conclude the proof, we need to recall the following concept.

[2] Recall that $\alpha! = \prod_{i=1}^d \alpha_j!$.

Definition 7.2.3 Convolution

Let $f, g \in L^\infty(\mathbb{R})$ and assume that g has compact support. Then the *convolution* of f and g is defined as

$$f \star g(x) := \int_{\mathbb{R}} f(x-y)g(y)\,dy. \tag{7.33}$$

Lemma 7.2.5 Point 3 of the proof of Theorem 7.2.1

Assume that $\sigma : \mathbb{R} \to \mathbb{R}$ is bounded and piecewise continous. Then, for any $\varphi \in C_c^\infty$ we have that $\sigma \star \varphi \in C_c^\infty(\mathbb{R})$ is in the closure of $\mathcal{N}_1^1(\sigma; 1)$ w.r.t. uniform convergence on compact sets.
Moreover, if $\sigma \star \varphi$ is a polynomial for any $\varphi \in C_c^\infty(\mathbb{R})$, then σ is a polynomial.

Proof: The fact that $\sigma \star \varphi \in C_c^\infty(\mathbb{R})$ is standard, and follows observing that $(\sigma \star \varphi)' = \sigma \star (\varphi')$. For the proof of the last part of the statement, we refer to [11, Lemma 3.15].

Let us build an explicit sequence $f_n \in \mathcal{N}_1^1(\sigma; 1)$ such that $f_n \to \sigma \star \varphi$ uniformly on compact sets. Let $a > 0$ be such that $\operatorname{supp} \varphi \subset [-a, a]$. Then, for any $n \in \mathbb{N}$ we let $y_j = -a + 2a/n$ and set

$$f_n(x) := \frac{1}{N} \sum_{j=0}^{N-1} \sigma(x - y_j)\varphi(y_j). \tag{7.34}$$

It follows easily that $f_n \in \mathcal{N}_1^1(\sigma; 1)$. Let us show the required convengerce on an arbitrary compact set $[-b, b]$, $b > 0$.

Since $y_{j+1} - y_j \le 1/n$, for any $x \in [-b, b]$ we have

$$|\sigma \star \varphi(x) - f_n(x)| \le \sum_{j=0}^{N-1} \left| \int_{y_j}^{y_{j+1}} \big(\sigma(x-y)\varphi(y) - \sigma(x-y_j)\varphi(y_j)\big)\,dy \right|. \tag{7.35}$$

Let us bound each term in the sum above.

Observe that there exists $C > 0$ such that

$$\big|\sigma(x-y)\varphi(y) - \sigma(x-y_j)\varphi(y_j)\big| \le C, \qquad \forall x, y \in \mathbb{R},\ j \in [\![0, n-1]\!]. \tag{7.36}$$

Then, for any $j \in [\![0, N-1]\!]$ we can estimate

$$\left| \int_{y_j}^{y_{j+1}} \big(\sigma(x-y)\varphi(y) - \sigma(x-y_j)\varphi(y_j)\big)\,dy \right| \le C(y_{j+1} - y_j) \le \frac{C}{n}. \tag{7.37}$$

Recall that, by assumption, there exists $\{z_1, \dots z_M\} \subset [-a, a]$ such that σ is continous on $[-a, a] \setminus \{z_1, \dots, z_M\}$. We will use the above estimate for the at most M intervals containing one of the points $x - z_i$, $i \in [\![1, M]\!]$. Let J_d be the set of such indexes and $J_c = [\![0, N-1]\!] \setminus J_d$.

Let us now fix $\varepsilon > 0$ and such that $\varepsilon \le \min\{\frac{|(x-z_i)-y_j|}{2} \mid i \in [\![1, M]\!] \text{ and } j \in [\![0, n]\!]\}$ and consider the compact set

$$K = [-a, a] \setminus \bigcup_{i=1}^{m} (z_i - \varepsilon, z_i + \varepsilon) \supset \bigcup_{j \in J_c} [y_j, y_{j+1}]. \tag{7.38}$$

Observe that σ is uniformly continuous on K. Then, for any $j \in J_c$ and $y \in [y_j, y_{j+1}]$ we have that $x - y \in K$ and $x - y_j \in K$. Thus,

$$\big|\sigma(x-y)\varphi(y) - \sigma(x-y_j)\varphi(y_j)\big| \le |\sigma(x-y) - \sigma(x-y_j)|\,|\varphi(y)| + |\sigma(x-y_j)|\,|\varphi(y) - \varphi(y_j)| \le \|\varphi\|_\infty \max_{\substack{s_1, s_2 \in K \\ |s_1 - s_2| \le \frac{2a}{n}}} |\sigma(s_1) - \sigma(s_2)| + \|sigma\|_\infty \max_{\substack{s_1, \\ |s_1}}$$

(7.39)

By uniform continuity of σ on K and of φ on $[-a, a]$ we have $\eta_n(x) \to 0$ as $n \to 0$ uniformly for $x \in [-b, b]$. Using (7.35) and (7.37), this shows that

$$|\sigma \star \varphi(x) - f_n(x)| \le \sum_{j \in J_d} \frac{C}{n} + \sum_{j \in J_c} \frac{\eta_n(x)}{n} \le \frac{CM}{n} + \eta_n(x) \to 0, \tag{7.40}$$

uniformly for $x \in [-b, b]$. ∎

Putting together the above results, we complete the proof of Theorem 7.2.1.

7.2.2 Deep neural networks

As a corollary to Theorem 7.2.1 we have the following.

Corollary 7.2.6 Universal approximation for deep neural netowrks

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an bounded piecewise continuous function. Then, for any $L \in \mathbb{N}$, $\mathcal{N}_d^1(\sigma; L)$ is a universal approximator of $C^0(\mathbb{R}^d)$ if and only if σ is not a polynomial.

The proof reduces to show that it is possible to approximate the identity with a deep network with $L-1$ layers. This is contained in the following.

Proposition 7.2.1

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be differentiable and non constant on an open set. The, for any $L \in \mathbb{N}$ and $\varepsilon > 0$, there exists a neural network $\phi \in \mathcal{N}_d^1(\sigma; L, d)$ such that

$$|\phi(x) - x| \leq \varepsilon, \qquad \forall x \in K. \tag{7.41}$$

Proof: We prove the statement for $L = 1$, the extension to $L > 1$ being straightforward. Let $s^\star \in \mathbb{R}$ be such that σ is differentiable in a neighborhood of $s^\star$ and $\varsigma := \sigma'(s^\star) \neq 0$. Let $b = (s^\star, \dots, s^\star) \in \mathbb{R}^d$ and, for $\lambda > 0$, define

$$\phi_\lambda(x) := \frac{\lambda}{\varsigma}\sigma\left(\frac{x}{\lambda} + b\right) - \frac{\lambda}{\varsigma}\sigma(b). \tag{7.42}$$

It is clear that $\phi_\lambda \in \mathcal{N}_d^1(\sigma; 1, d)$. Moreover, a Taylor development easily shows that

$$\phi_\lambda(x) - x = \lambda\frac{\sigma(x/\lambda + b) - \sigma(b)}{\varsigma} - x \to 0, \qquad \text{as } \lambda \to +\infty. \tag{7.43}$$

This completes the proof. ■

Proof of Corollary 7.2.6: We already know that $\mathcal{N}_d^1(\sigma; 1)$ is an universal approximator. That is, for any $f \in C(\mathbb{R}^d)$, for any K compact and any $\varepsilon > 0$ there exists a shallow netowrk $\phi_1 \in \mathcal{N}_d^1(\sigma; 1)$ such that

$$\|f - \phi_1\|_K \leq \frac{\varepsilon}{2}. \tag{7.44}$$

Concatenanting this with the deep network approximating the identity given by Proposition 7.2.1 completes the proof. ■

7.2.3 Further results

We conclude this chapter by mentioning the following remarkable result stating that, for an appropriate activation function, it is possible to approximate every function $f \in C(K)$, where $K \subset \mathbb{R}^d$, at every accuracy $\varepsilon > 0$ with a neural network of size $O(d^2)$. Remarkably, this size is independent of f, of K, and even of ε. We refer to [11, Section 3.2] for a proof, based on the Kolmogorov's superposition theorem.

Theorem 7.2.7 Universal approximation via superexpressive activations

There exists a continuous activation function $\sigma : \mathbb{R} \to \mathbb{R}$ such that for every compact $K \subset \mathbb{R}^d$, every $\varepsilon > 0$, and every $f \in C(K)$, there exists[a] $\phi \in \mathcal{N}_d^1(\sigma; 2, 2d^2 + d)$ such that

$$|f(x) - \phi(x)| \leq \varepsilon, \qquad \forall x \in K. \tag{7.45}$$

[a] That is, width $\phi = 2d^2 + 2$ and depth $\phi = 2$

Chapter 8

Residual neural networks, neural ODEs and control

In Section 7.1.1 we discussed some issues that might arise when training deep neural networks using backpropagation. To overcome these issues, a popular approach consists in considering *residual neural networks* (ResNets) [7]. These are defined as follows.

Definition 8.0.1 Residual neural network

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function and let $d \in \mathbb{N}$. An N-hidden layer ResNet $\Phi_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is identified by the set of parameters

$$\Theta = \left\{ (W_\ell^{(1)}, b_\ell^{(1)}, W_\ell^{(2)}, b_\ell^{(2)})_{\ell=1}^{L+1} \mid W_\ell^{(i)} \in \mathbb{R}^{d\times d},\ b_\ell^{(i)} \in \mathbb{R}^d, i = 1, 2 \right\}. \tag{8.1}$$

The corresponding function is then

$$\Phi_\theta(x) = \Phi^{L+1} \circ \Phi^L \circ \cdots \Phi^1(x), \tag{8.2}$$

where, each layer $\Phi^\ell : \mathbb{R}^d \to \mathbb{R}^d$ is given by

$$\Phi^\ell(x) : x^{(\ell-1)} \in \mathbb{R}^d \mapsto x^{(\ell-1)} + \phi^\ell(x^{(\ell-1)}) \in \mathbb{R}^d, \qquad \ell \in [\![1, L]\!], \tag{8.3}$$

for some standard 2-layer feed-forward ANN $\phi^\ell : \mathbb{R}^d \to \mathbb{R}^d$ defined by

$$\phi^\ell(x) = W_\ell^{(2)} \sigma(W_\ell^{(1)} x + b_\ell^{(1)}) + b_\ell^{(2)}. \tag{8.4}$$

The advantage of ResNets is that the skip connections (i.e., the addition of the input $x^{(\ell-1)}$ to the output of the layer ϕ^ℓ) help in mitigating issues like vanishing gradients during training evoked in Section 7.1.1. Indeed, adapting the back-propagation algorithm to ResNets, one can see that the gradients $\alpha^{(\ell)}$ defined in (7.10) satisfy

$$\alpha^{(\ell)} = \alpha^{(\ell+1)} + \left[\frac{\partial \phi^{\ell+1}}{\partial x^{(\ell)}} \right]^\top \alpha^{(\ell+1)}, \qquad \ell \in [\![1, L]\!], \tag{8.5}$$

which prevents them from vanishing too quickly.

8.1 Control formalism via neural ODEs

The key observation in [3, 6] is that ResNets can be interpreted as a forward Euler discretization of an ordinary differential equation (ODE). In particular, letting $u = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$ denote

$$F(x, u) := W^{(2)} \sigma(W^{(1)} x + b^{(1)}) + b^{(2)}. \tag{8.6}$$

Then, fixing $\eta = 1$, the ResNet defined by the parameters $\theta = \{(W_\ell^{(1)}, b_\ell^{(1)}, W_\ell^{(2)}, b_\ell^{(2)})\}_{\ell=1}^{L+1}$ can be rewritten as

$$x^{(\ell)} = x^{(\ell-1)} + \eta F(x^{(\ell-1)}, u_\ell), \qquad \ell \in [\![1, L]\!], \tag{8.7}$$

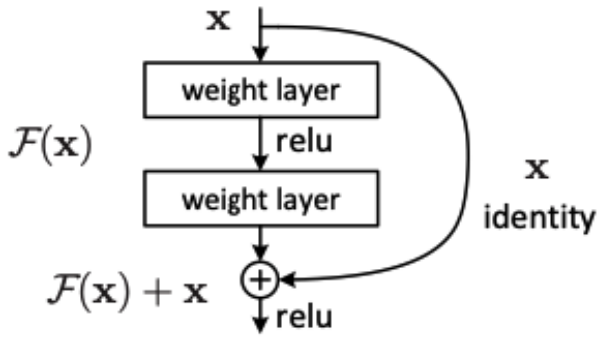

Figure 8.1: Schematic representation of a residual neural network layer. The input $x^{(\ell-1)}$ is transformed by a function $\mathcal{F}_\ell$ and then added to the original input to produce the output $x^{(\ell)}$. Image from [7]

where $u_\ell = (W_\ell^{(1)}, b_\ell^{(1)}, W_\ell^{(2)}, b_\ell^{(2)})$. Rearranging the terms, we have

$$\frac{x^{(\ell)} - x^{(\ell-1)}}{\eta} = F(x^{(\ell-1)}, u_\ell), \qquad \ell \in [\![1, L]\!]. \tag{8.8}$$

Thus, the ResNet can be seen as the Euler discretization of the ODE

$$\dot{x}(t) = F(x(t), u(t)), \qquad t \in [0, L], \tag{8.9}$$

with initial condition $x(0) = x \in \mathbb{R}^{d_0}$ and control u defined as $u(t) = u_\ell$ for $t \in [(\ell-1)\eta, \ell\eta)$. Rescaling the time variable as $t \mapsto t/L$ and letting $L \to +\infty$ we then obtain the continuous-time limit

$$\dot{x}(t) = F(x(t), u(t)), \qquad t \in [0, 1]. \tag{8.10}$$

This observation allows to interpret ResNets as *neural ODEs*, i.e., ODEs whose dynamics is parametrized by a neural network. More precisely, we have the following.

Definition 8.1.1 Neural ODEs control systems

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function, and let $F : \mathbb{R}^d \times U \to \mathbb{R}^d$ be defined as[a]

$$F(x, u) = W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)}, \tag{8.11}$$

for $u = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}) \in U$, where U is a suitable subset of $\mathbb{R}^{d_1 \times d} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d \times d_1} \times \mathbb{R}^d$.
A *neural ODE* is the controlled ODE

$$\dot{x}(t) = F(x(t), u(t)), \qquad t \in [0, 1], \tag{8.12}$$

with initial condition $x(0) = x \in \mathbb{R}^d$ and control $u \in L^\infty([0, 1]; U)$.
The *flow* associated to the neural ODE is the map $\Phi_u : \mathbb{R}^d \to \mathbb{R}^d$ defined as

$$\Phi_u(x) := x_u(1), \tag{8.13}$$

where x_u is the solution of the neural ODE with control u and initial condition x.

[a] Here, one could of course consider different and more general architectures for F.

We summarize the previous observation in the following result.

Theorem 8.1.1 Neural ODEs

Then, for any control $u \in L^\infty([0, 1]; U)$ and $\varepsilon > 0$ there exists a sufficiently small $\eta > 0$ such that the solution x_u of the neural ODE satisfies

$$|x_u(\ell/L) - x^{(\ell)}| \le \varepsilon, \qquad \forall \ell \in [\![1, L]\!], \tag{8.14}$$

where $x^{(\ell)}$ are the layers outputs of the ResNet defined by

$$x^{(\ell)} = x^{(\ell-1)} + \frac{1}{L} F(x^{(\ell-1)}, u(\ell-1)), \qquad \ell \in [\![1, L]\!]. \tag{8.15}$$

The above result says that a ResNet with $L \gg 1$ layers can be seen as the approximation of the flow of a neural ODE, that is,

$$\Phi_u \approx \Phi^L \circ \Phi^{L-1} \circ \cdots \circ \Phi^1. \tag{8.16}$$

The advantage of the neural ODE formalism is that it allows to leverage the rich theory of optimal control to study and train deep neural networks. In particular, we have that

- the *expressivity* of the ResNet can be studied via controllability properties of the corresponding neural ODE,
- the *training* of a ResNet with L layers can be seen as the optimal control problem for the corresponding neural ODE with piecewise constant controls with L pieces.

Since the control system given by the neural ODE is non-linear, we cannot apply directly the results of Part II. The results in the following sections discuss recent advances in this direction and are based on geometric control theory; we refer to [1, 15] for more details.

8.1.1 Training via optimal control

Let us consider the training problem detailed in Definition 6.0.1, for a set of data $\mathfrak{X} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$. Using the neural ODE formalism, the training of a ResNet with L layers is related to the following optimal control problem

$$\text{Find } u^\star \in L^\infty([0,1]; U) \text{ such that } u^\star = \arg\min_{u \in L^\infty([0,1];U)} \frac{1}{N} \sum_{i=1}^N \ell(\Phi_u(x_i), y_i), \tag{8.17}$$

where Φ_u is the flow of the neural ODE defined in Definition 8.1.1.

Under mild assumptions on $\mathcal{L}$ (namely that $\mathcal{L}(x, x) = 0$ for any $x \in \mathbb{R}^d$), we have that the solution to the above is obtained at zero cost if the following controllability property holds.

Definition 8.1.2 Simultaneous controllability

The neural ODE defined in Definition 8.1.1 is said to be *simultaneously controllable* on a set $\Omega \subset \mathbb{R}^d$ if for any $N \in \mathbb{N}$, any set of points $\{x_1, \dots, x_N\} \subset \Omega$ and any set of target points $\{y_1, \dots, y_N\} \subset \Omega$, there exists a control $u \in L^\infty([0,1]; U)$ such that the corresponding flow Φ_u satisfies

$$\Phi_u(x_i) = y_i, \qquad \forall i \in [\![1, N]\!]. \tag{8.18}$$

Observe that this definition requires to be able to steer *simultaneously* all points x_i to the corresponding points y_i using the same control u, and is thus stronger than the standard notion of controllability of Definition 4.1.3.

A result in this direction has been recently obtained in [15], where the authors prove simultaneous controllability for neural ODEs with specific activation functions.

Theorem 8.1.2 Simultaneous controllability for generic inputs

Assume that the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is such that

- σ is Lipschitz continuous and such that $\sigma'(s) \geq 0$;
- there exists $j \in \mathbb{N}_0$, $a_1, a_2, a_3 \in \mathbb{R}$, $a_2 \neq 0$, such that $\eta(s) := \sigma^{(j)}(s)$ is injective and such that

$$\eta'(s) = a_1 + a_2\eta(s) + a_3\eta(s)^2, \qquad \forall s \in \mathbb{R}. \tag{8.19}$$

Then, the neural ODE defined in Definition 8.1.1 is simultaneously controllable on the set

$$\Omega = \{x \in \mathbb{R}^d \mid x_i \neq x_j, \forall i \neq j\}. \tag{8.20}$$

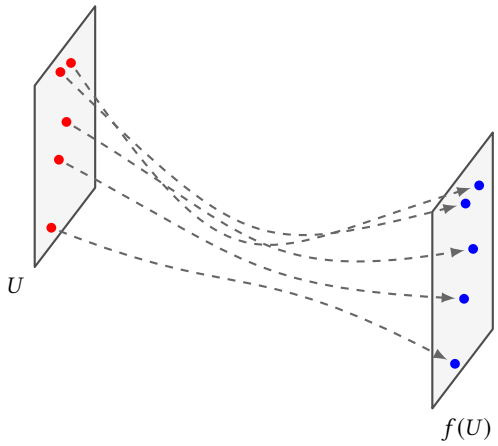

Figure 8.2: Schematic representation of simultaneous controllability. The same control u is used to steer all points x_i to the corresponding points y_i.

The assumptions on the activation function are satisfied, for example, by the sigmoid function $\sigma(s) = 1/(1+e^{-s})$ and by the hyperbolic tangent $\sigma(s) = \tanh(s)$.

This approach has two problems, however:

- The controllability result does not really provide a constructive (or easily implementable) way to compute the control u steering the points x_i to the points y_i;
- Striving for exactly zero training error might lead to overfitting issues, which is known to lead to poor generalization properties on unseen data.

To this end, an alternative approach studied in [14] consists in considering a relaxed version of the simultaneous controllability problem, where one looks for a control u minimizing the cost functional

$$J(u) = \frac{1}{N}\sum_{i=1}^{N} \|\Phi_u(x_i) - y_i|^2 + \frac{\beta}{2}\|u(t)\|_2^2. \tag{8.21}$$

8.1.2 Expressivity via controllability

Observe, that $u \mapsto \Phi_u$ defines a family of diffeomorphisms of $\mathbb{R}^d$ parametrized by the controls u. Since by Theorem 8.1.1 ResNets approximate the flow of neural ODEs, it is natural to study the expressivity of ResNets via the properties of the family of diffeomorphisms $\{\Phi_u \mid u \in L^\infty([0,1];U)\}$. More precisely, we have the following immediate consequence of Theorem 8.1.1.

Proposition 8.1.1

Let $\mathrm{ResNet}^d(\sigma; m)$ denote the set of all ResNets with activation function $\sigma : \mathbb{R} \to \mathbb{R}$, input/output dimension d, and width at most m. Then, the closure of $\mathrm{ResNet}^d(\sigma; d)$ w.r.t. uniform convergence on compact sets contains the set

$$\{\Phi_u \mid u \in L^\infty([0,1];U)\}. \tag{8.22}$$

In particular, if for any compact $K \subset \mathbb{R}^d$ the family $\{\Phi_u \mid u \in L^\infty([0,1];U)\}$ is dense in $C^0(K, \mathbb{R}^d)$, then $\mathrm{ResNet}^d(\sigma; d)$ is a universal approximator of $C^0(\mathbb{R}^d, \mathbb{R}^d)$.

Observe that the density of the family $\{\Phi_u \mid u \in L^\infty([0,1];U)\}$ in $C^0(K, \mathbb{R}^d)$ is a much stronger assumption that the notion of controllability defined in Definition 4.1.3 and the one of simultaneous controllability defined above. Indeed, here we need to be able to find a single control that steers any point $x \in K$ to any target point $y := f(x)$, for a given continuous function $f : K \to \mathbb{R}^d$.

The result we present below requires the following assumption on f.

Definition 8.1.3 Monotone analytic homotopy

A function $f : \mathbb{R}^d \to \mathbb{R}^d$ is monotone homotopic to the identity if there exists an analytic homotopy[a] $h : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ such that, for any $t \in [0,1]$, the map $h(t,\cdot)$ is monotone, i.e.,

$$x_i \leq y_i \qquad \forall i \in [\![1,d]\!] \implies h(t,x)_i \leq h(t,y)_i \qquad \forall i \in [\![1,d]\!]. \tag{8.23}$$

[a] That is, h is analytic and such that $h(0,x) = x$ and $h(1,x) = f(x)$.

We then have the following result in [15].

Theorem 8.1.3 Universal approximation of functions homotopic to the identity

Assume that σ satisfies the assumptions of Theorem 8.1.2. Then, for any compact set $K \subset \mathbb{R}^d$ and any continuous function $f : K \to \mathbb{R}^d$ that is monotone analytic homotopic to the identity, for any $\varepsilon > 0$ there exists a control $u \in L^\infty([0,1];U)$ such that the corresponding flow Φ_u satisfies

$$\|f - \Phi_u\|_K \leq \varepsilon. \tag{8.24}$$

In particular, $\mathrm{ResNet}^d(\sigma;d)$ is a universal approximator of the set of functions that are monotone analytic homotopic to the identity on compact sets.

The above theorem can be extended to general continuous functions assuming that the initial and final layers have different dimensions, i.e., $d_0, d_{L+1} \neq d$.

Theorem 8.1.4

Assume that σ satisfies the assumptions of Theorem 8.1.2 and let d' be such that $d = 2d' + 1$. Then, for any compact set $K \subset \mathbb{R}^d$ and any continuous function $f : K \to \mathbb{R}^{d'}$, for any $\varepsilon > 0$ there exists a control $u \in L^\infty([0,1];U)$, an injection $\alpha : \mathbb{R}^{d'} \to \mathbb{R}^{2d'+1}$, and a projection $\beta : \mathbb{R}^{2d'+1} \to \mathbb{R}^{d'}$, such that

$$\|f - \beta \circ \Phi_u \circ \alpha\|_K \leq \varepsilon. \tag{8.25}$$

In particular, letting $\mathrm{ResNet}^{d'}(\sigma; 2d'+1)$ denote the set of all ResNets with activation function $\sigma : \mathbb{R} \to \mathbb{R}$, input/output dimension d', and width at most $2d'+1$, we have that $\mathrm{ResNet}^{d'}(\sigma; 2d'+1)$ is a universal approximator of $C^0(\mathbb{R}^{d'}, \mathbb{R}^{d'})$.

Bibliography

[1] Andrei Agrachev and Andrey Sarychev. "Control on the Manifolds of Mappings with a View to the Deep Learning". In: *Journal of Dynamical and Control Systems* 28.4 (Oct. 2022), pp. 989–1008. ISSN: 1079-2724, 1573-8698. DOI: 10.1007/s10883-021-09561-2. URL: https://link.springer.com/10.1007/s10883-021-09561-2 (visited on 11/10/2025).

[2] 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.

[3] Ricky T. Q. Chen et al. "Neural Ordinary Differential Equations". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paperfiles/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

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

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

[6] Eldad Haber and Lars Ruthotto. "Stable Architectures for Deep Neural Networks". In: *Inverse Problems* 34.1 (Jan. 1, 2018), p. 014004. ISSN: 0266-5611, 1361-6420. DOI: 10.1088/1361-6420/aa9a90. URL: https://iopscience.iop.org/article/10.1088/1361-6420/aa9a90 (visited on 11/12/2025).

[7] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, June 2016, pp. 770–778. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.90. URL: http://ieeexplore.ieee.org/document/7780459/ (visited on 11/12/2025).

[8] Arnulf Jentzen, Benno Kuckuck, and Philippe von Wurstemberger. *Mathematical Introduction to Deep Learning: Methods, Implementations, and Theory*. Version 3. 2023. DOI: 10.48550/ARXIV.2310.20360. URL: https://arxiv.org/abs/2310.20360 (visited on 11/03/2025). Pre-published.

[9] Velimir Jurdjevic. *Geometric Control Theory*. Cambridge Studies in Advanced Mathematics 51. Cambridge: Cambridge university press, 1997. ISBN: 978-0-521-49502-8.

[10] Moshe Leshno et al. "Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function". In: *Neural Networks* 6.6 (Jan. 1993), pp. 861–867. ISSN: 08936080. DOI: 10.1016/S0893-6080(05)80131-5. URL: https://linkinghub.elsevier.com/retrieve/pii/S0893608005801315 (visited on 11/04/2025).

[11] Philipp Petersen and Jakob Zech. *Mathematical Theory of Deep Learning*. Apr. 7, 2025. DOI: 10.48550/arXiv.2407.18384. arXiv: 2407.18384 [cs]. URL: http://arxiv.org/abs/2407.18384 (visited on 11/04/2025). Pre-published.

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

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

[14] Alessandro Scagliotti. "Ensembles of Affine-Control Systems with Applications to Deep Learning". Trieste, Italy: Scuola Internazionale Superiore di Studi Avanzati, 2022.

[15] Paulo Tabuada and Bahman Gharesifard. "Universal Approximation Power of Deep Residual Neural Networks Through the Lens of Control". In: *IEEE Transactions on Automatic Control* 68.5 (May 2023), pp. 2715–2728. ISSN: 0018-9286, 1558-2523, 2334-3303. DOI: 10.1109/TAC.2022.3190051. URL: https://ieeexplore.ieee.org/document/9827563/ (visited on 11/03/2025).

[16] 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).

[17] Emmanuel Trélat. *Contrôle optimal: théorie & applications*. Paris: Vuibert, 2005. ISBN: 978-2-7117-7175-2.