

# Optimization and Optimal Control in Banach Spaces

Bernhard Schmitzer

December 7, 2018

## 1 Convex non-smooth optimization with proximal operators

**Remark 1.1** (Motivation). Convex optimization:

- easier to solve, global optimality,
- convexity is strong regularity property, even if functions are not differentiable, even in infinite dimensions,
- usually strong duality,
- special class of algorithms for non-smooth, convex problems; easy to implement and to parallelize. Objective function may assume value  $+\infty$ , i.e. well suited for implementing constraints.

So if possible: formulate convex optimization problems.

Of course: some phenomena can only be described by non-convex problems, e.g. formation of transport networks.

**Definition 1.2.** Throughout this section  $H$  is Hilbert space, possibly infinite dimensional.

### 1.1 Convex sets

**Definition 1.3** (Convex set). A set  $A \subset H$  is convex if for any  $a, b \in A$ ,  $\lambda \in [0, 1]$  one has  $\lambda \cdot a + (1 - \lambda) \cdot b \in A$ .

---

Comment: Line segment between any two points in  $A$  is contained in  $A$

---

**Sketch:** Positive example with ellipsoid, counterexample with ‘kidney’

---

Comment: Study of geometry of convex sets is whole branch of mathematical research. See lecture by Prof. Wirth in previous semester for more details. In this lecture: no focus on convex sets, will repeat all relevant properties where required.

---

**Proposition 1.4** (Intersection of convex sets). If  $\{C_i\}_{i \in I}$  is family of convex sets, then  $C := \bigcap_{i \in I} C_i$  is convex.

*Proof.* • Let  $x, y \in C$  then for all  $i \in I$  have  $x, y \in C_i$ , thus  $\lambda \cdot x + (1 - \lambda) \cdot y \in C_i$  for all  $\lambda \in [0, 1]$  and consequently  $\lambda \cdot x + (1 - \lambda) \cdot y \in C$ .  $\square$

**Definition 1.5** (Convex hull). The *convex hull*  $\text{conv } C$  of a set  $C$  is the intersection of all convex sets that contain  $C$ .

**Proposition 1.6.** Let  $C \subset H$ , let  $T$  be the set of all convex combinations of elements of  $C$ , i.e.,

$$T := \left\{ \sum_{i=1}^k \lambda_i x_i \mid k \in \mathbb{N}, x_1, \dots, x_k \in C, \lambda_1, \dots, \lambda_k > 0, \sum_{i=1}^k \lambda_i = 1 \right\}.$$

Then  $T = \text{conv } C$ .

*Proof.* • **Part I,  $\text{conv } C \subset T$ :**  $T$  is convex: any  $x, y \in T$  are (finite) convex combinations of points in  $C$ . Thus, so is any convex combination of  $x$  and  $y$ . Also,  $C \subset T$ . So  $\text{conv } C \subset T$ .

- **Part II,  $\text{conv } C \supset T$ :** Let  $S$  be convex and  $S \supset C$ . We will show that  $S \supset T$  and thus  $\text{conv } C \supset T$ , which with the previous step implies equality of the two sets.
- We show  $S \supset T$  by induction. By definition, any element in  $T$  can be represented as follows: For some  $k \in \mathbb{N}$ ,  $x_1, \dots, x_k \in C$ ,  $\lambda_1, \dots, \lambda_k > 0$ ,  $\sum_{i=1}^k \lambda_i = 1$  let

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

- When  $k = 1$  clearly  $s_k = x_1 \in C \subset S$ .
- Assume, we have shown that all linear combinations up to  $k - 1$  elements in  $T$  are also contained in  $S$ .
- For  $k > 1$  set  $\tilde{\lambda}_i = \lambda_i / (1 - \lambda_k)$  for  $i = 1, \dots, k - 1$ . Then

$$s_k = \lambda_k x_k + (1 - \lambda_k) \cdot \underbrace{\sum_{i=1}^{k-1} \tilde{\lambda}_i x_i}_{:= s_{k-1}}.$$

- We have  $x_k \in C \subset S$  and by assumption  $s_{k-1} \in S$ . Therefore,  $s_k \in S$ . □

**Proposition 1.7** (Carathéodory). Let  $H = \mathbb{R}^n$ . Every  $x \in \text{conv } C$  can be written as convex combination of at most  $n + 1$  elements of  $C$ .

*Proof.* Consider arbitrary convex combination  $x = \sum_{i=1}^k \lambda_i x_i$  for  $k > n + 1$ .

Claim: without changing  $x$  can change  $(\lambda_i)_i$  such that one  $\lambda_i$  becomes 0.

- The vectors  $\{x_2 - x_1, \dots, x_k - x_1\}$  are linearly dependent, since  $k - 1 > n$ .
- $\Rightarrow$  There are  $(\beta_2, \dots, \beta_k) \in \mathbb{R}^{k-1} \setminus \{0\}$  such that

$$0 = \sum_{i=2}^k \beta_i (x_i - x_1) = \sum_{i=2}^k \beta_i x_i - \underbrace{\sum_{i=2}^k \beta_i x_1}_{:= -\beta_1}.$$

- Define  $\tilde{\lambda}_i = \lambda_i - t^* \beta_i$  for  $t^* = \frac{\lambda_{i^*}}{\beta_{i^*}}$  and  $i^* = \text{argmin}_{i=1, \dots, k: \beta_i \neq 0} \frac{\lambda_i}{|\beta_i|}$ .

- $\tilde{\lambda}_i \geq 0$ :  $\tilde{\lambda}_i = \lambda_i \cdot \underbrace{\left(1 - \frac{\lambda_{i^*}/\beta_{i^*}}{\lambda_i/\beta_i}\right)}_{|\cdot| \leq 1}$
- $\tilde{\lambda}_{i^*} = 0$
- $\sum_{i=1}^k \tilde{\lambda}_i = \underbrace{\sum_{i=1}^k \lambda_i}_{=1} - t^* \underbrace{\sum_{i=1}^k \beta_i}_{=0} = 1$
- $\sum_{i=1}^k \tilde{\lambda}_i x_i = \underbrace{\sum_{i=1}^k \lambda_i x_i}_{=x} - t^* \underbrace{\sum_{i=1}^k \beta_i x_i}_{=0} = x$

□

## 1.2 Convex functions

**Definition 1.8** (Convex function). A function  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  is convex if for all  $x, y \in H$ ,  $\lambda \in [0, 1]$  one has  $f(\lambda \cdot x + (1 - \lambda) \cdot y) \leq \lambda \cdot f(x) + (1 - \lambda) \cdot f(y)$ . Set of convex functions over  $H$  is denoted by  $\text{Conv}(H)$ .

- $f$  is *strictly convex* if for  $x \neq y$  and  $\lambda \in (0, 1)$ :  $f(\lambda \cdot x + (1 - \lambda) \cdot y) < \lambda \cdot f(x) + (1 - \lambda) \cdot f(y)$ .
- $f$  is *concave* if  $-f$  is convex.
- The *domain* of  $f$ , denoted by  $\text{dom } f$  is the set  $\{x \in H | f(x) < +\infty\}$ .  $f$  is called *proper* if  $\text{dom } f \neq \emptyset$ .
- The *graph* of  $f$  is the set  $\{(x, f(x)) | x \in \text{dom } f\}$ .
- The *epigraph* of  $f$  is the set ‘above the graph’,  $\text{epi } f = \{(x, r) \in H \times \mathbb{R} | r \geq f(x)\}$ .
- The *sublevel set* of  $f$  with respect to  $r \in \mathbb{R}$  is  $S_r(f) = \{x \in H | f(x) \leq r\}$ .

---

**Sketch:** Strictly convex, graph, secant, epigraph, sublevel set

---

**Proposition 1.9.** (i)  $f$  convex  $\Rightarrow \text{dom } f$  convex.

(ii)  $[f \text{ convex}] \Leftrightarrow [\text{epi } f \text{ convex}]$ .

(iii)  $[(x, r) \in \text{epi } f] \Leftrightarrow [x \in S_r(f)]$ .

**Example 1.10.** (i) *characteristic or indicator function* of convex set  $C \subset H$ :

$$\iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{else.} \end{cases} \quad \text{Do not confuse with} \quad \chi_C(x) = \begin{cases} 1 & \text{if } x \in C \\ 0 & \text{else.} \end{cases}$$

(ii) any *norm* on  $H$  is convex: For all  $x, y \in H$ ,  $\lambda \in [0, 1]$ :

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

(iii) for  $H = \mathbb{R}^n$  the *maximum function*

$$\mathbb{R}^n \ni x \mapsto \max\{x_i | i = 1, \dots, n\}$$

is convex (follows from previous point, since it is also a norm).

(iv) *linear and affine functions* are convex.

**Example 1.11** (Optimization with constraints). Assume we want to solve an optimization problem with linear constraints, e.g.,

$$\min\{f(x) | x \in \mathbb{R}^n, Ax = b\}$$

where  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ ,  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$ . This can be formally rewritten as unconstrained problem:

$$\min\{f(x) + g(Ax) | x \in \mathbb{R}^n\} \quad \text{where} \quad g = \iota_{\{b\}}.$$

We will later discuss algorithms that are particularly suited for problems of this form where one only has to ‘interact’ with  $f$  and  $g$  separately, but not their combination.

As mentioned in the motivation: convexity is a strong regularity property. Here we give some examples of consequences of convexity.

**Definition 1.12.** A function  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  is (sequentially) continuous in  $x$  if for every convergent sequence  $(x_k)_k$  with limit  $x$  one has  $\lim_{k \rightarrow \infty} f(x_k) = f(x)$ . The set of points  $x$  where  $f(x) \in \mathbb{R}$  and  $f$  is continuous in  $x$  is denoted by  $\text{cont } f$ .

**Remark 1.13** (Continuity in infinite dimensions). If  $H$  is infinite dimensional, it is a priori not clear, whether closedness and sequential closedness coincide. But since  $H$  is a Hilbert space, it has an inner product, which induces a norm, which induces a metric. On metric spaces the notions of closedness and sequential closedness coincide and thus so do the corresponding notions of continuity.

**Proposition 1.14** (On convexity and continuity I). Let  $f \in \text{Conv}(H)$  be proper and let  $x_0 \in \text{dom } f$ . Then the following are equivalent:

- (i)  $f$  is locally Lipschitz continuous near  $x_0$ .
- (ii)  $f$  is bounded on a neighbourhood of  $x_0$ .
- (iii)  $f$  is bounded from above on a neighbourhood of  $x_0$ .

*Proof.* The implications (i)  $\Rightarrow$  (ii)  $\Rightarrow$  (iii) are clear. We show (iii)  $\Rightarrow$  (i).

- If  $f$  is bounded from above in an environment of  $x_0$  then there is some  $\rho \in \mathbb{R}_{++}$  such that  $\sup f(\overline{B(x_0, \rho)}) = \eta < +\infty$ .
- Let  $x \in H$ ,  $x \neq x_0$ , such that  $\alpha := \|x - x_0\|/\rho \in (0, 1]$

---

**Sketch:** Draw position of  $\tilde{x}$ .

---

- Let  $\tilde{x} = x_0 + \frac{1}{\alpha}(x - x_0) \in \overline{B(x_0, \rho)}$ . Then  $x = (1 - \alpha) \cdot x_0 + \alpha \cdot \tilde{x}$  and therefore by convexity of  $f$

$$\begin{aligned} f(x) &\leq (1 - \alpha) \cdot f(x_0) + \alpha \cdot f(\tilde{x}) \\ f(x) - f(x_0) &\leq \alpha \cdot (\eta - f(x_0)) = \|x - x_0\| \cdot \frac{\eta - f(x_0)}{\rho} \end{aligned}$$

---

**Sketch:** Draw position of new  $\tilde{x}$ .

---

- Now let  $\tilde{x} = x_0 + \frac{1}{\alpha}(x_0 - x) \in \overline{B(x_0, \rho)}$ . Then  $x_0 = \frac{\alpha}{1+\alpha} \cdot \tilde{x} + \frac{1}{1+\alpha} \cdot x$ . So:

$$\begin{aligned} f(x_0) &\leq \frac{1}{1+\alpha} \cdot f(x) + \frac{\alpha}{1+\alpha} \cdot f(\tilde{x}) \\ f(x_0) - f(x) &\leq \frac{\alpha}{1+\alpha} \cdot (f(\tilde{x}) - f(x_0) + f(x_0) - f(x)) \\ f(x_0) - f(x) &\leq \alpha \cdot (\eta - f(x_0)) = \|x - x_0\| \cdot \frac{\eta - f(x_0)}{\rho} \end{aligned}$$

We combine to get:

$$|f(x) - f(x_0)| \leq \|x - x_0\| \cdot \frac{\eta - f(x_0)}{\rho}$$

- Now need to extend to other ‘base points’ near  $x_0$ .
- For every  $x_1 \in \overline{B(x_0, \rho/4)}$  have  $\sup f(\overline{B(x_1, \rho/2)}) \leq \eta$  and  $f(x_1) \geq f(x_0) - \frac{\rho}{4} \cdot \frac{\eta - f(x_0)}{\rho} \geq 2f(x_0) - \eta$ . With arguments above get for every  $x \in \overline{B(x_1, \rho/2)}$  that

$$|f(x) - f(x_1)| \leq \|x - x_1\| \cdot \frac{\eta - f(x_1)}{\rho/2} \leq \|x - x_1\| \cdot \frac{4(\eta - f(x_0))}{\rho}.$$

- For every  $x_1, x_2 \in \overline{B(x_0, \rho/4)}$  have  $\|x_1 - x_2\| \leq \rho/2$  and thus

$$|f(x_1) - f(x_2)| \leq \|x_1 - x_2\| \cdot \frac{4(\eta - f(x_0))}{\rho}.$$

□

**Proposition 1.15** (On convexity and continuity II). If any of the conditions of Proposition 1.14 hold for some  $x_0 \in \text{dom } f$ , then  $f$  is locally Lipschitz continuous on  $\text{int dom } f$ .

---

*Proof. Sketch:* Positions of  $x_0, x, y$  and balls  $B(x_0, \rho), B(x, \alpha \cdot \rho)$

---

- By assumption there is some  $x_0 \in \text{dom } f$ ,  $\rho \in \mathbb{R}_{++}$  and  $\eta < \infty$  such that  $\sup f(\overline{B(x_0, \rho)}) \leq \eta$ .
- For any  $x \in \text{int dom } f$  there is some  $y \in \text{dom } f$  such that  $x = \gamma \cdot x_0 + (1 - \gamma) \cdot y$  for some  $\gamma \in (0, 1)$ .
- Further, there is some  $\alpha \in (0, \gamma)$  such that  $\overline{B(x, \alpha \cdot \rho)} \subset \text{dom } f$  and  $y \notin \overline{B(x, \alpha \cdot \rho)}$ .
- Then,  $\overline{B(x, \alpha \cdot \rho)} \subset \text{conv}(\overline{B(x_0, \rho)} \cup \{y\})$ .
- So for any  $z \in \overline{B(x, \alpha \cdot \rho)}$  there is some  $w \in B(x_0, \rho)$  and some  $\beta \in [0, 1]$  such that  $z = \beta \cdot w + (1 - \beta) \cdot y$ . Therefore,

$$f(z) \leq \beta \cdot f(w) + (1 - \beta) \cdot f(y) \leq \max\{\eta, f(y)\}.$$

- So  $f$  is bounded from above on  $\overline{B(x, \alpha \cdot \rho)}$  and thus by Proposition 1.14  $f$  is locally Lipschitz near  $x$ .  $\square$

**Remark 1.16.** One can show: If  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  is proper, convex and lower semicontinuous, then  $\text{cont } f = \text{int dom } f$ . ← VL1

**Proposition 1.17** (On convexity and continuity in finite dimensions). If  $f \in \text{Conv}(H = \mathbb{R}^n)$  then  $f$  is locally Lipschitz continuous at every point in  $\text{int dom } f$ .

*Proof.* • Let  $x_0 \in \text{int dom } f$ .

- If  $H$  is finite-dimensional then there is a finite set  $\{x_i\}_{i \in I} \subset \text{dom } f$  such that  $x_0 \in \text{int conv}(\{x_i\}_{i \in I}) \subset \text{dom } f$ .
- For example: along every axis  $i = 1, \dots, n$  pick  $x_{2i-1} = x + \varepsilon \cdot e_i$ ,  $x_{2i} = x - \varepsilon \cdot e_i$  for sufficiently small  $\varepsilon$  where  $e_i$  denotes the canonical  $i$ -th Euclidean basis vector.
- Since every point in  $\text{conv}(\{x_i\}_{i \in I})$  can be written as convex combination of  $\{x_i\}_{i \in I}$  we find  $\sup f(\text{conv}(\{x_i\}_{i \in I})) \leq \max_{i \in I} f(x_i) < +\infty$ .
- So  $f$  is bounded from above on an environment of  $x_0$  and thus Lipschitz continuous in  $x_0$  by the previous Proposition.  $\square$

---

Comment: Why is interior necessary in Proposition above?

---

**Example 1.18.** The above result does not extend to infinite dimensions.

- For instance, the  $H^1$ -norm is not continuous with respect to the topology induced by the  $L^2$ -norm.
- An unbounded linear functional is convex but not continuous.

**Definition 1.19** (Lower semi-continuity). A function  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  is called (sequentially, see Remark 1.13) *lower semicontinuous* in  $x \in H$  if for every sequence  $(x_n)_n$  that converges to  $x$  one has

$$\liminf_{n \rightarrow \infty} f(x_n) \geq f(x).$$

$f$  is called lower semicontinuous if it is lower semicontinuous on  $H$ .

**Example 1.20.**  $f(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 & \text{if } x > 0 \end{cases}$  is lower semicontinuous,  $f(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0 \end{cases}$  is not.

---

**Sketch:** Plot the two graphs.

---

Comment: Assuming continuity is sometimes impractically strong. Lower semi-continuity is a weaker assumption and also sufficient for well-posedness of minimization problems: If  $(x_n)_n$  is a convergent minimizing sequence of a lower semicontinuous function  $f$  with limit  $x$  then  $x$  is a minimizer.

---

**Proposition 1.21.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$ . The following are equivalent:

- (i)  $f$  is lower semicontinuous.

(ii)  $\text{epi } f$  is closed in  $H \times \mathbb{R}$ .

(iii) The sublevel sets  $S_r(f)$  are closed for all  $r \in \mathbb{R}$ .

*Proof.* (i)  $\Rightarrow$  (ii). Let  $(y_k, r_k)_k$  be a converging sequence in  $\text{epi } f$  with limit  $(y, r)$ . Then

$$r = \lim_{k \rightarrow \infty} r_k \geq \liminf_{k \rightarrow \infty} f(y_k) \geq f(y) \quad \Rightarrow \quad (y, r) \in \text{epi } f.$$

(ii)  $\Rightarrow$  (iii). For  $r \in \mathbb{R}$  let  $A_r : H \rightarrow H \times \mathbb{R}$ ,  $x \mapsto (x, r)$  and  $Q_r = \text{epi } f \cap (H \times \{r\})$ .  $Q_r$  is closed,  $A_r$  is continuous.

$$S_r(f) = \{x \in H \mid f(x) \leq r\} = \{x \in H \mid (x, y) \in Q_r\} = A_r^{-1}(Q_r) \quad \text{is closed.}$$

(iii)  $\Rightarrow$  (i). Assume (i) is false. Then there is a sequence  $(y_k)_k$  in  $H$  converging to  $y \in H$  such that  $\rho := \lim_{k \rightarrow \infty} f(y_k) < f(y)$ . Let  $r \in (\rho, f(y))$ . For  $k \geq k_0$  sufficiently large,  $f(y_k) \leq r < f(y)$ , i.e.  $y_k \in S_r(f)$  but  $y \notin S_r(f)$ . Contradiction.  $\square$

### 1.3 Subdifferential

**Definition 1.22.** The power set of  $H$  is the set of all subsets of  $H$  and denoted by  $2^H$ .

---

Comment: Meaning of notation.

---

**Definition 1.23** (Subdifferential). Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper. The *subdifferential* of  $f$  is the set-valued operator

$$\partial f : H \rightarrow 2^H, \quad x \mapsto \{u \in H \mid f(y) \geq f(x) + \langle y - x, u \rangle \text{ for all } y \in H\}$$

For  $x \in H$ ,  $f$  is *subdifferentiable* at  $x$  if  $\partial f(x) \neq \emptyset$ . Elements of  $\partial f(x)$  are called *subgradients* of  $f$  at  $x$ .

---

**Sketch:** Subgradients are slopes of affine functions that bound  $f$  from below and are equal to  $f$  in  $x$ .

---

**Definition 1.24.** The *domain*  $\text{dom } A$  of a set-valued operator  $A$  are the points where  $A(x) \neq \emptyset$ .

**Definition 1.25.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper.  $x$  is a *minimizer* of  $f$  if  $f(x) = \inf f(H)$ . The set of minimizers of  $f$  is denoted by  $\text{argmin } f$ .

The following is an adaption of first order optimality condition for differentiable functions to convex non-smooth functions.

**Proposition 1.26** (Fermat's rule). Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper. Then

$$\text{argmin } f = \{x \in H \mid 0 \in \partial f(x)\}.$$

*Proof.* Let  $x \in H$ . Then

$$[x \in \text{argmin } f] \Leftrightarrow [f(y) \geq f(x) = f(x) + \langle y - x, 0 \rangle \text{ for all } y \in H] \Leftrightarrow [0 \in \partial f(x)]. \quad \square$$

**Proposition 1.27** (Basic properties of subdifferential). Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$ .

- (i)  $\partial f(x)$  is closed and convex.
- (ii) If  $x \in \text{dom } \partial f$  then  $f$  is lower semicontinuous at  $x$ .

*Proof.* (i):

$$\partial f(x) = \bigcap_{y \in \text{dom } f} \{u \in H \mid f(y) \geq f(x) + \langle y - x, u \rangle\}$$

So  $\partial f(x)$  is the intersection of closed and convex sets. Therefore it is closed and convex.

(ii): Let  $u \in \partial f(x)$ . Then for all  $y \in H$ :  $f(y) \geq f(x) + \langle y - x, u \rangle$ . So, for any sequence  $(x_k)_k$  converging to  $x$  one finds

$$\liminf_{k \rightarrow \infty} f(x_k) \geq f(x) + \liminf_{k \rightarrow \infty} \langle x_k - x, u \rangle = f(x). \quad \square$$

**Definition 1.28** (Monotonicity). A set-valued function  $A : H \rightarrow 2^H$  is monotone if

$$\langle x - y, u - v \rangle \geq 0$$

for every tuple  $(x, y, u, v) \in H^4$  such that  $u \in A(x)$  and  $v \in A(y)$ .

**Proposition 1.29.** The subdifferential of a proper function is monotone.

*Proof.* Let  $u \in \partial f(x)$ ,  $v \in \partial f(y)$ . We get:

$$\begin{aligned} f(y) &\geq f(x) + \langle y - x, u \rangle, \\ f(x) &\geq f(y) + \langle x - y, v \rangle, \end{aligned}$$

and by combining:

$$0 \geq \langle y - x, u - v \rangle \quad \square$$

**Proposition 1.30.** Let  $I$  be a finite index set, let  $H = \bigotimes_{i \in I} H_i$  a product of several Hilbert spaces. Let  $f_i : H_i \rightarrow \mathbb{R} \cup \{\infty\}$  be proper and let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$ ,  $x = (x_i)_{i \in I} \mapsto \sum_{i \in I} f_i(x_i)$ . Then  $\partial f(x) = \bigotimes_{i \in I} \partial f_i(x_i)$ .

*Proof.*  $\bigotimes_{i \in I} \partial f_i(x_i) \subset \partial f(x)$ : For  $x \in H$  let  $p_i \in \partial f_i(x_i)$ . Then

$$f(x + y) = \sum_{i \in I} f_i(x_i + y_i) \geq \sum_{i \in I} f_i(x_i) + \langle y_i, p_i \rangle = f(x) + \langle y, p \rangle.$$

Therefore  $p = (p_i)_{i \in I} \in \partial f(x)$ .

$\partial f(x) \subset \bigotimes_{i \in I} \partial f_i(x_i)$ : Let  $p = (p_i)_{i \in I} \in \partial f(x)$ . For  $j \in I$  let  $y_j \in H_j$  and let  $y = (\tilde{y}_i)_{i \in I}$  where  $\tilde{y}_i = 0$  if  $i \neq j$  and  $\tilde{y}_j = y_j$ . We get

$$f(x + y) = \sum_{i \in I} f_i(x_i + \tilde{y}_i) = \sum_{i \in I \setminus \{j\}} f_i(x_i) + f_j(x_j + y_j) \geq f(x) + \langle y, p \rangle = \sum_{i \in I} f_i(x_i) + \langle y_j, p_j \rangle$$

This holds for all  $y_j \in H_j$ . Therefore,  $p_j \in \partial f_j(x_j)$ .  $\square$

**Example 1.31.** •  $f(x) = \frac{1}{2}\|x\|^2$ :  $f$  is Gâteaux differentiable (see below) with  $\nabla f(x) = x$ . We will show that this implies  $\partial f(x) = \{\nabla f(x)\} = \{x\}$ .

- $f(x) = \|x\|$ :
  - For  $x \neq 0$   $f$  is again Gâteaux differentiable with  $\nabla f(x) = \frac{x}{\|x\|}$ .
  - For  $x = 0$  we get  $f(y) \geq \langle y, p \rangle = f(0) + \langle y - 0, p \rangle$  for  $\|p\| \leq 1$  via the Cauchy-Schwarz inequality. So  $\overline{B(0, 1)} \subset \partial f(0)$ .
  - Assume some  $p \in \partial f(0)$  has  $\|p\| > 1$ . Then  $\frac{p}{\|p\|} \in \partial f(p)$ . We test:  $\left\langle p - 0, \frac{p}{\|p\|} - p \right\rangle = \|p\| - \|p\|^2 < 0$  which contradicts monotonicity of the subdifferential. Therefore  $\partial f(0) = \overline{B(0, 1)}$ .
- $H = \mathbb{R}$ ,  $f(x) = |x|$  is a special case of the above.

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 0, \\ [-1, 1] & \text{if } x = 0, \\ \{+1\} & \text{if } x > 0 \end{cases}$$

---

**Sketch:** Draw ‘graph’ of subdifferential.

---

- $H = \mathbb{R}^n$ ,  $f(x) = \|x\|_1$ . The  $\ell_1$  norm is not induced by an inner product. Therefore the above does not apply. We can use Proposition 1.30:

$$\partial f(x) = \bigotimes_{k=1}^n \partial \text{abs}(x_k)$$

---

**Sketch:** Draw subdifferential ‘graph’ for 2D.

---

**Proposition 1.32.** Let  $f, g : H \rightarrow \mathbb{R} \cup \{\infty\}$ . For  $x \in H$  one finds  $\partial f(x) + \partial g(x) \subset \partial(f+g)(x)$ .

*Proof.* Let  $u \in \partial f(x)$ ,  $v \in \partial g(x)$ . Then

$$f(x+y) + g(x+y) \geq f(x) + \langle u, y \rangle + g(x) + \langle v, y \rangle = f(x) + g(x) + \langle u+v, y \rangle .$$

Therefore,  $u+v \in \partial(f+g)(x)$ . □

**Remark 1.33.** The converse inclusion is not true in general and much harder to prove. A simple counter-example is  $f(x) = \|x\|^2$  and  $g(x) = -\|x\|^2/2$ . The subdifferential of  $g$  is empty but the subdifferential of  $f+g$  is not.

An application of the sub-differential is a simple proof of Jensen’s inequality.

**Proposition 1.34** (Jensen’s inequality). Let  $f : H = \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  be convex. Let  $\mu$  be a probability measure on  $H$  such that

$$\bar{x} = \int_H x \, d\mu(x) \in H$$

and  $\bar{x} \in \text{dom } \partial f$ . Then

$$\int_H f(x) \, d\mu(x) \geq f(\bar{x}) .$$

*Proof.* Let  $u \in \partial f(\bar{x})$ .

$$\int_H f(x) \, d\mu(x) \geq \int_H f(\bar{x}) + \langle x - \bar{x}, u \rangle \, d\mu(x) = f(\bar{x})$$

□

Let us examine the subdifferential of differentiable functions.

**Definition 1.35** (Gâteaux differentiability). A function  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  is *Gâteaux differentiable* in  $x \in \text{dom } f$  if there is a unique *Gâteaux gradient*  $\nabla f(x) \in H$  such that for any  $y \in H$  the directional derivative is given by

$$\lim_{\alpha \searrow 0} \frac{f(x+\alpha \cdot y) - f(x)}{\alpha} = \langle y, \nabla f(x) \rangle .$$

**Proposition 1.36.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper and convex, let  $x \in \text{dom } f$ . If  $f$  is Gâteaux differentiable in  $x$  then  $\partial f(x) = \{\nabla f(x)\}$ .

*Proof.*  $\nabla f(x) \in \partial f(x)$ :

- For fixed  $y \in H$  consider the function  $\phi : (0, \infty) \rightarrow \mathbb{R} \cup \{\infty\}$ ,  $\alpha \mapsto \frac{f(x+\alpha \cdot y) - f(x)}{\alpha}$ .

- $\phi$  is increasing: let  $\beta \in (0, \alpha)$ . Then  $x + \beta \cdot y = (1 - \beta/\alpha) \cdot x + \beta/\alpha \cdot (x + \alpha \cdot y)$ . So

$$\begin{aligned} f(x + \beta \cdot y) &\leq (1 - \beta/\alpha) \cdot f(x) + \beta/\alpha \cdot f(x + \alpha \cdot y), \\ \phi(\beta) &\leq \frac{(1 - \beta/\alpha) \cdot f(x) + \beta/\alpha \cdot f(x + \alpha \cdot y) - f(x)}{\beta} \\ &= \frac{\beta/\alpha \cdot (f(x + \alpha \cdot y) - f(x))}{\beta} = \phi(\alpha). \end{aligned}$$

- Therefore,

$$\langle y, \nabla f(x) \rangle = \lim_{\alpha \searrow 0} \frac{f(x + \alpha \cdot y) - f(x)}{\alpha} = \inf_{\alpha \in \mathbb{R}_{++}} \phi(\alpha) \leq f(x + y) - f(x).$$

(We set  $\alpha = 1$  to get the last inequality.)

← VL2

$\partial f(x) \subset \{\nabla f(x)\}$ :

- For  $u \in \partial f(x)$  we find for any  $y \in H$

$$\langle y, \nabla f(x) \rangle = \lim_{\alpha \searrow 0} \frac{f(x + \alpha \cdot y) - f(x)}{\alpha} \geq \lim_{\alpha \searrow 0} \frac{f(x) + \langle \alpha \cdot y, u \rangle - f(x)}{\alpha} = \langle y, u \rangle.$$

- This inequality holds for any  $y$  and  $-y$  simultaneously. Therefore  $u = \nabla f(x)$ . □

**Remark 1.37.** For differentiable functions in one dimension this implies monotonicity of the derivative: Let  $f \in C^1(\mathbb{R})$ . With Propositions 1.36 and 1.29 we get: if  $x \geq y$  then  $f'(x) \geq f'(y)$ .

## 1.4 Cones and support functions

Cones are a special class of sets with many applications in convex analysis.

**Definition 1.38.** A set  $C \subset H$  is a *cone* if for any  $x \in C$ ,  $\lambda \in \mathbb{R}_{++}$  one has  $\lambda \cdot x \in C$ . In short notation:  $C = \mathbb{R}_{++} \cdot C$ .

**Remark 1.39.** A cone need not contain 0, but for any  $x \in C$  it must contain the open line segment  $(0, x]$ .

**Proposition 1.40.** The intersection of a family  $\{C_i\}_{i \in I}$  of cones is cone. The *conical hull* of a set  $C \subset H$ , denoted by  $\text{cone } C$  is the smallest cone that contains  $C$ . It is given by  $\mathbb{R}_{++} \cdot C$ .

*Proof.* • Let  $C = \bigcap_{i \in I} C_i$ . If  $x \in C$  then  $x \in C_i$  for all  $i \in I$  and for any  $\lambda \in \mathbb{R}_{++}$  one has  $\lambda \cdot x \in C_i$  for all  $i \in I$ . Hence  $\lambda \cdot x \in C$  and  $C$  is also a cone.

• Let  $D = \mathbb{R}_{++} \cdot C$ . Then  $D$  is a cone,  $C \subset D$  and therefore  $\text{cone } C \subset D$ . Conversely, let  $y \in D$ . Then there are  $x \in C$  and  $\lambda \in \mathbb{R}_{++}$  such that  $y = \lambda \cdot x$ . So  $x \in \text{cone } C$ , therefore  $y \in \text{cone } C$  and thus  $D \subset \text{cone } C$ .  $\square$

**Proposition 1.41.** A cone  $C$  is convex if and only if  $C + C \subset C$ .

*Proof.*  $C$  **convex**  $\Rightarrow C + C \subset C$ : Let  $a, b \in C$ .  $\Rightarrow \frac{1}{2} \cdot a + \frac{1}{2} \cdot b \in C \Rightarrow a + b \in C \Rightarrow C + C \subset C$ .  
 $C + C \subset C \Rightarrow C$  **convex**: Let  $a, b \in C$ .  $\Rightarrow a + b \in C$  and  $\lambda \cdot a, (1 - \lambda) \cdot b \in C$  for all  $\lambda \in (0, 1)$ .  
 $\Rightarrow \lambda \cdot a + (1 - \lambda) \cdot b \in C$ .  $\Rightarrow [a, b] \in C \Rightarrow C$  convex.  $\square$

**Definition 1.42.** Let  $C \subset H$ . The *polar cone* of  $C$  is

$$C^\ominus = \{y \in H \mid \sup \langle C, y \rangle \leq 0\}.$$

---

**Sketch:** Draw a cone in 2D with angle  $< \pi/2$  and its polar cone.

---

**Proposition 1.43.** Let  $C$  be a linear subspace of  $H$ . Then  $C^\ominus = C^\perp$ .

*Proof.* • Since  $C$  is a linear subspace, if  $\langle x, y \rangle \neq 0$  for some  $y \in H$ ,  $x \in C$  then  $\sup \langle C, y \rangle = \infty$ .

• Therefore,  $C^\ominus = \{y \in H \mid \langle x, y \rangle = 0 \text{ for all } x \in C\}$ .  $\square$

**Definition 1.44.** Let  $C \subset H$  convex, non-empty and  $x \in H$ . The *tangent cone* to  $C$  at  $x$  is

$$T_C x = \begin{cases} \overline{\text{cone}(C - x)} & \text{if } x \in C, \\ \emptyset & \text{else.} \end{cases}$$

The *normal cone* to  $C$  at  $x$  is

$$N_C x = \begin{cases} (C - x)^\ominus = \{u \in H \mid \sup \langle C - x, u \rangle \leq 0\} & \text{if } x \in C, \\ \emptyset & \text{else.} \end{cases}$$

**Example 1.45.** Let  $C = \overline{B(0, 1)}$ . Then for  $x \in C$ :

$$T_C x = \begin{cases} \{y \in H \mid \langle y, x \rangle \leq 0\} & \text{if } \|x\| = 1, \\ H & \text{if } \|x\| < 1. \end{cases}$$

Note: the  $\leq$  in the  $\|x\| = 1$  case comes from the closure in the definition of  $T_C x$ . Without closure it would merely be  $<$ .

$$N_C x = \begin{cases} \mathbb{R}_+ \cdot x & \text{if } \|x\| = 1, \\ \{0\} & \text{if } \|x\| < 1. \end{cases}$$

**Example 1.46.** What are tangent and normal cone for the  $L_1$ -norm ball in  $\mathbb{R}^2$ ?

We start to see connections between different concepts introduced so far.

**Proposition 1.47.** Let  $C \subset H$  be a convex set. Then  $\partial \iota_C(x) = N_C x$ .

*Proof.* •  $x \notin C$ :  $\partial \iota_C(x) = \emptyset = N_C x$ .

•  $x \in C$ :

$$\begin{aligned} [u \in \partial \iota_C(x)] &\Leftrightarrow [\iota_C(y) \geq \iota_C(x) + \langle y - x, u \rangle \ \forall y \in C] \Leftrightarrow [0 \geq \langle y - x, u \rangle \ \forall y \in C] \\ &\Leftrightarrow [\sup \langle C - x, u \rangle \leq 0] \Leftrightarrow [u \in N_C x] \end{aligned}$$

□

---

Comment: This will become relevant, when doing constrained optimization, where parts of the objective are given by indicator functions.

---

Now we introduce the projection onto convex sets. It will play an important role in analysis and numerical methods for constrained optimization.

**Proposition 1.48** (Projection). Let  $C \subset H$  be non-empty, closed convex. For  $x \in H$  the problem

$$\inf \{\|x - p\| \mid p \in C\}$$

has a unique minimizer. This minimizer is called the *projection* of  $x$  onto  $C$  and is denoted by  $P_C x$ .

*Proof.* • We will need the following inequality for any  $x, y, z \in H$ , which can be shown by careful expansion:

$$\|x - y\|^2 = 2\|x - z\|^2 + 2\|y - z\|^2 - 4\|(x + y)/2 - z\|^2$$

- $C$  is non-empty,  $y \mapsto \|x - y\|$  is bounded from below, so the infimal value is a real number, denoted by  $d$ .
- Let  $(p_k)_{k \in \mathbb{N}}$  be a minimizing sequence. For  $k, l \in \mathbb{N}$  one has  $\frac{1}{2}(p_k + p_l) \in C$  by convexity and therefore  $\|x - \frac{1}{2}(p_k + p_l)\| \geq d$ .
- With the above inequality we find:

$$\|p_k - p_l\|^2 = 2\|p_k - x\|^2 + 2\|p_l - x\|^2 - 4\|\frac{p_k + p_l}{2} - x\|^2 \leq 2\|p_k - x\|^2 + 2\|p_l - x\|^2 - 4d^2$$

- So by sending  $k, l \rightarrow \infty$  we find that  $(p_k)_k$  is a Cauchy sequence which converges to a limit  $p$ . Since  $C$  is closed,  $p \in C$ . And since  $y \mapsto \|x - y\|$  is continuous,  $p$  is a minimizer.

- Uniqueness of  $p$ , quick answer: the optimization problem is equivalent to minimizing  $y \mapsto \|x - y\|^2$ , which is strictly convex. Therefore  $p$  must be unique.
- Uniqueness of  $p$ , detailed answer: assume there is another minimizer  $q \neq p$ . Then  $\frac{1}{2}(p+q) \in C$  and we find:

$$\|x - p\|^2 + \|x - q\|^2 - 2\|x - \frac{1}{2}(p+q)\|^2 = \frac{1}{2}\|p - q\|^2 > 0$$

So the sum of the objectives at  $p$  and  $q$  is strictly larger than twice the objective at the midpoint. Therefore, neither  $p$  nor  $q$  can be optimal.  $\square$

**Proposition 1.49** (Characterization of projection). Let  $C \subset H$  be non-empty, convex, closed. Then  $p = P_C x$  if and only if

$$[p \in C] \wedge [\langle y - p, x - p \rangle \leq 0 \text{ for all } y \in C].$$

---

**Sketch:** Illustrate inequality.

---

*Proof.* • It is clear that  $[p = P_C x] \Rightarrow [p \in C]$ , and that  $[p \notin C] \Rightarrow [p \neq P_C x]$ .

- So, need to show that for  $p \in C$  one has  $[p = P_C x] \Leftrightarrow [\langle y - p, x - p \rangle \leq 0 \text{ for all } y \in C]$ .
- For some  $y \in C$  and some  $\varepsilon \in \mathbb{R}_{++}$  consider:

$$\begin{aligned} \|x - (p + \varepsilon \cdot (y - p))\|^2 - \|x - p\|^2 &= \|p + \varepsilon \cdot (y - p)\|^2 - \|p\|^2 - 2\varepsilon \langle x, y - p \rangle \\ &= \varepsilon^2 \|y - p\|^2 - 2\varepsilon \langle x - p, y - p \rangle \end{aligned}$$

If  $\langle x - p, y - p \rangle > 0$  then this is negative for sufficiently small  $\varepsilon$  and thus  $p$  cannot be the projection. Conversely, if  $\langle x - p, y - p \rangle \leq 0$  for all  $y \in C$ , then for  $\varepsilon = 1$  we see that  $p$  is indeed the minimizer of  $y \mapsto \|x - y\|^2$  over  $C$  and thus the projection.  $\square$

**Corollary 1.50** (Projection and normal cone). Let  $C \subset H$  be non-empty, closed, convex. Then  $[p = P_C x] \Leftrightarrow [x \in p + N_C p]$ .

*Proof.*  $[p = P_C x] \Leftrightarrow [p \in C \wedge \sup \langle C - p, x - p \rangle \geq 0] \Leftrightarrow [x - p \in N_C p]$ .  $\square$

---

Comment: This condition is actually useful for computing projections.

---

**Example 1.51** (Projection onto  $L_1$ -ball in  $\mathbb{R}^2$ ). Let  $C = \{(x, y) \in \mathbb{R}^2 \mid |x| + |y| \leq 1\}$ . We find:

$$N_C(x, y) = \begin{cases} \emptyset & \text{if } |x| + |y| > 1, \\ \{0\} & \text{if } |x| + |y| < 1, \\ \overline{\text{cone}} \text{conv}\{(1, 1), (-1, 1)\} & \text{if } (x, y) = (0, 1), \\ \overline{\text{cone}} \text{conv}\{(1, 1), (1, -1)\} & \text{if } (x, y) = (1, 0), \\ \overline{\text{cone}}\{(1, 1)\} & \text{if } x + y = 1, x \in (0, 1), \\ \dots & \end{cases}$$

---

**Sketch:** Draw normal cones attached to points in  $C$ .

---

Now compute projection of  $(a, b) \in \mathbb{R}^2$ . W.l.o.g. assume  $(a, b) \in \mathbb{R}_+^2 \setminus C$ . Then

$$P_C(a, b) = \begin{cases} (0, 1) & \text{if } [a + b \geq 1] \wedge [b - a \geq 1], \\ (1, 0) & \text{if } [a + b \geq 1] \wedge [a - b \geq 1], \\ ((1 + a - b)/2, (1 - a + b)/2) & \text{else.} \end{cases}$$

---

Comment: Do computation in detail.

---

Comment: Result is very intuitive, but not so trivial to prove rigorously due to non-smoothness of problem. Comment: Eistüte.

---

We now establish a sequence of results that will later allow us to analyze the subdifferential via cones and prepare results for the study of the Fenchel–Legendre conjugate.

**Proposition 1.52.** Let  $K \subset H$  be a non-empty, closed, convex cone. Let  $x, p \in H$ . Then

$$[p = P_K x] \Leftrightarrow [p \in K, x - p \perp p, x - p \in K^\ominus].$$

*Proof.* • By virtue of Corollary 1.50 (Characterization of projection with normal cone inclusion) we need to show

$$[x - p \in N_K p] \Leftrightarrow [p \in K, x - p \perp p, x - p \in K^\ominus].$$

- $\Rightarrow$ : Let  $x - p \in N_K p$ . Then  $p \in K$ . By definition have  $\sup \langle K - p, x - p \rangle \leq 0$ . Since  $2p, 0 \in K$  ( $K$  is closed) this implies  $\langle p, x - p \rangle = 0$ . Further, since  $K$  is convex, we have (Prop. 1.41)  $K + K \subset K$ , and in particular  $K + p \subset K$ . Therefore  $\sup \langle K + p - p, x - p \rangle \leq \sup \langle K - p, x - p \rangle \leq 0$  and thus  $x - p \in K^\ominus$ .

---

**Sketch:** Recall that  $K + p \subset K$ . Counter-example for non-convex  $K$ .

---

- $\Leftarrow$ : Since  $p \perp x - p$  have  $\sup \langle K - p, x - p \rangle = \sup \langle K, x - p \rangle \leq 0$  since  $x - p \in K^\ominus$ . Then, since  $p \in K$  have  $x - p \in N_K p$ .  $\square$

**Proposition 1.53.** Let  $K \subset H$  be a non-empty, closed, convex cone. Then  $K^{\ominus\ominus} = K$ .

*Proof.* •  $K \subset K^{\ominus\ominus}$ : Recall:  $K^\ominus = \{u \in H \mid \sup \langle K, u \rangle \leq 0\}$ .

- Let  $x \in K$ . Then  $\langle x, u \rangle \leq 0$  for all  $u \in K^\ominus$ . Therefore  $\sup \langle x, K^\ominus \rangle \leq 0$  and so  $x \in K^{\ominus\ominus}$ . Therefore:  $K \subset K^{\ominus\ominus}$ .
- $K^{\ominus\ominus} \subset K$ : Let  $x \in K^{\ominus\ominus}$ , set  $p \in P_K x$ . Then by Proposition 1.52 (Projection onto closed, convex cone):  $x - p \perp p, x - p \in K^\ominus$ .
- $[x \in K^{\ominus\ominus}] \wedge [x - p \in K^\ominus] \Rightarrow \langle x, x - p \rangle \leq 0$ .
- $\|x - p\|^2 = \langle x, x - p \rangle - \langle p, x - p \rangle \leq 0 \Rightarrow x = p \Rightarrow x \in K$ . Therefore  $K^{\ominus\ominus} \subset K$ .  $\square$

For subsequent results we need the following Lemma that once more illustrates that convexity implies strong regularity.

**Proposition 1.54.** Let  $C \subset H$  be convex. Then the following hold:

- (i)  $\forall x \in \text{int } C, y \in \overline{C}: [x, y] \subset \text{int } C$ .

(ii)  $\overline{C}$  is convex.

(iii)  $\text{int } C$  is convex.

(iv) If  $\text{int } C \neq \emptyset$  then  $\text{int } C = \text{int } \overline{C}$  and  $\overline{C} = \overline{\text{int } C}$ .

← VL3

*Proof.* • **(i):** Assume  $x \neq y$  (otherwise the result is trivial). Then for  $z \in [x, y]$  there is some  $\alpha \in (0, 1]$  such that  $z = \alpha \cdot x + (1 - \alpha) \cdot y$ .

- Since  $x \in \text{int } C$  there is some  $\varepsilon \in \mathbb{R}_{++}$  such that  $B(x, \varepsilon \cdot (2 - \alpha)/\alpha) \subset C$ .
- Since  $y \in \overline{C}$ , one has  $y \in C + B(0, \varepsilon)$ .
- By convexity of  $C$ :

$$\begin{aligned} B(z, \varepsilon) &= \alpha \cdot x + (1 - \alpha) \cdot y + B(0, \varepsilon) \\ &\subset \alpha \cdot x + (1 - \alpha) \cdot (C + B(0, \varepsilon)) + B(0, \varepsilon) \\ &= \alpha \cdot B(x, \varepsilon \cdot \frac{2-\alpha}{\alpha}) + (1 - \alpha) \cdot C \\ &\subset \alpha \cdot C + (1 - \alpha) \cdot C = C \end{aligned}$$

- Therefore  $z \in \text{int } C$ .
- **(ii):** Let  $x, y \in \overline{C}$ . By definition there are sequences  $(x_k)_k, (y_k)_k$  in  $C$  that converge to  $x$  and  $y$ . For  $\lambda \in [0, 1]$  the sequence  $(\lambda \cdot x_k + (1 - \lambda) \cdot y_k)_k$  converges to  $\lambda \cdot x + (1 - \lambda) \cdot y \in \overline{C}$ .
- **(iii):** Let  $x, y \in \text{int } C$ . Then  $y \in \overline{C}$ . By **(i)** therefore  $(x, y) \in \text{int } C$ .
- **(iv):** By definition  $\text{int } C \subset \text{int } \overline{C}$ . Show converse inclusion. Let  $y \in \text{int } \overline{C}$ . Then there is  $\varepsilon \in \mathbb{R}_{++}$  such that  $B(y, \varepsilon) \subset \overline{C}$ . Let  $x \in \text{int } C$ ,  $x \neq y$ . Then there is some  $\alpha \in \mathbb{R}_{++}$  such that  $y + \alpha \cdot (y - x) \in B(y, \varepsilon) \subset \overline{C}$ .
- Since  $y \in (x, y + \alpha \cdot (y - x))$  it follows from **(i)** that  $y \in \text{int } C$ .
- Similarly, it is clear that  $\overline{\text{int } C} \subset \overline{C}$ . We show the converse inclusion. Let  $x \in \text{int } C$ ,  $y \in \overline{C}$ . For  $\alpha \in (0, 1]$  let  $y_\alpha = (1 - \alpha) \cdot y + \alpha \cdot x$ . Then  $y_\alpha \in \text{int } C$  by **(i)** and thus  $y = \lim_{\alpha \rightarrow 0} y_\alpha \in \overline{\text{int } C}$ .  $\square$

**Example 1.55.** Let  $H = \mathbb{R}$ ,  $C = \mathbb{Q} \cup [0, 1]$ .  $\text{int } C = (0, 1) \neq \emptyset$  but  $C$  is not convex. We find  $\text{int } C = (0, 1) \neq \text{int } \overline{C} = \text{int } \mathbb{R} = \mathbb{R}$  and  $\overline{C} = \mathbb{R} \neq \overline{\text{int } C} = [0, 1]$ .

We can characterize the tangent and normal cones of a convex set, depending on the base point position.

**Proposition 1.56.** Let  $C \subset H$  be convex with  $\text{int } C \neq \emptyset$  and  $x \in C$ . Then

$$[x \in \text{int } C] \Leftrightarrow [T_C x = H] \Leftrightarrow [N_C x = \{0\}].$$

*Proof.* •  $[x \in \text{int } C] \Leftrightarrow [T_C x = H]$ : Let  $D = C - x$ . Then  $0 \in D$ ,  $[x \in \text{int } C] \Leftrightarrow [0 \in \text{int } D]$  and  $T_C x = \text{cone } \overline{D}$ .

- One can show: if  $D \subset H$  is convex with  $\text{int } D \neq \emptyset$  and  $0 \in D$ , then  $[0 \in \text{int } D] \Leftrightarrow [\text{cone } \overline{D} = H]$ .

- Sketch: assume  $0 \in \text{int } D$ . Then  $\overline{\text{cone } D} = \text{cone } D = H$  since there is some  $\varepsilon > 0$  such that for any  $u \in H \setminus \{0\}$  one has  $\varepsilon \frac{u}{\|u\|} \in D$ . The converse conclusion is more tedious. It relies on Proposition 1.54. See [Bauschke, Combettes; Prop. 6.17] for details.
- $[T_C x = H] \Leftrightarrow [N_C x = \{0\}]$ : Recall  $N_C x = (C - x)^\ominus = \{u \in H : \sup \langle C - x, u \rangle \leq 0\}$ . We can extend the supremum to  $\text{cone}(C - x)$  and we can then extend it to the closure  $\overline{\text{cone}(C - x)} = T_C x$  without changing whether it will be  $\leq 0$  (why?). So  $N_C x = \{u \in H : \sup \langle T_C x, u \rangle \leq 0\} = (T_C x)^\ominus$ .
- Now, if  $T_C x = H$  then  $N_C x = \{0\}$ .
- Conversely, since for  $x \in C$ ,  $T_C x$  is a non-empty, closed, convex cone, one has  $(T_C x)^{\ominus\ominus} = T_C x$  (Prop. 1.53) and therefore  $T_C x = (N_C x)^\ominus$ . So if  $N_C x = \{0\}$  then  $T_C x = H$ .  $\square$

---

Comment: Observation: subdifferential describes affine functions that touch graph in one point and always lie below graph. Similarly: for convex sets there are hyperplanes, that touch set in one point and separate the set from the opposite half-space. These are called ‘supporting hyperplanes’. The study of the subdifferential is thus related to the study of supporting hyperplanes. Supporting hyperplanes, in turn, are again closely related to normal cones, as we will learn.

---

**Definition 1.57.** Let  $C \subset H$ ,  $x \in C$  and let  $u \in H \setminus \{0\}$ . If

$$\sup \langle C, u \rangle \leq \langle x, u \rangle$$

then the set  $\{y \in H : \langle y, u \rangle = \langle x, u \rangle\}$  is a *supporting hyperplane* of  $C$  at  $x$  and  $x$  is a *support point* at  $C$  with *normal vector*  $u$ . The set of support points of  $C$  is denoted by  $\text{spts } C$ .

**Proposition 1.58.** Let  $C \subset H$ ,  $C \neq \emptyset$  and convex. Then:

$$\text{spts } C = \{x \in C : N_C x \neq \{0\}\}$$

*Proof.* Let  $x \in C$ . Then:

$$[x \in \text{spts } C] \Leftrightarrow [\exists u \in H \setminus \{0\} : \sup \langle C - x, u \rangle \leq 0] \Leftrightarrow [0 \neq u \in N_C x]$$

**Proposition 1.59.** Let  $C \subset H$  convex,  $\text{int } C \neq \emptyset$ . Then

$$\text{bdry } C \subset \text{spts } \overline{C} \quad \text{and} \quad C \cap \text{bdry } C \subset \text{spts } C.$$

*Proof.* • If  $C = H$  the result is clear. (Why?) So assume  $C \neq H$ .

- Let  $x \in \text{bdry } C \subset \overline{C}$ . So  $x \in \overline{C} \setminus \text{int } C = \overline{C} \setminus \text{int } \overline{C}$  (Prop. 1.54).
- Consequence of Prop. 1.56:  $\exists u \in N_{\overline{C}} x \setminus \{0\}$ .
- Consequence of Prop. 1.58:  $x \in \text{spts } \overline{C}$ . Therefore  $\text{bdry } C \subset \text{spts } \overline{C}$ .
- Show  $\text{spts } C = C \cap \text{spts } \overline{C}$ : For this use  $\sup \langle \overline{C}, u \rangle = \sup \langle C, u \rangle$  (why?).
- Let  $x \in \text{spts } C$ :  $\Rightarrow x \in C \subset \overline{C}$ ,  $\exists u \neq 0$  s.t.  $\sup \langle C, u \rangle \leq \langle x, u \rangle$ .  $\Rightarrow x \in C \cap \text{spts } \overline{C}$ .
- Let  $x \in \text{spts } \overline{C} \cap C$ :  $\Rightarrow x \in C$ ,  $\exists u \neq 0$  s.t.  $\sup \langle \overline{C}, u \rangle \leq \langle x, u \rangle$ .  $\Rightarrow x \in \text{spts } C$ .
- So:  $C \cap \text{bdry } C \subset C \cap \text{spts } \overline{C} = \text{spts } C$ .  $\square$

**Example 1.60.** Let  $H = \mathbb{R}$ ,  $C = [-1, 1)$ . Then  $\text{int } C = (-1, 1)$ ,  $\overline{C} = [-1, 1]$ ,  $\text{bdry } C = \{-1, 1\}$ ,  $\text{spts } C = \{-1\}$ ,  $\text{spts } \overline{C} = \{-1, 1\}$ .

An application of the previous results is to show that the subdifferential of a convex function is non-empty in a point of its domain where the function is continuous.

**Proposition 1.61.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper and convex and let  $x \in \text{dom } f$ . If  $x \in \text{cont } f$  then  $\partial f(x) \neq \emptyset$ .

*Proof.* • Since  $f$  is proper and convex,  $\text{epi } f$  is non-empty and convex.

- Since  $x \in \text{cont } f$ ,  $f$  is bounded in an environment of  $x$ . Let  $\varepsilon > 0$ ,  $\eta < +\infty$  such that  $f(y) < f(x) + \eta$  for  $\|x - y\| < \varepsilon$ . Therefore,  $\text{int epi } f \neq \emptyset$  because it contains  $B(x, \varepsilon/2) \times (f(x) + 2\eta, \infty)$ .
- Further: consider sequence  $(y_k = (x, f(x) - 1/k))_{k=1}^\infty$ . Clearly  $y_k \notin \text{epi } f$  but  $\lim_{k \rightarrow \infty} y_k = (x, f(x)) \in \text{epi } f$ . Therefore  $(x, f(x)) \in \text{bdry epi } f$ .
- So by Proposition 1.59 there is some  $(u, r) \in N_{\text{epi } f}(x, f(x)) \setminus \{(0, 0)\}$ .
- By definition of normal cone: For every  $(v, s) \in \text{epi } f$  have:

$$\left\langle \begin{pmatrix} v \\ s \end{pmatrix} - \begin{pmatrix} x \\ f(x) \end{pmatrix}, \begin{pmatrix} u \\ r \end{pmatrix} \right\rangle \leq 0$$

- So in particular for  $y \in \text{dom } f$  have  $(y, f(y)) \in \text{epi } f$  and therefore:

$$\langle y - x, u \rangle + (f(y) - f(x)) \cdot r \leq 0$$

- If  $r < 0$  we could divide by  $r$  and get that  $u/|r| \in \partial f(x)$ . So need to show  $r < 0$ .
- Show that  $r \leq 0$ : For any  $\delta > 0$  have:

$$[(x, f(x) + \delta) \in \text{epi } f] \Leftrightarrow \left[ \left\langle \begin{pmatrix} x \\ f(x) + \delta \end{pmatrix} - \begin{pmatrix} x \\ f(x) \end{pmatrix}, \begin{pmatrix} u \\ r \end{pmatrix} \right\rangle \leq 0 \right] \Leftrightarrow [\delta \cdot r \leq 0] \Leftrightarrow [r \leq 0]$$

- Assume  $r = 0$ : Then must have  $u \neq 0$ . Then there is some  $\rho > 0$  such that  $\|\rho \cdot u\| < \varepsilon$  and therefore  $(x + \rho \cdot u, f(x) + \eta) \in \text{epi } f$ . Then:

$$\left[ \left\langle \begin{pmatrix} x + \rho \cdot u \\ f(x) + \eta \end{pmatrix} - \begin{pmatrix} x \\ f(x) \end{pmatrix}, \begin{pmatrix} u \\ 0 \end{pmatrix} \right\rangle \leq 0 \right] \Leftrightarrow [\rho \cdot \langle u, u \rangle \leq 0]$$

This is a contradiction, therefore  $r \neq 0$ . □

**Corollary 1.62.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  convex, proper, lower semicontinuous. Then

$$\text{int dom } f = \text{cont } f \subset \text{dom } \partial f \subset \text{dom } f$$

*Proof.* • The first inclusion was cited in Remark 1.16 (see e.g. [Bauschke, Combettes; Corollary 8.30]).

- The second inclusion is shown in Prop. 1.61.

- The third inclusion follows from contraposition of  $[x \notin \text{dom } f] \Rightarrow [\partial f(x) = \emptyset]$ . □ ← VL4

Finally, we show that closed, convex sets can be expressed solely in terms of their supporting hyperplanes.

For notational convenience introduce ‘support function’.

**Definition 1.63.** Let  $C \subset H$ . The *support function* of  $C$  is

$$\sigma_C : H \mapsto [-\infty, \infty], \quad u \mapsto \sup \langle C, u \rangle .$$

---

**Sketch:** Definition.

We will later learn that each convex, lower semicontinuous and 1-homogeneous function is the support function of a suitable auxiliary set.

---

**Sketch:** Following remark.

---

**Remark 1.64.** If  $C \neq \emptyset$ ,  $u \in H \setminus \{0\}$  and  $\sigma_C(u) < +\infty$ , then  $\{x \in H : \langle x, u \rangle \leq \sigma_C(u)\}$  is smallest closed half-space with outer normal  $u$  that contains  $C$ . If  $x \in C$  and  $\sigma_C(u) = \langle x, u \rangle$  then  $x \in \text{spts } C$  and  $\{y \in H : \langle y, u \rangle = \sigma_C(u) = \langle x, u \rangle\}$  is a supporting hyperplane of  $C$  at  $x$ .

**Proposition 1.65.** Let  $C \subset H$  and set for  $u \in H$

$$A_u = \{x \in H \mid \langle x, u \rangle \leq \sigma_C(u)\} .$$

Then  $\overline{\text{conv } C} = \bigcap_{u \in H} A_u$ .

*Proof.* • If  $C = \emptyset$  then  $\sigma_C(u) = -\infty$  and  $A_u = \emptyset$  for all  $u \in H$ . Hence, the result is trivial.

- Otherwise,  $\sigma_C(u) > -\infty$ . Let  $D = \bigcap_{u \in H} A_u$ .
- Each  $A_u$  is closed, convex and contains  $C$ . Therefore  $D$  is closed, convex and  $\text{conv } C \subset D$ . Since  $D$  is closed, also  $\overline{\text{conv } C} \subset D$ .
- Now, let  $x \in D$ , set  $p = P_{\overline{\text{conv } C}} x$ .
- Then  $\langle x - p, y - p \rangle \leq 0$  for all  $y \in \overline{\text{conv } C}$  and thus  $\sigma_{\overline{\text{conv } C}}(x - p) = \sup \langle \overline{\text{conv } C}, x - p \rangle = \langle p, x - p \rangle$ .
- Moreover,  $x \in D \subset A_{x-p}$ . So  $\langle x, x - p \rangle \leq \sigma_C(x - p)$ .
- Since  $C \subset \overline{\text{conv } C}$  we get  $\sigma_C \leq \sigma_{\overline{\text{conv } C}}$ .
- Now:  $\|x - p\|^2 = \langle x, x - p \rangle - \langle p, x - p \rangle \leq \sigma_C(x - p) - \sigma_{\overline{\text{conv } C}}(x - p) \leq 0$ . Therefore  $x = p \in \overline{\text{conv } C}$  and thus  $D \subset \overline{\text{conv } C}$ . □

**Corollary 1.66.** Any closed convex subset of  $H$  is the intersection of all closed half-spaces of which it is a subset.

## 1.5 The Fenchel–Legendre conjugate

**Remark 1.67** (Motivation). Previous result (Cor. 1.66): closed, convex set is intersection of all half-spaces that contain set.

Analogous idea: is convex, lower semicontinuous function  $f$  pointwise supremum over all affine lower bounds  $x \mapsto \langle x, u \rangle - a_u$ ? How to get minimal offset  $a_u$  for given slope  $u$ ?

$$\begin{aligned} a_u &= \inf\{r \in \mathbb{R} \mid f(x) \geq \langle x, u \rangle - r \text{ for all } x \in H\} \\ &= \inf\{r \in \mathbb{R} \mid r \geq \sup_{x \in H} \langle x, u \rangle - f(x)\} \\ &= \sup_{x \in H} \langle x, u \rangle - f(x) \end{aligned}$$

For given slopes and offsets  $(u, a_u)$ , how do we reconstruct  $f$ ? Pointwise-supremum ( $\equiv$  intersection of all half-spaces containing  $\text{epi } f$ ):

$$f(x) = \sup_{u \in H} \langle x, u \rangle - a_u$$

Note: same formula for obtaining  $a_u$  and reconstructing  $f$ . Write  $a_u = f^*(u)$  and call this *Fenchel–Legendre conjugate*. Reconstruction of  $f$  is then bi-conjugate  $f^{**}$ . When is  $f^{**} = f$  and what happens if  $f^{**} \neq f$ ?

The Fenchel–Legendre conjugate and the bi-conjugate are fundamental in convex analysis and optimization. We start by a formal definition of  $f^*$ , by studying some examples and showing some basic properties of  $f^*$ . We return to a systematic study of  $f^{**}$  in second half of this subsection.

**Definition 1.68** (Fenchel–Legendre conjugate). Let  $f : H \mapsto [-\infty, \infty]$ . The *Fenchel–Legendre conjugate* of  $f$  is

$$f^* : H \mapsto [-\infty, \infty], \quad u \mapsto \sup_{x \in H} \langle x, u \rangle - f(x).$$

The *biconjugate* of  $f$  is  $(f^*)^* = f^{**}$ .

**Example 1.69.** (i)  $f(x) = \frac{1}{2}\|x\|^2$ :

$$f^*(u) = \sup_{x \in H} \langle x, u \rangle - \frac{1}{2}\|x\|^2 = - \left( \inf_{x \in H} \frac{1}{2}\|x\|^2 - \langle x, u \rangle \right) = - \inf_{x \in H} \tilde{f}(x)$$

Convex optimization problem. Fermat's rule (Prop. 1.26):  $y$  is optimizer if  $0 \in \partial \tilde{f}(y)$ . Minkowski sum of subdifferentials (Prop. 1.32):  $y - u \in \partial \tilde{f}(y)$ .  $\Rightarrow$  sufficient optimality condition:  $y = u$ , so  $u$  is minimizer.  $\Rightarrow f^*(u) = \frac{1}{2}\|u\|^2$ ,  $f$  is *self-conjugate*.

(ii)  $f(x) = \|x\|$ :

$$f^*(u) = \sup_{x \in H} \langle x, u \rangle - \|x\|$$

If  $\|u\| > 1$  consider sequence  $x_k = u \cdot k$ . Then

$$f^*(u)|_{[\|u\|>1]} \geq \limsup_{k \rightarrow \infty} (\|u\|^2 - \|u\|) \cdot k = \infty$$

If  $\|u\| \leq 1$  then by Cauchy-Schwarz:

$$f^*(u)|_{\|u\| \leq 1} \leq \sup_{x \in H} (\|u\| \cdot \|x\| - \|x\|) \leq 0$$

And by setting  $x = 0$  get  $f^*(u)|_{\|u\| \leq 1} \geq 0$ . We summarize

$$f^*(u) = \begin{cases} +\infty & \text{if } \|u\| > 1, \\ 0 & \text{if } \|u\| \leq 1 \end{cases} = \iota_{\overline{B(0,1)}}(u)$$

(iii) special case:  $H = \mathbb{R}$ ,  $f(x) = |x|$ :  $f^* = \iota_{[-1,1]}$

(iv)  $H = \mathbb{R}^n$ ,  $f(x) = \|x\|_1 = \sum_{k=1}^n |x_k|$ :

$$f^*(u) = \sup_{x \in H} \langle u, x \rangle - f(x) = \sup_{x \in H} \sum_{k=1}^n u_k \cdot x_k - |x_k| = \sum_{k=1}^n \sup_{s \in \mathbb{R}} u_k \cdot s - |s| = \sum_{k=1}^n \text{abs}^*(u_k)$$

(v)  $f(x) = 0$ :

$$f^*(u) = \sup_{x \in H} \langle u, x \rangle = \begin{cases} 0 & \text{if } u = 0, \\ +\infty & \text{else.} \end{cases}$$

From Examples 1.69 we learn a result on conjugation.

**Proposition 1.70.** Let  $(H_k)_{k=1}^n$  be a tuple of Hilbert spaces,  $f_k : H_k \rightarrow [-\infty, \infty]$ , let  $H = \bigotimes_{k=1}^n H_k$ ,  $f : H \rightarrow [-\infty, \infty]$ ,  $((x_k)_k) \mapsto \sum_{k=1}^n f_k(x_k)$ . Then  $f^*((u_k)_k) = \sum_{k=1}^n f_k^*(u_k)$ .

*Proof.* The proof is completely analogous to Example 1.69, (iv).  $\square$

A few simple ‘transformation rules’:

**Proposition 1.71.** Let  $f : H \rightarrow [-\infty, \infty]$ ,  $\gamma \in \mathbb{R}_{++}$ .

- (i) Let  $h : x \mapsto f(\gamma \cdot x)$ . Then  $h^*(u) = f^*(u/\gamma)$ .
- (ii) Let  $h : x \mapsto \gamma \cdot f(x)$ . Then  $h^*(u) = \gamma \cdot f^*(u/\gamma)$ .
- (iii) Let  $h : x \mapsto f(-x)$ . Then  $h^*(u) = f^*(-u)$ .
- (iv) Let  $h : x \mapsto f(x) - a$  for  $a \in \mathbb{R}$ . Then  $h^*(u) = f^*(u) + a$ . (Adding offset to function adds same offset to all affine lower bounds.)
- (v) Let  $h : x \mapsto f(x - y)$  for  $y \in H$ . Then  $h^*(u) = f^*(u) + \langle u, y \rangle$ . (Shifting the effective origin of a function requires adjustment of all offsets  $\equiv$  axis intercept at origin.)

*Proof.* All points follow from direct computation.  $\square$

**Proposition 1.72** (Fenchel–Young inequality). Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper. Then for all  $x, u \in H$ :

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

*Proof.* • Let  $x, u \in H$ .

- Since  $f$  is proper, have  $f^* > -\infty$  (why?).
- So if  $f(x) = \infty$ , the inequality holds trivially.
- Otherwise:  $f^*(u) = \sup_{y \in H} \langle u, y \rangle - f(y) \geq \langle u, x \rangle - f(x)$ .  $\square$

Now we establish some basic properties of the conjugate. We need an auxiliary Lemma.

**Proposition 1.73.** Let  $(f_i)_{i \in I}$  be an arbitrary set of functions  $H \rightarrow [-\infty, \infty]$ . Set  $f : H \rightarrow [-\infty, \infty]$ ,  $x \mapsto \sup_{i \in I} f_i(x)$ . Then:

- (i)  $\text{epi } f = \bigcap_{i \in I} \text{epi } f_i$
- (ii) If all  $f_i$  are lower semicontinuous, so is  $f$ .
- (iii) If all  $f_i$  are convex, so is  $f$ .

*Proof.* • **(i):**  $[(x, r) \in \text{epi } f] \Leftrightarrow [\mathbb{R} \ni r \geq f(x)] \Leftrightarrow [\mathbb{R} \ni r \geq f_i(x) \text{ for all } i \in I] \Leftrightarrow [(x, r) \in \text{epi } f_i \text{ for all } i \in I] \Leftrightarrow [(x, r) \in \bigcap_{i \in I} \text{epi } f_i]$ .

- **(ii):** If all  $f_i$  are lower semicontinuous, all  $\text{epi } f_i$  are closed (Prop. 1.21). Then  $\text{epi } f = \bigcap_{i \in I} \text{epi } f_i$  is closed, i.e.  $f$  is lower semicontinuous.
- **(iii):** If all  $f_i$  are convex, all  $\text{epi } f_i$  are convex (Prop. 1.9). Then  $\text{epi } f = \bigcap_{i \in I} \text{epi } f_i$  is convex (Prop. 1.4), i.e.  $f$  is convex.  $\square$

**Proposition 1.74** (Basic properties of conjugate). Let  $f : H \rightarrow [-\infty, \infty]$ . Then  $f^*$  is convex and lower semicontinuous.

*Proof.* • The result is trivial if  $f(x) = -\infty$  for some  $x \in H$ . So assume  $f > -\infty$  from now on.

- Can write conjugate as:  $f^*(u) = \sup_{x \in \text{dom } f} \langle u, x \rangle - f(x)$ .
- So conjugate is pointwise supremum over family of convex, lower semicontinuous functions:  $(y \mapsto \langle y, x \rangle - f(x))_{x \in \text{dom } f}$ .
- By Proposition 1.73 have:  $f^*$  is convex and lower semicontinuous.  $\square$

Now, we return to the initial motivation and start to study the bi-conjugate  $f^{**}$ . We first give some related background.

**Definition 1.75.** Let  $f : H \rightarrow [-\infty, \infty]$ .

- The *lower semicontinuous envelope* or *closure* of  $f$  is given by

$$\bar{f} : x \mapsto \sup\{g(x) \mid g : H \rightarrow [-\infty, \infty], g \text{ is lsc}, g \leq f\}.$$

- The *convex lower semicontinuous envelope* of  $f$  is given by

$$\overline{\text{conv } f} : x \mapsto \sup\{g(x) \mid g : H \rightarrow [-\infty, \infty], g \text{ is convex, lsc}, g \leq f\}.$$

**Proposition 1.76.**  $\bar{f}$  is lower semicontinuous and  $\overline{\text{conv } f}$  is convex, lower continuous.

*Proof.* This follows directly from Prop. 1.73. □

**Proposition 1.77.** Let  $f : H \rightarrow [-\infty, \infty]$ . Then  $\text{epi } \overline{\text{conv } f} = \overline{\text{conv epi } f}$ .

*Proof.* • The claim is trivial when  $f = +\infty \Leftrightarrow \text{epi } f = \emptyset$ . So now assume  $f$  is proper: then both sets are non-empty.

- Set  $F = \overline{\text{conv } f}$  and  $D = \overline{\text{conv epi } f}$ .
- Since  $F \leq f \Rightarrow \text{epi } f \subset \text{epi } F$ . Since  $\text{epi } F$  is convex, have  $\text{conv epi } f \subset \text{epi } F$ . Since  $\text{epi } F$  is also closed (why?), have  $D = \overline{\text{conv epi } f} \subset \text{epi } F$ .
- Show converse inclusion. Let  $(x, \zeta) \in \text{epi } F \setminus D$ . Since  $D \neq \text{emptyset}$  is closed and convex, the projection onto  $D$  is well defined. Let  $(p, \pi) = P_D(x, \zeta)$ . Characterization of projection:

$$\left\langle \begin{pmatrix} x - p \\ \zeta - \pi \end{pmatrix}, \begin{pmatrix} y - p \\ \eta - \pi \end{pmatrix} \right\rangle \leq 0 \quad \text{for all } (y, \eta) \in D \quad (*)$$

- For some  $(y, \eta) \in D$ , send  $\eta \rightarrow \infty$  (which is still in  $D$ , why?). We deduce:  $\zeta - \pi \leq 0$ .
- Assume  $\zeta = \pi$ . Then  $(*)$  implies  $\langle x - p, y - p \rangle \leq 0$  for all  $(y, \eta) \in D = \overline{\text{conv epi } f}$ .
- Note:  $[\exists \eta \in \mathbb{R} \text{ s.t. } (y, \eta) \in \text{conv epi } f] \Leftrightarrow [y \in \text{conv dom } f]$ . So:  $\langle x - p, y - p \rangle \leq 0$  for all  $y \in \text{conv dom } f$  and therefore for all  $y \in \overline{\text{conv dom } f}$ .
- Also note:  $\text{dom } F \subset \overline{\text{conv dom } f} = E$ : Define function

$$g(x) = \begin{cases} F(x) & \text{if } x \in E, \\ +\infty & \text{else.} \end{cases}$$

Since  $E$  is closed and convex, and  $F$  is lsc and convex,  $g$  is lsc and convex. Since  $F \leq f$  and  $g(x) = F(x)$  for  $x \in \text{dom } f \subset E$ , have  $g \leq f$ . Since  $F$  is the convex lower semicontinuous envelope of  $f$  we must therefore have  $g \leq F$  and therefore  $\text{dom } F \subset E$ .

- So we can set  $y = x$  in projection characterization and obtain:  $\|x - p\|^2 \leq 0$ . Therefore  $x = p$  which contradicts  $(x, \zeta) \notin D$ .
- Now assume  $\zeta < \pi$ . Set  $u = \frac{x-p}{\pi-\zeta}$  and let  $y \in \text{dom } f$ ,  $\eta = f(y)$  (i.e.  $(y, \eta) \in \text{epi } f \subset D$ ). Then from characterization of projection get

$$\langle u, y - p \rangle + \pi \leq f(y).$$

So  $f$  is lower bounded by affine function  $g : y \mapsto \langle u, y - p \rangle + \pi$ . Therefore,  $g \leq F$ .

- Since  $(x, \zeta) \in \text{epi } F$  get

$$\pi \leq \frac{\|x - p\|^2}{\pi - \zeta} + \pi = g(x) \leq F(x) \leq \zeta$$

This is a contradiction and therefore there cannot be any  $(x, \zeta) \in \text{epi } F \setminus D$ . □ ← VL5

Now some basic properties of the biconjugate.

**Proposition 1.78.** Let  $f : H \rightarrow [-\infty, \infty]$ . Then  $f^{**} \leq f$  and  $f^{**}$  is the pointwise supremum over all continuous affine lower bounds on  $f$ .

*Proof.* • If  $f = \pm\infty$  then  $f^* = \mp\infty \Rightarrow$  claim is true. Now assume  $f$  is proper.

- We find:

$$\begin{aligned} f^*(u) &= \sup_{y \in H} \langle u, y \rangle - f(y) \\ f^{**}(x) &= \sup_{u \in H} \langle u, x \rangle - \left( \sup_{y \in H} \langle u, y \rangle - f(y) \right) = \sup_{u \in H} \inf_{y \in H} \langle u, x \rangle - \langle u, y \rangle + f(y) \\ &\leq \sup_{u \in H} \langle u, x - x \rangle + f(x) = f(x) \quad (\text{set } y = x \text{ in infimum}) \end{aligned}$$

- By Prop. 1.72 (Fenchel–Young):  $f(x) \geq \langle u, x \rangle - f^*(u)$  for all  $x, u \in H$ . So  $f^{**}(x) = \sup_{u \in H} \langle u, x \rangle - f^*(u)$  is the pointwise supremum over a family of continuous affine lower bounds on  $f$ .
- So  $f^{**}$  is pointwise supremum over family of convex, lsc functions  $\Rightarrow f^{**}$  is convex lsc (Prop. 1.73).
- On the other hand, let  $g(x) = \langle v, x \rangle - r \leq f(x)$  for some  $(v, r) \in H \times \mathbb{R}$  be a continuous affine lower bound. Then:

$$\begin{aligned} f^*(v) &= \sup_{x \in H} \langle v, x \rangle - \underbrace{f(x)}_{\geq g(x)} \leq \sup_{x \in H} \langle v, x \rangle - \langle v, x \rangle + r = r \\ f^{**}(x) &= \sup_{u \in H} \langle u, x \rangle - f^*(u) \geq \langle v, x \rangle - \underbrace{f^*(v)}_{\leq r} \geq \langle v, x \rangle - r = g(x) \end{aligned}$$

So  $f^{**}$  is larger (or equal) than any continuous affine lower bound on  $f$ .  $\square$

We now prove the main result of this subsection.

**Proposition 1.79.** Assume  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  has a continuous affine lower bound. Then  $f^{**} = \overline{\text{conv } f}$ .

*Proof.* • Let  $F = \overline{\text{conv } f}$ . By Prop. 1.77 have  $\text{epi } F = \overline{\text{conv epi } f}$  and by Prop. 1.65  $\text{epi } F$  is the intersection of all closed halfspaces that contain  $\text{epi } f$ .

- Let  $(v, r) \in H \times \mathbb{R}$  be the outward normal of a closed halfspace that contains  $\text{epi } f$ . Such a halfspace must exist:  $f$  has affine lower bound  $\Rightarrow \text{epi } f$  contained in some (closed) halfspace. Halfspace is closed, convex  $\Rightarrow \overline{\text{conv epi } f}$  also contained in halfspace.
- If  $r > 0$  then  $\text{epi } F = \emptyset$  and then  $f = +\infty = f^{**}$  and we are done.
- So assume that  $\text{epi } F \neq \emptyset$  and therefore  $r \leq 0$  for all closed halfspaces that contain  $\text{epi } f$ .
- Similarly,  $f^{**}$  is the pointwise supremum over all continuous affine lower bounds on  $f$ . Therefore,  $\text{epi } f^{**}$  is the intersection of all closed halfspaces that contain  $\text{epi } f$  and for which the outward normal  $(v, r)$  has  $r < 0$ .

- Therefore,  $\text{epi } F \subset \text{epi } f^{**}$  which implies  $f^{**} \leq F$ . (Also follows from  $f^{**}$  convex, lsc and  $f^{**} \leq f$ , why?)
- Let  $(u, a) \in H \times \mathbb{R}$  such that  $x \mapsto \langle u, x \rangle - a$  is a continuous affine lower bound of  $f$ . Then it is also a lower bound on  $f^{**}$  and finally  $F$ .
- Assume  $(z, \zeta) \in \text{epi } f^{**} \setminus \text{epi } F$ .
- Then there must be a closed halfspace in  $H \times \mathbb{R}$  with horizontal outward normal (i.e.  $r = 0$ ) that contains  $\text{epi } F$ , but not  $(z, \zeta)$ . That is, there is some  $(v, y) \in H^2$  such that  $\langle x - y, v \rangle \leq 0$  for all  $x \in \text{dom } F$  but  $\langle z - y, v \rangle > 0$ .

**Sketch:**  $\text{epi } F, (z, \zeta), (y, v) \in H \times H, (u, a) \in H \times \mathbb{R}$

- For  $s \geq 0$  let  $g_s(x) = \langle u, x \rangle - a + s \cdot \langle x - y, v \rangle$ . Recall that  $g_0$  is a continuous affine lower bound on  $f$ .
- For  $x \in \text{dom } f \subset \text{dom } F$  (follows from  $F \leq f$ ) have  $g_s(x) = g_0(x) + s \cdot \langle x - y, v \rangle \leq f(x)$ . So for  $s \geq 0$ ,  $g_s$  is a continuous affine lower bound on  $f$ , and thus on  $f^{**}$ .
- But for  $s \rightarrow \infty$  have  $g_s(z) = g_0(z) + s \cdot \langle z - y, v \rangle \rightarrow \infty > \zeta \geq f^{**}(z)$ .
- This is a contradiction, thus points like  $(z, \zeta)$  cannot exist and  $\text{epi } f^{**} = \text{epi } F$ .  $\square$

We obtain the famous Fenchel–Moreau Theorem as a corollary.

**Corollary 1.80** (Fenchel–Moreau). Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper. Then

$$[f \text{ is convex, lsc}] \quad \Leftrightarrow \quad [f^{**} = f] \quad \Rightarrow \quad [f^* \text{ is proper}].$$

*Proof.* •  **$\Leftarrow$  of equivalence:** If  $f = f^{**}$  then  $f$  is the conjugate of  $f^*$ . Therefore, it is convex and lsc.

- **$\Rightarrow$  of equivalence:**  $f$  is convex, lsc.  $\Rightarrow$   $\text{epi } f$  is convex, closed.  $\Rightarrow$  it is intersection of all closed halfspaces that contain  $\text{epi } f$ . If  $f$  has no continuous affine lower bound then all these halfspaces must have ‘horizontal’ normals ( $r = 0$ )  $\Rightarrow f(H) \subset \{-\infty, +\infty\}$ , which contradicts assumptions. So  $f$  must have continuous affine lower bound.
- By previous result  $f^{**} = \overline{\text{conv } f}$  which equals  $f$  since  $f$  convex, lsc.
- **$f^*$  is proper:** we have just shown that  $f$  has continuous affine lower bound, say  $f(x) \geq \langle x, v \rangle - a$  for some  $(v, a) \in H \times \mathbb{R}$ . Recall: this implies  $f^*(v) \leq a$ . Conversely,  $f$  is proper, i.e.  $f(x_0) < \infty$  for some  $x_0$  and then  $f^*(u) \geq \langle x_0, u \rangle - f(x_0)$ .  $\square$

Comment: We showed in proof: A convex lsc function must have a continuous affine lower bound. This is not true for general convex (but not lsc) functions. Recall: unbounded linear functions are convex.

A few applications: The following result is helpful to translate knowledge from  $\partial f$  or  $f^*$  onto the other. It gives the ‘extreme cases’ of the Fenchel–Young inequality.

**Proposition 1.81.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be convex, lsc, proper. Let  $x, u \in H$ . Then:

$$[u \in \partial f(x)] \quad \Leftrightarrow \quad [f(x) + f^*(u) = \langle x, u \rangle] \quad \Leftrightarrow \quad [x \in \partial f^*(u)]$$

---

Comment: Intuitive interpretation: conjugate  $f^*(u)$  computes minimal offset  $a$  such that  $y \mapsto \langle u, y \rangle - a$  is lower bound on  $f$ . If  $u \in \partial f(x)$  then  $y \mapsto \langle u, y - x \rangle + f(x)$  is affine lower bound for  $f$  that touches graph in  $x$ . So offset  $\langle u, x \rangle - f(x)$  is minimal for slope  $u$ .

---

*Proof.* • Consider first equivalence.

- $\Rightarrow$ : By Prop. 1.72 (Fenchel–Young):  $f^*(u) \geq \langle u, x \rangle - f(x)$ .
- Have  $f(y) \geq f(x) + \langle u, y - x \rangle$  for all  $y \in H$ . Get:

$$f^*(u) = \sup_{y \in H} \langle u, y \rangle - f(y) \leq \sup_{y \in H} \langle u, y \rangle - \langle u, y - x \rangle - f(x) = \langle u, x \rangle - f(x)$$

- So  $f^*(u) + f(x) = \langle u, x \rangle$ .
- $\Leftarrow$ :

$$f^*(u) = \langle x, u \rangle - f(x) = \sup_{y \in H} \langle y, u \rangle - f(y) \geq \langle y, u \rangle - f(y) \text{ for all } y \in H$$

So  $f(y) \geq \langle u, y - x \rangle + f(x)$  for all  $y \in H$ .  $\Rightarrow u \in \partial f(x)$ .

- For second equivalence, apply first equivalence to  $f^*$  and use that  $f^{**} = f$ . □

Now we can relate one-homogeneous functions and indicator functions:

**Definition 1.82.** A function  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  is *positively 1-homogeneous* if  $f(\lambda \cdot x) = \lambda \cdot f(x)$  for all  $x \in H$ ,  $\lambda \in \mathbb{R}_{++}$ .

**Proposition 1.83.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$ . Then  $f$  is a convex, lsc, positively 1-homogeneous function if and only if  $f = (\iota_C)^* = \sigma_C$  for some closed, convex, non-empty  $C \subset H$ .

---

Comment: Relation between indicator functions and support functions:  $\iota_C^* = \sigma_C$ .

---

*Proof.* •  $\Leftarrow$ :  $\iota_C^*$  is lsc and convex. Moreover, for  $x \in H$ ,  $\lambda \in \mathbb{R}_{++}$

$$\iota_C^*(\lambda \cdot x) = \sigma_C(\lambda \cdot x) = \sup_{y \in C} \langle y, \lambda \cdot x \rangle = \lambda \sup_{y \in C} \langle y, x \rangle = \lambda \cdot \sigma_C(x).$$

So  $\iota_C^*$  is positively 1-homogeneous.

- $\Rightarrow$ : Observe:  $f(0) = 0$  (why?). So

$$f^*(u) = \sup_{x \in H} \langle u, x \rangle - f(x) \geq 0 \quad (\text{set } x = 0 \text{ in sup}).$$

- If, for fixed  $u \in H$  there is some  $x \in H$  such that  $\langle u, x \rangle - f(x) > 0$ , then

$$f^*(u) \geq \limsup_{k \rightarrow \infty} \langle u, k \cdot x \rangle - f(k \cdot x) = \limsup_{k \rightarrow \infty} k \cdot (\langle u, x \rangle - f(x)) = \infty.$$

- So  $f^*(H) \subset \{0, +\infty\}$  and therefore  $f^* = \iota_C$  for some  $C \subset H$ . Since  $f^*$  is convex, lsc  $\Rightarrow C$  is convex, closed (why?).
- Since  $f$  is convex, lsc, proper ( $f(0) = 0$ ) have  $f = f^{**} = \iota_C^*$ . □

This allows us to describe subdifferential of 1-homogeneous functions.

**Corollary 1.84.** If  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  is convex, lsc, positively 1-homogeneous, then  $f = \sigma_C$  where  $C = \partial f(0)$ .

*Proof.* • By assumption,  $f = \sigma_C$  for some closed, convex  $C \subset H$ ,  $f^* = \iota_C$ .

• Then  $[u \in \partial f(0)] \Leftrightarrow [0 \in \partial f^*(u) = \partial \iota_C(u)] \Leftrightarrow [u \in C]$ . □

**Example 1.85.** Go through Examples 1.69 and study biconjugates. Note the relation between positively 1-homogeneous functions and indicator functions.

## 1.6 Convex variational problems

**Remark 1.86** (Motivation). We want to find minimizers of functionals. Standard argument: minimizing sequence + compactness: Weierstrass provides cluster point. Lower semicontinuity: cluster point is minimizer.

Problem: compactness in infinite dimensions is far from trivial. Example: orthonormal sequences  $(x_k)_{k \in \mathbb{N}}$ ,  $\langle x_i, x_j \rangle = \delta_{i,j}$  (e.g. ‘traveling bumps’ in  $L^2(\mathbb{R})$  or canonical ‘basis vectors’ in  $\ell^2(\mathbb{N})$ ).  $\Rightarrow$  closed unit ball in infinite-dimensional Hilbert spaces is not compact.

Recall: avoided this problem for proof of existence of projection via Cauchy sequence, but this argument will not work in general.  $\Rightarrow$  we need a different tool.

**Definition 1.87** (Weak convergence on Hilbert space). A sequence  $(x_k)_k$  in  $H$  is said to *converge weakly* to some  $x \in H$ , we write  $x_k \rightharpoonup x$ , if for all  $u \in H$

$$\lim_{k \rightarrow \infty} \langle u, x_k \rangle = \langle u, x \rangle .$$

---

Comment: For now only use weak convergence for Hilbert spaces. More general and detailed discussion will follow later.

---

**Remark 1.88.** Weak convergence corresponds to weak topology. Weak topology is coarsest topology in which all maps  $x \mapsto \langle u, x \rangle$  for all  $u \in H$  are continuous (this implies precisely that  $\langle u, x_k \rangle \rightarrow \langle u, x \rangle$  for weakly converging sequences  $x_k \rightharpoonup x$ ). So, subbasis is given by all open halfspaces. Weak topology still yields Hausdorff space (e.g. for any two distinct points  $x, y \in H$  can find open halfspace  $A$  such that  $x \in A, y \notin A$ ). Need Hausdorff property for uniqueness of limits.

In general it is easier to obtain compactness with respect to the weak topology due to the following theorem.

**Theorem 1.89** (Banach–Alaoglu). The closed unit ball of  $H$  is weakly compact.

**Corollary 1.90.** Weakly closed, bounded subsets of  $H$  are weakly compact.

*Proof.* Let  $C \subset H$  be weakly closed and bounded. Then there is some  $\rho \in \mathbb{R}_{++}$  such that  $C \subset \overline{B(0, \rho)}$ , which is weakly compact by Banach–Alaoglu.  $C$  is a weakly closed subset of a weakly compact set, therefore it is weakly compact.  $\square$

**Example 1.91** (Orthonormal sequence and Bessel’s inequality). Let  $(x_k)_{k \in \mathbb{N}}$  be an orthonormal sequence in  $H$ , i.e.  $\langle x_i, x_j \rangle = \delta_{i,j}$  for all  $i, j \in \mathbb{N}$ , and let  $u \in H$ . Then for all  $N \in \mathbb{N}$

$$\begin{aligned} 0 &\leq \left\| u - \sum_{k=1}^N x_k \langle x_k, u \rangle \right\|^2 = \|u\|^2 - 2 \left\langle u, \sum_{k=1}^N x_k \langle x_k, u \rangle \right\rangle + \left\| \sum_{k=1}^N x_k \langle x_k, u \rangle \right\|^2 \\ &= \|u\|^2 - 2 \sum_{k=1}^N \langle u, x_k \rangle^2 + \sum_{k=1}^N \langle u, x_k \rangle^2 = \|u\|^2 - \sum_{k=1}^N \langle u, x_k \rangle^2 . \end{aligned}$$

So  $\|u\|^2 \geq \sum_{k=1}^N \langle u, x_k \rangle^2$  for all  $N$  (which then also holds in the limit  $N \rightarrow \infty$ ) and  $\langle u, x_k \rangle \rightarrow 0$  as  $k \rightarrow \infty$ . Therefore  $x_k \rightharpoonup 0$ . (But clearly not  $x_k \rightarrow 0$ .)

The previous example shows that weak convergence does in general not imply strong convergence. We require an additional condition.

**Proposition 1.92.** Let  $(x_k)_{k \in \mathbb{N}}$  be a sequence in  $H$  and let  $x \in H$ . Then the following are equivalent:

$$[x_k \rightarrow x] \quad \Leftrightarrow \quad [x_k \rightharpoonup x \text{ and } \|x_k\| \rightarrow \|x\|]$$

*Proof.* •  $\Rightarrow$ : For every  $u \in H$  have  $y \mapsto \langle u, y \rangle$  is continuous. Therefore, if  $x_k \rightarrow x$  one finds  $\langle u, x_k \rangle \rightarrow \langle u, x \rangle$  for all  $u \in H$ , therefore  $x_k \rightharpoonup x$ . The norm function is also (strongly) continuous, therefore it also implies  $\|x_k\| \rightarrow \|x\|$ .

•  $\Leftarrow$ :

$$\|x_k - x\|^2 = \underbrace{\|x_k\|^2}_{\rightarrow \|x\|^2} - 2 \underbrace{\langle x_k, x \rangle}_{\rightarrow \langle x, x \rangle} + \|x\|^2 \rightarrow 0 \quad \square$$

**Remark 1.93.** In the previous example we find indeed  $\lim_{k \rightarrow \infty} \|x_k\| = 1 \neq \|0\|$ . Therefore, the sequence cannot converge strongly.

**Theorem 1.94** (Characterization of infinite-dimensional Hilbert spaces). The following are equivalent:

- (i)  $H$  is finite-dimensional.
- (ii) The closed unit ball  $\overline{B(0, 1)}$  is compact.
- (iii) The weak topology of  $H$  coincides with its strong topology.
- (iv) The weak topology of  $H$  is metrizable.

**Remark 1.95.** Note that item (iv) implies that for the weak topology we can in general not equate sequential closedness and closedness, as for the strong topology (cf. Remark 1.13). We will now show that it remains at least equivalent for convex sets (and functions). ← VL6

**Proposition 1.96.** Let  $C \subset H$  be convex. Then the following are equivalent:

- (i)  $C$  is weakly sequentially closed.
- (ii)  $C$  is sequentially closed.
- (iii)  $C$  is closed.
- (iv)  $C$  is weakly closed.

*Proof.* • (i)  $\Rightarrow$  (ii): Let  $(x_k)_k$  be a sequence in  $C$  that converges strongly to some  $x \in H$ . Prop. 1.92:  $[x_k \rightarrow x] \Rightarrow [x_k \rightharpoonup x]$ . Therefore,  $x \in C$  since  $C$  is weakly sequentially closed. Therefore,  $C$  is (strongly) sequentially closed.

- (ii)  $\Leftrightarrow$  (iii): The two are equivalent because the strong topology is metrizable (cf. Remark 1.13).
- (iii)  $\Rightarrow$  (iv): For this need convexity.  $C$  is closed and convex. Therefore,  $C$  is the intersection of all closed halfspaces that contain  $C$ .
- A subbasis for the open sets of the weak topology are open halfspaces. So subbasis for weakly closed sets are closed halfspaces.  $C$  can be written as intersection of weakly closed sets.  $\Rightarrow C$  is weakly closed.

- (iv)  $\Rightarrow$  (i): Sequential closedness is implied by ‘full’ closedness. (Proof: Let  $C$  be weakly closed. Let  $(x_k)_k$  be a sequence in  $C$  with  $x_k \rightharpoonup x$  for some  $x \in H$ . Assume  $x \notin C$ . Then there is some weakly open  $U$  such that  $x \in U$ ,  $U \cap C = \emptyset$ . But since  $x_k \rightharpoonup x$ , for sufficiently large  $k$  one must have  $x_k \in U$  which is a contradiction.)  $\square$

**Corollary 1.97.** For a convex function  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  the notions of weak, strong, sequential and ‘full’ lower semicontinuity coincide.

*Proof.* When  $f$  is convex, all its sublevel sets are convex and for these all corresponding notions of closedness coincide.  $\square$

**Corollary 1.98.** The norm  $x \mapsto \|x\|$  is (sequentially) weakly lower semicontinuous.

**Remark 1.99.** Note: the norm is not (sequentially) weakly continuous in infinite dimensions. Recall an orthonormal sequence  $(x_k)_{k \in \mathbb{N}}$ . Then  $x_k \rightharpoonup 0$  but  $\|x_k\| \rightarrow 1$ .

**Corollary 1.100.** The closed unit ball  $\overline{B(0,1)}$  is weakly closed. But in infinite dimensions the (strongly) open unit ball  $B(0,1)$  is not weakly open.

*Proof.* •  $\overline{B(0,1)}$  is a convex set. Therefore the notion of strong and weak closure coincide.

- Consider once more an orthonormal sequence  $(x_k)_{k \in \mathbb{N}}$ . Then  $x_k \notin B(0,1)$  for all  $k$ , but  $x_k \rightharpoonup 0 \in B(0,1)$ .  $\square$

So in the following we resort to weak topology to obtain minimizers via compactness. We do not have to worry too much about the new notion of lower semicontinuity. But since (strongly) open balls are no longer weakly open, we will face some subtleties when we try to extract converging subsequences from minimizing sequences: we do not know whether weak compactness implies weak sequential compactness. This is provided by the following theorem:

**Theorem 1.101** (Eberlein–Šmulian). For subsets of  $H$  weak compactness and weak sequential compactness are equivalent.

Now we give a prototypical theorem for the existence of minimizers.

**Proposition 1.102.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be convex, lower semicontinuous. Let  $C \subset H$  be closed, convex such that for some  $r \in \mathbb{R}$  the set  $C \cap S_r(f)$  is non-empty and bounded. Then  $f$  has a minimizer over  $C$ .

*Proof.* • The sets  $C$  and  $S_r(f)$  are closed and convex. So  $D = C \cap S_r(f)$  is closed and convex and by assumption bounded.

- $D$  closed, convex  $\Rightarrow D$  is weakly closed (Prop. 1.96).
- $D$  bounded, weakly closed  $\Rightarrow$  weakly compact (Cor. 1.90 of Banach–Alaoglu).
- $D$  weakly compact  $\Rightarrow$  weakly sequentially compact (Thm. 1.101, Eberlein–Šmulian).
- Since  $D = C \cap S_r(f)$  is non-empty, we can confine minimization of  $f$  over  $C$  to minimization of  $f$  over  $D$ .
- Let  $(x_k)_{k \in \mathbb{N}}$  be minimizing sequence of  $f$  over  $D$ . Since  $D$  is weakly sequentially compact, there is a subsequence of  $(x_k)_k$  that converges to some  $x \in D$  in the weak topology.

- Since  $f$  is convex and lower semicontinuous, it is weakly sequentially lower semicontinuous (Cor. 1.97). Therefore,  $x$  is a minimizer.  $\square$

A useful criterion to check whether the sublevel sets of a function are bounded is coerciveness.

**Definition 1.103** (Coerciveness). A function  $f : H \rightarrow [-\infty, \infty]$  is *coercive* if  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$ .

**Proposition 1.104.** Let  $f : H \rightarrow [-\infty, \infty]$ . Then  $f$  is coercive if and only if its sublevel sets  $S_r(f)$  are bounded for all  $r \in \mathbb{R}$ .

*Proof.* • Assume  $S_r(f)$  is unbounded for some  $r \in \mathbb{R}$ . Then we can find a sequence  $(x_k)$  in  $S_r(f)$  with  $\|x_k\| \rightarrow \infty$  but  $\limsup_{k \rightarrow \infty} f(x_k) \leq r$ .

- Assume  $S_r(f)$  is bounded for every  $r \in \mathbb{R}$ . Let  $(x_k)_k$  be an unbounded sequence with  $\lim_{k \rightarrow \infty} \|x_k\| \rightarrow \infty$ . Then for any  $s \in \mathbb{R}$  there is some  $N \in \mathbb{N}$  such that  $x_k \notin S_s(f)$  for  $k \geq N$ . Hence,  $\liminf_{k \rightarrow \infty} f(x_k) \geq s$ . Since this holds for any  $s \in \mathbb{R}$ , have  $\lim_{k \rightarrow \infty} f(x_k) = \infty$ .  $\square$

Once existence of minimizers is ensured, uniqueness is simpler to handle. ‘Mere’ convexity is not sufficient for uniqueness. We require additional assumptions. Strict convexity is sufficient.

**Proposition 1.105.** Consider the setting of Prop. 1.102. If  $f$  is strictly convex then there is a unique minimizer.

*Proof.* Assume  $x$  and  $y \in C$  are two distinct minimizers. Then  $f(x) = f(y)$ . Then  $z = (x + y)/2 \in C$  and  $f(z) < \frac{1}{2}f(x) + \frac{1}{2}f(y) = f(x) = f(y)$ . So neither  $x$  nor  $y$  can be minimizers.  $\square$

## 1.7 Proximal operators

**Definition 1.106.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  convex, lsc and proper. Then the map  $\text{Prox}_f : H \rightarrow H$  is given by

$$x \mapsto \underset{y \in H}{\operatorname{argmin}} \left( \frac{1}{2} \|x - y\|^2 + f(y) \right).$$

The minimizer exists and is unique, so the map is well-defined.

**Remark 1.107** (Motivation). Interpretation: near point  $x$  try to minimize  $f$ , but penalize if we move too far from  $x$ . Intuitively: do small step in direction where  $f$  decreases, similarly to gradient descent, but  $\text{Prox}_f$  is also defined for non-smooth  $f$ .

The proximal operator will be our basic tool for optimization. Later we will show that we can optimize  $f + g$  by only knowing the proximal operators of  $f$  and  $g$  separately. This is the basis for the *proximal splitting* strategy. One tries to decompose the objective into components such that the proximal operator for each component is easy to compute.

*Proof that  $\text{Prox}_f$  is well-defined.* • Since  $f$  is convex, lsc, proper  $\Rightarrow f^*$  is proper. Therefore  $f$  has a continuous affine lower bound, which we denote by  $\tilde{f} : y \mapsto \langle u, y \rangle - r$ .

- For fixed  $x \in H$  let  $g : y \mapsto \frac{1}{2} \|x - y\|^2$ . By ‘completing the square’ we get

$$\tilde{f}(y) + g(y) = \frac{1}{2} \|x - y\|^2 + \langle u, y \rangle - r = \frac{1}{2} \|y - v\|^2 + C$$

for some  $v \in H, C \in \mathbb{R}$ . So sublevel sets of  $\tilde{f} + g$  are bounded.

- Since  $\tilde{f} \leq f$  have  $S_r(f + g) \subset S_r(\tilde{f} + g)$ , so sublevel sets of  $f + g$  are bounded.
- Since  $f$  is proper and  $g$  is finite, there is some  $r \in \mathbb{R}$  such that  $S_r(f + g)$  is non-empty.
- Using Prop. 1.102 with  $C = H$  and  $f = f + g$  we find that  $f + g$  has a minimizer over  $H$ .
- Since  $f$  is convex and  $g$  is strictly convex,  $f + g$  is strictly convex. Prop. 1.105  $\Rightarrow$  this minimizer is unique.  $\square$

Characterization of proximal operator.

**Proposition 1.108.** Let  $f$  be convex, lsc, proper, let  $x \in H$ . Then

$$[p = \text{Prox}_f(x)] \Leftrightarrow [\langle y - p, x - p \rangle + f(p) \leq f(y) \text{ for all } y \in H] \Leftrightarrow [x - p \in \partial f(p)]$$

*Proof.* • The second equivalence is trivial. We prove the first.

- $\Rightarrow$ : Assume  $p = \text{Prox}_f(x)$ , let  $y \in H$ . For  $\alpha \in [0, 1]$  let  $p_\alpha = \alpha \cdot y + (1 - \alpha) \cdot p$ .
- Then  $f(p_\alpha) + \frac{1}{2} \|x - p_\alpha\|^2 \geq f(p) + \frac{1}{2} \|x - p\|^2$ .
- By convexity of  $f$ :  $f(p_\alpha) \leq \alpha \cdot f(y) + (1 - \alpha) \cdot f(p)$ .
- We get:

$$\alpha \cdot f(y) + (1 - \alpha) \cdot f(p) + \frac{1}{2} \|x - p_\alpha\|^2 \geq f(p) + \frac{1}{2} \|x - p\|^2$$

- Setting  $g(\alpha) = \alpha \cdot f(y) + (1 - \alpha) \cdot f(p) + \frac{1}{2}\|x - p_\alpha\|^2$  this translates to  $g(\alpha) \geq g(0)$  for  $\alpha \in [0, 1]$ .
- Note that  $g$  is differentiable, so we must have  $\frac{d}{d\alpha}g(\alpha)|_{\alpha=0} \geq 0$ . This implies:

$$f(y) - f(p) + \langle x - p, y - p \rangle \geq 0.$$

- $\Leftarrow$ : For fixed  $x$  let  $g : y \mapsto \frac{1}{2}\|x - y\|^2$ . Then  $\partial g(y) = \{y - x\}$ . Then

$$\begin{aligned} [x - p \in \partial f(p)] &\Leftrightarrow [0 \in p - x + \partial f(p) = \partial g(p) + \partial f(p)] \\ &\Rightarrow (\partial f + \partial g \subset \partial(f + g), \text{Prop. 1.32}) \quad [0 \in \partial(g + f)(p)] \\ &\Leftrightarrow [p \in \operatorname{argmin}(g + f)] \quad \Leftrightarrow [p = \operatorname{Prox}_f(x)] \quad \square \end{aligned}$$

---

Comment: Since we did not prove any results of the form  $\partial(f + g) = \partial f + \partial g$  we had to ‘manually’ do the  $\Rightarrow$ -argument.

---

**Example 1.109** (Projections). Projections are special cases of proximal operators. Let  $C \subset H$  be non-empty, closed, convex. We find

$$P_C x = \operatorname{argmin}_{p \in C} \frac{1}{2}\|x - p\|^2 = \operatorname{argmin}_{p \in H} \frac{1}{2}\|x - p\|^2 + \iota_C(p) = \operatorname{Prox}_{\iota_C}(x).$$

Then the characterization for projections (Prop. 1.49) is a special case of Prop. 1.108:

$$\begin{aligned} [p = \operatorname{Prox}_{\iota_C}(x)] &\Leftrightarrow [\langle y - p, x - p \rangle + \iota_C(p) \leq \iota_C(y) \text{ for all } y \in H] \\ &\Leftrightarrow [p \in C \wedge \langle y - p, x - p \rangle \leq 0 \text{ for all } y \in C] \quad \Leftrightarrow [p = P_C x] \end{aligned}$$

Similarly, the characterization of projections via the normal cone (Cor. 1.50) is a special case of the characterization of the proximal operator via the subdifferential: Recall  $\partial \iota_C(y) = N_C y$  (Prop. 1.47). Then:

$$[x \in p + \partial \iota_C(p)] \quad \Leftrightarrow \quad [x \in p + N_C p]$$

So, conversely we may think of the proximal operator as a generalization of projections with ‘soft walls’: instead of paying an infinite penalty when we leave  $C$ , the penalty is now controlled by a more general function  $f$ .

A few more examples, that are not projections:

**Example 1.110.** Let  $\lambda > 0$ .

(i)  $f(y) = \frac{\lambda}{2}\|y\|^2$ :  $[p = \operatorname{Prox}_f(x)] \Leftrightarrow [x - p \in \partial f(p) = \{\lambda \cdot p\}] \Leftrightarrow [p = x/(1 + \lambda)]$ . So  $\operatorname{Prox}_f(x) = x/(1 + \lambda)$ .

(ii)  $f(y) = \lambda \cdot \|y\|$ : Recall:

$$\partial f(y) = \lambda \cdot \begin{cases} \frac{y}{\|y\|} & \text{if } y \neq 0, \\ B(0, 1) & \text{if } y = 0. \end{cases}$$

If  $x \in \lambda \cdot \overline{B(0, 1)}$  we find that  $p = 0$  is a solution to  $x \in p + \partial f(p)$ . Otherwise, we need to solve  $x = p + \frac{\lambda p}{\|p\|}$  for some  $p \neq 0$ . We deduce that  $p = \rho \cdot x$  for some  $\rho \in \mathbb{R} \setminus \{0\}$  (since  $p$  and  $x$  must be linearly dependent) and get:

$$[x = \rho \cdot x + \frac{\lambda \rho x}{\|\rho \cdot x\|}] \Leftrightarrow [1 = \rho + \frac{\lambda}{\|x\|}] \Leftrightarrow [\rho = 1 - \frac{\lambda}{\|x\|}] \Leftrightarrow [p = x - \frac{\lambda x}{\|x\|}]$$

We summarize:

$$\text{Prox}_f(x) = \begin{cases} 0 & \text{if } x \in \overline{B(0, \lambda)} \\ x - \frac{\lambda x}{\|x\|} & \text{else.} \end{cases}$$

Interpretation: if  $\|x\| > \lambda$  we move towards the origin with stepsize  $\lambda$ , otherwise, go directly to origin.

**Example 1.111** (Comparison with explicit gradient descent). Assume  $f$  is Gâteaux differentiable. Then  $\partial f(x) = \{\nabla f(x)\}$ . Consider a naive discrete gradient descent with stepsize  $\lambda > 0$  for some initial  $x^{(0)} \in H$ :

$$x^{(\ell+1)} := x^{(\ell)} - \lambda \nabla f(x^{(\ell)})$$

For comparison consider repeated application of the proximal operator on some initial  $y^{(0)} \in H$ :

$$y^{(\ell+1)} := \text{Prox}_{\lambda f}(y^{(\ell)})$$

We find  $y^{(\ell)} \in y^{(\ell+1)} + \lambda \partial f(y^{(\ell+1)}) = \{y^{(\ell+1)} + \lambda \nabla f(y^{(\ell+1)})\}$ , so

$$y^{(\ell+1)} = y^{(\ell)} - \lambda \nabla f(y^{(\ell+1)}).$$

This is called an *implicit* gradient descent, since the new iterate depends on the gradient at the position of the new iterate, and it is thus only implicitly defined. For comparison, the above rule for  $x^{(\ell+1)}$  is called *explicit*.

Usually, the explicit gradient scheme is much easier to implement, but the proximal operator has several important advantages:

- The proximal scheme also works, when  $f$  is not differentiable. (But it must be convex.)
- The proximal scheme can be started from any point in  $H$ , even from outside of  $\text{dom } f$ .
- The proximal scheme tends to converge more robustly.

As an illustration of the latter point return to two previous examples:

(i)  $f(x) = \frac{1}{2}\|x\|^2$ . Then  $\nabla f(x) = x$  and we get

$$x^{(\ell+1)} = x^{(\ell)} - \lambda x^{(\ell)} = x^{(0)} (1 - \lambda)^\ell.$$

This converges geometrically to  $x^{(\ell)} \rightarrow 0$  for  $|1 - \lambda| < 1 \Leftrightarrow \lambda \in (0, 2)$ . For  $\lambda > 1$  the solution oscillates around the minimizer, for  $\lambda > 2$  the sequence diverges.

For comparison we get

$$y^{(\ell+1)} = y^{(\ell)} / (1 + \lambda) = y^{(0)} (1 + \lambda)^{-\ell}$$

This converges geometrically for all  $\lambda > 0$ . For very small positive  $\lambda$  we have  $(1 + \lambda)^{-1} \approx 1 - \lambda$  and the implicit and explicit scheme act similarly (for the first few iterations). Intuitively, this stems from the fact that if  $f$  is continuously differentiable and the stepsize is small, then  $\nabla f(x^{(\ell)}) \approx \nabla f(x^{(\ell+1)})$ .

(ii)  $f(x) = \|x\|$ . Then  $\nabla f(x) = x/\|x\|$  for  $x \neq 0$  and we obtain

$$x^{(\ell+1)} = x^{(\ell)} - \frac{\lambda x^{(\ell)}}{\|x^{(\ell)}\|}$$

For  $\|x\| > \lambda$  this is the same effect as the proximal operator, but for  $\|x\| < \lambda$  it does not jump to the origin and terminate, but oscillates around the minimizer.

The examples indicate that  $\text{Prox}_f(x)$  moves from  $x$  towards a minimum of  $f$ . We also observe that a prefactor  $\lambda$  acts like a stepsize. We establish a few corresponding results.

**Proposition 1.112.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be convex, lsc, proper. Let  $\lambda \in \mathbb{R}_{++}$ .

- (i)  $[x \in \argmin f] \Leftrightarrow [x = \text{Prox}_f(x)]$ .
- (ii)  $[x \notin \argmin f] \Rightarrow [f(\text{Prox}_f(x)) < f(x)]$ .
- (iii) Let  $p = \text{Prox}_f(x)$ ,  $C = S_{f(p)}(f)$ . Then  $p = P_C x$ .
- (iv) The function  $\lambda \mapsto \|x - \text{Prox}_{\lambda f}(x)\|$  is increasing.
- (v) The function  $\lambda \mapsto f(\text{Prox}_{\lambda f}(x))$  is decreasing. ← VL7

*Proof.* • **(i):**  $[x = \text{Prox}_f(x)]$  (Characterization of Prox, Prop. 1.108)  $\Leftrightarrow [x \in x + \partial f(x)] \Leftrightarrow [0 \in \partial f(x)] \Leftrightarrow$  (Fermat's rule, Prop. 1.26)  $[x \in \argmin f]$ .

- **(ii):** By assumption  $x \notin \argmin f$ . Let  $p = \text{Prox}_f(x)$ . By (i)  $p \neq x$  and then

$$\frac{1}{2}\|x - p\|^2 + f(p) < \frac{1}{2}\|x - x\|^2 + f(x) = f(x)$$

which implies  $f(x) - f(p) > \frac{1}{2}\|x - p\|^2 > 0$ .

- **(iii):** By construction  $p \in C$ . Let  $p' = P_C x$ . So  $p' \in C = S_{f(p)} \Rightarrow f(p') \leq f(p)$ . Assume  $p' \neq p$ . Then  $\|x - p\| > \|x - p'\|$  ( $p'$  is point that minimizes distance to  $x$  among all points in  $C$ ). Then

$$\frac{1}{2}\|x - p'\|^2 + f(p') < \frac{1}{2}\|x - p\|^2 + f(p)$$

and therefore  $p'$  is a better candidate for  $\text{Prox}_f(x)$  than  $p$ . Therefore we must have  $p' = p$ .

- **(iv):** We use the monotonicity of the subdifferential for this (Prop. 1.29). Let  $0 < \lambda_1 \leq \lambda_2$ . Let  $p_i = \text{Prox}_{\lambda_i f}(x)$  and set  $u_i = x - p_i$  for  $i = 1, 2$ .
- Let  $\Delta u = u_2 - u_1$ ,  $\Delta p = p_2 - p_1$ . From  $x = p_i + u_i$  we get  $\Delta u = -\Delta p$ .
- By characterization of the proximal operator we find:  $u_i \in \lambda_i \partial f(p_i)$ .

---

**Sketch:**  $x, p_1, u_1$ , then transition from  $p_1$  to  $p_2$  'towards'  $x$  and change of  $u_1$  to  $u_2$  as dictated by  $\lambda_2 \geq \lambda_1$  and monotonicity of subdifferential.

---

- By monotonicity of the subdifferential:

$$\begin{aligned} 0 &\leq \left\langle \frac{u_2}{\lambda_2} - \frac{u_1}{\lambda_1}, p_2 - p_1 \right\rangle \\ 0 &\leq \left\langle u_2 - \frac{\lambda_2}{\lambda_1} u_1, \Delta p \right\rangle = \left\langle \Delta u - \frac{\lambda_2 - \lambda_1}{\lambda_1} u_1, \Delta p \right\rangle = -\|\Delta p\|^2 - \frac{\lambda_2 - \lambda_1}{\lambda_1} \langle u_1, \Delta p \rangle \end{aligned}$$

We deduce  $\langle u_1, \Delta p \rangle \leq 0$ . Then

$$\begin{aligned} \|x - p_2\|^2 &= \|x - p_1 - (p_2 - p_1)\|^2 = \|x - p_1 - \Delta p\|^2 \\ &= \|x - p_1\|^2 - 2 \langle u_1, \Delta p \rangle + \|\Delta p\|^2 \geq \|x - p_1\|^2. \end{aligned}$$

- **(v):** Use notation from previous point. Assume  $f(p_2) > f(p_1)$ , let  $C = S_{f(p_2)}(f)$ . Then  $p_1 \neq p_2$  and  $p_1 \in C$ . By (iii) have  $p_2 = P_C x$ , therefore  $\|x - p_2\| < \|x - p_1\|$ , which contradicts (iv). Therefore we must have  $f(p_2) \leq f(p_1)$ .  $\square$

It turns out that there is a surprisingly simple relation between the proximal operators for  $f$  and  $f^*$ . This can be used to compute one via the other, in case one seems easier to implement.

**Proposition 1.113** (Moreau decomposition). Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be convex, lsc and proper,  $x \in H$ . Then  $\text{Prox}_f(x) + \text{Prox}_{f^*}(x) = x$ .

*Proof.* Let  $p \in H$ . Then:

$$\begin{aligned} [p = \text{Prox}_f(x)] &\Leftrightarrow [x - p \in \partial f(p)] \Leftrightarrow (\text{Prop. 1.81}) [p \in \partial f^*(x - p)] \\ &\Leftrightarrow [x - (x - p) \in \partial f^*(x - p)] \Leftrightarrow [x - p = \text{Prox}_{f^*}(x)] \end{aligned} \quad \square$$

**Example 1.114** (Moreau decomposition for projections). Let  $C$  be a closed subspace of  $H$ . Then  $\iota_C$  is convex, lsc. Consider the conjugate

$$\iota_C^*(x) = \sup_{y \in H} \langle x, y \rangle - \iota_C(y) = \sup_{y \in C} \langle x, y \rangle = \begin{cases} 0 & \text{if } x \perp y \text{ for all } y \in C, \\ +\infty & \text{else} \end{cases} = \iota_{C^\perp}(x)$$

So  $\iota_C^*$  is the indicator of the orthogonal complement of  $C$ . Then  $\text{Prox}_{\iota_C} = P_C$  and  $\text{Prox}_{\iota_C^*} = P_{C^\perp}$  and the Moreau decomposition yields:

$$x = P_C x + P_{C^\perp} x$$

which is the orthogonal decomposition of  $x$ . So we may interpret the Moreau decomposition as a generalization in the same sense that the proximal operator generalizes the projection.

**Example 1.115.** In an implicit descent scheme  $x^{(\ell+1)} = \text{Prox}_f(x^{(\ell)})$  we now find that  $x^{(\ell+1)} + \text{Prox}_{f^*}(x^{(\ell)}) = x^{(\ell)}$ ,  $\Rightarrow x^{(\ell+1)} = x^{(\ell)} - \text{Prox}_{f^*}(x^{(\ell)})$ , so  $\text{Prox}_{f^*}$  gives the ‘implicit gradient steps’  $\Delta x^{(\ell+1)}$ .

Let  $f(x) = \|x\|$ . Then  $f^* = \iota_{\overline{B(0,1)}}$  and

$$\text{Prox}_f(x) = \begin{cases} 0 = x - x & \text{if } x \in \overline{B(0,1)}, \\ x - \frac{x}{\|x\|} & \text{else.} \end{cases} \quad \text{Prox}_{f^*}(x) = \begin{cases} x & \text{if } x \in \overline{B(0,1)}, \\ \frac{x}{\|x\|} & \text{else.} \end{cases}$$

Interpretation: if  $x^{(\ell)} \in \overline{B(0,1)}$  then  $\Delta x^{(\ell+1)} = -x^{(\ell)}$  (i.e. we jump directly to the origin). Otherwise we move by  $-\frac{x^{(\ell)}}{\|x^{(\ell)}\|}$ .

## 1.8 Proximal algorithm

Now we discuss the simplest possible algorithm built from the proximal operator: simple iteration of the proximal operator of the objective. We have already discussed this in the context of simple examples (Example 1.111) and shown some preliminary results that support our intuition (Prop. 1.112).

**Proposition 1.116** (Proximal algorithm). Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper, convex, lsc with  $\operatorname{argmin} f \neq \emptyset$ . For some  $\gamma \in \mathbb{R}_{++}$  and  $x^{(0)} \in H$  set

$$x^{(\ell+1)} = \operatorname{Prox}_{\gamma f}(x^{(\ell)}).$$

Then

- (i)  $(x^{(\ell)})_\ell$  is a minimizing sequence of  $f$ .
- (ii)  $(x^{(\ell)})_\ell$  converges weakly to some point in  $\operatorname{argmin} f$ .

For the proof we need to gather some auxiliary definitions and results.

**Definition 1.117.** Let  $C \subset H$  be non-empty. Let  $(x_k)_k$  be a sequence in  $H$ .  $(x_k)_k$  is *Fejér monotone* with respect to  $C$  if for all  $y \in C$  and  $k \in \mathbb{N}$

$$\|x_{k+1} - y\| \leq \|x_k - y\|.$$

**Proposition 1.118** (Basic consequences of Fejér monotonicity). If  $(x_k)_k$  is Fejér monotone with respect to some  $C$  then:

- (i)  $(x_k)_k$  is bounded.
- (ii) For all  $y \in C$  the sequence  $(\|x_k - y\|)_k$  converges.
- (iii) Let  $d_C(z) := \inf_{y \in C} \|y - z\|$ . The sequence  $(d_C(x_k))_k$  is decreasing and converges.

*Proof.* • **(i):** Let  $y \in C$ . Then by definition  $\|x_k - y\| \leq \|x_0 - y\|$ , so  $(x_k)_k$  is in  $\overline{B(y, \|x_0 - y\|)}$ .

- **(ii):** By definition, the sequence  $(\|x_k - y\|)_k$  is decreasing and bounded from below. Therefore  $\lim_{k \rightarrow \infty} \|x_k - y\| = \inf_k \|x_k - y\|$ .
- **(iii):** We find:  $d_C(x_{k+1}) = \inf_{y \in C} \|x_{k+1} - y\| \leq \inf_{y \in C} \|x_k - y\| = d_C(x_k)$ . Therefore, the sequence is decreasing. Also clearly  $d_C(x_k) \geq 0$  for all  $k$ . Therefore the sequence  $(d_C(x_k))_k$  is converging.  $\square$

**Lemma 1.119.** Let  $(x_k)_k$  be a bounded sequence in  $H$ . Then  $(x_k)_k$  converges weakly if and only if it has at most one weak sequential cluster point.

*Proof.* • Assume  $(x_k)_k$  converges weakly. Since the weak topology is Hausdorff, it has a unique limit which is its only cluster point.

- Assume  $(x_k)_k$  has at most one weak sequential cluster point. Since  $(x_k)_k$  is bounded, by Banach–Alaoglu and Eberlein–Šmulian (Theorems 1.89 and 1.101) it has at least one weak sequential cluster point. So it has precisely one. Let this cluster point be  $x$ .
- Assume  $x_k$  does not converge weakly to  $x$ . Then there is a weakly open environment  $U$  of  $x$  such that  $H \setminus U$  contains an infinite number of elements of the sequence.

- $U$  is weakly open  $\Rightarrow H \setminus U$  is weakly closed  $\Rightarrow$  weakly sequentially closed.
- Apply Banach–Alaoglu and Eberlein–Šmulian to the sequence in  $H \setminus U$  to get a cluster point in  $H \setminus U$ .
- This cluster point cannot be  $x$  which contradicts the assumption of a unique cluster point. Hence,  $x_k$  must converge weakly to  $x$ .  $\square$

**Remark 1.120.** One can in fact show a slightly stronger result:  $[(x_k)_k \text{ converges weakly}] \Leftrightarrow [(x_k)_k \text{ is bounded and has at most one cluster point}]$ . See [Bauschke, Combettes; Lemma 2.38].

**Lemma 1.121.** Let  $(x_k)_k$  be a sequence in  $H$ , let  $C \subset H$  nonempty. Suppose that for every  $y \in C$  the sequence  $(\|x_k - y\|)_k$  converges (to a finite value) and that every weak sequential cluster point of  $(x_k)_k$  lies in  $C$ . Then  $(x_k)_k$  converges weakly to a point in  $C$ .

*Proof.* • By assumption  $(x_k)_k$  is bounded. Therefore by Lemma 1.119 it suffices to show that  $(x_k)_k$  can have at most one weak sequential cluster point.

- Let  $x$  and  $y$  be two weak sequential cluster points of  $(x_k)_k$ , i.e.  $x_{i_k} \rightharpoonup x$  and  $x_{j_k} \rightharpoonup y$ .
- By assumption  $x, y \in C$ . Therefore  $(\|x_k - x\|)_k$  and  $(\|x_k - y\|)_k$  converge.
- Therefore, by

$$\|x_k - y\|^2 - \|x_k - x\|^2 - \|y\|^2 + \|x\|^2 = 2 \langle x_k, x - y \rangle$$

we find that  $(\langle x_k, x - y \rangle)_k$  converges, call the limit  $r \in \mathbb{R}$ :  $\lim_{k \rightarrow \infty} \langle x_k, x - y \rangle = r \in \mathbb{R}$ .

- Further, by weak convergence of the two extracted subsequences we find

$$\lim_{k \rightarrow \infty} \langle x_{i_k}, x - y \rangle = \langle x, x - y \rangle, \quad \lim_{k \rightarrow \infty} \langle x_{j_k}, x - y \rangle = \langle y, x - y \rangle.$$

- Both sequences are subsequences of the converging sequence  $(\langle x_k, x - y \rangle)_k$ . Therefore, their limits must therefore coincide and equal  $r$ :

$$\underbrace{\lim_{k \rightarrow \infty} \langle x_{i_k}, x - y \rangle}_{=\langle x, x - y \rangle} = \underbrace{\lim_{k \rightarrow \infty} \langle x_{j_k}, x - y \rangle}_{=\langle y, x - y \rangle} = \lim_{k \rightarrow \infty} \langle x_k, x - y \rangle = r$$

Then

$$\|x - y\|^2 = \langle x - y, x - y \rangle = r - r = 0$$

and therefore the two cluster points must coincide.  $\square$

**Corollary 1.122.** Let  $(x_k)_k$  be Fejér monotone with respect to  $C$  and every weak sequential cluster point of  $(x_k)_k$  is in  $C$ . Then  $(x_k)_k$  converges weakly to some  $x \in C$ .

*Proof.* • From Prop. 1.118 (ii) the sequence  $(\|x_k - y\|)_k$  converges for all  $y \in C$  (to a finite value).

- Then the result follows from Lemma 1.121.  $\square$

Finally, we can give the proof for the convergence of the proximal minimization scheme.

*Proof of Prop. 1.116.* • Let  $z \in \operatorname{argmin} f$ . From  $x^{(\ell+1)} = \operatorname{Prox}_{\gamma f}(x^{(\ell)})$  we deduce  $x^{(\ell)} - x^{(\ell+1)} \in \gamma \partial f(x^{(\ell+1)})$ . Therefore:

$$\begin{aligned} f(x^{(\ell)}) &\geq f(x^{(\ell+1)}) + \left\langle x^{(\ell)} - x^{(\ell+1)}, x^{(\ell)} - x^{(\ell+1)} \right\rangle / \gamma, \\ f(z) &\geq f(x^{(\ell+1)}) + \left\langle z - x^{(\ell+1)}, x^{(\ell)} - x^{(\ell+1)} \right\rangle / \gamma. \end{aligned}$$

- The first inequality implies that  $(f(x^{(\ell)}))_\ell$  is decreasing.
- The second inequality implies:

$$\begin{aligned} \|x^{(\ell+1)} - z\|^2 &= \|(x^{(\ell+1)} - x^{(\ell)}) - (z - x^{(\ell)})\|^2 \\ &= \|x^{(\ell+1)} - x^{(\ell)}\|^2 + \|z - x^{(\ell)}\|^2 - 2 \left\langle x^{(\ell+1)} - x^{(\ell)}, z - (x^{(\ell+1)} - x^{(\ell+1)}) - x^{(\ell)} \right\rangle \\ &= \|x^{(\ell)} - z\|^2 - \|x^{(\ell+1)} - x^{(\ell)}\|^2 + 2 \left\langle z - x^{(\ell+1)}, x^{(\ell)} - x^{(\ell+1)} \right\rangle \\ &\leq \|x^{(\ell)} - z\|^2 + 2\gamma(f(z) - f(x^{(\ell+1)})) \end{aligned}$$

- Therefore,  $(x^{(\ell)})_\ell$  is Fejér monotone with respect to  $\operatorname{argmin} f$ .
- Summing the above inequality over  $\ell = 0, \dots, N$  we obtain:

$$\begin{aligned} \sum_{\ell=0}^N \left[ f(x^{(\ell+1)}) - f(z) \right] &\leq \frac{1}{2\gamma} \sum_{\ell=0}^N \left[ \|x^{(\ell)} - z\|^2 - \|x^{(\ell+1)} - z\|^2 \right] \\ &= \frac{1}{2\gamma} \left( \|x^{(0)} - z\|^2 - \|x^{(N+1)} - z\|^2 \right) < \infty \end{aligned}$$

- **(i):** So  $(f(x^{(\ell)}) - f(z))_\ell$  is monotone decreasing, nonnegative (since  $z$  is minimizer) and the sum over its elements is bounded. Therefore  $\lim_{\ell \rightarrow \infty} f(x^{(\ell)}) = f(z)$  and  $(x^{(\ell)})_\ell$  is a minimizing sequence.
- **(ii):** Let  $x$  be a weak sequential cluster point of  $(x^{(\ell)})_\ell$ . Since  $f$  is convex and lsc, it is weakly sequentially lsc (Cor. 1.97). Therefore,  $x \in \operatorname{argmin} f$ .
- Now apply Cor. 1.122. □

**Remark 1.123.** Observe that the proximal algorithm converges for all step sizes  $\gamma > 0$ , unlike the explicit gradient step scheme. ← VL8

## 1.9 The Douglas–Rachford algorithm

Now we introduce the first true proximal splitting method that minimizes the sum  $f + g$  of two convex lsc proper functions by only applying the proximal operators of  $f$  and  $g$  separately. The algorithm will therefore be much more practical and easier to implement than the simple proximal algorithm. But its convergence analysis is more involved.

**Proposition 1.124** (Douglas–Rachford algorithm). Let  $f$  and  $g$  be convex, lsc, proper such that  $\exists z \in H$  with  $0 \in \partial f(z) + \partial g(z)$ . Further, let  $\lambda \in (0, 2)$  and  $\gamma \in \mathbb{R}_{++}$ . For some  $x^{(0)} \in H$  set by iteration for  $\ell = 0, 1, \dots$ :

$$\begin{aligned} y^{(\ell)} &= \text{Prox}_{\gamma g}(x^{(\ell)}), \\ z^{(\ell)} &= \text{Prox}_{\gamma f}(2y^{(\ell)} - x^{(\ell)}), \\ x^{(\ell+1)} &= x^{(\ell)} + \lambda \cdot (z^{(\ell)} - y^{(\ell)}) \end{aligned}$$

Then there exists some  $x \in H$  such that

- (i)  $\text{Prox}_{\gamma g}(x) \in \text{argmin}(f + g)$ .
- (ii)  $y^{(\ell)} - z^{(\ell)} \rightarrow 0$  strongly.
- (iii)  $x^{(\ell)} \rightharpoonup x$  weakly.

**Remark 1.125.** One can in addition show that  $y^{(\ell)} \rightharpoonup \text{Prox}_{\gamma g}(x) \in \text{argmin } f$ .

We start by going through an explicit example.

**Example 1.126.** Let  $H = \mathbb{R}^2$ .  $f(x) = \frac{1}{2}\|x\|^2$ ,  $C = \{x \in H \mid x_1 = 1\}$ ,  $g(x) = \iota_C(x)$ . Then  $\text{Prox}_{\gamma f}(x) = \frac{1}{1+\gamma}x$ . Further  $\text{Prox}_{\gamma g}(x) = P_C x$ . Recall  $[y = P_C x] \Leftrightarrow [y \in C \wedge x \in y + N_C y]$ . For the normal cone we get:

$$N_C x = \begin{cases} \emptyset & \text{if } x \notin C, \\ \mathbb{R} \cdot (1, 0) & \text{else.} \end{cases}$$

This implies

$$y_1 = 1, \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

We obtain  $\lambda = x_1 - 1$ ,  $y_1 = 1$ ,  $y_2 = x_2$ . So  $P_C(x_1, x_2) = (1, x_2)$ . For the iterations we get:

$$\begin{aligned} y^{(\ell)} &= (1, x_2^{(\ell)}), & v^{(\ell)} &= 2y^{(\ell)} - x^{(\ell)} = (2 - x_1^{(\ell)}, x_2^{(\ell)}) \\ z^{(\ell)} &= \frac{1}{1+\gamma}(2 - x_1^{(\ell)}, x_2^{(\ell)}), & x^{(\ell+1)} &= \left( x_1^{(\ell)} \cdot \left(1 - \frac{\lambda}{1+\gamma}\right) + \frac{\lambda(1-\gamma)}{1+\gamma}, x_2^{(\ell)} \cdot \left(1 - \frac{\lambda\gamma}{1+\gamma}\right) \right) \end{aligned}$$

Iterations for  $x_1^{(\ell)}$  and  $x_2^{(\ell)}$  separate. Both are affine maps. For the slopes we find:

$$\begin{aligned} 1 - \frac{\lambda}{1+\gamma} &< 1, & 1 - \frac{\lambda}{1+\gamma} &> -1, \\ (1 - \frac{\lambda\gamma}{1+\gamma}) &< 1, & (1 - \frac{\lambda\gamma}{1+\gamma}) &> -1. \end{aligned}$$

So by the Banach fixed-point theorem both coordinates converge to a unique limit. Determine the fixed point of  $x_1^{(\ell)}$ :

$$r = r \cdot \left(1 - \frac{\lambda}{1+\gamma}\right) + \frac{\lambda(1-\gamma)}{1+\gamma} \quad \Leftrightarrow \quad r = 1 - \gamma$$

For the fixed-point of  $x_2^{(\ell)}$  we immediately find  $r = 0$ . So we deduce  $x^{(\ell)} \rightarrow x = (1 - \gamma, 0)$ . Note in particular that this is not optimal (in fact  $g(x) = +\infty$ ). Then  $y^{(\ell)} = P_C x^{(\ell)} \rightarrow y = (1, 0)$  which is indeed the minimizer. Further,  $v^{(\ell)} = 2y^{(\ell)} - x^{(\ell)} \rightarrow 2y - x = (1 + \gamma, 0)$  such that  $z^{(\ell)} = \text{Prox}_{\gamma f}(v^{(\ell)}) = \frac{1}{1+\gamma} v^{(\ell)} \rightarrow (1, 0) = y$ .

**Remark 1.127** (Interpretation of Douglas–Rachford algorithm). The goal is to find a minimizer  $y = z$  and an offset vector  $\Delta x$  such that  $x = y + \Delta x$  ‘lies on the  $f$  side of the minimizer’ and  $v = y - \Delta x$  ‘lies on the  $g$  side’, i.e.  $y = \text{Prox}_{\gamma g}(y + \Delta x) = \text{Prox}_{\gamma f}(y - \Delta x)$ . Unless  $\text{argmin } f \cap \text{argmin } g \neq \emptyset$  there can be no such point for  $\Delta x = 0$ , which is why mere alternating application of  $\text{Prox}_{\gamma f}$  and  $\text{Prox}_{\gamma g}$  is in general too simplistic and a more sophisticated combination of proximal operators is required.

---

**Sketch:** Compare interpretation with example:  $x_2^{(\ell)}$  iteration simply approaches 0. For  $x_1^{(\ell)}$  iterations the sign of the  $x_1^{(\ell)}$ -update depends on whether  $x_1^{(\ell)}$  was too close to 0 or too close to 1.

---

Now we start proving Prop. 1.124. We need considerable auxiliary definitions and results. First note, that if  $\text{argmin } f$  contains more than one element, then  $\text{Prox}_f$  has more than one fixed-point (Prop. 1.112 (i)). So  $\text{Prox}_f$  cannot be a contraction and therefore convergence proofs cannot rely e.g. on the Banach fixed-point theorem. We need a refined notion of contraction.

**Definition 1.128.** A map  $T : H \mapsto H$  is called

- (i) *nonexpansive* if it is Lipschitz continuous with constant 1. That is

$$\|T(x) - T(y)\| \leq \|x - y\| \quad \text{for all } x, y \in H;$$

- (ii) *firmly nonexpansive* if

$$\|T(x) - T(y)\|^2 + \|(T(x) - x) - (T(y) - y)\|^2 \leq \|x - y\|^2 \quad \text{for all } x, y \in H.$$

**Proposition 1.129.** Let  $f : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper, convex lsc. Then

- (i)  $\text{Prox}_f$  is firmly nonexpansive.
- (ii)  $\text{id} - \text{Prox}_f$  is firmly nonexpansive.
- (iii)  $2 \text{Prox}_f - \text{id}$  is nonexpansive.

*Proof.* • (i) Use monotonicity of subdifferential. Let  $x, y \in H$ , set  $p = \text{Prox}_f(x)$ ,  $q = \text{Prox}_f(y)$ . Denote  $\Delta x = x - y$  and  $\Delta p = p - q$ .

- From  $x - p \in \partial f(p)$ ,  $y - q \in \partial f(q)$  and monotonicity of the subdifferential we get:

$$\begin{aligned} [(\langle x - p - (y - q), p - q \rangle \geq 0] &\Leftrightarrow [\langle \Delta x - \Delta p, \Delta p \rangle \geq 0] \Leftrightarrow [2\|\Delta p\|^2 - 2\langle \Delta p, \Delta x \rangle \leq 0] \\ &\Leftrightarrow [\|\Delta p\|^2 + \|\Delta p\|^2 - 2\langle \Delta p, \Delta x \rangle + \|\Delta x\|^2 \leq \|\Delta x\|^2] \Leftrightarrow [\|\Delta p\|^2 + \|\Delta p - \Delta x\|^2 \leq \|\Delta x\|^2] \end{aligned}$$

- (ii) By the Moreau decomposition (Prop. 1.113),  $\text{id} - \text{Prox}_f = \text{Prox}_{f^*}$ . Since  $f^*$  is convex, lsc, proper,  $\text{Prox}_{f^*}$  is firmly nonexpansive by (i).
- (iii) Use the above notation. Then we need to bound:

$$\|(2p - x) - (2q - y)\|^2 = \|2\Delta p - \Delta x\|^2 = \underbrace{4\|\Delta p\|^2 - 4\langle \Delta p, \Delta x \rangle}_{\leq 0, \text{ see (i)}} + \|\Delta x\|^2 \leq \|\Delta x\|^2 \quad \square$$

We need a result to identify fixed-points of nonexpansive maps in the context of weak convergence.

**Proposition 1.130.** Let  $T : H \rightarrow H$  be nonexpansive. Let  $(x_k)_k$  be a bounded sequence in  $H$  and let  $x \in H$ . If  $x_k \rightharpoonup x$  and  $x_k - T(x_k) \rightarrow 0$  then  $x = T(x)$ .

*Proof.*

$$\begin{aligned} \|x - T(x)\|^2 &= \|(x - x_k) - (T(x) - x_k)\|^2 \\ &= \|x - x_k\|^2 + \|T(x) - x_k\|^2 - 2\langle x - x_k, T(x) - x_k \rangle \\ &= \|T(x) - x_k\|^2 - \|x - x_k\|^2 - 2\langle x - x_k, T(x) - x \rangle \\ &= \|(T(x) - T(x_k)) - (x_k - T(x_k))\|^2 - \|x - x_k\|^2 - 2\langle x - x_k, T(x) - x \rangle \end{aligned}$$

(use nonexpansiveness of  $T$  to bound  $\|T(x) - T(x_k)\| \leq \|x - x_k\|$ )

$$\leq \|x_k - T(x_k)\|^2 - 2\langle T(x) - T(x_k), x_k - T(x_k) \rangle - 2\langle x - x_k, T(x) - x \rangle$$

Now, recall  $x_k \rightharpoonup x$ ,  $x_k - T(x_k) \rightarrow 0$  and  $(x_k)_k$  bounded, i.e.  $\|x_k\| \leq C_1$  for some  $C_1 < +\infty$ . Further, since  $T$  is nonexpansive:  $\|T(x_k)\| = \|T(x_k) - T(0) + T(0)\| \leq \|T(x_k) - T(0)\| + \|T(0)\| \leq C + \|T(0)\| := C_2$ . Then

$$\begin{aligned} \|x_k - T(x_k)\|^2 &\rightarrow 0, \\ \limsup_k |\langle T(x) - T(x_k), x_k - T(x_k) \rangle| &\leq \limsup_k (\|T(x)\| + C_2) \cdot \|x_k - T(x_k)\| = 0 \end{aligned}$$

(here have used Cauchy-Schwarz and  $\|T(x) - T(x_k)\| \leq \|T(x)\| + \|T(x_k)\| \leq \|T(x)\| + C_2$ )

$$\langle x - x_k, T(x) - x \rangle \rightarrow \langle x - x, T(x) - x \rangle = 0.$$

Therefore, by going to the limit in the above upper bound on  $\|x - T(x)\|^2$  we get  $\|x - T(x)\| \leq 0$ , i.e.  $T(x) = x$ .  $\square$

**Definition 1.131.** In the following, for an operator  $T : H \rightarrow H$  denote by

$\text{Fix } T = \{x \in H \mid T(x) = x\}$  the set of fixed-points of  $T$ .

$\text{zer } T = \{x \in H \mid T(x) = 0\}$  the set of ‘roots’ of  $T$ .

Analogously, for  $T : H \rightarrow 2^H$  let

$$\text{zer } T = \{x \in H \mid 0 \in T(x)\}.$$

We will shortly show that the Douglas–Rachford iteration can be compactly rewritten as an iteration as analyzed in the following Proposition.

**Proposition 1.132.** Let  $T : H \rightarrow H$  be nonexpansive, let  $\text{Fix } T \neq \emptyset$ ,  $\lambda \in (0, 1)$  and  $x^{(0)} \in H$ . Set

$$x^{(\ell+1)} = x^{(\ell)} + \lambda (T(x^{(\ell)}) - x^{(\ell)}).$$

Then:

- (i)  $(x^{(\ell)})_\ell$  is Fejér monotone with respect to  $\text{Fix } T$ .
- (ii)  $(T(x^{(\ell)}) - x^{(\ell)})_\ell$  converges strongly to 0.
- (iii)  $(x^{(\ell)})_\ell$  converges weakly to a point in  $\text{Fix } T$ .

*Proof.* • For the proof we use the following equality which can be verified by expansion: For  $\lambda \in [0, 1]$ ,  $x, y \in H$ :

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

- (i) Let  $y \in \text{Fix } T$ . Then

$$\begin{aligned} \|x^{(\ell+1)} - y\|^2 &= \|(1 - \lambda)(x^{(\ell)} - y) + \lambda(T(x^{(\ell)}) - y)\|^2 \\ &= (1 - \lambda)\|x^{(\ell)} - y\|^2 + \lambda\|T(x^{(\ell)}) - T(y)\|^2 - \lambda(1 - \lambda)\|x^{(\ell)} - T(x^{(\ell)})\|^2 \\ &\leq \|x^{(\ell)} - y\|^2 - \lambda(1 - \lambda)\|x^{(\ell)} - T(x^{(\ell)})\|^2 \end{aligned}$$

So  $(x^{(\ell)})_\ell$  is Fejér monotone with respect to  $\text{Fix } T$ .

- (ii) From the above bound we find:

$$\sum_{\ell=0}^N \lambda(1 - \lambda)\|x^{(\ell)} - T(x^{(\ell)})\|^2 \leq \|x^{(0)} - y\|^2 - \|x^{(N+1)} - y\|^2 < \infty$$

Therefore  $\|x^{(\ell)} - T(x^{(\ell)})\| \rightarrow 0 \Rightarrow x^{(\ell)} - T(x^{(\ell)}) \rightarrow 0$ . (Here use that  $\lambda(1 - \lambda) > 0$ .)

- (iii)  $(x^{(\ell)})_\ell$  is bounded due to Fejér monotonicity (Prop. 1.118). From (ii) we have  $T(x^{(\ell)}) - x^{(\ell)} \rightarrow 0$ . So by the previous result, Prop. 1.130, we obtain that any weak sequential cluster point of  $(x^{(\ell)})_\ell$  is in  $\text{Fix } T$ .
- It follows then from Cor. 1.122 that  $(x^{(\ell)})_\ell$  converges weakly to a point in  $\text{Fix } T$ . □

**Proposition 1.133.** Let  $f, g : H \rightarrow \mathbb{R} \cup \{\infty\}$  be proper, convex, lsc. Let

$$R_f = 2 \text{Prox}_f - \text{id}, \quad R_g = 2 \text{Prox}_g - \text{id}.$$

Then  $\text{zer}(\partial f + \partial g) = \text{Prox}_g(\text{Fix } R_f R_g)$ .

---

Comment: We will shortly show that fixed-points of  $(x^{(\ell)})_\ell$  in the Douglas–Rachford iterations are precisely the fixed-points of  $R_f R_g$ .

---

*Proof.* We find:

$$\begin{aligned}
[0 \in \partial f(x) + \partial g(x)] &\Leftrightarrow [\exists u \in H : [u \in \partial f(x)] \wedge [-u \in \partial g(x)]] \\
&\Leftrightarrow [\exists y \in H : [x - y \in \partial f(x)] \wedge [y - x \in \partial g(x)]] \\
&\Leftrightarrow [\exists y \in H : [x = \text{Prox}_f(2x - y)] \wedge [x = \text{Prox}_g(y)]] \\
&\Leftrightarrow [\exists y \in H : [x = \text{Prox}_f(R_g(y))] \wedge [x = \text{Prox}_g(y)]]
\end{aligned}$$

$$\begin{aligned}
(\Rightarrow: R_g(y) = 2 \text{Prox}_g(y) - y = 2x - y \Rightarrow y = 2x - R_g(y) = 2 \text{Prox}_f(R_g(y)) - R_g(y) = R_f(R_g(y)), \\
(\Leftarrow: y = R_f(R_g(y)) = 2 \text{Prox}_f(R_g(y)) - 2 \text{Prox}_g(y) + y \Rightarrow x = \text{Prox}_g(y) = \text{Prox}_f(R_g(y)))
\end{aligned}$$

$$\Leftrightarrow [\exists y \in H : [y = R_f(R_g(y))] \wedge [x = \text{Prox}_g(y)]] \quad \square$$

Now we are ready to assemble the proof of Prop. 1.124.

*Proof of Prop. 1.124.* • Let

$$R_f = 2 \text{Prox}_{\gamma f} - \text{id}, \quad R_g = 2 \text{Prox}_{\gamma g} - \text{id}, \quad T = R_f R_g.$$

- Conversely  $\text{Prox}_{\gamma g} = \frac{1}{2}(R_g + \text{id})$ . Then we can rewrite the Douglas–Rachford iterations as follows:

$$\begin{aligned}
x^{(\ell+1)} &= x^{(\ell)} + \lambda \cdot \left( \text{Prox}_{\gamma f}(2y^{(\ell)} - x^{(\ell)}) - y^{(\ell)} \right) \\
&= x^{(\ell)} + \lambda \cdot \left( \text{Prox}_{\gamma f}(2 \text{Prox}_{\gamma g}(x^{(\ell)}) - x^{(\ell)}) - \text{Prox}_{\gamma g}(x^{(\ell)}) \right) \\
&= x^{(\ell)} + \lambda \cdot \left( \text{Prox}_{\gamma f}(R_g(x^{(\ell)})) - \frac{1}{2}R_g(x^{(\ell)}) - \frac{1}{2}x^{(\ell)} \right) \\
&= x^{(\ell)} + \lambda \cdot \left( \frac{1}{2}R_f(R_g(x^{(\ell)})) - \frac{1}{2}x^{(\ell)} \right) = x^{(\ell)} + \frac{\lambda}{2} \cdot \left( T(x^{(\ell)}) - x^{(\ell)} \right)
\end{aligned}$$

Note that  $z^{(\ell)} - y^{(\ell)} = \frac{1}{2}(T(x^{(\ell)}) - x^{(\ell)})$ .

- Apply Prop. 1.133 to  $\gamma f$  and  $\gamma g$  to obtain

$$\text{zer}(\partial f + \partial g) = \text{zer}(\partial \gamma f + \partial \gamma g) = \text{Prox}_{\gamma g}(\text{Fix } T).$$

Since by assumption  $\text{zer}(\partial f + \partial g) \neq \emptyset$  this implies also  $\text{Fix } T \neq \emptyset$ .

- In view of (i) we note that for every  $x \in \text{Fix } T$  have therefore  $\text{Prox}_{\gamma g}(x) \in \text{argmin}(f + g)$ .
- Due to Prop. 1.129(iii) the operators  $R_f$  and  $R_g$  are nonexpansive. Then so is their composition  $T = R_f R_g$ .
- (ii) From Prop. 1.132(ii) we find  $z^{(\ell)} - y^{(\ell)} = \frac{1}{2}(T(x^{(\ell)}) - x^{(\ell)}) \rightarrow 0$  strongly.
- From Prop. 1.132 (iii) we get: there is some  $x \in \text{Fix } T$  such that  $x^{(\ell)} \rightharpoonup x$  weakly. Since  $\text{Prox}_{\gamma g}(x) \in \text{argmin}(f + g)$  (see above) this establishes (i) and (iii). □ ← VL9

## 1.10 Primal-Dual Methods

In this subsection we study an alternative approach to optimizing objectives of the form  $f + g$  that is intimately linked to conjugation.

**Definition 1.134.** • For convex functions  $f, g : H \rightarrow \mathbb{R} \cup \{\infty\}$  let

$$P(x) = f(x) + g(x), \quad D(y) = -f^*(-y) - g^*(y), \quad L(x, y) = f(x) - g^*(y) + \langle x, y \rangle.$$

- The problem  $\inf_{x \in H} P(x)$  is called *primal problem*, the problem  $\sup_{y \in H} D(y)$  is called *dual problem* and  $L$  is called *Lagrangian*.
- Note that since  $D$  is concave, the dual problem is also a convex optimization problem.
- For all  $(x, y) \in H^2$  one has

$$P(x) \geq L(x, y) \geq D(y).$$

This follows quickly from the Fenchel–Young inequality (Prop. 1.72).

- In particular:

$$P(x) \geq \inf_{x' \in H} P(x') \geq \sup_{y' \in H} D(y') \geq D(y)$$

- So for every feasible pair  $(x, y) \in H^2$  of primal and dual problem the value  $\Delta(x, y) = P(x) - D(y)$  is an upper bound on the combined suboptimality of  $x$  and  $y$  with respect to primal and dual problem. Therefore  $\Delta(x, y)$  is called the *duality gap*.
- If  $\Delta(x, y) = 0$  then  $x$  and  $y$  must be optimizers of primal and dual problem respectively.

We give a simple variant of the famous Fenchel–Rockafellar duality.

**Proposition 1.135** (Duality). Assume there exists some  $x_0 \in H$  such that  $f(x_0) < \infty$ ,  $g(x_0) < \infty$  and  $f$  is continuous in  $x_0$ . Then

$$\inf_{x \in H} \{f(x) + g(x)\} = \max_{y \in H} \{-f^*(-y) - g^*(y)\}.$$

In particular, a maximizer for the dual problem exists.

---

Comment: Sometimes one tries to show that a given optimization problem is the dual problem of some auxiliary problem to use the above Proposition for showing that a solution exists.

---

For the proof we need an auxiliary result on separating points from convex sets via hyperplanes.

**Proposition 1.136.** Let  $C \subset H$  be convex,  $0 \notin \text{int } C$ ,  $\text{int } C \neq \emptyset$ . Then there is some  $z \in H \setminus \{0\}$  such that  $\langle z, x \rangle \geq 0$  for all  $x \in C$ .

---

Comment: This means, the hyperplane with normal  $z$  through the origin separates  $C$  from 0.

---

*Proof.* • Let  $D = \overline{C}$ . By Prop. 1.54 we have  $\text{int } D = \text{int } C \neq \emptyset$  and in particular  $0 \notin \text{int } C = \text{int } D$ .

- Since  $D$  is closed and convex, it is the intersection of all closed halfspaces that contain  $D$  (Cor. 1.66).

- If  $0 \notin D$  there must be a closed halfspace  $A_u = \{x \in H \mid \langle u, x \rangle - r \geq 0\}$ ,  $u \neq 0$ , such that  $D \subset A_u$  and  $0 \notin A_u$ .
- $0 \notin A_u \Rightarrow \langle 0, x \rangle - r < 0 \Rightarrow r > 0$ .
- $x \in D \subset A_u \Rightarrow 0 \leq \langle u, x \rangle - r < \langle u, x \rangle$ . So setting  $z = u$  we have found an appropriate  $z$ .
- Alternatively, must have  $0 \in D$ , but  $0 \notin \text{int } D$ . Then by Prop. 1.56 there is some  $u \in N_D 0$  with  $u \neq 0$ . By definition  $\sup \langle u, D \rangle \leq 0$ . Then set  $z = -u$  above:

$$\langle -u, x \rangle \geq \inf \langle -u, D \rangle = -\sup \langle u, D \rangle \geq 0 \quad \square$$

*Proof of Prop. 1.135.* • The inequality  $\inf_{x \in H} P(x) \geq \sup_{y \in H} D(y)$  is clear by Def. 1.134. We need to show the converse inequality.

- Denote by  $m = \inf_{x \in H} P(x)$ . We have  $m \leq f(x_0) + g(x_0) < \infty$ . If  $m = -\infty$  the converse inequality is trivial. Hence, assume  $m \in \mathbb{R}$ .
- Assume we had some  $z \in H$  such that for all  $a, b \in H$  one has

$$f(a) + g(b) + \langle z, a - b \rangle \geq m.$$

- Then we find:

$$\begin{aligned} \sup_{y \in H} -f^*(-y) - g^*(y) &= \sup_{y \in H} \inf_{a, b \in H} [f(a) - \langle a, -y \rangle + g(b) - \langle b, y \rangle] \\ &\geq \inf_{a, b \in H} [f(a) + g(b) + \langle z, a - b \rangle] \geq m \end{aligned}$$

- Since also  $\sup_{y \in H} -f^*(-y) - g^*(y) \leq m$ ,  $z$  must therefore be a dual maximizer and primal and dual problem have the same optimal value.
- Now we show existence of a suitable  $z$ .
- Let

$$\begin{aligned} A &= \{(a, \lambda) \in H \times \mathbb{R} \mid \lambda > f(a)\}, \\ B &= \{(b, \mu) \in H \times \mathbb{R} \mid \mu \leq m - g(b)\}. \end{aligned}$$

- $A$  and  $B$  are convex. Since  $f$  is continuous in  $x_0$  and  $f(x_0) < \infty$ , we have  $\text{int } A \neq \emptyset$ .
- Assume there were some  $(a, \lambda) \in A \cap B$ . Then we would find

$$f(a) + g(a) < \lambda + (m - \lambda) = m$$

which is a contradiction. Therefore  $A \cap B = \emptyset$ .

- This implies  $0 \notin A - B$  and in particular  $0 \notin \text{int}(A - B)$ .  $A - B$  is convex. Also,  $\emptyset \neq \text{int } A - B \subset \text{int}(A - B)$ . So by Prop. 1.136 there is some  $(u, r) \in H \times \mathbb{R} \setminus \{(0, 0)\}$  such that  $\langle (u, r), (a, \lambda) \rangle \geq 0$  for all  $(a, \lambda) \in A - B$ . This implies

$$\langle (u, r), (a, \lambda) \rangle \geq \langle (u, r), (b, \mu) \rangle \quad \Leftrightarrow \quad \langle u, a \rangle + r \cdot \lambda \geq \langle u, b \rangle + r \cdot \mu$$

for all  $(a, \lambda) \in A$ ,  $(b, \mu) \in B$ .

- Since we can send  $\lambda \rightarrow \infty$  and  $\mu \rightarrow -\infty$  (and remain in  $A, B$  respectively) we find that  $r \geq 0$ .
- If  $r = 0$ , we can violate the inequality by setting  $b = x_0$  (where  $g$  is finite) and  $a = x_0 - \varepsilon \cdot u$  for sufficiently small  $\varepsilon > 0$  (which works since  $f$  is continuous in  $x_0$ ). So must have  $r > 0$ .
- Set now  $z = u/r$ . The above inequality yields

$$\langle z, a - b \rangle + \lambda - \mu \geq 0$$

for all  $\lambda > f(a)$ ,  $\mu \leq m - g(b)$ , i.e. for  $\lambda - \mu > f(a) + g(b) - m$ . Therefore, the given  $z$  is as needed above and the proof is complete.  $\square$

**Remark 1.137.** There are considerably more general variants of Prop. 1.135 on Banach spaces and their dual spaces. Then the auxiliary separation result, Prop. 1.136, must usually be provided by the Hahn–Banach theorem.

**Proposition 1.138.** A pair  $(x, y) \in H^2$  are solutions to the primal and dual problem if and only if  $x \in g^*(y)$  and  $-y \in \partial f(x)$ .

*Proof.*

$$\begin{aligned} [(x, y) \text{ are solutions}] &\Leftrightarrow [P(x) = D(y)] \Leftrightarrow [f(x) + g(x) = -f^*(-y) - g^*(y)] \\ &\Leftrightarrow [(f(x) + f^*(-y) - \langle x, -y \rangle) + (g(x) + g^*(y) - \langle x, y \rangle) = 0] \end{aligned}$$

(By Fenchel–Young (Prop. 1.72) this is 0 if and only if both parantheses are 0, which, by Prop. 1.81, is equivalent to:)

$$\Leftrightarrow [-y \in \partial f(x)] \wedge [x \in \partial g^*(y)] \quad \square$$

We provide a few observations that provide some intuition for the dual problem and the optimality conditions.

**Remark 1.139.** • Recall:  $[x \in \operatorname{argmin}(f + g)] \Leftrightarrow [0 \in \partial(f + g)(x)] \Leftarrow [0 \in \partial f(x) + \partial g(x)]$ .

- Now start from the stronger condition for the existence of a minimizer:

$$\begin{aligned} [\exists x \in H : 0 \in \partial f(x) + \partial g(x)] &\Leftrightarrow [\exists x, y \in H : -y \in \partial f(x) \wedge y \in \partial g(x)] \\ &\Leftrightarrow [\exists x, y \in H : x \in \partial f^*(-y) \wedge x \in \partial g^*(y)] \\ &\Leftrightarrow [\exists x, y \in H : -x \in (\partial(f^*(-\cdot)))(y) \wedge x \in \partial g^*(y)] \\ &\Leftrightarrow [\exists y \in H : 0 \in (\partial(f^*(-\cdot)))(y) + \partial g^*(y)] \end{aligned}$$

- So: dual optimal  $y$  is slope of  $g$  at  $x$  and negative slope of  $f$  at  $x$ .

**Remark 1.140.** • Reconsider sets from proof of Prop. 1.135:

$$\begin{aligned} A &= \{(a, \lambda) \in H \times \mathbb{R} \mid \lambda > f(a)\}, \\ B &= \{(b, \mu) \in H \times \mathbb{R} \mid \mu \leq m - g(b)\}. \end{aligned}$$

- We look for a hyperplane that separates  $A$  and  $B$  with normal vector  $(z, r)$ . We have shown that one must have  $r > 0$ , so w.l.o.g. set  $r = 1$  (by rescaling  $z$  appropriately).

- When  $x$  is a primal minimizer, set  $\lambda = f(a)$ . Then  $(a, \lambda) \in \overline{A} \cap B$ . So the sets almost touch in  $(x, \lambda)$ .
- $(z, 1)$  is the normal of a hyperplane that separates  $A$  and  $B$  (and a dual solution). We then see that  $z$  must be compatible with the ‘slopes’ of  $A$  and  $B$  at  $(x, \lambda)$ .
- A sufficient condition for  $(z, 1)$  being the normal of a separating hyperplane is that  $-z \in \partial f(x)$  and  $z \in \partial g(x)$ .

Finally, we give another proximal splitting algorithm, specialized for primal-dual problem pairs as above.

**Proposition 1.141.** Assume that  $f$  and  $g$  are proper, convex, lsc and that primal and dual problem have solutions. For  $\tau \in (0, 1)$  and  $x^{(0)}, y^{(0)} \in H$  set

$$\begin{aligned} x^{(\ell+1)} &= \text{Prox}_{\tau f}(x^{(\ell)} - \tau \cdot y^{(\ell)}), \\ y^{(\ell+1)} &= \text{Prox}_{\tau g^*}(y^{(\ell)} + \tau \cdot (2x^{(\ell+1)} - x^{(\ell)})). \end{aligned}$$

Then  $(x^{(\ell)})_\ell$  and  $(y^{(\ell)})_\ell$  converge weakly to solutions of the primal and dual problem, respectively.

**Remark 1.142.** For more details on such algorithms and generalizations, see for instance [Chambolle, Pock: A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging, 2011].

**Remark 1.143.** The strategy for the convergence proof of Prop. 1.141 is as follows: we show that the optimality condition of Prop. 1.138 can be written as zeros of a monotone operator acting on the pair  $(x, y)$ . These zeros can be identified with fixed-points of carefully constructed firmly nonexpansive operators and corresponding metrics. If we choose the right metric, this operator can be identified with the iterations of Prop. 1.141. We then generalize the original proximal optimization algorithm (Prop. 1.116) to firmly nonexpansive operators.

**Proposition 1.144.** Let  $A : H \rightarrow 2^H$  be monotone. Let  $M : H \rightarrow H$  be linear, continuous, self-adjoint and positive definite, i.e.  $\langle x, Mx \rangle \geq C\|x\|^2$  for some  $C \in \mathbb{R}_{++}$ . Let the operator  $T : K \rightarrow H$  be given by

$$[y = T(x)] \Leftrightarrow [Mx \in My + A(y)]$$

where  $K \subset H$  is the set such that the above inclusion has a solution  $y$  for fixed  $x \in K$ .

- (i) This inclusion has at most one solution, i.e.  $T$  is well-defined on  $K$ .
- (ii)  $\text{Fix } T = \text{zer } A$ .
- (iii)  $T$  is firmly nonexpansive with respect to the inner product induced by  $M$ ,  $\langle x, y \rangle_M = \langle x, My \rangle$ .

*Proof.* • (i) For fixed  $x \in H$  and  $y_1, y_2 \in H$  assume:

$$[Mx \in My_1 + A(y_1)] \wedge [Mx \in My_2 + A(y_2)]$$

By monotonicity of  $A$  we get:

$$0 \leq \langle M(x - y_1) - M(x - y_2), y_1 - y_2 \rangle = -\langle M(y_1 - y_2), (y_1 - y_2) \rangle \leq -C\|y_1 - y_2\|^2$$

Therefore  $y_1 = y_2$  and thus,  $T$  is well-defined on  $K$ .

- (ii)

$$[x \in \text{zer}(A)] \Leftrightarrow [0 \in A(x)] \Leftrightarrow [Mx \in Mx + A(x)] \Leftrightarrow [T(x) = x] \Leftrightarrow [x \in \text{Fix } T]$$

- (iii) Let  $p = T(x)$ ,  $q = T(y)$ ,  $\Delta x = x - y$ ,  $\Delta p = p - q$ . Then

$$M(x - p) \in A(p), \quad M(y - q) \in A(q).$$

By monotonicity:

$$\begin{aligned} [\langle M(x - p) - M(y - q), p - q \rangle \geq 0] &\Leftrightarrow [\langle \Delta x - \Delta p, \Delta p \rangle_M \geq 0] \\ &\Leftrightarrow [\|\Delta p\|_M^2 - \langle \Delta p, \Delta x \rangle_M \leq 0] \\ &\Leftrightarrow [\|p - q\|_M^2 + \|(p - x) - (q - y)\|_M^2 \leq \|x - y\|_M^2] \end{aligned}$$

□

Now we generalize Prop. 1.116 to arbitrary firmly nonexpansive operators.

← VL10

**Proposition 1.145.** Assume  $T : H \rightarrow H$  is firmly nonexpansive and  $\text{Fix } T \neq \emptyset$ . For  $x^{(0)} \in H$  set

$$x^{(\ell+1)} = T(x^{(\ell)}).$$

Then  $(x^{(\ell)})_\ell$  converges weakly to some point  $x \in \text{Fix } T$ .

*Proof.* • Let  $z \in \text{Fix } T$ . Then by firm nonexpansiveness:

$$\begin{aligned} \|x^{(\ell+1)} - z\|^2 &= \|T(x^{(\ell)}) - T(z)\|^2 \leq \|x^{(\ell)} - z\|^2 - \|(T(x^{(\ell)}) - x^{(\ell)}) - (T(z) - z)\|^2 \\ &= \|x^{(\ell)} - z\|^2 - \|T(x^{(\ell)}) - x^{(\ell)}\|^2 \end{aligned}$$

- So  $(x^{(\ell)})_\ell$  is Fejér monotone with respect to  $\text{Fix } T$ .
- Further  $\sum_{\ell=0}^N \|T(x^{(\ell)}) - x^{(\ell)}\|^2 \leq \|x^{(0)} - z\|^2$ . Therefore  $T(x^{(\ell)}) - x^{(\ell)} \rightarrow 0$  strongly.
- By Prop. 1.130 every weak sequential cluster point of  $(x^{(\ell)})_\ell$  is a fixed-point of  $T$ .
- By Cor. 1.122  $x^{(\ell)} \rightharpoonup x$  for some  $x \in \text{Fix } T$ . □

*Proof of Proposition 1.141.* • Define set valued operator  $A : H \times H \rightarrow 2^H \times 2^H$  as follows:

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \partial f(x) + y \\ \partial g^*(y) - x \end{pmatrix}$$

- By Prop. 1.138 the primal and dual optimizers are given precisely by  $\text{zer } A$ , and therefore by assumption  $\text{zer } A \neq \emptyset$ .
- Note that  $2^H \times 2^H$  can be identified with a subset of  $2^{(H \times H)}$ . So formally  $A$  can be interpreted as  $H \times H \rightarrow 2^{H \times H}$ .
- $A$  is monotonous: let  $[a_i \in \partial f(x_i) \wedge b_i \in \partial g^*(y_i)] \Leftrightarrow (a_i + y_i, b_i - x_i) \in A(x_i, y_i)$  for  $i = 1, 2$ . Denote by  $\Delta x, \Delta y, \Delta a, \Delta b$  all pairwise differences. Then by monotonicity of  $\partial f$  and  $\partial g^*$ :

$$\begin{aligned} \langle (a_2 + y_2, b_2 - x_2) - (a_1 + y_1, b_1 - x_1), (x_2 - x_1, y_2 - y_1) \rangle &= \\ \langle (\Delta a + \Delta y, \Delta b - \Delta x), (\Delta x, \Delta y) \rangle &= \langle \Delta a, \Delta x \rangle + \langle \Delta y, \Delta x \rangle + \langle \Delta b, \Delta y \rangle - \langle \Delta x, \Delta y \rangle \geq 0 \end{aligned}$$

- Now set  $M : H^2 \rightarrow H^2$  as

$$M = \begin{pmatrix} \frac{1}{\tau} \text{id} & -\text{id} \\ -\text{id} & \frac{1}{\tau} \text{id} \end{pmatrix}$$

$M$  is continuous, linear and symmetric. Furthermore, it is positive definite, since for  $x, y \in H$ :

$$\langle (x, y), M(x, y) \rangle = \frac{1}{\tau} \|x\|^2 + \frac{1}{\tau} \|y\|^2 - 2 \langle x, y \rangle \geq \left(\frac{1}{\tau} - 1\right) \cdot (\|x\|^2 + \|y\|^2) + \|x - y\|^2$$

- Now analyze operator  $T$  constructed from  $A$  and  $M$  via Prop. 1.144: From  $(a, b) = T(x, y)$  we obtain the following inclusion condition:

$$\begin{aligned} M \begin{pmatrix} x - a \\ y - b \end{pmatrix} &\in A \begin{pmatrix} a \\ b \end{pmatrix} \\ \Leftrightarrow \begin{pmatrix} \frac{1}{\tau}x - \frac{1}{\tau}a - y + b \\ \frac{1}{\tau}y - \frac{1}{\tau}b - x + a \end{pmatrix} &\in \begin{pmatrix} \partial f(a) + b \\ \partial g^*(b) - a \end{pmatrix} \\ \Leftrightarrow \begin{pmatrix} x - \tau y \\ y + \tau(2a - x) \end{pmatrix} &\in \begin{pmatrix} a + \partial \tau f(a) \\ b + \partial \tau g^*(b) \end{pmatrix} \\ \Leftrightarrow a = \text{Prox}_{\tau f}(x - \tau y) \wedge b = \text{Prox}_{\tau g^*}(y + \tau(2a - x)) \end{aligned}$$

So the iterations of the algorithm, Prop. 1.141, correspond to the induced operator  $T$ , i.e.  $(x^{(\ell+1)}, y^{(\ell+1)}) = T(x^{(\ell)}, y^{(\ell)})$ . This also implies that the domain of  $T$  is  $H^2$ .

- By Prop. 1.144 (ii) have  $\text{Fix } T = \text{zer } A \neq \emptyset$ .
- Now consider the Hilbert space  $H \times H$ , equipped with the inner product induced by  $M$ . Note that the topology induced by  $M$  is the same as the product topology on  $H \times H$ . Therefore, both topologies induce the same weak topology.
- On this space  $T$  is firmly nonexpansive by Prop. 1.144 (iii). So by Prop. 1.145 the sequence  $(x^{(\ell)}, y^{(\ell)})_\ell$  converges weakly to some  $(x, y) \in \text{Fix } T = \text{zer } A$ , which is therefore a pair of solutions to primal and dual problem.
- Note: the iteration would converge to a fixed-point for every positive definite  $M$ . We chose  $M$  carefully such that computing  $x^{(\ell+1)}$  does not depend on  $y^{(\ell+1)}$  and thus the two updates can be computed subsequently and separately.  $\square$

**Remark 1.146.** • Prop. 1.145 studies convergence to fixed-points of general firmly nonexpansive operators, it is a generalization of Prop. 1.116, which only treated the special case of the proximal operator.

- Prop. 1.144 defines a firmly nonexpansive operator from a monotone operator.  $T$  is usually called *resolvent* of  $A$ . This is a generalization of the relation between the subdifferential and the proximal operator, Prop. 1.108.
- Generalize minimization problems to finding zeros of monotone operators via their resolvents. Douglas–Rachford algorithm can also be generalized to finding zero of sum of two monotone operators. In fact, this was application of the algorithm in first publication.

### 1.11 Dykstra's projection algorithm

As an illustrative example for convex analysis and optimization via splitting we discuss Dykstra's famous projection algorithm for the projection onto the intersection of two convex sets by computing only projections onto the two sets separately.

**Definition 1.147** (Dykstra's projection algorithm). Let  $C, D \subset H$  be closed, convex sets with  $C \cap D \neq \emptyset$ . Then the projection onto  $C \cap D$  is well-defined. For  $r \in H$  let  $x^{(0)} = r$ ,  $p^{(0)} = 0$ ,  $q^{(0)} = 0$ . Then the algorithm is given by

$$\begin{aligned} y^{(\ell)} &= P_C(x^{(\ell)} + p^{(\ell)}) \\ p^{(\ell+1)} &= p^{(\ell)} + x^{(\ell)} - y^{(\ell)} \\ x^{(\ell+1)} &= P_D(y^{(\ell)} + q^{(\ell)}) \\ q^{(\ell+1)} &= q^{(\ell)} + y^{(\ell)} - x^{(\ell+1)} \end{aligned}$$

It can be shown that  $x^{(\ell)} \rightarrow P_{C \cap D} r$  and  $y^{(\ell)} \rightarrow P_{C \cap D} r$ .

**Remark 1.148** (Reformulation).

- Consider:

$$\begin{aligned} p^{(k+1)} + q^{(k+1)} &= p^{(k+1)} - p^{(0)} + q^{(k+1)} - q^{(0)} = \sum_{\ell=0}^k [p^{(\ell+1)} - p^{(\ell)} + q^{(\ell+1)} - q^{(\ell)}] \\ &= \sum_{\ell=0}^k [x^{(\ell)} - y^{(\ell)} + y^{(\ell)} - x^{(\ell+1)}] = x^{(0)} - x^{(k+1)} = r - x^{(k+1)} \end{aligned}$$

- So:  $x^{(\ell)} + p^{(\ell)} = r - q^{(\ell)}$  and  $y^{(\ell)} + q^{(\ell)} = x^{(\ell+1)} + q^{(\ell+1)} = r - p^{(\ell+1)}$ . The algorithm becomes:

$$\begin{aligned} y^{(\ell)} &= P_C(r - q^{(\ell)}) & p^{(\ell+1)} &= p^{(\ell)} + x^{(\ell)} - y^{(\ell)} = r - q^{(\ell)} - y^{(\ell)} \\ x^{(\ell+1)} &= P_D(r - p^{(\ell+1)}) & q^{(\ell+1)} &= q^{(\ell)} + y^{(\ell)} - x^{(\ell+1)} \end{aligned}$$

Convex analysis and optimization provide a good perspective to get intuition on the algorithm and to prove convergence.

**Proposition 1.149** (Dual projection problem). Let  $C \subset H$  be closed, convex, non-empty,  $r \in H$ . Then

$$\min_{x \in C} \frac{1}{2} \|x - r\|^2 = \max_{z \in H} -\frac{1}{2} \|z - r\|^2 + \frac{1}{2} \|r\|^2 - \sigma_C(z)$$

$p$  and  $z$  are solutions to the left and right side if and only if  $z = r - p$  and  $z \in N_C p$ .

*Proof.* • We apply duality. Let  $f(z) = \frac{1}{2} \|z - r\|^2 - \frac{1}{2} \|r\|^2$ ,  $g(z) = \sigma_C(z)$ . Then the right side equals  $-\inf_{z \in H} f(z) + g(z)$ .

- $f$  is finite and continuous,  $g(0) = 0$  (since  $C$  non-empty). Therefore by Prop. 1.135 (dual problem):

$$-\inf_{z \in H} f(z) + g(z) = \min_{x \in H} f^*(-x) + g^*(x)$$

- $g = \sigma_C = \iota_C^*$ . Since  $C$  is non-empty, closed, convex  $\Rightarrow g^* = \iota_C^{**} = \iota_C$ .
- Direct computation:  $f^*(x) = \frac{1}{2}\|x + r\|^2$ . So:

$$\min_{x \in H} f^*(-x) + g^*(x) = \min_{x \in C} \frac{1}{2}\|x - r\|^2$$

- Check primal-dual optimality conditions, Prop. 1.138:

$$-p \in \partial f(z) = \{z - r\} \Leftrightarrow z = r - p, \quad p \in \partial g(z) \Leftrightarrow z \in \partial g^*(p) = \partial \iota_C(p) = N_C p$$

- Minimizer  $p$  for projection exists, here also provided by duality.
- For given minimizer  $p$  find via optimality conditions that  $z = r - p$  is dual solution.  $\square$

**Remark 1.150** (Dual interpretation of Dykstra's projection algorithm).

- Analogous to Prop. 1.149, consider dual problem for projection onto intersection. We introduce two auxiliary primal variables:

$$\min_{x \in H} \frac{1}{2}\|x - r\|^2 + \underbrace{\iota_{C \cap D}(x)}_{=\iota_C(x) + \iota_D(x)} = \min_{x, y \in H} \underbrace{\frac{1}{2}\|x - r\|^2 + \iota_{\{0\}}(x - y)}_{:=f(x, y)} + \underbrace{\iota_C(x) + \iota_D(y)}_{:=g(x, y)}$$

- Now dualize analogous to simple projection. Careful: role of primal and dual problem is switched due to 'constraint qualification'. Get:

$$\begin{aligned} f^*(v, w) &= \sup_x \langle x, v + w \rangle - \frac{1}{2}\|r - x\|^2 = \frac{1}{2}\|v + w + r\|^2 - \frac{1}{2}\|r\|^2 \\ g^*(v, w) &= \sup_{x, y} \langle x, v \rangle + \langle y, w \rangle - \iota_C(x) + \iota_D(y) = \sigma_C(v) + \sigma_D(w) \end{aligned}$$

- So:

$$\min_{x, y \in H} f(x, y) + g(x, y) = \sup_{v, w \in H} -\frac{1}{2}\|v + w - r\|^2 + \frac{1}{2}\|r\|^2 - \sigma_C(v) - \sigma_D(w) \quad (*)$$

- Compare with Prop. 1.149 Solve (\*) over  $v$  while fixing  $w$  is (up to constant) equivalent to solving the dual projection problem  $P_C(r - w)$  and consequently the optimal  $v$  is given by  $v = r - w - P_C(r - w)$ . Conversely, fix  $w$  and optimize over  $v$ , get dual to projection  $P_C(r - v)$  and solution is given by  $w = r - v - P_C(r - v)$ .
- Let now  $v^{(0)} = w^{(0)} = 0$ . Define recursively for  $\ell = 0, 1, \dots$ :

$$\begin{aligned} y^{(\ell)} &= P_C(r - w^{(\ell)}) \\ v^{(\ell+1)} &= r - w^{(\ell)} - y^{(\ell)} \\ x^{(\ell+1)} &= P_D(r - v^{(\ell+1)}) \\ w^{(\ell+1)} &= r - v^{(\ell+1)} - x^{(\ell+1)} \end{aligned}$$

- If we set  $x^{(0)} = r$  ( $x^{(0)}$  does not appear in this algorithm), then we find:

$$r - w^{(\ell)} = x^{(\ell)} + v^{(\ell)}, \quad r - v^{(\ell+1)} = y^{(\ell)} + w^{(\ell)}$$

and thus alternating optimization of this dual is equivalent to Dykstra's projection algorithm.

- Not shown in lecture: alternating optimization of (\*) actually converges.
- So get intuition on mechanism of algorithm, as well as potential path to convergence proof.

**Remark 1.151** (Douglas–Rachford interpretation of Dykstra’s algorithm).

- Dykstra’s algorithm can be interpreted as extreme special case of the Douglas–Rachford algorithm. Set:  $f(x) = \frac{1}{2}\|x - r\|^2 + \iota_C(x)$ ,  $g(x) = \frac{1}{2}\|x - r\|^2 + \iota_D(x)$ . Then

$$P_{C \cap D}(r) = \underset{x}{\operatorname{argmin}} f(x) + g(x) \quad (*)$$

- We can attempt to solve this with Douglas–Rachford algorithm. Need proximal operators of  $f$  and  $g$ . Find:

$$\operatorname{Prox}_f(x) = \underset{y}{\operatorname{argmin}} \frac{1}{2}\|x - y\|^2 + \frac{1}{2}\|y - r\|^2 + \iota_C(y) = P_C\left(\frac{x+r}{2}\right)$$

Analogous:  $\operatorname{Prox}_g(x) = P_D\left(\frac{x+r}{2}\right)$ .

- Apply Douglas–Rachford to (\*) for  $\gamma = 1$  and extreme case  $\lambda = 2$  (switch roles of  $f$  and  $g$ , names of  $x$  and  $z$ , and slightly alter the numbering scheme):

$$\begin{aligned} y^{(\ell)} &= P_C\left(\frac{r+z^{(\ell)}}{2}\right) \\ v^{(\ell+1)} &= 2y^{(\ell)} - z^{(\ell)} \\ x^{(\ell+1)} &= P_D\left(\frac{r+v^{(\ell+1)}}{2}\right) \\ z^{(\ell+1)} &= z^{(\ell)} + 2x^{(\ell+1)} - 2y^{(\ell)} \end{aligned}$$

- Compare this line-by-line with Dykstra’s algorithm, as given Remark 1.148.
- First, identify  $x^{(\ell)}$  and  $y^{(\ell)}$  in both algorithms.
- Identifying arguments for  $P_C$  yields:  $2(r - q^{(\ell)}) = r + z^{(\ell)} \Rightarrow z^{(\ell)} = r - 2q^{(\ell)}$ . Compare with fourth line of DR (use Dykstra rule for  $q^{(\ell+1)}$  update):

$$z^{(\ell)} + 2x^{(\ell+1)} - 2y^{(\ell)} = r - 2(q^{(\ell)} + y^{(\ell)} - x^{(\ell+1)}) = r - 2q^{(\ell+1)} = z^{(\ell+1)}$$

- Identify arguments for  $P_D$ , get:  $2(r - p^{(\ell+1)}) = r + v^{(\ell+1)} \Rightarrow v^{(\ell+1)} = r - 2p^{(\ell+1)}$ . Compare with second line of DR (use Dykstra rule for  $p^{(\ell+1)}$  update):

$$v^{(\ell+1)} = r - 2p^{(\ell+1)} = r - 2r + 2q^{(\ell)} + 2y^{(\ell)} = 2y^{(\ell)} - z^{(\ell)}.$$

- So Dykstra can be identified with a particular instance of Douglas–Rachford. However, the convergence proof for DR does not apply directly, since we set  $\lambda = 2$ .

← VL11

## 2 Optimization in Banach spaces

**Remark 2.1** (Motivation). • We go through some fundamental aspects of analysis in Banach spaces. Goal: ensure that minimization problems are well-defined (e.g. minimizers exist, sufficient and necessary criteria for optimality).

- For numerical solution we need to discretize, i.e. approximate by finite-dimensional problem. Must ensure ‘quality’ of approximations. Will introduce  $\Gamma$ -convergence, essentially notion of convergence for minimization problems.
- Discrete problems always finite dimensional. Still: must analyze infinite dimensional problems, to ensure that limit of solutions as we choose finer and finer discretizations is reasonable.

**Remark 2.2** (Convexity in Banach spaces). The notions of convex sets, convex hull, convex functions, lower semicontinuity and cones can be defined on Banach spaces just as in the previous part, since these do not rely on the inner product. For definitions such as subdifferential, normal cone, conjugation we need to be more careful, but generalizations exist. Some more details later.

**Definition 2.3.** Throughout this subsection  $V$  is a real vector space.

### 2.1 Foundations

**Definition 2.4.** • A finite set  $\{x_1, \dots, x_n\} \subset V$  is linearly independent if for all  $(\alpha_i)_{i=1}^n \in \mathbb{R}^n$  one has  $\sum_{i=1}^n \alpha_i \cdot x_i = 0 \Rightarrow \alpha_i = 0$ .

- The span of a finite set  $\{x_1, \dots, x_n\}$  is  $\text{span}\{x_1, \dots, x_n\} = \{\sum_{i=1}^n \alpha_i \cdot x_i \mid (\alpha_i)_{i=1}^n \in \mathbb{R}^n\}$ .
- A basis of  $V$  is a linearly independent finite set  $X \subset V$  such that  $\text{span } X = V$ .
- The dimension of  $V$  is the cardinality of any basis. If no basis exists, the dimension is  $\infty$ .
- On an  $\infty$ -dimensional normed vector space a *Schauder basis* is a sequence  $(x_k)_k$  in  $V$  such that for any  $x \in V$  there is a unique sequence  $(\lambda_k)_k$  such that

$$\lim_{n \rightarrow \infty} \left\| x - \sum_{k=1}^n \lambda_k \cdot x_k \right\| = 0.$$

We recall a few basic facts about topologies.

**Proposition 2.5.** Let  $\{T_i\}_{i \in I}$  be a set of topologies over a set  $X$ , where  $I$  is an arbitrary index set. Then their *intersection*  $T$ , where  $t \in T$  iff  $t \in T_i$  for all  $i \in I$ , is a topology.

**Definition 2.6** (Induced topology). Let  $(f_i : X \rightarrow Y_i)_{i \in I}$  be a family of maps from  $X$  to topological spaces  $Y_i$ . Then the intersection  $T$  of all topologies  $(T_j)_{j \in J}$  such that all maps  $(f_i)_{i \in I}$  are continuous is the *induced topology*. Since  $T$  only contains sets that are contained in all other  $T_j$ ,  $T$  also referred to as *coarsest topology* in which the family  $(f_i)_{i \in I}$  is continuous.

**Remark 2.7** (Metric topology).

- Let  $(X, d)$  be a metric space. The metric topology on  $X$  is the topology induced by the family of maps  $(y \mapsto d(x, y))_{x \in X}$  from  $X$  to  $\mathbb{R}$  where  $\mathbb{R}$  is equipped with the standard topology.

- For  $x \in X$ ,  $\varepsilon > 0$  the set  $B(x, \varepsilon) = \{y \in X | d(x, y) < \varepsilon\}$  is called the *open ball* of radius  $\varepsilon$  around  $x$ .  $B(x, \varepsilon)$  is open in the metric topology.
- Any open set in the metric topology can be written as union of open balls.
- The metric topology is Hausdorff. That is, for every distinct pair  $x_1, x_2 \in X$  there are open sets  $A_1, A_2 \subset X$  with  $x_i \in A_i$ ,  $A_1 \cap A_2 = \emptyset$ .
- Let  $A \subset X$  be open. Then for every  $x \in A$  there is some  $\varepsilon > 0$  such that  $B(x, \varepsilon) \subset A$ .
- Convergence in the metric is equivalent to convergence in the metric topology.

We list a few exemplary Banach spaces that serve as toy cases throughout the lecture.

Let  $\Omega \subset \mathbb{R}^n$  be open, bounded, non-empty. For a ‘multiindex’  $a \in \{0, 1, 2, \dots\}^n$  let  $|a| = a_1 + \dots + a_n$  and  $D^a f := \frac{\partial^{|a|}}{\partial x_1^{a_1} \dots \partial x_n^{a_n}} f$ .

**Definition 2.8.** The space of  $k$  times continuously differentiable functions on  $\overline{\Omega}$  is denoted by  $C^k(\overline{\Omega})$ . It is a Banach space when equipped with the norm

$$\|f\|_{C^k(\overline{\Omega})} := \max_{a: |a| \leq k} \max_{x \in \overline{\Omega}} |D^a f(x)|.$$

**Remark 2.9.** The maximizer in the definition of the norm exists since  $D^a f$  is continuous for  $|a| \leq k$  and  $\overline{\Omega}$  is compact (since  $\Omega$  is bounded).

We give a prototypical result for the relation between different function spaces.

**Proposition 2.10** (Relation between  $C^k([0, 1])$ ). For any integer  $k > 0$  the space  $C^0([0, 1])$  is the completion of  $C^k([0, 1])$  with respect to the norm  $\|\cdot\|_{C^0([0, 1])}$ . More precisely,

- (i) any sequence  $(f_i)_i$  in  $C^k([0, 1])$  that is Cauchy with respect to the norm  $\|\cdot\|_{C^0([0, 1])}$  has a limit in  $C^0([0, 1])$ ,
- (ii) and any  $f \in C^0([0, 1])$  can be reached as limit of such a sequence.

---

Comment: Result sometimes allows to ‘temporarily’ restrict an optimization problem to a space with higher regularity, since the regularity is only lost ‘in the limit’.

---



---

Comment: Approximating sequences are in general not Cauchy in  $C^k([0, 1])$ .

---

*Proof.* • (i) follows directly from  $C^k([0, 1]) \subset C^0([0, 1])$ . We turn to (ii).

- By the famous Weierstrass approximation theorem any  $f \in C^0([0, 1])$  can be approximated to any given precision  $\varepsilon > 0$  in the norm  $\|\cdot\|_{C^0([0, 1])}$  by a polynomial, see e.g. [Narici, Beckenstein: Topological Vector Spaces; Section 16.5].
- So any  $f$  can be written as limit of a convergent (in  $\|\cdot\|_{C^0([0, 1])}$ ) sequence  $(f_i)_i$  of polynomials; therefore  $(f_i)_i$  is Cauchy with respect to this norm. And clearly  $f_i \in C^k([0, 1])$ .  $\square$

The following family of spaces will often serve as instructive examples.

**Definition 2.11.** For  $p \in [1, \infty]$  let  $\ell^p = \{x = (x_1, x_2, \dots) \in \mathbb{R}^{\mathbb{N}} | \|x\|_{\ell^p} < \infty\}$  where

$$\|x\|_{\ell^p} = \begin{cases} (\sum_{i=1}^{\infty} |x_i|^p)^{1/p} & \text{if } p < \infty, \\ \sup_i |x_i| & \text{if } p = \infty. \end{cases}$$

The following inequalities are often useful when one must derive upper bounds. They would also allow to prove that  $\ell^p$  is a Banach space.

**Proposition 2.12** (Hölder inequality for  $\ell^p$ ). For  $p, q \in [1, \infty]$  with  $\frac{1}{p} + \frac{1}{q} = 1$ ,  $x \in \ell^p$ ,  $y \in \ell^q$  we have  $\sum_{i=1}^{\infty} |x_i y_i| \leq \|x\|_{\ell^p} \cdot \|y\|_{\ell^q}$ . For  $p, q \in (1, \infty)$  there is equality if and only if  $\left(\frac{|x_i|}{\|x\|_{\ell^p}}\right)^p = \left(\frac{|y_i|}{\|y\|_{\ell^q}}\right)^q$ .

---

Comment: Generalization of Cauchy-Schwarz inequality

---

**Proposition 2.13** (Minkowski inequality for  $\ell^p$ ). For  $p \in [1, \infty]$  find  $x, y \in \ell^p \Rightarrow x + y \in \ell^p$  with  $\|x + y\|_{\ell^p} \leq \|x\|_{\ell^p} + \|y\|_{\ell^p}$ . For  $p \in (1, \infty)$  there is equality if and only if  $x = q \cdot y$  for some  $q \geq 0$ .

Other prominent examples that are also very common in applications are  $L^p$  spaces and Sobolev spaces.

## 2.2 Compactness and separability

We introduce the topological dual of a Banach space, which can be interpreted as an ‘approximation’ for an inner product on Banach spaces. We study related questions on compactness and see how far we can adapt notions of convex duality to this setting.

In Hilbert spaces we have shown that projections onto closed convex sets, i.e. points of minimal distance, exist (and are unique). This is in general no longer true in Banach spaces, due to a lack of a notion of orthogonality.

To gain some intuition, we first give an example where projections exist and then give a counterexample.

**Example 2.14.** • Let  $X = C^0([0, 1])$ , equipped with the norm  $\|\cdot\| = \|\cdot\|_{C^0([0,1])}$ .

- Let  $Y = \{f \in X \mid \int_0^1 f(x) dx = 0\}$ .  $Y$  is a closed subspace of  $X$ .
- Consider now the projection problem from  $X$  onto  $Y$ . For  $f \in X$  study  $\inf_{g \in Y} \|f - g\|$ .
- Intuition: for given  $f \notin Y$ , how do we change ‘mean’ of  $f$  with minimal perturbation in the norm?  $\Rightarrow$  change each point of  $f$  by same value.
- Let  $g \in Y$  and set  $h = f - g$ . From  $\int_0^1 g dx = 0$  we find that we need to find  $h \in X$  with  $\int_0^1 f dx = \int_0^1 h dx$  that has minimal norm.
- This implies that  $\|h\| \geq |\int_0^1 f dx|$  (otherwise  $|\int_0^1 h dx| \leq \int_0^1 \|h\| dx < |\int_0^1 f dx|$ ).
- Try constant function  $h(x) = \int_0^1 f(y) dy$ . This satisfies integral constraint and we find  $\|h\| = |\int_0^1 f(y) dy|$ . So  $h$  is optimal and  $g(x) = f(x) - \int_0^1 f(y) dy$ .
- Note:  $h$  is unique. Assume,  $h$  were not constant. Then  $\|h\| > |\int_0^1 h dx|$ .

Now, by giving a counterexample, we show that projections onto closed convex sets do not always exist in Banach spaces.

**Proposition 2.15.** For a Banach space  $(X, \|\cdot\|)$  and a closed subspace  $Y \subset X$  and some fixed  $x \in X$ , there is not always a minimizer of  $\inf_{y \in Y} \|x - y\|$ .

*Proof.* • The proof is a slight modification of the example above.

- Let  $X = \{f \in C^0([0, 1]) \mid f(0) = 0\}$  with norm  $\|\cdot\| = \|\cdot\|_{C^0([0, 1])}$ .  $X$  is a closed subspace of  $C^0([0, 1])$  (why? check Cauchy sequences) and therefore  $X$  is a Banach space.
- Let  $Y = \{g \in X \mid \int_0^1 g(x)dx = 0\}$ .
- Analogous to above: fix  $f \in X \setminus Y$ , rewrite problem. Let  $g \in Y$ , set  $h = f - g$ . Solve  $s = \inf\{\|h\| \mid h \in X \mid \int_0^1 h dx = \int_0^1 f dx\}$ . Almost as above, but now have additional constraint  $h(0) = 0$ , since we may not change  $f(0) = 0$ .
- So constant shift no longer works, infimal norm of  $h$  cannot be smaller than above, i.e.  $s \geq |\int_0^1 f dx|$ . Since feasible  $h$  cannot be constant, must have  $\|h\| > |\int_0^1 f dx|$  for all feasible  $h$ .
- Now show that infimum  $s = |\int_0^1 f dx|$  which then implies that no minimizer exists. Do this by ‘approximating’ constant shift as good as possible, while obeying the  $h(0) = 0$  constraint.
- Set  $h_i(x) = (\int_0^1 f(y)dy) \cdot (1 + 1/i) \cdot x^{1/i}$ .  $h_i(0) = 0$ ,  $\int_0^1 h_i(x)dx = \int_0^1 f(y)dy$  and  $\|h_i\| = |h_i(1)| = |\int_0^1 f(y)dy| \cdot (1 + 1/i)$ . So  $(h_i)_i$  is a minimizing sequence.

---

**Sketch:**  $h_i$ , approximation of constant shift.

---

Comment:  $(h_i)_i$  is not Cauchy. Its pointwise limit is  $h_\infty(x) = 1$  for  $x \in (0, 1]$ ,  $h(0) = 0$ , which is not in  $X \subset C^0([0, 1])$ .

---

□

So the projection does not exist, but we can find a sequence  $(h_i)_i$  that is approximately orthogonal to the subspace. Such a sequence exists in general.

**Proposition 2.16** (Almost orthogonal element). For a Banach space  $X$  let  $Y$  be a subspace,  $Y \neq X$ . For any  $x \in X$  and  $\theta > 1$  there are some  $x_\theta \in X$ ,  $y_\theta \in Y$  such that  $x = x_\theta + y_\theta$  and  $\text{dist}(x, Y) = \text{dist}(x_\theta, Y) \leq \|x_\theta\| \leq \theta \cdot \text{dist}(x, Y)$ .

*Proof.* •  $\text{dist}(x, Y) = \inf_{y \in Y} \|x - y\|$  is finite and non-negative.

- Let  $y_\theta \in Y$  such that  $\|x - y_\theta\| < \theta \cdot \text{dist}(x, Y)$  (take from a minimizing sequence) and set  $x_\theta = x - y_\theta$ .
- $\text{dist}(x_\theta, Y) = \inf_{y \in Y} \|x - y_\theta - y\| = \inf_{y \in Y} \|x - y\| = \text{dist}(x, Y)$ . □

**Remark 2.17.** If  $Y$  is not closed, then may have  $\text{dist}(x, Y) = 0$  even when  $x \notin Y$ . Then  $\|x_\theta\| \rightarrow 0$  as  $\theta \searrow 1$ .

**Corollary 2.18.** If  $Y$  is closed and  $Y \neq X$  then for any  $\theta > 1$  there is some  $x_\theta$  with  $\|x_\theta\| = 1$  and  $\text{dist}(x_\theta, Y) \geq \frac{1}{\theta}$ .

We have learned that an important ingredient for existence of minimizers is compactness. Now study (strong) compactness on Banach spaces.

**Proposition 2.19.** On a metric space  $(X, d)$  the notions of compactness and sequential compactness are equivalent.

*Proof.* • **compactness  $\Rightarrow$  sequential compactness:** Let  $A \subset X$  be compact, let  $(x_k)_k$  be a sequence in  $A$ . Assume  $(x_k)_k$  has no cluster point. Then  $\forall y \in A \exists \delta_y > 0$  such that  $N_y = \{k \in \mathbb{N} : x_k \in B(y, \delta_y)\}$  is finite.

- The sets  $B(y, \delta_y)$  for  $y \in A$  form an open cover of  $A$ . Since  $A$  is compact, there is a finite subcover for some  $(y_1, \dots, y_n)$ :

$$A \subset \bigcup_{y \in A} B(y, \delta_y) \Rightarrow A \subset \bigcup_{i=1}^n B(y_i, \delta_{y_i}) \Rightarrow \mathbb{N} = \bigcup_{i=1}^n N_{y_i}$$

- This implies that  $\mathbb{N}$  is finite. So  $(x_k)_k$  must have at least one cluster point and thus  $A$  is sequentially compact.
- **sequential compactness  $\Rightarrow$  compactness:** For any  $\varepsilon > 0$  can cover  $A$  with finitely many  $\varepsilon$ -balls (otherwise, we could define sequence in  $A$  without cluster points via  $x_k \in A \setminus \bigcup_{i=1}^{k-1} B(x_i, \varepsilon)$ ).
- For an open cover  $A \subset \bigcup_{i \in I} U_i$ ,  $\exists \varepsilon_0 > 0$  such that  $\forall x \in A \exists i_x \in I$  such that  $B(x, \varepsilon_0) \subset U_{i_x}$ . Prove this by contradiction.
- Assume  $\forall k \in \mathbb{N} \exists x_k \in A$  such that  $\forall i \in I B(x_k, 1/k) \not\subset U_i$ . Since  $A$  is sequentially compact,  $(x_k)_k$  has cluster point  $x \in A$ . Let  $(x_{k_j})_j$  be subsequence converging to  $x$ .  $\exists \delta > 0$  such that  $B(x, \delta) \subset U_i$  for some  $i \in I$ .
- $\exists j \in \mathbb{N}$  such that  $\frac{1}{k_j} < \frac{\delta}{2}$  and  $d(x_{k_j}, x) < \frac{\delta}{2} \Rightarrow B(x_{k_j}, 1/k_j) \subset U_i$ , which is a contradiction.
- So for this  $\varepsilon_0$  chose  $x_1, \dots, x_n \in A$  such that

$$A \subset \bigcup_{k=1}^n B(x_k, \varepsilon_0) \subset \bigcup_{k=1}^n U_{i_{x_k}}. \quad \square$$

**Corollary 2.20.** Projections onto compact sets exist in Banach spaces.

*Proof.* Can extract cluster point from any minimizing sequence of  $\text{dist}(x, Y)$ . Is minimizer (and therefore, projection) since  $y \mapsto d(x, y)$  is continuous (by construction of metric topology).  $\square$

**Proposition 2.21.** For a Banach space  $X$  we find:  $[\overline{B(0, 1)} \text{ compact}] \Leftrightarrow [X \text{ is finite-dimensional}]$

*Proof.* •  $\Rightarrow$ :  $\overline{B(0, 1)} \subset \bigcup_{y \in B(0, 1)} B(y, \frac{1}{2}) \Rightarrow$  (finite subcover from compactness)  $\overline{B(0, 1)} \subset \bigcup_{i=1}^n B(y_i, \frac{1}{2})$ .

- $Y = \text{span}\{y_1, \dots, y_n\}$  is closed subspace. Assume  $Y \neq X$ .
- By Corollary 2.18 for  $\theta > 1$  there is some  $x_\theta$  with  $\|x_\theta\| = 1$  and  $\text{dist}(x_\theta, \{y_1, \dots, y_n\}) \geq \text{dist}(x_\theta, Y) \geq \frac{1}{\theta}$ .
- But since  $\|x_\theta\| = 1 \Rightarrow x_\theta \in \overline{B(0, 1)} \subset \bigcup_{i=1}^n B(y_i, \frac{1}{2}) \Rightarrow \text{dist}(x_\theta, \{y_1, \dots, y_n\}) < \frac{1}{2}$ . For  $\theta < 2$  this is a contradiction. So  $Y = X$  and  $X$  is finite-dimensional.
- $\Leftarrow$ : If  $X$  is finite dimensional, identify it with  $\mathbb{R}^n$ . All norms are equivalent on  $\mathbb{R}^n$ . Therefore compactness of  $\overline{B(0, 1)}$  follows from Heine–Borel.  $\square$

So  $\overline{B(0,1)}$  in  $C^0([0,1])$  is not compact. However, one can show that  $\overline{B(0,1)}$  of  $C^1([0,1])$  is compact with respect to the  $C^0([0,1])$  topology.

**Definition 2.22** (Equicontinuity). A family of functions  $f_i : \Omega \rightarrow \mathbb{R}$ ,  $i \in I$  is equicontinuous if for any  $x \in \Omega$  and  $\varepsilon > 0$  there is some  $\delta > 0$  such that  $|f_i(y) - f_i(x)| < \varepsilon$  for all  $y \in \Omega$  with  $\|y - x\| < \delta$ ,  $i \in I$ .

**Proposition 2.23** (Arzelà–Ascoli). A set  $A \subset C^0([0,1])$  is (sequentially, equivalent, why?) pre-compact (the closure of  $A$  is compact) if and only if it is bounded and equicontinuous.

**Corollary 2.24.** The set  $A = B(0,1)$  of  $C^1([0,1])$  is pre-compact with respect to the  $C^0([0,1])$  topology.

*Proof.* For every  $f \in A$ ,  $x \in [0,1]$  have  $|f(x)| \leq 1$  and  $|f'(x)| \leq 1$ . Therefore  $A$  is bounded in  $C^0([0,1])$  and  $A$  is equicontinuous:  $|f(x) - f(y)| \leq |x - y|$ .  $\square$

**Definition 2.25** (Separable metric space). A topological space  $X$  is *separable* if it contains a countable, dense subset  $A$ .

**Remark 2.26.**  $A$  dense in  $X$  means that any point in  $X$  can be reached as the limit of a sequence in  $A$ , or equivalently that any non-empty open set in  $X$  has non-empty intersection with  $A$ . Intuitively, separability is a bound on the cardinality of the space. Even if  $X$  is uncountable, it can be ‘reasonably approximated’ by a countable set of elements. On separable spaces many proofs are constructive and one can avoid the axiom of choice.

**Example 2.27.** (i) The set  $\mathbb{R}$  with the usual topology is separable, as  $\mathbb{Q}$  is dense in  $\mathbb{R}$ .

(ii) The set  $\mathbb{R}$  with the discrete topology (all sets are open) is not separable, since the only set that is dense in this space is  $\mathbb{R}$  itself, which is not countable.

More examples:

**Proposition 2.28.** A compact metric space  $(X, d)$  is separable.

*Proof.* • For  $n \in \mathbb{N}$  have  $X \subset \bigcup_{x \in X} B(x, \frac{1}{n})$  as an open cover. Therefore, there is a finite subcover  $X \subset \bigcup_{i=1}^{k_n} B(x_{n,i}, \frac{1}{n})$ .

• The countable set  $A = \bigcup_{n=1}^{\infty} \{x_{n,1}, \dots, x_{n,k_n}\}$  is dense in  $X$ : For  $x \in X$  and  $\varepsilon > 0$ , set  $n > 1/\varepsilon$ , then  $x \in B(x_{n,i}, \frac{1}{n}) \subset B(x_{n,i}, \varepsilon)$  for some  $i \in \{1, \dots, k_n\}$ .

• So we can generate a sequence in  $A$  that converges to  $x$ .  $\square$

**Proposition 2.29.** An infinite-dimensional Banach space with a Schauder basis is separable.

**Remark 2.30.** The converse implication is not true in general, see e.g. [Narici, Beckenstein: Topological Vector Spaces; Section 11.1] (which is primarily a very interesting brief historical summary of the mathematical life of Stefan Banach).

*Proof.* • Let  $(x_i)_i$  be a Schauder basis of  $X$ . In particular it is countable. W.l.o.g. we can assume that  $\{\|x_i\|\}_i$  is bounded by some  $C < \infty$ .

• Let  $A_n = \{\sum_{i=1}^n s_i x_i \mid (s_i)_i \in \mathbb{Q}^n\}$ . Since  $\mathbb{Q}$  is countable and  $A_n$  is a finite union of countable sets  $\{\mathbb{Q} \cdot x_i\}$ ,  $A_n$  is countable.

- Let  $A = \bigcup_{n=1}^{\infty} A_n$ . Since  $A$  is a countable union of countable sets,  $A$  is countable.
- Fix now  $x \in X$  and some  $\varepsilon > 0$ .
- By definition there is a (unique) sequence  $(\alpha_i)_i$  in  $\mathbb{R}$  such that  $\lim_{n \rightarrow \infty} \|x - \sum_{i=1}^n \alpha_i x_i\| \rightarrow 0$ , in particular there is some  $n$  such that  $\|x - \sum_{i=1}^n \alpha_i x_i\| < \varepsilon/2$ .
- Let now  $\beta_i \in \mathbb{Q}$  such that  $|\alpha_i - \beta_i| < \frac{\varepsilon}{2^{i+1}C}$ . Let  $z_n = \sum_{i=1}^n \alpha_i x_i$ ,  $y_n = \sum_{i=1}^n \beta_i x_i \in A$ .

$$\|x - y_n\| \leq \|x - z_n\| + \|z_n - y_n\| < \frac{\varepsilon}{2} + \sum_{i=1}^n |\alpha_i - \beta_i| \cdot \|x_i\| < \frac{\varepsilon}{2} + \varepsilon \sum_{i=1}^n 2^{-i-1} < \varepsilon \quad \square$$

**Proposition 2.31.**  $C^0([0, 1])$  is separable.

We use a small auxiliary Lemma for the proof that is often helpful when working on compact spaces.

**Lemma 2.32.** Let  $(X, d)$  be a compact metric space. A continuous function  $f : X \rightarrow \mathbb{R}$  is uniformly continuous, i.e.  $\forall \varepsilon > 0 \exists \delta > 0$  such that  $|f(x) - f(y)| < \varepsilon$  when  $d(x, y) < \delta$ .

*Proof.* • For every  $\varepsilon > 0$ ,  $x \in X \exists \delta_x > 0$  such that  $|f(x) - f(y)| < \varepsilon/2$  if  $y \in B(x, \delta_x)$ .

- The sets  $(B(x, \delta_x/2))_{x \in X}$  form an open cover of  $X$  and  $X$  is compact  $\Rightarrow$  there is a finite subcover with midpoints  $\{x_1, \dots, x_n\}$ . Let  $\delta = \min\{\delta_{x_1}, \dots, \delta_{x_n}\} > 0$ .
- Now let  $x, y \in X$ ,  $d(x, y) < \delta/2$ . Then  $x \in B(x_i, \delta_{x_i}/2)$  for some  $i$  and therefore  $y \in B(x_i, \delta_{x_i})$ .
- So  $|f(x) - f(y)| \leq |f(x) - f(x_i)| + |f(y) - f(x_i)| < \varepsilon$ .  $\square$

*Proof of Proposition 2.31.* • By Lemma 2.32 any  $f \in C^0([0, 1])$  is uniformly continuous. So for any  $\varepsilon > 0$  can find  $n \in \mathbb{N}$  such that  $|f(x) - f(y)| < \varepsilon$  for  $|x - y| < 2^{-n}$ . So we can uniformly approximate  $f$  by piecewise affine interpolation between values at points  $i \cdot 2^{-n}$  for  $i \in \{0, \dots, 2^n\}$ .

- Define set of ‘tent functions’ of scale  $n$  for  $i \in \{0, \dots, 2^n\}$ :

$$f_{n,i}(x) = \begin{cases} 0 & \text{if } |x - i \cdot 2^{-n}| \geq 2^{-n}, \\ 1 - |2^n x - i| & \text{else} \end{cases}$$

---

**Sketch:** Tent functions.

---

- So piecewise affine interpolation with resolution  $n$  can be written as superposition of functions  $f_{n,i}$ . The functions  $(f_{n,i})_{n,i}$  are an ‘overcomplete’ Schauder basis of  $C^0([0, 1])$ . The decomposition may not be unique, since the  $f_{n,i}$  are not all linearly independent.
- Could re-establish uniqueness, by iteratively removing linearly dependent elements. But reasoning of Prop. 2.29 does not depend on uniqueness of the decomposition. So separability of  $C^0([0, 1])$  follows.  $\square$

**Remark 2.33.** (i) Since  $C^k([0, 1])$  can be parametrized by  $C^0([0, 1])$  and a finite number of integration constants, this argument extends to  $C^k([0, 1])$ .

- (ii) For spaces  $X$  that can be written as completions of  $C^k([0, 1])$  with respect to coarser norms by construction  $C^k([0, 1])$  is dense in  $X$  and thus  $X$  is then also separable. This covers many spaces of integrable functions and Sobolev spaces.

---

Comment: Many ‘practical’ spaces remain separable, even if they ‘look’ very high dimensional. For this need ‘sufficiently coarse’ topology. (Recall  $\mathbb{R}$  with discrete topology is not separable.)

---

**Proposition 2.34.** The spaces  $\ell^p$  for  $p \in [1, \infty)$  have a Schauder basis and are separable.

*Proof.* • Let  $e_i \in \ell^p$  with  $e_{i,j} = \delta_{i,j}$ . Claim:  $(e_i)_{i \in \mathbb{N}}$  is a Schauder basis of  $\ell^p$ .

- Let  $x \in \ell^p$ . Set  $z_i = \sum_{j=1}^i e_j x_j$ ,  $z_{i,j} = x_j \delta_{j \leq i}$ . Find:

$$\|x - z_i\|_{\ell^p}^p = \sum_{j=1}^{\infty} |x_j - z_{i,j}|^p = \sum_{j=1}^{\infty} |x_j - x_j \delta_{j \leq i}|^p = \sum_{j=i+1}^{\infty} |x_j|^p = \|x\|_{\ell^p}^p - \sum_{j=1}^i |x_j|^p \rightarrow 0$$

as  $i \rightarrow \infty$ . So  $z_i \rightarrow x$ .

- This decomposition is unique. Let  $(y_i)_i$  be another sequence such that  $\lim_{i \rightarrow \infty} v_i \rightarrow x$  for  $v_i = \sum_{j=1}^i y_j \cdot e_j$  with  $|y_{i_0} - x_{i_0}| > \delta$  for some  $i_0$ . Then  $\|v_i - x\| \geq \delta$  for all  $i \geq i_0$  and thus this sequence cannot converge to  $x$ .  $\square$

**Proposition 2.35.** The space  $\ell^\infty$  is not separable.

*Proof.* • Let  $A \subset \ell^\infty$  be countable, i.e.  $A = (a_k)_{k \in \mathbb{N}}$  where for each  $k \in \mathbb{N}$  have  $a_k = (a_{k,1}, a_{k,2}, \dots) \in \ell^\infty$ .

- Define sequence  $(b_k)_k$  by

$$b_k = \begin{cases} a_{k,k} + 1 & \text{if } |a_{k,k}| \leq 1, \\ 0 & \text{if } |a_{k,k}| > 1. \end{cases}$$

- $\sup_{k \in \mathbb{N}} |b_k| \leq 2$ , i.e.  $\|b\|_{\ell^\infty} \leq 2$  and thus  $b \in \ell^\infty$ .
- $\|b - a_k\|_{\ell^\infty} \geq |b_k - a_{k,k}| \geq 1$  for all  $k \in \mathbb{N}$ . Therefore,  $b$  cannot be approximated by a sequence in  $A$  and thus no countable set can be dense in  $\ell^\infty$ .  $\square$

**Proposition 2.36.** An infinite-dimensional Hilbert space  $H$  is separable if and only if it has a orthonormal Schauder basis.

*Proof.* •  $\Leftarrow$ : If  $H$  has a orthonormal Schauder basis, separability follows from Prop. 2.29.

- $\Rightarrow$ : Let  $\{a_k\}_k$  be a countable set that is dense in  $H$ . Apply Gram-Schmidt orthonormalization to  $(a_k)_k$  to generate orthonormal sequence  $(x_i)_i$ . (Start with smallest  $k$  such that  $a_k \neq 0$ , set  $x_1 = a_k / \|a_k\|$ ,  $i = 1$ . Then increase  $k$  until  $a_k \notin \text{span}\{x_j\}_{j=1}^i$ . Add orthonormal component of  $a_k$  as new basis vector to  $(x_i)_i$ , increase  $i$ . Since  $H$  is infinite-dimensional and  $\{a_k\}_k$  is dense,  $i$  will tend to  $\infty$ . Note that  $i \leq k$  throughout the process and that  $a_k \in \text{span}\{x_j\}_{j=1}^k$  at all steps.)
- $(x_i)_i$  is orthonormal by construction. Show that it is Schauder basis.

- By construction  $a_k = \sum_{i=1}^k x_i \langle x_i, a_k \rangle$ . For any  $z \in H$  by density there is a subsequence  $(a_{k_j})_j$  that converges to  $z$ . So

$$\begin{aligned} \left\| z - \sum_{i=1}^{k_j} x_i \langle x_i, z \rangle \right\| &\leq \|z - a_{k_j}\| + \left\| \sum_{i=1}^{k_j} x_i \langle x_i, a_{k_j} - z \rangle \right\| \\ &\leq \|z - a_{k_j}\| + \left( \left\langle \sum_{i=1}^{k_j} x_i \langle x_i, a_{k_j} - z \rangle, \sum_{i=1}^{k_j} x_i \langle x_i, a_{k_j} - z \rangle \right\rangle \right)^{1/2} \end{aligned}$$

(Bessel inequality, cf. Example 1.91)

$$\leq 2\|z - a_{k_j}\| \rightarrow 0 \quad \text{as } j \rightarrow \infty.$$

- By orthonormality of  $(x_i)_i$  the coefficients  $\langle x_i, z \rangle$  are unique. □

**Corollary 2.37.** Every separable Hilbert space  $H$  is isomorphic to  $\ell^2$ , i.e. there is a bijection  $\phi : H \rightarrow \ell^2$  such that  $\langle x, y \rangle_H = \langle \phi(x), \phi(y) \rangle_{\ell^2}$ .

← VL12

## 2.3 Linear transformations and topological dual

**Definition 2.38** (Linear transformations and functionals).

- For two normed spaces  $X, Y$ ,  $D \subset X$  a map  $T : D \rightarrow Y$  is called *transformation* from  $X$  to  $Y$  with domain  $D$ .
- A transformation  $T : D \rightarrow \mathbb{R}$  is called *functional*.
- $T : X \rightarrow Y$  is *linear* if  $T(ax + by) = aT(x) + bT(y)$  for all  $x, y \in X$ ,  $a, b \in \mathbb{R}$ .
- The *operator norm* of a transformation  $T : X \rightarrow Y$  is defined as

$$\|T\| = \sup_{x \in X \setminus \{0\}} \frac{\|T(x)\|_Y}{\|x\|_X}.$$

$T$  is called *bounded* if  $\|T\| < \infty$ .

- The set of bounded linear transformations from  $X$  to  $Y$  is denoted by  $L(X, Y)$ .

**Proposition 2.39.** Let  $T$  be a linear transformation from  $X$  to  $Y$ .

- (i)  $[T \text{ continuous in } 0] \Leftrightarrow [T \text{ continuous on } X]$
- (ii)  $[T \text{ bounded}] \Leftrightarrow [T \text{ continuous}]$

*Proof.* •  $X$  and  $Y$  normed spaces  $\Rightarrow$  “ $\varepsilon$ - $\delta$ -notion” of continuity is sufficient.

- **(i):** Let  $x \in X$ ,  $\varepsilon > 0$ .  $[\exists \delta > 0 : \|T(z)\|_Y < \varepsilon \text{ if } \|z\|_X < \delta] \Leftrightarrow [\exists \delta > 0 : \|T(x) - T(y)\|_Y = \|T(x - y)\|_Y < \varepsilon \text{ if } y \in B_X(x, \delta)]$
- **(ii):**  $\Rightarrow$ :  $[T \text{ bounded}] \Rightarrow [\forall x \in X : \|T(x)\|_Y \leq \|T\| \cdot \|x\|_X \Rightarrow [T \text{ continuous in } 0] \Leftrightarrow [T \text{ continuous}].$
- $\Leftarrow$ : let  $\varepsilon > 0$ .  $[T \text{ continuous in } 0] \Rightarrow [\exists \delta > 0 : \|T(x)\|_Y \leq \varepsilon \text{ if } \|x\|_X < \delta].$
- for any  $y \in X \setminus \{0\}$  find:

$$\frac{\|T(y)\|_Y}{\|y\|_X} = \frac{\frac{2\|y\|_X}{\delta} \|T(\frac{\delta}{2\|y\|_X} y)\|_Y}{\|y\|_X} \leq \frac{2\varepsilon}{\delta} < \infty$$

- This bound is uniform for all  $y \in Y \setminus \{0\}$ , therefore  $\|T\| < \frac{2\varepsilon}{\delta}$ . □

**Proposition 2.40.**  $L(X, Y)$  is a vector space and the operator norm is indeed a norm on  $L(X, Y)$ .

*Proof.* • For  $S, T \in L(X, Y)$ ,  $a, b \in \mathbb{R}$  clearly  $(aS + bT) : x \mapsto aS(x) + bT(x)$  is a linear transformation.

- $\|T\| \geq 0$  by definition.  $[\|T\| = 0] \Leftrightarrow [T(x) = 0 \text{ for all } x \in X] \Leftrightarrow [T = 0].$
- $\|a \cdot T\| = |a| \cdot \|T\|$  by homogeneity of  $\|\cdot\|_Y$ .
- $\|S + T\| = \sup_{x \in X \setminus \{0\}} \frac{\|S(x) + T(x)\|_Y}{\|x\|_X} \leq \sup_{x \in X \setminus \{0\}} \frac{\|S(x)\|_Y}{\|x\|_X} + \sup_{y \in X \setminus \{0\}} \frac{\|T(y)\|_Y}{\|y\|_X} = \|S\| + \|T\|.$

- Since  $\|\cdot\|$  is subadditive, whenever  $S$  and  $T$  are bounded, so is  $S + T$ . So  $L(X, Y)$  is a vector space.  $\square$

**Proposition 2.41.** If  $Y$  is a Banach space, then so is  $L(X, Y)$  when equipped with the operator norm.

*Proof.* • Let  $(T_n)_n$  be a Cauchy sequence in  $L(X, Y)$ .  $\Rightarrow$  for any  $\varepsilon > 0 \exists N \in \mathbb{N}$  such that  $\|T_m - T_n\| < \varepsilon$  for  $m, n > N$ .

- For  $m, n > N$ ,  $x \in X$  get  $\|T_m(x) - T_n(x)\|_Y < \varepsilon \|x\|_X$ . So  $(T_n(x))_n$  is Cauchy in  $Y$ . Since  $Y$  is Banach, sequence has limit.
- Set  $T : x \mapsto \lim_{n \rightarrow \infty} T_n(x)$ .
- $T$  is linear:

$$T(ax + by) = \lim_{n \rightarrow \infty} T_n(ax + by) = \lim_{n \rightarrow \infty} aT_n(x) + bT_n(y)$$

(sum of Cauchy sequences is Cauchy)

$$= aT(x) + bT(y).$$

- $T$  is bounded: let  $x \in X \setminus \{0\}$ ,  $n > N$ .

$$\begin{aligned} \|T(x)\|_Y &\leq \|T(x) - T_n(x)\|_Y + \|T_n(x)\|_Y = \lim_{m \rightarrow \infty} \|T_m(x) - T_n(x)\|_Y + \|T_n(x)\|_Y \\ &\leq \varepsilon \cdot \|x\|_X + \|T_n\| \cdot \|x\|_X \end{aligned}$$

where we used  $\|T_m(x) - T_n(x)\|_Y < \varepsilon \|x\|_X$  for  $m, n > N$ . Divide by  $\|x\|_X \neq 0$  to get uniform bound on  $\|T(x)\|_Y / \|x\|_X$  and thus that  $T$  has finite norm.  $\square$

**Definition 2.42** (Dual space).

- For a normed space  $X$  the Banach space of functionals  $L(X, \mathbb{R})$  equipped with the operator norm is called the *topological dual* space of  $X$  and denoted by  $X^*$ .
- Every  $t \in X^*$  is a bounded linear functional on  $X$ . The application  $t(x)$  is often also denoted as  $\langle t, x \rangle_{X^* \times X}$ .
- The map  $X^* \times X \rightarrow \mathbb{R}$  via  $(t, x) \mapsto \langle t, x \rangle_{X^* \times X}$  is called *duality pairing*. The subscript  $X^* \times X$  is dropped when the context is clear.
- From linearity of  $t$  and since  $X^*$  is a vector space, the duality pairing is bilinear.

**Proposition 2.43.** The duality pairing  $X^* \times X \rightarrow \mathbb{R}$ ,  $(t, x) \mapsto t(x)$  is jointly continuous in the product topology of the (strong / norm) topologies on  $X^*$  and  $X$ .

*Proof.* • Let  $(s, x) \in X^* \times X$ ,  $\varepsilon > 0$ . Set  $\delta = \min\{1, \frac{\varepsilon}{\|s\|_{X^*} + \|x\|_X + 1}\}$ . Then  $\delta \leq \frac{\varepsilon}{\|s\|_{X^*} + \|x\|_X + 1}$

- Now let  $t \in B_{X^*}(s, \delta)$ ,  $y \in B_X(x, \delta)$ . Get

$$\begin{aligned} |s(x) - t(y)| &\leq |s(x) - s(y)| + |s(y) - t(y)| \leq \|s\| \|x - y\|_X + \|s - t\| \|y\|_X \\ &\leq \|s\| \delta + \delta (\|x\|_X + \delta) \leq \varepsilon \end{aligned}$$

$\square$

**Definition 2.44** (Weak topology).

- The topology induced on  $X$  by the family of maps  $X^*$  (see Def. 2.6) is called *weak topology* on  $X$ .
- It is the coarsest topology in which all maps  $t \in X^*$  are continuous.
- We denote convergence in the weak topology by  $x_n \rightharpoonup x$ .
- $[x_n \rightharpoonup x] \Leftrightarrow [t(x_n) \rightarrow t(x) \text{ for all } t \in X^*]$ .

**Definition 2.45** (Weak\* topology).

- For fixed  $x \in X$  consider the map  $f_x : X^* \rightarrow \mathbb{R}, t \mapsto \langle t, x \rangle$ .
- The weak\* topology on  $X^*$  is the topology induced by the family of maps  $\{f_x | x \in X\}$ .
- Weak\* convergence is denoted by  $t_n \xrightarrow{*} t$ .
- $[t_n \xrightarrow{*} t] \Leftrightarrow [\langle t_n, x \rangle \rightarrow \langle t, x \rangle \text{ for all } x \in X]$ .

**Example 2.46.** • By the Riesz representation theorem any bounded linear functional  $t$  on a Hilbert space  $H$  can be identified with a unique  $y \in H$  such that  $t(x) = \langle x, y \rangle_H$ . So  $H^*$  can be identified with  $H$  itself and the duality pairing is given by the inner product.

- The identification is not necessarily unique. Let  $H, J$  be Hilbert spaces with  $J \subset H$ , but  $J$  is equipped with different inner product. Then  $J^*$  can be identified with  $J$  via inner product  $\langle \cdot, \cdot \rangle_J$  or with subspace of  $H$  via inner product  $\langle \cdot, \cdot \rangle_H$ .

The following Proposition simplifies study of dual spaces on Banach spaces with Schauder bases.

**Proposition 2.47.** Let  $(z_n)_n$  be a Schauder basis on a Banach space  $X$ . Then any element of  $X^*$  can be identified with a unique real sequence  $(\lambda_n)_n$ .

*Proof.* • For  $x \in X$  let  $(\alpha_i)_i$  be the unique sequence such that  $x = \lim_{n \rightarrow \infty} x_n$  where  $x_n = \sum_{i=1}^n \alpha_i \cdot z_i$ .

- Let  $t \in X^*$ . By continuity of  $t$  find  $t(x) = \lim_{n \rightarrow \infty} t(x_n) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \alpha_i \cdot t(z_i)$ .
- Can represent  $t$  by sequence  $(\lambda_i = t(z_i))_i$ :  $t(x) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \alpha_i \cdot \lambda_i$ .
- Representation is unique: let  $t, \hat{t}$  be represented by two sequences  $(\lambda_n)_n, (\hat{\lambda}_n)_n$ . Assume  $\lambda_i \neq \hat{\lambda}_i$ . Then  $t(z_i) = \lambda_i \neq \hat{\lambda}_i = \hat{t}(z_i)$ , so  $t \neq \hat{t}$ .  $\square$

**Remark 2.48.** The above proposition does not specify which sequences  $(\lambda_n)_n$  represent some  $t \in X^*$ . Doing this helps to fully characterize  $X^*$ .

**Proposition 2.49.** For  $p \in [1, \infty)$  the dual space of  $\ell^p$  is isomorphic (exists bijection that preserves norm / metric) to  $\ell^q$  where  $q \in (1, \infty]$  such that  $\frac{1}{p} + \frac{1}{q} = 1$ .

*Proof.* • First treat  $p > 1$ . By Propositions 2.34 ( $\ell^p$  has Schauder basis  $(e_i)_i$ ) and 2.47 ( $\ell^p$ )\* can be identified with subset of real sequences.

- Let  $y \in \ell^q$ . Define  $t(x) = \lim_{n \rightarrow \infty} \sum_{i=1}^n x_i y_i$ . Hölder inequality: sequence converges absolutely, limit exists and is finite.  $\Rightarrow$  This defines linear functional.

- $|t(x)| \leq \|x\|_{\ell^p} \|y\|_{\ell^q} \Rightarrow$  functional is bounded. For every  $y \in \ell^q$  can find  $x \in \ell^p$  s.t.  $|t(x)| = \|x\|_{\ell^p} \|y\|_{\ell^q}$  (for construction of  $x_n$  see Prop. 2.12), so operator norm of  $t$  equals  $\ell^q$  norm of  $y$ . So can identify  $\ell^q$  with subset of  $(\ell^p)^*$ .
- Assume  $y \notin \ell^q$ . So  $\sum_{i=1}^n |y_i|^q$  is unbounded as  $n \rightarrow \infty$ .
- Let  $\hat{y}_n = (\hat{y}_{n,1}, \hat{y}_{n,2}, \dots)$  be real sequence where

$$\hat{y}_{n,i} = \begin{cases} y_i & \text{if } i \leq n, \\ 0 & \text{else.} \end{cases}$$

$\|\hat{y}_n\|_{\ell^q} < \infty$ ,  $(\hat{y}_n)_n$  is unbounded sequence in  $\ell^q$ .

- Let  $\hat{x}_n \in \ell^p$  such that  $\sum_{i=1}^n \hat{x}_{n,i} \hat{y}_{n,i} = \|\hat{x}_n\|_{\ell^p} \|\hat{y}_n\|_{\ell^q}$  with  $\hat{x}_{n,i} = 0$  for  $i > n$ .
- Consider  $\frac{1}{\|\hat{x}_n\|_{\ell^p}} \sum_{i=1}^n y_i \hat{x}_{n,i} = \frac{1}{\|\hat{x}_n\|_{\ell^p}} \sum_{i=1}^n \hat{y}_{n,i} \hat{x}_{n,i} = \|\hat{y}_n\|_{\ell^q} \rightarrow \infty$  as  $n \rightarrow \infty$ . So  $y$  does not represent a bounded linear functional on  $\ell^p$ .
- So can identify dual of  $\ell^p$  with  $\ell^q$  for  $p \in (1, \infty)$ .
- Now:  $p = 1, q = \infty$ . Let  $x \in \ell^1$ . If  $y \in \ell^\infty$  then  $|\sum_{i=1}^\infty x_i y_i| \leq \|x\|_{\ell^1} \|y\|_{\ell^\infty}$ . So  $y$  induces bounded, linear functional on  $\ell^1$  as above.
- Assume  $y \notin \ell^\infty$ . Then exists unbounded subsequence  $(y_{n_k})_k$ . Set  $\hat{x}_k = e_{n_k}$ , so  $\|\hat{x}_k\|_{\ell^p} = 1$ . Then  $|\sum_{i=1}^\infty y_i \hat{x}_{k,i}| = |y_{n_k}| \rightarrow \infty$ , therefore  $y$  does not represent bounded functional.  $\square$

**Remark 2.50.** Since  $\ell^\infty$  has no Schauder basis, cannot use this trick to study dual space of  $\ell^\infty$ . Will later see indirectly that it cannot be identified with  $\ell^1$ .

**Remark 2.51.** • Let  $Y$  be a subspace of a Banach space  $X$ .

- Any bounded linear functional on  $X$  is bounded linear functional on  $Y$ . So  $X^*$  is subset of  $Y^*$ .
- But  $Y^*$  may be strictly larger: there may be linear functionals on  $X$  that are bounded on  $Y$  but not on  $X$ .

**Example 2.52.**  $\ell^1$  is a (strict) subspace of  $\ell^2$  (why? careful: not true for ‘big  $L^p$ ’ spaces!) and  $\ell^2 = (\ell^2)^*$  is a strict subspace of  $\ell^\infty = (\ell^1)^*$ . (In the presence of a canonical isomorphism between two isomorphic spaces, we sometimes simply treat two isomorphic spaces as one. Here:  $(\ell^1)^* = \ell^\infty$ .)

**Definition 2.53.** (i) The dual of the dual of a Banach space is called *bidual* space.

(ii) When a Banach space can be identified with its bidual, a space is called *reflexive*.

**Remark 2.54.** Any  $x \in X$  defines a bounded linear functional on  $X^*$  via  $t \mapsto \langle t, x \rangle_{X^* \times X}$ , see Def. 2.42. So  $X$  can be identified with subspace of  $X^{**}$ . On a reflexive space, any bounded linear functional can be identified with some  $x \in X$ .

**Example 2.55.** (i) Hilbert spaces are reflexive.

(ii)  $\ell^p$  for  $p \in (1, \infty)$  are reflexive.

(iii) Will soon see:  $\ell^1$  is not reflexive, since  $(\ell^\infty)^* \neq \ell^1$ .

Now that we have introduced dual spaces and the weak and weak\* topologies, we can collect a few facts on corresponding compactness.

**Theorem 2.56** (Banach–Alaoglu). Let  $X$  be a normed space and  $X^*$  the induced topological dual space.

- (i) The closed unit ball of  $X^*$ ,  $\overline{B_{X^*}(0, 1)}$  is compact in the weak\* topology.
- (ii) If  $X$  is separable, then  $\overline{B_{X^*}(0, 1)}$  is sequentially weak\* compact.

We also quote a generalization of the Eberlein–Šmulian theorem.

**Theorem 2.57** (Eberlein–Šmulian). Compactness and sequential compactness are equivalent in the weak topology of a Banach space.

---

Comment: On Hilbert spaces  $H = H^*$ , weak and weak\* topology coincide. Therefore, we were able to use both Banach–Alaoglu and Eberlein–Šmulian in Section 1.

---

And a related result:

**Theorem 2.58.**  $X$  is reflexive if and only if  $\overline{B_X(0, 1)}$  is (sequentially) weakly compact.

With this result we can see that  $(\ell^\infty)^* \neq \ell^1$ .

**Corollary 2.59.**  $[l^1 \text{ is not reflexive}] \Leftrightarrow [(\ell^\infty)^* \neq \ell^1]$

*Proof.* • The equivalence in the statement follows from  $(\ell^1)^* = \ell^\infty$  (Prop. 2.49).

- By Theorem 2.58 it suffices to show that  $\overline{B_{\ell^1}(0, 1)}$  is not weakly compact.
- Sequence  $(e_n)_n$  of canonical Schauder basis vectors lies in  $\ell^1$ . Let  $(e_{n_k})_k$  be any subsequence.
- Let  $z \in \ell^\infty$  be given by

$$z_i = \begin{cases} (-1)^k & \text{if } i = n_k \text{ for some } k \in \mathbb{N}, \\ 0 & \text{else.} \end{cases}$$

- Then  $\langle z, e_{n_k} \rangle_{\ell^\infty \times \ell^1} = z_{n_k} = (-1)^k$ . So the sequence  $(\langle z, e_{n_k} \rangle_{\ell^\infty \times \ell^1})_k$  is not converging in  $\mathbb{R}$  and thus  $(e_{n_k})_k$  is not weakly converging.
- Since this holds for any subsequence of  $(e_n)_n$ , the sequence has no cluster point. Thus  $\overline{B_{\ell^1}(0, 1)}$  is not weakly sequentially compact, which by Thm. 2.57 implies that it is not weakly compact. □ ← VL13

## 2.4 Hahn–Banach theorem and convex duality on Banach spaces

---

Comment: Hahn–Banach theorem is fundamental in functional analysis. For us useful to prove existence of minimizers for problems formulated on a dual space.

---

**Theorem 2.60** (Hahn–Banach). Let  $X$  be real vector space. Let  $f : X \rightarrow \mathbb{R}$  be positively 1-homogeneous and sub-additive, i.e.

$$f(\alpha x) = \alpha f(x), \quad f(x + y) \leq f(x) + f(y) \quad \text{for all } x, y \in X, \alpha \in \mathbb{R}_+.$$

Let  $Y \subset X$  be a subspace and let  $t : Y = \text{dom}(t) \rightarrow \mathbb{R}$  be a linear functional on  $Y$  that is majorized on  $Y$  by  $f$ , i.e.  $t(x) \leq f(x)$  for  $x \in Y$ . Then, there is a linear extension  $T : X \rightarrow \mathbb{R}$  of  $t$  that is majorized by  $f$  on  $X$ , i.e.

$$T(x) = t(x) \quad \text{for } x \in Y \quad \text{and} \quad T(x) \leq f(x) \quad \text{for } x \in X.$$

---

Comment: The proof relies fundamentally on the axiom of choice, in form of Zorn’s Lemma.

---

A few applications:

**Proposition 2.61.** Let  $Y$  be a subspace of a normed space  $X$  and  $t \in L(Y, \mathbb{R})$ . Then there exists some  $T \in L(X, \mathbb{R})$ , such that  $T|_Y = t$  and  $\|T\| = \|t\|$ . Here  $T|_Y$  denotes the restriction of  $T$  to  $Y$ .

*Proof.* • Apply Hahn–Banach to  $t$  defined on  $Y$  with  $f(x) = \|t\| \|x\|$ .

- Get linear  $T : X \rightarrow \mathbb{R}$  with  $T|_Y = t$  and  $T(x) \leq f(x) = \|t\| \|x\|$ . So  $\|T\| = \|t\|$  and in particular  $T \in L(X, \mathbb{R})$ .  $\square$

---

Comment: Above proposition is ‘boring’ if we know how to project onto  $Y$ . Then set  $T = t \circ P_Y$ . But as we have seen, this projection does not always exist.

---

**Proposition 2.62.** Let  $X$  be normed space,  $x \in X$ . Then there exists some  $T \in X^* \setminus \{0\}$  such that  $T(x) = \|T\| \|x\|$ .

*Proof.* • Trivial if  $x = 0$ . So assume  $x \neq 0$ .

- Set subspace  $Y = \text{span}\{x\}$ ,  $t : Y \rightarrow \mathbb{R}$ ,  $\alpha \cdot x \mapsto \alpha \|x\|$ . Note:  $\|t\|_{Y^*} = 1$ .  $f(z) = \|z\|$ .
- Apply Hahn–Banach, get  $T : X \rightarrow \mathbb{R}$ .  $T(\alpha x) = t(\alpha x) = \alpha \|x\|$ ,  $T(z) \leq \|z\| \Rightarrow \|T\| = 1 \Rightarrow T(x) = \|x\| = \|T\| \|x\|$ .  $\square$

**Remark 2.63.** • Conversely, for fixed  $t \in X^*$  there is not always  $x \in X \setminus \{0\}$  such that  $t(x) = \|t\| \|x\|$ .

- Example:  $(y_i)_i \in \ell^\infty \sim (\ell^1)^*$  where  $y_i = (1 - 1/i)$ .  $\|y\|_{\ell^\infty} = \sup_i |y_i| = 1$ .
- Let  $x \in \ell^1$ ,  $x \neq 0$ . Then  $\|x\|_{\ell^1} = \sum_{i=1}^\infty |x_i| > \sum_{i=1}^\infty y_i x_i = y(x)$ .

Between a normed space and its topological dual we can now introduce notions analogous to orthogonality.

**Definition 2.64.** Let  $X$  be a normed space and  $X^*$  its topological dual space.

- (i)  $t \in X^*$  is called *aligned* with  $x \in X$  if  $t(x) = \langle t, x \rangle_{X^* \times X} = \|t\|_{X^*} \|x\|_X$ .

- (ii)  $x \in X, t \in X^*$  are *orthogonal* if  $t(x) = 0$ .
- (iii) The *orthogonal complement* of  $Y \subset X$  is  $Y^\perp = \{t \in X^* : t(x) = 0 \forall x \in Y\}$ .
- (iv) The *orthogonal complement* of  $Z \subset X^*$  is  ${}^\perp Z = \{x \in X : t(x) = 0 \forall t \in Z\}$ .

**Proposition 2.65.**  ${}^\perp[Y^\perp] = Y$  for any closed subspace  $Y \subset X$ .

---

Comment: Compare to polar cone, Def. 1.42 and Prop. 1.43.

---

*Proof.* •  ${}^\perp[Y^\perp] \supset Y$ :  $[x \in Y] \Rightarrow [t(x) = 0 \forall t \in Y^\perp] \Rightarrow [x \in {}^\perp[Y^\perp]]$ .

- ${}^\perp[Y^\perp] \subset Y$ : For  $y \notin Y$  define  $t \in L(V = Y \oplus \text{span}\{y\}, \mathbb{R})$  by  $t(x + \lambda y) = \lambda$  for every  $v = x + \lambda y \in V$  (the decomposition is unique). We get

$$\|t\| = \sup_{\substack{x \in Y, \lambda \in \mathbb{R}: \\ x + \lambda y \neq 0}} \frac{|t(x + \lambda y)|}{\|x + \lambda y\|} = \sup_{\substack{x \in Y \setminus \{0\}, \\ \lambda \in \mathbb{R} \setminus \{0\}}} \frac{|\lambda|}{\|x + \lambda y\|} = \sup_{x \in Y \setminus \{0\}} \frac{1}{\|x + y\|} < \infty$$

since  $\text{dist}(y, Y) > 0$  ( $Y$  is closed).

- Use Hahn–Banach via Prop. 2.62 to extend  $t$  to  $T \in L(X, \mathbb{R})$ .
- $T(x) = t(x) = 0$  for all  $x \in Y$ .  $\Rightarrow T \in Y^\perp$ .
- $T(y) = t(y) = 1 \neq 0$ . So  $y \notin {}^\perp[Y^\perp]$ . □

Now we very briefly generalize a few concepts of convex analysis from Hilbert spaces to Banach spaces.

**Definition 2.66** (Subdifferential). Let  $X$  be a normed space. For a function  $f : X \rightarrow \mathbb{R} \cup \{\infty\}$  the subdifferential of  $f$  at  $x \in X$  is given by

$$\partial f(x) = \{t \in X^* \mid f(y) \geq f(x) + \langle t, y - x \rangle_{X^* \times X} \text{ for all } y \in X\}.$$

**Definition 2.67** (Fenchel–Legendre conjugates). Let  $X$  be a normed space. For a proper function  $f : X \rightarrow \mathbb{R} \cup \{\infty\}$  the Fenchel–Legendre conjugate  $f^* : X^* \rightarrow \mathbb{R} \cup \{\infty\}$  is given by

$$f^*(t) = \sup_{x \in X} \langle t, x \rangle_{X^* \times X} - f(x).$$

The pre-conjugate of a proper function  $g : X^* \rightarrow \mathbb{R} \cup \{\infty\}$  is given by

$${}^*g(x) = \sup_{t \in X^*} \langle t, x \rangle_{X^* \times X} - g(t).$$

In complete analogy to Prop. 1.73 (pointwise suprema over families of convex, lsc functions remain convex, lsc.), Prop. 1.74 (basic properties of conjugate), Prop. 1.72 (Fenchel–Young) and Prop. 1.81 (‘extreme points’ of Fenchel–Young) we can show:

**Proposition 2.68.** For a normed space  $X$  let  $f : X \rightarrow \mathbb{R} \cup \{\infty\}$ ,  $g : X^* \rightarrow \mathbb{R} \cup \{\infty\}$  be proper. Then

- (i)  $f^*$  and  ${}^*g$  are convex, lsc.

(ii)  $f^*(t) + f(x) \geq \langle t, x \rangle$  and  $g(t) + {}^*g(x) \geq \langle t, x \rangle$  for all  $(t, x) \in X^* \times X$ .

(iii)  $[t \in \partial f(x)] \Leftrightarrow [f^*(t) + f(x) = \langle t, x \rangle]$ .

**Remark 2.69.** Since  $X^{**} \not\sim X$  in general, one has to be somewhat careful with statements about  ${}^*g$  and in particular  ${}^*(f^*)$ . The situation is a bit simpler on reflexive spaces.

We have already seen that the Hahn–Banach theorem can be invoked to imply existence of many particular elements of the dual space. We now give a geometric variant, that we can then use to prove an adaption of the Fenchel–Rockafellar theorem.

**Theorem 2.70** (Hahn–Banach: separation form). Let  $X$  be a normed space,  $C \subset X$  convex,  $\text{int } C \neq \emptyset$ ,  $x \notin \text{int } C$ . Then  $\exists t \in X^* \setminus \{0\}$  such that  $t(y - x) \geq 0$  for all  $y \in C$ .

For the proof we need the following auxiliary result.

**Proposition 2.71** (Minkowski functional). Let  $X$  be a normed space. Let  $C \subset X$  be convex,  $0 \in \text{int } C$ . The *Minkowski functional* of  $C$  is defined as

$$p_C : X \rightarrow \mathbb{R}, \quad x \mapsto \inf\{r \in \mathbb{R}_+ \mid x \in rC\}.$$

$p_C$  is nonnegative, positively 1-homogeneous, continuous and subadditive (this implies convexity).

*Proof.* •  $p_C$  is indeed real valued: since  $0 \in \text{int } C \exists \eta > 0$  such that  $\overline{B(0, \eta)} \subset C$  and thus  $\forall x \in X \setminus \{0\}$  get  $x \in \|x\| \overline{B(0, 1)} = \frac{\|x\|}{\eta} \overline{B(0, \eta)} \subset \frac{\|x\|}{\eta} C \Rightarrow p_C(x) \leq \frac{\|x\|}{\eta}$ .

- nonnegative and positively 1-homogeneous are immediate.
- subadditivity: assume  $x, y \neq 0$ , otherwise the result is trivial. Let  $x \in r \cdot C$ ,  $y \in s \cdot C$ , set  $a = x/r$ ,  $b = y/s$ , then  $a, b \in C$ . Then

$$x + y = r a + s b = (r + s) \left( \frac{r}{r+s} a + \frac{s}{r+s} b \right) \in (r + s) \cdot C$$

So for any  $x, y \in X$ :

$$\begin{aligned} p_C(x) + p_C(y) &= \inf \left\{ r + s \mid r, s \in \mathbb{R}_+, \underbrace{x \in rC, y \in sC}_{\Rightarrow x+y \in (r+s)C} \right\} \\ &\geq \inf \{r + s \in \mathbb{R}_+ \mid x + y \in (r + s)C\} = p_C(x + y) \end{aligned}$$

- continuity: for  $x \in X$ ,  $\varepsilon > 0$  set  $\delta = \eta \cdot \varepsilon$ . For  $y \in B(x, \delta)$  find

$$\begin{aligned} p_C(y) &= p_C(x + (y - x)) \leq p_C(x) + p_C(y - x) \leq p_C(x) + \frac{\delta}{\eta} \leq p_C(x) + \varepsilon, \\ p_C(x) &= p_C(y + (x - y)) \leq p_C(y) + p_C(x - y) \leq p_C(y) + \varepsilon \end{aligned}$$

So  $|p_C(x) - p_C(y)| \leq \varepsilon$ . □

*Proof of Theorem 2.70.* • W.l.o.g. assume  $0 \in \text{int } C$  (otherwise, simply translate).

- Define  $t : \text{span}\{x\} \rightarrow \mathbb{R}$ ,  $t(\lambda x) = \lambda$ .

- Set  $f(z) = p_{\text{int } C}(z)$ . Since  $x \notin \text{int } C \Rightarrow f(x) \geq 1$ . For  $\lambda \geq 0$ :  $t(\lambda x) = \lambda \leq \lambda f(x) = f(\lambda x)$ . For  $\lambda \leq 0$ :  $t(\lambda x) = \lambda \leq 0 \leq f(\lambda x)$ . Moreover,  $f(y) \leq 1$  if  $y \in \text{int } C$  and by continuity for  $y \in C$ .
- Now apply Hahn–Banach to  $t$ , majorized by  $f$ . Get  $T \in X^*$  ( $T$  bounded since  $T(z) \leq f(z)$  which is 1-homogeneous and continuous) with  $T(x) = t(x) = 1$  (so  $T \neq 0$ ),  $T(y) \leq f(y) \leq 1$  for  $y \in C$ . So  $T(y - x) = T(y) - T(x) \leq 0$ . Use  $-T$  to obtain sought-after functional.  $\square$

With this result we can now proof the analogue to the Fenchel–Rockafellar theorem, Prop. 1.135.

**Proposition 2.72.** Let  $X$  be a normed space. Let  $f, g : X \rightarrow \mathbb{R} \cup \{\infty\}$  be convex. Assume that there exists some  $x_0 \in X$  such that  $f(x_0) < \infty$ ,  $g(x_0) < \infty$  and  $f$  is continuous in  $x_0$ . Then

$$\inf_{x \in X} \{f(x) + g(x)\} = \max_{t \in X^*} \{-f^*(-t) - g^*(t)\}.$$

In particular, a maximizer for the dual problem exists.

*Proof.* The proof is completely equivalent to Prop. 1.135 except that we replace Prop. 1.136 by the above Hahn–Banach separation theorem.  $\square$

**Example 2.73.** • Recall subspace projection problem on Banach space. Let  $Y$  be closed subspace of normed space  $X$ ,  $x \in X \setminus Y$ .

$$\text{dist}(x, Y) = \inf_{y \in Y} (\|x - y\| + \iota_Y(y))$$

- Let  $f(y) = \|x - y\|$ ,  $g(y) = \iota_Y(y)$ . Brief calculation yields:

$$f^*(t) = \langle t, x \rangle + \iota_{\overline{B(0,1)}}(t), \quad g^*(t) = \iota_{Y^\perp}(t)$$

- By above duality we find:

$$\text{dist}(x, Y) = \max \left\{ \langle t, x \rangle \mid t \in \overline{B(0,1)} \cap Y^\perp \right\}$$

## 2.5 $\Gamma$ -convergence

**Remark 2.74** (Motivation). Gamma convergence is a notion of convergence for minimization problems and their solutions. We give a few intuitive examples where we ignore technical regularity considerations.

**Example 2.75** (Phase transitions). • Let  $\Omega \subset \mathbb{R}^n$  be a spatial domain in which we want to find the optimal configuration of two immiscible phases 0 and 1. We describe the domain of the two phases by a set  $A \subset \Omega$ , where  $x \in A$  indicates that  $x$  belongs to phase 1, otherwise phase 0.

- A function  $s : \Omega \rightarrow \mathbb{R}$  indicates the affinity of each point to the two phases ( $s(x)$  large  $\Leftrightarrow x$  prefers phase 0). The total volume of both phases is fixed. So  $|A| = \int_A dx = c$  for a constant  $c$ .
- In addition, there is surface tension between the two phases, which encourages the interface to have small length / surface area. The total surface tension energy is assumed to be proportional to the perimeter of  $A$ , denoted by  $\text{Per}(A)$ .
- The energetically best configuration will be given by a minimizer to the following optimization problem for sets  $A$ :

$$\min \left\{ \int_A s(x) dx + \text{Per}(A) \mid A \subset \Omega, |A| = c \right\}$$

- This is an optimization problem over sets. Could try to define sufficiently regular classes of sets for which  $\text{Per}$  is well defined, topology for sets, with notions of compactness etcetera.
- Alternative: try to approximate problem. Let  $u : \Omega \rightarrow [0, 1]$  indicate ‘concentration’ of phases.  $u(x) = 1 \Leftrightarrow$  phase 1 and vice versa. Ideally,  $u$  only takes values 0 and 1.
- But approximate perimeter term via gradient of  $u$ . So allow  $u$  to take values in  $[0, 1]$ .
- Let  $W(x)$  be a ‘double well’  $W(0) = W(1) = 0$ ,  $W(x) > 0$  for  $x \in (0, 1)$ .

---

**Sketch:**  $W$

---

Let  $\varepsilon > 0$ . Approximation for perimeter of ‘region’ described by  $u$  is then given by

$$\int_{\Omega} \left( \frac{W(u(x))}{\varepsilon} + \varepsilon |\nabla u(x)|^2 \right) dx.$$

- The full problem can then be approximated by

$$\min \left\{ \int_{\Omega} \left( s(x) u(x) + \frac{W(u(x))}{\varepsilon} + \varepsilon |\nabla u(x)|^2 \right) dx \mid u : \Omega \rightarrow [0, 1], \int_{\Omega} u(x) dx = c \right\}.$$

- Correspondences:  $s$ -term, volume constraint. As  $\varepsilon \rightarrow 0$  double well penalizes  $u(x) \notin \{0, 1\}$ , so wants to ‘jump’. But gradient term wants smooth transition. Transition between 0-phase and 1-phase will be determined by trade-off with optimal profile (for small  $\varepsilon$ ). For transition region: has ‘width’  $\mathcal{O}(\varepsilon)$ , ‘length’  $L = \text{Per}(A)$ , cost of  $W$ -term  $\mathcal{O}(L \cdot \varepsilon / \varepsilon = L)$ , gradient term:  $|\nabla u| = \mathcal{O}(1/\varepsilon)$ , so cost  $\mathcal{O}(L \cdot \varepsilon \cdot \varepsilon^{-2} \cdot \varepsilon = L)$ . Choose  $W$  careful, to get precisely  $L$  in limit.

**Example 2.76** (Homogenization). • Stationary heat equation on domain  $\Omega$  with spatially varying thermal conductivity  $a : \Omega \rightarrow \mathbb{R}_+$  and source  $f : \Omega \rightarrow \mathbb{R}$  and homogeneous boundary conditions. Stationary heat energy distribution given by solution to

$$\begin{aligned} -\operatorname{div}(a \cdot \nabla u) &= f && \text{on } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

- (Weak) solutions to this equation can be identified with minimizers to

$$\min \left\{ \int_{\Omega} \left( \frac{a}{2} |\nabla u|^2 - u f \right) dx \mid u : \Omega \rightarrow \mathbb{R}, u(x) = 0 \text{ for } x \in \partial\Omega \right\}.$$

(of course need to identify reasonable subspace for candidates  $u$ ).

- Now assume  $a$  contains ‘microscopic periodic’ structure at scale  $\varepsilon$ . Solution will probably also contain microscopic structure.

---

**Sketch:**  $a_\varepsilon$  with periodic cells with inhomogeneous content. Possibly breaks spatial symmetry.

---

- As we look at solution from far away, can no longer directly ‘see’ microscopic structure. Only see ‘average’ of  $u$  over areas of much larger scale than  $\varepsilon$ . Corresponds to sending  $\varepsilon \rightarrow 0$ .

---

**Sketch:** Oscillating  $u$ , local average.

---

Want to show: in limit get effective homogeneous equation

$$\begin{aligned} -\operatorname{div}(A \nabla u) &= f && \text{on } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

and optimization problem

$$\min \left\{ \int_{\Omega} \left( \frac{1}{2} \langle \nabla u, A \nabla u \rangle - u f \right) dx \mid u : \Omega \rightarrow \mathbb{R}, u(x) = 0 \text{ for } x \in \partial\Omega \right\}.$$

with matrix  $A$  (can now contain spatial symmetry breaking).

- How do we formalize transition between minimization problems?

**Example 2.77** (Dimension reduction). Consider again stationary heat equation on domain  $\Omega \times [0, \varepsilon]$ , i.e. a ‘thin film’. As  $\varepsilon \rightarrow 0$  the temperature in every ‘column’  $\{x\} \times [0, \varepsilon]$  for  $x \in \Omega$  will probably be approximately constant, compared to variations along the film (for suitable boundary conditions and assumptions on the source  $f$ ). So we want to approximate the  $n + 1$ -dimensional problem on  $\Omega \times [0, \varepsilon]$  by an  $n$ -dimensional problem on  $\Omega$ .

**Example 2.78** (Discretization). In theory often analyze infinite-dimensional optimization problems. Numerically, can only solve finite-dimensional problems. Need to ensure that our solutions to finite-dimensional approximations ‘converge’ to a solution to the underlying infinite-dimensional problem in a suitable sense, as we increase the dimension of the numerical approximations.

← VL14

**Remark 2.79.** Throughout this subsection let  $X$  be a topological space. We will focus on sequential notion of  $\Gamma$ -convergence. As before, it is possible to introduce more general topological definitions.

**Definition 2.80** ((Sequential)  $\Gamma$ -convergence). A sequence  $f_k : X \rightarrow \mathbb{R} \cup \{\infty\}$ ,  $k \in \mathbb{N}$  is said to (sequentially)  $\Gamma$ -converge to  $f : X \rightarrow \mathbb{R} \cup \{\infty\}$  if for all  $x \in X$  we have

- (i) (lim inf inequality) for every sequence  $(x_k)_k$  converging to  $x$

$$f(x) \leq \liminf_{k \rightarrow \infty} f_k(x_k);$$

- (ii) (lim sup inequality) there exists a sequence  $(x_k)_k$  converging to  $x$  such that

$$f(x) \geq \limsup_{k \rightarrow \infty} f_k(x_k).$$

The function  $f$  is called the  $\Gamma$ -limit of  $(f_k)_k$  and we write  $f = \Gamma\text{-}\lim_k f_k$ .

**Remark 2.81** (Recovery sequence). Let  $(x_k)_k$  satisfy the lim sup inequality. It must also satisfy the lim inf inequality and therefore

$$f(x) \geq \limsup_{k \rightarrow \infty} f_k(x_k) \geq \liminf_{k \rightarrow \infty} f_k(x_k) \geq f(x).$$

So  $\lim_{k \rightarrow \infty} f_k(x_k) = f(x)$  and  $(x_k)_k$  is referred to as *recovery sequence*.

**Remark 2.82** (Motivation of definition). The main motivation of  $\Gamma$ -convergence is to study the ‘limits’ of sequences of minimization problems. The sequence may be thought of as approximations to the limit problem. The two conditions serve two purposes:

- (i) (lim inf inequality): ‘the approximations add no new minimizers’: whenever  $(x_k)_k$  is a converging sequence of minimizers of  $(f_k)_k$ ,  $x_k \rightarrow x$ , then  $f(x) \leq \liminf_k f_k(x_k)$ . So the limit problem is potentially ‘even better’.
- (ii) (lim sup inequality): ‘the approximations do not remove minimizers’: for any  $x$  we can find a recovery sequence  $(x_k)_k$  such that  $f$  at  $x$  can be approximated by  $(f_k(x_k))_k$ . Note: not every minimizer of  $f$  can be written as limit of minimizers of  $(f_k)_k$ , but as limit of ‘almost minimizers’.

---

Comment: As seen in intro: sometimes limit is approximation of sequence, e.g. in homogenization.

---

**Remark 2.83.** In the above examples it is not always clear, whether the sequence of problems and the limit are indeed defined on the same space. Usually must find one big ‘summary’ space  $X$  that contains all subproblems.

- (i) phase transitions: limit problem on sets  $A \subset \Omega$ , sequence on differentiable functions  $u : \Omega \rightarrow [0, 1]$ . Possible solution: reasonable extension of function space, to handle both limit and sequences.
- (ii) dimension reduction: limit problem on functions  $u : \Omega \rightarrow \mathbb{R}$ , sequence on functions  $u : \Omega \times [0, \varepsilon] \rightarrow \mathbb{R}$ . Rescale sequence problems to functions  $u : \Omega \times [0, 1]$  (remove  $\varepsilon$  from space), limit problem on functions  $u : \Omega \rightarrow [0, 1] \rightarrow \mathbb{R}$ , add constraint  $u(x, \cdot) = \text{const}(x)$  for every  $x \in \Omega$ .

**Definition 2.84** ((Sequential)  $\Gamma$ -limits). For a sequence of functions  $(f_k)_k$ ,  $f_k : X \rightarrow \mathbb{R} \cup \{\infty\}$  and some  $x \in X$  the quantities

$$\begin{aligned}(\Gamma\text{-}\liminf_k f_k)(x) &= \inf\{\liminf_k f_k(x_k) \mid (x_k)_k : x_k \rightarrow x\}, \\(\Gamma\text{-}\limsup_k f_k)(x) &= \inf\{\limsup_k f_k(x_k) \mid (x_k)_k : x_k \rightarrow x\}\end{aligned}$$

are called sequential  $\Gamma$ -lower and upper limit of  $(f_k)_k$  at  $x$ , where the infima are over all sequences  $(x_k)_k$  converging to  $x$ . If both limits coincide, we call  $(\Gamma\text{-}\liminf_k f_k)(x) = (\Gamma\text{-}\limsup_k f_k)(x) = (\Gamma\text{-}\lim_k f_k)(x)$  the sequential  $\Gamma$ -limit of  $(f_k)_k$  at  $x$ .

**Proposition 2.85.** If the sequence  $(f_k)_k$   $\Gamma$ -converges to  $f$  the  $\Gamma$ -limit of  $(f_k)_k$  exists for all  $x \in X$  and equals  $f(x)$ .

*Proof.* • For all sequences  $(x_k)_k$  with limit  $x$  have by lim inf condition  $f(x) \leq \liminf_k f_k(x_k)$ . Therefore, this condition holds after taking infimum over all such sequences and therefore get  $f(x) \leq (\Gamma\text{-}\liminf_k f_k)(x)$ .

• Let  $(x_k)_k$  be recovery sequence at  $x$ . Then

$$f(x) \geq \limsup_k f_k(x_k) \geq (\Gamma\text{-}\limsup_k f_k)(x) \geq (\Gamma\text{-}\liminf_k f_k)(x) \geq f(x).$$

Therefore, both limits must coincide. □

**Remark 2.86.** This implies that  $\Gamma$ -limit  $f$  of sequence  $(f_k)_k$  is unique (if it exists) since  $\Gamma$ -lower and upper limit do not depend on  $f$ . Can conversely show that if  $\Gamma$ -lower and upper limit coincide at all points  $x$  then their value determines the  $\Gamma$ -limit function  $f$ .

**Proposition 2.87** (Stability under continuous perturbations). Let  $(f_k)_k$  be a sequence of functions with  $\Gamma$ -limit  $f$ . Let  $g$  be a (sequentially) continuous function. Then  $f + g = \Gamma\text{-}\lim_k f_k + g$ .

*Proof.* • Let  $(x_k)_k$  be a sequence converging to  $x$ . Then

$$\liminf_k (f_k(x_k) + g(x_k)) = \liminf_k f_k(x_k) + \lim_k g(x_k) \geq f(x) + g(x).$$

• Let  $(x_k)_k$  be a recovery sequence for  $x$ . Then

$$\limsup_k (f_k(x_k) + g(x_k)) = \lim_k f_k(x_k) + \lim_k g(x_k) \leq f(x) + g(x).$$

---

Comment: lim inf argument would work if  $g$  is lsc, lim sup argument does not. □

---

Some examples.

**Example 2.88.** • Set

$$g(t) = \begin{cases} +1 & \text{if } t = 1, \\ -1 & \text{if } t = -1, \\ 0 & \text{else.} \end{cases}$$

- Set  $f_k(t) = g(kt)$  and

$$f(t) = \begin{cases} 0 & \text{if } t \neq 0, \\ -1 & \text{if } t = 0. \end{cases}$$

- Proof that  $f$  is  $\Gamma$ -limit of  $f_k$ : let  $x \neq 0$ ,  $x_k \rightarrow x$ .  $\exists N \in \mathbb{N}$  s.t.  $|x_k| > 1/k$  for all  $k \geq N$ . So  $f_k(x_k) = g(x_k k) = 0 = f(x)$ . So any sequence will satisfy  $\liminf$  and  $\limsup$  condition.
- Let  $x = 0$ . Since  $f(x) = -1 \leq f_k(t)$  for all  $k$  and  $t$ , the  $\liminf$  condition is clear. As recovery sequence set  $x_k = -1/k$ . Then  $f_k(x_k) = g(-1) = -1 = f(x)$ .
- Note:  $\lim_{k \rightarrow \infty} f_k(x) = 0$  for all  $x$ . So the  $\Gamma$ -limit does not necessarily coincide with the pointwise limit.

**Example 2.89** (Constant sequence). • Let

$$f_k(x) = f(x) = \begin{cases} +1 & \text{if } x \leq 0, \\ 0 & \text{if } x > 0. \end{cases}$$

- Since  $f$  is not lower semicontinuous,  $f$  is not the  $\Gamma$ -limit of  $(f_k)_k$ : Let  $x_k = 1/k$ . Then  $f_k(x_k) = 0$  but  $f(0) = 1$ .
- A function  $f$  is the  $\Gamma$ -limit of the constant sequence  $(f)_k$  if and only if  $f$  is (sequentially) lower semicontinuous.

As illustration we prove a slightly more general result.

**Proposition 2.90.** Let  $(X, d)$  be a metric space. If  $f : X \rightarrow \mathbb{R} \cup \{\infty\}$  is a  $\Gamma$ -limit of some sequence  $(f_k)_k$  then  $f$  is sequentially lower semicontinuous.

*Proof.* • Let  $(x_k)_k$  be a sequence in  $X$  with limit  $x \in X$ .

- For fixed  $k$  let  $(x_{k,j})_j$  be recovery sequence in  $X$  with limit  $x_k$ . By the  $\limsup$  condition we obtain

$$f(x_k) = \lim_{j \rightarrow \infty} f_j(x_{k,j}).$$

- In particular there is some  $N_k$  such that  $f(x_k) \geq f_j(x_{k,j}) - \frac{1}{k}$  and  $d(x_{k,j}, x_k) < 1/k$  for  $j > N_k$ .
- Let  $(j_k)_k$  be an increasing sequence of indices with  $j_k > N_k$  for all  $k$ . Let

$$z_j = \begin{cases} x_{k,j_k} & \text{if } j = j_k \text{ for some } k, \\ x & \text{else.} \end{cases}$$

- Then  $d(x, z_j) \leq d(x, x_k) + d(x_k, x_{k,j_k}) \leq d(x, x_k) + 1/k$  if  $j = j_k$  for some  $k$  and  $d(x, z_j) = 0$  else. Therefore  $z_j \rightarrow x$ .
- Now:

$$\liminf_k f(x_k) \geq \liminf_k f_{j_k}(x_{k,j_k}) - \frac{1}{k} = \liminf_k f_{j_k}(x_{k,j_k}) \geq \liminf_j f_j(z_j) \geq f(x). \quad \square$$

Now give simple prototypical results that show how sequence of minimizers of  $(f_k)_k$  is related to minimizers of  $\Gamma$ -limit  $f$ .

**Proposition 2.91.** Let  $(f_k)_k$  be a sequence of functions  $X \rightarrow \mathbb{R} \cup \{\infty\}$  as well as  $f : X \rightarrow \mathbb{R} \cup \{\infty\}$ .

- (i) If  $(f_k)_k$  and  $f$  satisfy the lim inf inequality, Def. 2.80(i), and  $K$  is a sequentially compact set then

$$\inf_K f \leq \liminf_{k \rightarrow \infty} \inf_K f_k.$$

- (ii) If  $(f_k)_k$  and  $f$  satisfy the lim sup inequality, Def. 2.80(ii), and  $U$  is an open set then

$$\inf_U f \geq \limsup_{k \rightarrow \infty} \inf_U f_k.$$

*Proof.* • **(i):** Let  $(\tilde{x}_k)_k$  be a sequence in  $K$  such that  $\liminf_k \inf_K f_k = \liminf_k f_k(\tilde{x}_k)$  and let  $(\tilde{x}_{k_j})_j$  be a convergent subsequence (exists due to sequential compactness of  $K$ ) such that  $\liminf_k f_k(\tilde{x}_k) = \lim_j f_{k_j}(\tilde{x}_{k_j})$  with limit  $\tilde{x}_{k_j} \rightarrow \bar{x} \in K$ .

- Set

$$x_k = \begin{cases} \tilde{x}_{k_j} & \text{if } k = k_j, \\ \bar{x} & \text{if } k \neq k_j \forall j. \end{cases}$$

Then  $x_k \rightarrow \bar{x}$ .

- With the lim inf condition we get

$$\inf_K f \leq f(\bar{x}) \leq \liminf_k f_k(x_k) \leq \liminf_j f_{k_j}(\tilde{x}_{k_j}) = \lim_j f_{k_j}(\tilde{x}_{k_j}) = \liminf_k \inf_K f_k.$$

- **(ii):** Fix  $\delta > 0$ . Find  $x \in U$  such that  $f(x) \leq \inf_U f + \delta$ . Let  $(x_k)_k$  be a recovery sequence for  $x$ . Since  $U$  open, have eventually  $x_k \in U$  for  $k$  sufficiently large. Then

$$\inf_U f + \delta \geq f(x) \geq \limsup_k f_k(x_k) \geq \limsup_k \inf_U f_k.$$

- Result follows since true for all  $\delta > 0$ . □

Define (sequentially) compact notion of coerciveness. Recall: Def. 1.103 defined coerciveness via boundedness.

**Definition 2.92** (Variants of coerciveness). • A function  $f : X \rightarrow \mathbb{R} \cup \{\infty\}$  is sequentially coercive if for every  $r \in \mathbb{R}$  the sublevel set  $S_r(f)$  is sequentially precompact.

- A function  $f$  is mildly sequentially coercive if there exists a non-empty sequentially compact set  $K \subset X$  such that  $\inf_X f = \inf_K f$ .
- A family of functions  $(f_i)_{i \in I}$  is equi-mildly sequentially coercive if there exists a non-empty sequentially compact set  $K \subset X$  such that  $\inf_X f_i = \inf_K f_i$  for all  $i \in I$ .

**Proposition 2.93** (Convergence of minimizers). Let  $(f_k)_k, f_k : X \rightarrow \mathbb{R} \cup \{\infty\}$  be a sequence of equi-mildly sequentially coercive functions and let  $f = \Gamma\text{-}\lim_k f_k$ . Then

$$\min_X f = \min_K f = \lim_k \inf_X f_k.$$

Moreover, if  $(x_k)_k$  is a precompact sequence such that  $\lim_k f_k(x_k) = \lim_k \inf_X f_k$  then every cluster point of  $(x_k)_k$  is a minimizer of  $f$ .

*Proof.* • Let  $K$  be the sequentially compact set on which  $\inf_K f_k = \inf_X f_k$  for all  $k$ .

- Apply Prop. 2.91(i) to  $K$  and (ii) to  $U = X$  to obtain:

$$\inf_X f \leq \inf_K f \leq \lim_k \inf_K \inf_K f_k = \lim_k \inf_k \inf_X f_k \leq \lim_k \sup_k \inf_X f_k \leq \inf_X f$$

- So  $\inf_X f = \lim_k \inf_X f_k$ . Since  $\inf_X f = \inf_K f$ , by sequential compactness of  $K$ , a minimizer for  $f$  exists.
- Let  $(x_{k_j})_j$  be convergent subsequence of  $(x_k)_k$  with limit  $\bar{x}$ . ‘Replace’ all indices  $k \neq k_j$  for all  $j$  by  $\bar{x}$ , as in proof of Prop. 2.91. Call this sequence  $(\tilde{x}_k)_k$ . Get via  $\lim \inf$  condition:

$$\inf_X f \leq f(\bar{x}) \leq \lim_k \inf_k f_k(\tilde{x}_k) \leq \lim_j \inf_j f_{k_j}(x_{k_j}) = \lim_k f_k(x_k) = \lim_k \inf_k \inf_X f_k = \inf_X f$$

So  $\bar{x}$  is minimizer.

- Now let  $(x_k)_k$  be sequence in  $K$  such that  $\lim_k f_k(x_k) = \lim_k \inf_X f_k$ , for instance choose  $x_k$  such that  $f_k(x_k) \leq \inf_X f_k + \frac{1}{k}$ . This is possible since  $(f_k)_k$  is equi-mildly sequentially coercive. Then  $(x_k)_k$  has cluster point, which must be minimizer of  $f$ , thus  $\inf_X f = \inf_K f$ .

← VL15

□

### 3 Optimal transport

---

Comment: References:

Villani: Topics in Optimal Transportation, 2003,

Villani: Optimal Transport: Old and New, 2009

---

**Remark 3.1.** Throughout this section  $(\Omega, d)$  is a compact metric space. Most result extend to non-compact spaces (with appropriate modifications) but the arguments become more technical.

#### 3.1 Reminders on measure theory

---

Comment: Reference: Ambrosio, Fusco, Pallara: Functions of Bounded Variation and Free Discontinuity Problems, Chapters 1 & 2.

---

**Definition 3.2** ( $\sigma$ -algebra). A collection  $\mathcal{E} \subset 2^X$  of subsets of a set  $X$  is called  $\sigma$ -algebra if

- (i)  $\emptyset \in \mathcal{E}$ ;  $[A \in \mathcal{E}] \Rightarrow [X \setminus A \in \mathcal{E}]$ ;
- (ii) (closed under countable unions) for a sequence  $A_n \in \mathcal{E} \Rightarrow \bigcup_{n=0}^{\infty} A_n \in \mathcal{E}$ .

---

Comment: Closed under finite unions, intersections and countable intersections.  $A \cap B = X \setminus ((X \setminus A) \cup (X \setminus B))$ .

---

Comment: Elements of  $\mathcal{E}$ : ‘measurable sets’. Pair  $(X, \mathcal{E})$ : ‘measure space’.

---

**Example 3.3.** Borel algebra: smallest  $\sigma$  algebra containing all open sets of a topological space.

Comment: Intersection of two  $\sigma$ -algebras is again  $\sigma$ -algebra. ‘smallest’ is well-defined.

---

**Definition 3.4** (Positive measure and vector measure). For measure space  $(X, \mathcal{E})$  a function  $\mu : \mathcal{E} \mapsto [0, +\infty]$  is called ‘positive measure’ if

- (i)  $\mu(\emptyset) = 0$ ;
- (ii) for pairwise disjoint sequence  $A_n \in \mathcal{E} \Rightarrow \mu(\bigcup_{n=0}^{\infty} A_n) = \sum_{n=0}^{\infty} \mu(A_n)$

For measure space  $(X, \mathcal{E})$  and  $\mathbb{R}^m$ ,  $m \geq 1$ , a function  $\mu : \mathcal{E} \mapsto \mathbb{R}^m$  is called ‘measure’ if  $\mu$  satisfies (i) and (ii) with absolute convergence.

---

Comment: Measures are vector space, measures are finite, positive measures may be infinite.

---

**Example 3.5.** Examples for measures:

- (i) counting measure:  $\#(A) = |A|$  if  $A$  finite,  $+\infty$  else.
- (ii) Dirac measure:  $\delta_x(A) = 1$  if  $x \in A$ , 0 else.
- (iii) Lebesgue measure  $\mathcal{L}([a, b]) = b - a$  for  $b \geq a$ .
- (iv) Scaled measures: positive measure  $\mu$ , function  $f \in L^1(\mu)$ , new measure  $\nu = f \cdot \mu$ .  $\nu(A) := \int_A f(x) d\mu(x)$ .

**Definition 3.6** (Total variation). For finite measure  $\mu$  on  $(X, \mathcal{E})$  the total variation  $|\mu|$  of  $A \in \mathcal{E}$  is

$$|\mu|(A) = \sup \left\{ \sum_{n=0}^{\infty} |\mu(A_n)| \mid A_n \in \mathcal{E}, \text{ pairwise disjoint, } \bigcup_{n=0}^{\infty} A_n = A \right\}.$$

$|\mu|$  is finite, positive measure on  $(X, \mathcal{E})$ .

**Definition 3.7** (Negligible sets). A set  $N \subset X$  is  $\mu$ -negligible if  $\exists A \in \mathcal{E}$  with  $N \subset A$  and  $\mu(A) = 0$ . Two functions  $f, g : X \rightarrow Y$  are identical ‘ $\mu$ -almost everywhere’ when  $\{x \in X | f(x) \neq g(x)\}$  is  $\mu$ -negligible.

**Example 3.8.** Null sets are Lebesgue-negligible sets.

**Definition 3.9** (Measurable functions, push-forward). Let  $(X, \mathcal{E}), (Y, \mathcal{F})$  be measurable spaces. A function  $f : X \rightarrow Y$  is ‘measurable’ if  $f^{-1}(A) \in \mathcal{E}$  for  $A \in \mathcal{F}$ .

For measure  $\mu$  on  $(X, \mathcal{E})$  the ‘push-forward’ of  $\mu$  under  $f$  to  $(Y, \mathcal{F})$ , we write  $f_{\#}\mu$ , is defined by  $f_{\#}\mu(A) = \mu(f^{-1}(A))$  for  $A \in \mathcal{F}$ .

Change of variables formula:

$$\int_X g(f(x)) d\mu(x) = \int_Y g(y) df_{\#}\mu(y)$$

---

**Sketch:** Varying densities.

---

**Example 3.10** (Marginal). Let  $\text{proj}_i : X \times X \rightarrow X$ ,  $\text{proj}_i(x_0, x_1) = x_i$ . Marginals of measure  $\gamma$  on  $X \times X$ :

$$\text{proj}_{0\#}\gamma(A) = \gamma(A \times X), \quad \text{proj}_{1\#}\gamma(A) = \gamma(X \times A).$$

---

**Sketch:** Discuss pre-images of  $\text{proj}_i$ .

---

**Definition 3.11** (Absolute continuity, singularity). Let  $\mu$  be positive measure,  $\nu$  measure on measurable space  $(X, \mathcal{E})$ .  $\nu$  is ‘absolutely continuous’ w.r.t.  $\mu$ , we write  $\nu \ll \mu$ , if  $[\mu(A) = 0] \Rightarrow [\nu(A) = 0]$ .

**Sketch:** Density  $\ll$  Lebesgue, density  $\not\ll$  density when support different, Dirac measures  $\not\ll$  Lebesgue, mixed measures  $\not\ll$  density, mixed measures  $\ll$  mixed measures when Diracs coincide.

Positive measures  $\mu, \nu$  are ‘mutually singular’, we write  $\mu \perp \nu$ , if  $\exists A \in \mathcal{E}$  such that  $\mu(A) = 0$ ,  $\nu(X \setminus A) = 0$ . For general measures replace  $\mu, \nu$  by  $|\mu|, |\nu|$ .

**Definition 3.12** ( $\sigma$ -finite). A positive measure  $\mu$  is called  $\sigma$ -finite if  $X = \bigcup_{n=0}^{\infty} A_n$  for sequence  $A_n \in \mathcal{E}$  with  $\mu(A_n) < +\infty$ .

**Example 3.13.** Lebesgue measure is not finite but  $\sigma$ -finite.

**Theorem 3.14** (Radon–Nikodym, Lebesgue decomposition [Ambrosio et al., Theorem 1.28]). Let  $\mu$  be  $\sigma$ -finite positive measure.  $\nu$  general measure.

Radon–Nikodym: For  $\nu \ll \mu$  there is a function  $f \in L^1(\mu)$  such that  $\nu = f \cdot \mu$ .  $f$  is unique  $\mu$ -almost everywhere. It is called ‘density of  $\nu$  with respect to  $\mu$ ’ and usually denoted by  $f = \frac{d\nu}{d\mu}$ .

Lebesgue decomposition: there exist unique measures  $\nu_a, \nu_s$  such that

$$\nu = \nu_a + \nu_s, \quad \nu_a \ll \mu, \quad \nu_s \perp \mu.$$

Note:  $\nu_a = f \cdot \mu$  for some  $f \in L^1(\mu)$ .

**Corollary 3.15.** A real-valued measure  $\nu$  can be decomposed into  $\nu = \nu_+ - \nu_-$  with  $\nu_+, \nu_-$  mutually singular positive measures.

*Proof.* Since  $\nu \ll |\nu|$  there exists  $f \in L^1(|\nu|)$  with  $\nu = f \cdot |\nu|$ . Set  $A_+ = f^{-1}((0, +\infty))$ ,  $A_- = f^{-1}((-\infty, 0))$  and set  $\nu_{\pm}(B) = |\nu|(B \cap A_{\pm})$ .  $\square$

---

Comment:  $f$  is only unique  $|\nu|$ -almost everywhere.

---

**Definition 3.16** (Support of measure). Let  $(\Omega, d)$  be a compact metric space with its Borel  $\sigma$ -algebra and  $\mu \in \mathcal{M}_+(\Omega)$ . The support of  $\mu$ , denoted  $\text{spt } \mu$  is the smallest closed set  $A \subset \Omega$  such that  $\mu(A) = \mu(\Omega)$ . For  $x \in \text{spt } \mu$  one has  $\mu(B_r(x)) > 0$  for any  $r > 0$ .

**Definition 3.17** (Radon measures). Let  $(\Omega, d)$  be compact metric space, let  $\mathcal{E}$  be Borel- $\sigma$ -algebra. A finite measure (positive or vector valued) is called a ‘Radon measure’. Write:

- $\mathcal{M}_+(\Omega)$ : positive Radon measures,
- $\mathcal{P}(\Omega) \subset \mathcal{M}_+(\Omega)$ : Radon probability measures (total mass = 1),
- $\mathcal{M}(\Omega)^m$ : (vector valued) Radon measures.

**Theorem 3.18** (Regularity [Ambrosio et al., Proposition 1.43]). For positive Radon measures on  $(\Omega, \mathcal{E})$  one has for  $A \in \mathcal{E}$

$$\mu(A) = \sup \{ \mu(B) \mid B \in \mathcal{E}, B \subset A, B \text{ compact} \} = \inf \{ \mu(B) \mid B \in \mathcal{E}, A \subset B, B \text{ open} \} .$$

**Theorem 3.19** (Duality [Ambrosio et al., Theorem 1.54]). Let  $(\Omega, d)$  be compact metric space. Let  $C(\Omega)^m$  be space of continuous functions from  $\Omega$  to  $\mathbb{R}^m$ , equipped with sup-norm. The topological dual of  $C(\Omega)^m$  can be identified with the space  $\mathcal{M}(\Omega)^m$  equipped with the total variation norm  $\|\mu\|_{\mathcal{M}} := |\mu|(\Omega)$ . Duality pairing for  $\mu \in \mathcal{M}(\Omega)^m$ ,  $f \in C(\Omega)^m$ :

$$\mu(f) = \langle \mu, f \rangle_{\mathcal{M} \times C} = \int_{\Omega} f(x) d\mu(x)$$

---

Comment: Notation:  $C(\Omega) \equiv C^0(\Omega)$  from Def. 2.8 (examples for Banach spaces).

---

**Corollary 3.20.** Two measures  $\mu, \nu \in \mathcal{M}(\Omega)^m$  with  $\mu(f) = \nu(f)$  for all  $f \in C(\Omega)^m$  coincide.

**Remark 3.21.** • By Theorem 2.56(i) (Banach–Alaoglu)  $\overline{B_{\mathcal{M}}(0, 1)}$  is weak\* compact.

- Analogous to Prop. 2.31 ( $C([0, 1])$  is separable) can show that  $C(\Omega)$  is separable. So by Theorem 2.56(ii) (‘sequential’ Banach–Alaoglu)  $\overline{B_{\mathcal{M}}(0, 1)}$  is sequentially weak\* compact.

### 3.2 Monge formulation of optimal transport

---

Comment: Gaspard Monge: French mathematician and engineer, 18<sup>th</sup> century. Studied problem of optimal allocation of resources to minimize transport cost.

---

**Sketch:** Bakeries and cafes

---

**Example 3.22** (According to Villani). Every morning in Paris bread must be transported from bakeries to cafes for consumption. Every bakery produces prescribed amount of bread, every cafe orders prescribed amount. Assume: total amounts identical. Look for most economical way to distribute bread.

Mathematical model:

- $\Omega \subset \mathbb{R}^2$ : area of Paris

- $\mu \in \mathcal{P}(\Omega)$ : distribution of bakeries and produced amount of bread,
- $\nu \in \mathcal{P}(\Omega)$ : distribution of cafes and consumed amount of bread
- Cost function  $c : \Omega \times \Omega \rightarrow \mathbb{R}_+$ .  $c(x, y)$  gives cost of transporting 1 unit of bread from bakery at  $x$  to cafe at  $y$ .
- Describe transport by map  $T : \Omega \rightarrow \Omega$ . Bakery at  $x$  will deliver bread to cafe at  $T(x)$ . Consistency condition:  $T_{\#}\mu = \nu$ .

---

Comment: Each cafe receives precisely ordered amount of bread.

---

- Total cost of transport map

$$C_M(T) = \int_{\Omega} c(x, T(x)) d\mu(x)$$

---

Comment: For bakery at location  $x$  pay  $c(x, T(x)) \cdot \mu(x)$ . Sum (i.e. integrate) over all bakeries.

---

**Definition 3.23.** Monge optimal transport problem: find  $T$  that minimizes  $C_M$ .

Problems:

- Do maps  $T$  with  $T_{\#}\mu = \nu$  exist? Can not split mass.

---

**Sketch:** Splitting of mass.

---

- Does minimal  $T$  exist? Non-linear, non-convex constraint and objective.

---

Comment:  $\Rightarrow$  problem remained unsolved for long time.

---

### 3.3 Kantorovich formulation of optimal transport

---

Comment: Leonid Kantorovich: Russian mathematician, 20<sup>th</sup> century. Founding father of linear programming, proposed modern formulation of optimal transport. (Nobel prize in economics 1975.)

---

Do not describe transport by map  $T$ , but by positive measure  $\pi \in \mathcal{M}_+(\Omega \times \Omega)$ .

**Definition 3.24** (Coupling / Transport Plan). Let  $\mu, \nu \in \mathcal{P}(\Omega)$ . Set of ‘couplings’ or ‘transport plans’  $\Pi(\mu, \nu)$  is given by

$$\Pi(\mu, \nu) = \left\{ \pi \in \mathcal{P}(\Omega \times \Omega) \mid \text{proj}_0 \pi = \mu, \text{proj}_1 \pi = \nu \right\}.$$

**Example 3.25.**  $\Pi(\mu, \nu) \neq \emptyset$ , contains at least product measure  $\mu \otimes \nu \in \Pi(\mu, \nu)$ .  $(\mu \otimes \nu)(A \times B) = \mu(A) \cdot \nu(B)$  for measurable  $A, B \subset \Omega$ .

**Definition 3.26.** For compact metric space  $(\Omega, d)$ ,  $\mu, \nu \in \mathcal{P}(\Omega)$ ,  $c \in C(\Omega \times \Omega)$  the Kantorovich optimal transport problem is given by

$$\mathcal{C}(\mu, \nu) = \inf \left\{ \int_{\Omega \times \Omega} c(x, y) d\pi(x, y) \mid \pi \in \Pi(\mu, \nu) \right\}$$

---

Comment: Linear (continuous) objective, affine constraint set.

---

Comment: Language of measures covers finite dimensional and infinite dimensional case.

---

**Proposition 3.27.** Minimizers of Kantorovich problem (Def. 3.26) exist.

For proof use following result.

**Proposition 3.28.** The set  $\Pi(\mu, \nu)$  is weak\* sequentially closed.

*Proof.* • Let  $(\pi_n)_n$  be sequence in  $\Pi(\mu, \nu)$ , with  $\pi_n \xrightarrow{*} \pi \in \mathcal{M}(\Omega \times \Omega)$ .

- Positivity:  $\pi$  is a positive measure. Otherwise find function  $\phi \in C(\Omega \times \Omega)$ ,  $\phi \geq 0$ , with  $\int_{\Omega \times \Omega} \phi d\pi < 0$  (use Cor. 3.15 and Thm. 3.18 for construction) which contradicts weak\* convergence since  $\int \phi d\pi_n \geq 0 \forall n$ .
- Unit mass:  $\pi(\Omega \times \Omega) = \int_{\Omega \times \Omega} 1 d\pi = \lim_{n \rightarrow \infty} \int_{\Omega \times \Omega} 1 d\pi_n = \lim_{n \rightarrow \infty} \pi_n(\Omega \times \Omega) = 1$ .
- Marginal constraint: For every  $\phi \in C(\Omega)$

$$\begin{aligned} \int_{\Omega} \phi d\text{proj}_0 \# \pi &= \int_{\Omega \times \Omega} \phi \circ \text{proj}_0 d\pi \\ &= \lim_{n \rightarrow \infty} \int_{\Omega \times \Omega} \phi \circ \text{proj}_0 d\pi_n = \lim_{n \rightarrow \infty} \int_{\Omega} \phi d\text{proj}_0 \# \pi_n = \int_{\Omega} \phi d\mu \end{aligned}$$

So  $\text{proj}_0 \# \pi = \mu$ . Analogous:  $\text{proj}_1 \# \pi = \nu$ . □

*Proof of Proposition 3.27.* • Let  $\pi_n$  be minimizing sequence. Since  $\pi_n \in \mathcal{P}(\Omega \times \Omega)$  have  $\|\pi_n\|_{\mathcal{M}} = 1$ . By Banach-Alaoglu (Thm. 2.56)  $\exists$  converging subsequence. After extraction of subsequence have convergent minimizing sequence  $\pi_n \xrightarrow{*} \pi$ .

- By Prop. 3.28  $\pi \in \Pi(\mu, \nu)$ .
- Since  $c \in C(\Omega \times \Omega)$  and  $\pi_n \xrightarrow{*} \pi$  have

$$\int_{\Omega \times \Omega} c d\pi = \lim_{n \rightarrow \infty} \int_{\Omega \times \Omega} c d\pi_n.$$

Therefore,  $\pi$  is minimizer. □

---

Comment: For proof under more general conditions see for instance [Villani, 2009, Chapter 4].

---

Relation to Monge problem:

**Proposition 3.29** (Kantorovich is a relaxation of the Monge problem). Assume  $T : \Omega \rightarrow \Omega$  is a feasible transport map for the Monge problem between  $\mu$  and  $\nu$ , Definition 3.23. In particular  $T_{\#} \mu = \nu$ .

Let

$$(\text{id}, T) : \Omega \rightarrow \Omega \times \Omega, \quad x \mapsto (x, T(x)).$$

Then  $\pi = (\text{id}, T)_{\#} \mu \in \Pi(\mu, \nu)$  and

$$\int_{\Omega \times \Omega} c d\pi = \int_{\Omega} c(x, T(x)) d\mu(x).$$

*Proof.* • Clearly  $\pi \in \mathcal{P}(\Omega \times \Omega)$ : non-negative, unit mass.

- $\text{proj}_0 \# \pi = \text{proj}_0 \# (\text{id}, T) \# \mu = (\text{proj}_0 \circ (\text{id}, T)) \# \mu = (\text{id}) \# \mu = \mu$ .
- $\text{proj}_1 \# \pi = \text{proj}_1 \# (\text{id}, T) \# \mu = (\text{proj}_1 \circ (\text{id}, T)) \# \mu = T \# \mu = \nu$ .
- Equality of cost:

$$\int_{\Omega \times \Omega} c(x, y) d((\text{id}, T) \# \pi)(x, y) = \int_{\Omega} (c \circ (\text{id}, T))(x) d\mu(x) = \int_{\Omega} c(x, T(x)) d\mu(x). \quad \square$$

**Remark 3.30.** Under suitable conditions (e.g.  $\Omega$  compact subset of  $\mathbb{R}^d$  with Euclidean distance,  $c(x, y) = \|x - y\|^2$ ,  $\partial\Omega$  Lebesgue-negligible,  $\mu$  Lebesgue-absolutely continuous) one can show that the optimal coupling indeed corresponds to an optimal Monge map and thus both problems are equivalent. Proof is beyond scope of lecture. But shows: difficult non-convex problems can sometimes be rewritten as equivalent convex problems in higher dimensions.

### 3.4 Duality

Now we study the corresponding dual problem.

**Proposition 3.31.** Given the setting of Definition 3.26 one finds

$$\mathcal{C}(\mu, \nu) = \sup \left\{ \int_{\Omega} \alpha d\mu + \int_{\Omega} \beta d\nu \mid \alpha, \beta \in C(\Omega), \alpha(x) + \beta(y) \leq c(x, y) \text{ for all } (x, y) \in \Omega^2 \right\}$$

*Proof.* • Problem of Prop. 3.31 can be written as

$$\mathcal{C}(\mu, \nu) = - \inf \{ f(\alpha, \beta) + g(A(\alpha, \beta)) \mid (\alpha, \beta) \in C(\Omega)^2 \}$$

with

$$\begin{aligned} f : C(\Omega)^2 &\rightarrow \mathbb{R}, & (\alpha, \beta) &\mapsto - \int_{\Omega} \alpha d\mu - \int_{\Omega} \beta d\nu \\ g : C(\Omega^2) &\rightarrow \mathbb{R} \cup \{\infty\}, & \psi &\mapsto \begin{cases} 0 & \text{if } \psi(x, y) \leq c(x, y) \text{ for all } (x, y) \in \Omega^2 \\ +\infty & \text{else.} \end{cases} \\ A : C(\Omega)^2 &\rightarrow C(\Omega^2), & [A(\alpha, \beta)](x, y) &= \alpha(x) + \beta(y). \end{aligned}$$

- $f, g$  are convex, lsc.  $A$  is bounded, linear.
- Let  $(\alpha, \beta)$  be two constant, finite functions with  $\alpha(x) + \beta(y) < \min\{c(x', y') \mid (x', y') \in \Omega^2\}$ . Then  $f(\alpha, \beta) < \infty$ ,  $g(A(\alpha, \beta)) < \infty$  and  $g$  is continuous at  $A(\alpha, \beta)$ . Thus, by the Fenchel–Rockafellar theorem (an extension of Prop. 2.72 to account for the linear transformation  $A$ )

$$\mathcal{C}(\mu, \nu) = \min \{ f^*(-A^* \pi) + g^*(\pi) \mid \pi \in \mathcal{M}(\Omega^2) \}.$$

- One obtains:

$$\begin{aligned} f^*(-\rho, -\sigma) &= \sup \left\{ - \int_{\Omega} \alpha d\rho - \int_{\Omega} \beta d\sigma + \int_{\Omega} \alpha d\mu + \int_{\Omega} \beta d\nu \mid (\alpha, \beta) \in C(\Omega)^2 \right\} \\ &= \begin{cases} 0 & \text{if } \rho = \mu, \sigma = \nu, \\ +\infty & \text{else.} \end{cases} \end{aligned}$$

(Reasoning similar than for positivity of limit  $\pi$  in proof of Prop. 3.28.)

$$\begin{aligned} g^*(\pi) &= \sup \left\{ \int_{\Omega^2} \psi \, d\pi \mid \psi \in C(\Omega^2), \psi(x, y) \leq c(x, y) \text{ for all } (x, y) \in \Omega^2 \right\} \\ &= \begin{cases} \int_{\Omega^2} c \, d\pi & \text{if } \pi \in \mathcal{M}_+(\Omega^2), \\ +\infty & \text{else.} \end{cases} \end{aligned}$$

← VL16

- Adjoint of  $A$ :

$$\begin{aligned} \langle A^* \pi, (\alpha, \beta) \rangle_{\mathcal{M}(\Omega)^2 \times C(\Omega)^2} &= \langle \pi, A(\alpha, \beta) \rangle_{\mathcal{M}(\Omega \times \Omega) \times C(\Omega \times \Omega)} \\ &= \int_{\Omega \times \Omega} [\alpha(x) + \beta(y)] \, d\pi(x, y) \\ &= \int_{\Omega \times \Omega} [\alpha \circ \text{proj}_0 + \beta \circ \text{proj}_1] \, d\pi \\ &= \int_{\Omega} \alpha \, d(\text{proj}_0 \# \pi) + \int_{\Omega} \beta \, d(\text{proj}_1 \# \pi) \end{aligned}$$

$$\Rightarrow A^* \pi = (\text{proj}_0 \# \pi, \text{proj}_1 \# \pi).$$

- Summarize:

$$\begin{aligned} f^*(-A^* \pi) + g^*(\pi) &= f^*(-\text{proj}_0 \# \pi, -\text{proj}_1 \# \pi) + g^*(\pi) \\ &= \begin{cases} \int_{\Omega^2} c \, d\pi & \text{if } \pi \in \Pi(\mu, \nu), \\ +\infty & \text{else.} \end{cases} \end{aligned} \quad \square$$

**Remark 3.32** (Economic Interpretation of Kantorovich Duality). Bakeries and cafes hire a third-party company to do the transportation and agree to split the transport cost. When picking up bread at bakery  $x$  in the morning, the company charges an advance payment  $\alpha(x)$  per unit of bread for the transport. Upon delivery at a cafe at  $y$  it charges a final payment  $\beta(y)$  per unit of bread from the cafe.

The total payment to the company will be  $\int_{\Omega} \alpha \, d\mu + \int_{\Omega} \beta \, d\nu$  and they will try to maximize their profit. It is left to the company to decide which bread to deliver where, the bakeries and cafes do not know which bread goes where. (Of course, ‘internally’ the company will probably also need to minimize the total transport cost, i.e. to find the *global minimum* of  $\int_{\Omega \times \Omega} c \, d\pi$ , or maybe they have more efficient means of transportation.)

For every bakery  $x$  and cafe  $y$  the prices that the company charges must satisfy  $\alpha(x) + \beta(y) \leq c(x, y)$ . Otherwise, bakery  $x$  and cafe  $y$  could hire another driver who transports bread between them at lower cost  $c(x, y)$ .

**Proposition 3.33** (Primal-dual optimality conditions). A pair  $(\pi, (\alpha, \beta)) \in \mathcal{M}(\Omega \times \Omega) \times C(\Omega)^2$  is optimal for  $\mathcal{C}(\mu, \nu)$  in formulations Def. 3.26 and Prop. 3.31 if and only if  $\pi \in \Pi(\mu, \nu)$ ,  $\alpha(x) + \beta(y) \leq c(x, y)$  for all  $(x, y) \in \Omega^2$  and  $\text{spt } \pi \subset \{(x, y) \in \Omega^2 \mid \alpha(x) + \beta(y) = c(x, y)\}$ .

*Proof.* • Continue to use notation from proof of Prop. 3.31.

- For any  $(\pi, (\alpha, \beta)) \in \mathcal{M}(\Omega \times \Omega) \times C(\Omega)^2$  know by the Fenchel–Young inequality that

$$f(\alpha, \beta) + f^*(-A^* \pi) \geq \langle -A^* \pi, (\alpha, \beta) \rangle \quad g(A(\alpha, \beta)) + g^*(\pi) \geq \langle \pi, A(\alpha, \beta) \rangle$$

This implies

$$f(\alpha, \beta) + g(A(\alpha, \beta)) \geq -f^*(-A^*\pi) - g^*(\pi)$$

with equality if and only if they are primal and dual optimal.

- Precisely in this case, both Fenchel–Young inequality must be equalities, which is equivalent to

$$[-A^*\pi \in \partial f(\alpha, \beta)] \quad \wedge \quad [\pi \in \partial g(A(\alpha, \beta))]$$

- Now determine the subdifferentials.  $f(\alpha, \beta) = -\langle \mu, \alpha \rangle - \langle \nu, \beta \rangle$  is linear, therefore  $\partial f(\alpha, \beta) = \{(-\mu, -\nu)\}$ .
- Let now  $\psi \in C(\Omega^2)$  and set  $S := \{\pi \in \mathcal{M}_+(\Omega^2) \mid \psi = c \pi\text{-a.e.}\}$ . We claim that

$$\partial g(\psi) = \begin{cases} S & \text{if } \psi \leq c \\ \emptyset & \text{else.} \end{cases}$$

- If  $\psi \not\leq c$  then  $g(\psi) = \infty$  and thus  $\partial g(\psi) = \emptyset$  by definition. Let now  $\psi \leq c$ .
- Show that  $\partial g(\psi) \subset S$ . Let  $\pi \in \mathcal{M}(\Omega^2)$ ,  $\pi \not\geq 0$ .  $\Rightarrow \exists \phi \in C(\Omega^2)$ ,  $\phi \geq 0$ ,  $\langle \pi, \phi \rangle < 0$  such that

$$g(\psi - \phi) = 0 < -\langle \pi, \phi \rangle = g(\psi) + \langle \pi, -\phi \rangle$$

$\Rightarrow \pi \notin \partial g(\psi)$ . (How to construct / show existence of such  $\phi$ ? Can use decomposition, Cor. 3.15 and regularity, Thm. 3.18)

- $\langle \pi, c - \psi \rangle > 0$ :  $\Rightarrow \exists \phi \in C(\Omega^2)$ ,  $\phi \geq 0$  such that  $\langle \pi, \phi \rangle > 0$  and  $\psi + \phi \leq c$ . But then

$$g(\psi + \phi) = 0 < \langle \pi, \phi \rangle = g(\psi) + \langle \pi, \phi \rangle$$

$\Rightarrow \pi \notin \partial g(\psi)$ .

- So  $\partial g(\psi) \subset S$ . Now show converse inclusion. Let  $\pi \in S$ .
- Let  $B = \{(x, y) \in \Omega^2 \mid \psi(x, y) = c(x, y)\}$ . Then  $\pi(B) = 0$ .
- For  $\phi \in C(\Omega^2)$  consider the decomposition  $\phi = \phi_+ - \phi_-$  into positive and negative part, with  $\phi_{\pm} \geq 0$ . Then:

$$g(\psi + \phi) = g(\psi + \phi_+) \geq \begin{cases} 0 & \text{if } \phi_+(x) = 0 \forall x \in B, \\ +\infty & \text{else} \end{cases} \geq \langle \pi, \phi \rangle = g(\psi) + \langle \pi, \phi \rangle$$

- So  $S = \partial g(\psi)$ .
- With this we find that

$$[-A^*\pi \in \partial f(\alpha, \beta)] \quad \wedge \quad [\pi \in \partial g(A(\alpha, \beta))]$$

is equivalent to

$$-A^*\pi = -(\text{proj}_{0\sharp}\pi, \text{proj}_{1\sharp}\pi) = -(\mu, \nu), \quad A(\alpha, \beta) \leq c, \quad A(\alpha, \beta) = c \text{ } \pi\text{-a.e.}, \quad \pi \geq 0. \square$$

**Remark 3.34.**     • The optimality conditions can also be shown directly. A necessary condition for optimality is  $\pi \in \Pi(\mu, \nu)$  and  $A(\alpha, \beta) \leq c$ . Then one finds:

$$\int_{\Omega} \alpha \, d\mu + \int_{\Omega} \beta \, d\nu = \int_{\Omega^2} (\alpha(x) + \beta(y)) \, d\pi(x, y) \leq \int_{\Omega^2} c \, d\pi$$

with equality if and only if  $A(\alpha, \beta) = c$   $\pi$ -a.e..

**Remark 3.35** (Outlook on dual problem).     • Could use duality to apply proximal primal dual algorithm to solve problem: very simple proximal steps ( $f$  linear,  $g^*$  almost linear, pointwise). Not very efficient on large problems. But all efficient numerical methods heavily rely on simultaneous primal and dual perspective.

- Primal-dual perspective also essential for proving equivalence of Kantorovich and Monge formulation.

### 3.5 Discretization and $\Gamma$ -convergence

In this subsection discuss discretization of optimal transport problem. Approximate original problem by finite-dimensional problem which (in principle) we can solve. Furthermore, show  $\Gamma$ -convergence to original problem, as discretization is refined. Show that optimal plan can be extracted as cluster point of discrete optimal plans.

**Definition 3.36** (Discretization of domain). • For each  $n \in \mathbb{N}$  let  $\{x_{n,i}\}_{i=1}^n \subset \Omega$ ,  $\{T_{n,i}\}_{i=1}^n$ ,  $T_{n,i} \subset \Omega$ ,

- with  $x_{n,i} \in T_{n,i} \subset B(x_{n,i}, r_n)$  for some maximal radius  $r_n$  with  $(r_n)_n$  decreasing and  $r_n \rightarrow 0$  as  $n \rightarrow \infty$ ,
- $\Omega = \bigcup_{i=1}^n T_{n,i}$  and  $T_{n,i} \cap T_{n,j} = \emptyset$  if  $i \neq j$  ('collectively exhaustive and mutually exclusive').

**Example 3.37.**  $\{T_{n,i}\}_{i=1}^n$  and  $\{x_{n,i}\}_{i=1}^n$  might be cells and centroids of subsequently finer triangulations or of Cartesian grids over  $\Omega$ .

**Definition 3.38** (Discretization of marginals). For given  $\mu, \nu \in \mathcal{P}(\Omega)$  set

$$\begin{aligned} \hat{\mu}_n &= (\hat{\mu}_{n,i})_{i=1}^n, & \hat{\mu}_{n,i} &= \mu(T_{n,i}), & \mu_n &= \sum_{i=1}^n \delta_{x_{n,i}} \cdot \hat{\mu}_{n,i}, \\ \hat{\nu}_n &= (\hat{\nu}_{n,i})_{i=1}^n, & \hat{\nu}_{n,i} &= \nu(T_{n,i}), & \nu_n &= \sum_{i=1}^n \delta_{x_{n,i}} \cdot \hat{\nu}_{n,i}. \end{aligned}$$

**Proposition 3.39.**  $\mu_n \in \mathcal{P}(\Omega)$  and  $\mu_n \xrightarrow{*} \mu$  as  $n \rightarrow \infty$ . Analogously  $\nu_n \in \mathcal{P}(\Omega)$  and  $\nu_n \xrightarrow{*} \nu$ .

*Proof.* • Only proof for  $\mu$ . Result for  $\nu$  completely analogous.

- $\hat{\mu}_{n,i} \geq 0 \Rightarrow \mu_n$  is non-negative. Total mass: (use  $T_{n,i}$  collectively exhaustive, mutually exclusive)

$$\mu_n(\Omega) = \sum_{i=1}^n \hat{\mu}_{n,i} = \sum_{i=1}^n \mu(T_{n,i}) = \mu\left(\bigcup_{i=1}^n T_{n,i}\right) = \mu(\Omega) = 1$$

- Let  $\phi \in C(\Omega)$ .  $(\Omega, d)$  compact  $\Rightarrow \phi$  uniformly continuous (Lemma 2.32).  $\Rightarrow \forall \varepsilon > 0 \exists N \in \mathbb{N}$  such that for  $n \geq N$  have  $|\phi(x) - \phi(x_{n,i})| < \varepsilon$  for  $x \in T_{n,i} \subset B(x_{n,i}, r_n)$ .

$$\begin{aligned} \left| \int_{\Omega} \phi d\mu - \int_{\Omega} \phi d\mu_n \right| &\leq \sum_{i=1}^n \left| \int_{T_{n,i}} \phi d\mu - \int_{T_{n,i}} \phi d\mu_n \right| = \sum_{i=1}^n \left| \int_{T_{n,i}} \phi d\mu - \phi(x_{n,i}) \cdot \hat{\mu}_{n,i} \right| \\ &\leq \sum_{i=1}^n \left| \int_{T_{n,i}} (\phi(x) - \phi(x_{n,i})) d\mu(x) \right| \leq \sum_{i=1}^n \varepsilon \hat{\mu}_{n,i} = \varepsilon. \end{aligned}$$

Since  $\varepsilon > 0$  was arbitrary find  $\langle \mu_n, \phi \rangle_{\mathcal{M} \times C} \rightarrow \langle \mu, \phi \rangle_{\mathcal{M} \times C}$ . This is true for all  $\phi \in C(\Omega)$ .  $\square$

---

Comment: In general  $\mu_n \not\rightarrow \mu$  in the norm topology.

---

**Proposition 3.40** (Discretization of couplings). For two non-negative vectors  $(a_i)_{i=1}^n, (b_i)_{i=1}^n \in \mathbb{R}_+^n$  let

$$\hat{\Pi}_n(a, b) = \left\{ (\hat{\pi}_{i,j})_{i,j} \in \mathbb{R}_+^{n \times n} : \sum_{j'=1}^n \hat{\pi}_{i,j'} = a_i, \sum_{i'=1}^n \hat{\pi}_{i',j} = b_j \forall i, j \in \{1, \dots, n\} \right\}.$$

Then  $[\pi \in \Pi(\mu_n, \nu_n)] \Leftrightarrow [\pi = \sum_{i,j=1}^n \delta_{(x_{n,i}, x_{n,j})} \cdot \hat{\pi}_{i,j} \text{ for some } \hat{\pi} \in \hat{\Pi}_n(\hat{\mu}_n, \hat{\nu}_n)]$ .

*Proof.* •  $\Rightarrow$ :  $[\pi \in \Pi(\mu_n, \nu_n)] \Rightarrow [\text{for measurable } A \subset \Omega: \pi(A \times \Omega) = \mu_n(A) = \mu_n(A \cap \{x_{n,i}\}_{i=1}^n)] \Rightarrow [\text{spt } \pi \subset (\{x_{n,i}\}_{i=1}^n \times \Omega)]$ . Likewise:  $[\text{spt } \pi \subset (\Omega \times \{x_{n,i}\}_{i=1}^n)] \Rightarrow [\text{spt } \pi \subset (\{x_{n,i}\}_{i=1}^n \times \{x_{n,i}\}_{i=1}^n)]$ .

- So  $\pi = \sum_{i,j=1}^n \delta_{(x_{n,i}, x_{n,j})} \cdot \hat{\pi}_{i,j}$  for some  $\hat{\pi} \in \mathbb{R}_+^{n \times n}$ . (+ since  $\pi$  is non-negative).
- For  $i \in \{1, \dots, n\}$ :  $\hat{\mu}_{n,i} = \mu_n(\{x_{n,i}\}) = \pi(\{x_{n,i}\} \times \Omega) = \sum_{j=1}^n \hat{\pi}_{i,j}$ .
- For  $j \in \{1, \dots, n\}$ :  $\hat{\nu}_{n,j} = \nu_n(\{x_{n,j}\}) = \pi(\Omega \times \{x_{n,j}\}) = \sum_{i=1}^n \hat{\pi}_{i,j}$ .
- So  $\hat{\pi} \in \hat{\Pi}_n(\hat{\mu}_n, \hat{\nu}_n)$ .
- $\Leftarrow$ : Let  $A \subset \Omega$  be measurable. Find:

$$\pi(A \times \Omega) = \sum_{\substack{i=1: \\ x_{n,i} \in A}}^n \sum_{j=1}^n \hat{\pi}_{i,j} = \sum_{\substack{i=1: \\ x_{n,i} \in A}}^n \hat{\mu}_{n,i} = \mu_n(A).$$

Likewise,  $\pi(\Omega \times A) = \nu_n(A)$ .

- Clearly  $\pi \in \mathcal{M}_+(\Omega \times \Omega)$  and  $\pi \in \mathcal{P}(\Omega \times \Omega)$ . So  $\pi \in \Pi(\mu_n, \nu_n)$ . □

**Corollary 3.41** (Discretized transport problem).

$$\mathcal{C}(\mu_n, \nu_n) = \min \left\{ \sum_{i,j=1}^n c(x_{n,i}, x_{n,j}) \cdot \hat{\pi}_{i,j} \mid \hat{\pi} \in \hat{\Pi}_n(\hat{\mu}_n, \hat{\nu}_n) \right\}$$

---

Comment: This is a finite-dimensional problem. Discretization of marginals suffices such that problem becomes finite-dimensional and can be solved numerically. No discretization of derivatives etc. required.

---

Now we show  $\Gamma$ -convergence of discretized problems to original transport problem.

**Proposition 3.42.** Let

$$\begin{aligned} F : \mathcal{M}(\Omega \times \Omega) &\rightarrow \mathbb{R} \cup \{\infty\}, & \pi &\mapsto \int_{\Omega \times \Omega} c \, d\pi + \iota_{\Pi(\mu, \nu)}(\pi) \\ F_n : \mathcal{M}(\Omega \times \Omega) &\rightarrow \mathbb{R} \cup \{\infty\}, & \pi &\mapsto \int_{\Omega \times \Omega} c \, d\pi + \iota_{\Pi(\mu_n, \nu_n)}(\pi) \end{aligned}$$

Then  $F_n$   $\Gamma$ -converges sequentially to  $F$  in the weak\* topology as  $n \rightarrow \infty$ .

*Proof.* • **(lim inf)**: Let  $\pi_n \xrightarrow{*} \pi$ . Since  $c$  is continuous  $\Rightarrow \int_{\Omega \times \Omega} c \, d\pi_n \rightarrow \int_{\Omega \times \Omega} c \, d\pi$ .

- For every subsequence of  $(\pi_n)_n$  for which  $\pi_n \notin \Pi(\mu_n, \nu_n)$  we have  $\iota_{\Pi(\mu_n, \nu_n)}(\pi_n) = \infty$  and thus the lim inf condition is trivial.
- So assume  $\pi_n \in \Pi(\mu_n, \nu_n)$  for all  $n$ . Then by weak\* convergence for any  $\phi \in C(\Omega)$ :

$$\int_{\Omega} \phi d\mu \leftarrow \int_{\Omega} \phi d\mu_n = \int_{\Omega \times \Omega} \phi(x) d\pi_n(x, y) \rightarrow \int_{\Omega \times \Omega} \phi(x) d\pi(x, y)$$

So  $\text{proj}_{0\#} \pi = \mu$  and likewise  $\text{proj}_{1\#} \pi = \nu$ .

- By weak\* convergence,  $\pi$  is also non-negative and has unit mass (as all  $\pi_n$  are). So  $\pi \in \Pi(\mu, \nu)$  and thus

$$\liminf_n F_n(\pi_n) = \liminf_n \int_{\Omega \times \Omega} c d\pi_n = \lim_n \int_{\Omega \times \Omega} c d\pi_n = \int_{\Omega \times \Omega} c d\pi = F(\pi).$$

- **(lim sup):** Let  $\pi$  be fixed. If  $\pi \notin \Pi(\mu, \nu)$  then  $F(\pi) = \infty$  and any sequence  $(\pi_n)_n$  satisfies the lim sup condition.
- So consider  $\pi \in \Pi(\mu, \nu)$ . Define  $\hat{\pi}_n = (\hat{\pi}_{n,i,j})_{i,j=1}^n \in \mathbb{R}_+^{n \times n}$  via  $\hat{\pi}_{n,i,j} = \pi(T_{n,i} \times T_{n,j})$  and  $\pi_n = \sum_{i,j=1}^n \delta_{(x_{n,i}, x_{n,j})} \cdot \hat{\pi}_{n,i,j}$ .
- $\sum_{j=1}^n \hat{\pi}_{n,i,j} = \pi(T_{n,i} \times \Omega) = \mu(T_{n,i}) = \hat{\mu}_{n,i}$ . Similarly  $\sum_{i=1}^n \hat{\pi}_{n,i,j} = \nu_{n,j}$ . So  $\hat{\pi}_n \in \hat{\Pi}_n(\hat{\mu}_n, \hat{\nu}_n)$  and thus  $\pi_n \in \Pi(\mu_n, \nu_n)$  (Prop. 3.40).
- Analogous to  $\mu_n \xrightarrow{*} \mu$  (Prop. 3.39) show that  $\pi_n \xrightarrow{*} \pi$ .
- So  $\limsup_n F_n(\pi_n) = \limsup_n \int_{\Omega \times \Omega} c d\pi_n = \int_{\Omega \times \Omega} c d\pi = F(\pi)$ .  $\square$

**Proposition 3.43.** The minimal values of  $F_n$  converge to the minimal value of  $F$ . Any sequence  $(\pi_n)_n$  of minimizers of  $F_n$  is weak\* sequentially precompact and any cluster point  $\pi$  is a minimizer of  $F$ .

*Proof.* • For all  $n \in \mathbb{N}$  have  $\Pi(\mu_n, \nu_n) \subset \overline{B_{\mathcal{M}}(0, 1)}$ . Which is weak\* sequentially precompact by Banach–Alaoglu (cf. Remark 3.21). Therefore  $(F_n)_n$  are equi-mildly weak\* sequentially coercive.

- Also any sequence of minimizers  $(\pi_n)_n$  lies in  $\overline{B_{\mathcal{M}}(0, 1)}$  and therefore is weak\* sequentially precompact.
- The result then follows from Prop. 2.93.  $\square$

### 3.6 Wasserstein spaces

**Definition 3.44** (Wasserstein distance). Let  $(\Omega, d)$  be a compact metric space. For  $p \in [1, \infty)$  let  $W_p : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \rightarrow \mathbb{R}$ ,

$$W_p(\mu, \nu) = \left( \inf \left\{ \int_{\Omega \times \Omega} d(x, y)^p d\pi(x, y) \mid \pi \in \Pi(\mu, \nu) \right\} \right)^{1/p}$$

---

Comment: On non-compact spaces one usually restricts the Wasserstein space to measures with finite moment of order  $p$ , i.e.,  $\int_{\Omega} d(x, x_0)^p d\mu < +\infty$  for some arbitrary reference point  $x_0 \in \Omega$ . This is a sufficient condition to keep  $W_p$  finite.

---

**Example 3.45.** Dirac measures are isometric embedding of  $\Omega$  into  $\mathcal{P}(\Omega)$ :  $W_p(\delta_x, \delta_y) = d(x, y)$ , since  $\Pi(\delta_x, \delta_y) = \{\delta_{(x,y)}\}$ .

To prove that  $W_p$  is indeed a distance we will rely on the following powerful theorem which is often useful to dissect and reassemble measures with certain sought-after properties.

**Theorem 3.46** (Disintegration [Theorem 5.3.1, Ambrosio, Gigli, Savaré: Gradient Flows in Metric Spaces and in the Space of Probability Measures, 2008]). Let  $\tilde{\Omega}$ ,  $\Omega$  be compact metric spaces, let  $f : \tilde{\Omega} \rightarrow \Omega$  be measurable and  $\pi \in \mathcal{P}(\tilde{\Omega})$ . Set  $\mu = f_{\#}\pi \in \mathcal{P}(\Omega)$ . Then there is a family  $(\pi_y)_{y \in \Omega}$  in  $\mathcal{P}(\tilde{\Omega})$ , unique  $\mu$ -a.e., such that  $\pi_y(f^{-1}(\{y\})) = 1$  and for  $\phi \in C(\tilde{\Omega})$  one has

$$\int_{\tilde{\Omega}} \phi \, d\pi = \int_{\Omega} \left( \int_{\tilde{\Omega}} \phi \, d\pi_y \right) d\mu(y).$$

---

**Sketch:** Table, disintegration.

---

Comment: Disintegration formalizes the notion of conditional probability. It is easiest to visualize in a discrete case when  $\tilde{\Omega} = \Omega \times \Omega$  and  $f = \text{proj}_0$ . Then  $\pi$  can be interpreted as table and any  $\pi_y$  will be the restriction of  $\pi$  to row  $y$ , renormalized to mass 1 (if the row is non-empty).  $\pi_y$  gives the probabilities of picking a given column under the condition that row  $y$  has already been selected.

---

**Example 3.47** (Disintegration of transport plan). Let  $\pi \in \Pi(\mu, \nu)$ . Let  $(\gamma_x)_{x \in \Omega}$  be the disintegration of  $\pi$  with respect to  $\text{proj}_0$ . That is, for any  $\phi \in C(\Omega \times \Omega)$  have

$$\int_{\Omega \times \Omega} \phi(x, y) \, d\pi(x, y) = \int_{\Omega} \left( \int_{\Omega} \phi(x, y) \, d\gamma_x(y) \right) d\mu(x).$$

$\gamma_x$  can be interpreted as describing where mass particles starting in  $x$  are going. Note that it is only uniquely defined  $\mu$ -a.e..

---

Comment: By the disintegration theorem  $\gamma_x$  would be in  $\mathcal{P}(\Omega \times \Omega)$ . But since  $\gamma_x(\text{proj}_0^{-1}(\{x\})) = \gamma_x(\{x\} \times \Omega) = 1$  we can interpret  $\gamma_x$  as element of  $\mathcal{P}(\Omega)$ .

---

**Theorem 3.48.**  $W_p$  is a metric on  $\mathcal{P}(\Omega)$ .

*Proof.* •  $W_p$  is non-negative (since  $d(x, y)^p \geq 0$ ), symmetric (since  $d(x, y)^p$  is symmetric) and finite (since  $\Omega$  is compact, i.e.,  $d$  is bounded).

- Let  $T : \Omega \rightarrow \Omega \times \Omega$ ,  $T(x) = (x, x)$  be the ‘diagonal’ embedding of  $\Omega$  into  $\Omega \times \Omega$ .  $W_p(\mu, \mu) = 0$ , since  $\pi = T_{\#}\mu \in \Pi(\mu, \mu)$  and  $\int d^p \, d\pi = 0$ : Note that  $(\text{proj}_i \circ T)(x) = x$  and that  $f_{\#}(g_{\#}\rho) = (f \circ g)_{\#}\rho$ . Hence,  $\text{proj}_{i\#}T_{\#}\mu = \mu$ . Further,

$$\int_{\Omega \times \Omega} d^p \, d\pi = \int_{\Omega \times \Omega} d^p \, d(T_{\#}\mu) = \int_{\Omega} d^p \circ T \, d\mu = 0.$$

- Let  $W_p(\mu, \nu) = 0$ . Then there must be some  $\pi \in \Pi(\mu, \nu)$  with  $\int_{\Omega \times \Omega} d(x, y)^p \, d\pi(x, y) = 0$ , which implies  $d(x, y) = 0$   $\pi$ -a.e., i.e.,  $x = y$   $\pi$ -a.e.. So for  $\phi \in C(\Omega)$

$$\int_{\Omega \times \Omega} \phi(x) \, d\pi(x, y) = \int_{\Omega \times \Omega} \phi(y) \, d\pi(x, y)$$

and thus  $\text{proj}_{0\#}\pi = \text{proj}_{1\#}\pi$  which implies  $\mu = \nu$ .

- Towards triangle inequality: Let  $\mu, \nu, \rho \in \mathcal{P}(\Omega)$ , let  $\pi_{01}, \pi_{12}$  be optimal couplings for  $W_p(\mu, \nu)$  and  $W_p(\nu, \rho)$ . Let  $(\gamma_{01,y})_{y \in \Omega}$  be the disintegration of  $\pi_{01}$  with respect to  $\text{proj}_1$ . That is, for any  $\phi \in C(\Omega \times \Omega)$  have

$$\int_{\Omega \times \Omega} \phi(x, y) d\pi_{01}(x, y) = \int_{\Omega} \left( \int_{\Omega} \phi(x, y) d\gamma_{01,y}(x) \right) d\nu(y).$$

Similarly, let  $(\gamma_{12,y})_{y \in \Omega}$  be the disintegration of  $\pi_{12}$  with respect to  $\text{proj}_0$ .

- Define a new measure  $\pi \in \mathcal{P}(\Omega \times \Omega)$  via

$$\int_{\Omega \times \Omega} \phi(x, z) d\pi(x, z) := \int_{\Omega} \left( \int_{\Omega \times \Omega} \phi(x, z) d\gamma_{01,y}(x) d\gamma_{12,y}(z) \right) d\nu(y).$$

---

**Sketch:** Some intuition for  $\pi$ .

---

- Claim:  $\pi \in \Pi(\mu, \rho)$ . For  $\phi \in C(\Omega)$  get

$$\begin{aligned} \int_{\Omega \times \Omega} \phi(x) d\pi(x, z) &= \int_{\Omega} \left( \int_{\Omega \times \Omega} \phi(x) d\gamma_{01,y}(x) d\gamma_{12,y}(z) \right) d\nu(y) \\ &= \int_{\Omega} \left( \int_{\Omega} \phi(x) d\gamma_{01,y}(x) \right) d\nu(y) = \int_{\Omega \times \Omega} \phi(x) d\pi_{01}(x, y) = \int_{\Omega} \phi d\mu \end{aligned}$$

Second marginal analogously.

- Triangle inequality:

$$\begin{aligned} W_p(\mu, \rho) &\leq \left( \int_{\Omega \times \Omega} d(x, z)^p d\pi(x, z) \right)^{1/p} = \left( \int_{\Omega} \left( \int_{\Omega \times \Omega} d(x, z)^p d\gamma_{01,y}(x) d\gamma_{12,y}(z) \right) d\nu(y) \right)^{1/p} \\ &\leq \left( \int_{\Omega} \left( \int_{\Omega \times \Omega} (d(x, y) + d(y, z))^p d\gamma_{01,y}(x) d\gamma_{12,y}(z) \right) d\nu(y) \right)^{1/p} \\ &\stackrel{\text{Minkowski ineq.}}{\leq} \left( \int_{\Omega} \left( \int_{\Omega \times \Omega} d(x, y)^p d\gamma_{01,y}(x) d\gamma_{12,y}(z) \right) d\nu(y) \right)^{1/p} \\ &\quad + \left( \int_{\Omega} \left( \int_{\Omega \times \Omega} d(y, z)^p d\gamma_{01,y}(x) d\gamma_{12,y}(z) \right) d\nu(y) \right)^{1/p} \\ &= \left( \int_{\Omega \times \Omega} d(x, y)^p d\pi_{01}(x, y) \right)^{1/p} + \left( \int_{\Omega \times \Omega} d(y, z)^p d\pi_{12}(x, y) \right)^{1/p} \\ &= W_p(\mu, \nu) + W_p(\nu, \rho). \end{aligned} \quad \square$$

**Theorem 3.49** ( $W_p$  metrizes weak\* convergence). Let  $(\Omega, d)$  be a compact metric space.  $W_p$  metrizes the weak\* convergence on  $\mathcal{P}(\Omega)$ . That is, for a sequence  $(\mu_n)_n$  and some  $\mu$  in  $\mathcal{P}(\Omega)$  one has:

$$[W_p(\mu_n, \mu) \rightarrow 0] \quad \Leftrightarrow \quad [\mu_n \xrightarrow{*} \mu]$$

**Remark 3.50.** Careful! This equivalence only holds when  $(\Omega, d)$  is compact. On non-compact spaces convergence in  $W_p$  is equivalent to [weak\* convergence  $\wedge$  convergence of  $p$ -th moments  $\int_{\Omega} d(x_0, x)^p d\mu_n(x) \rightarrow \int_{\Omega} d(x_0, x)^p d\mu(x)$  for some arbitrary reference point  $x_0 \in \Omega$ ]. See [Villani, 2009, Theorem 6.9].

*Proof.* •  $\Rightarrow$ : assume  $W_p(\mu_n, \mu) \rightarrow 0$ . Let  $(\pi_n)_n$  be a corresponding sequence of optimal transport plans. Let  $\tilde{\mu}$  be a cluster point of  $(\mu_n)_n$  and let  $\pi \in \Pi(\tilde{\mu}, \mu)$  a corresponding cluster point of  $(\pi_n)_n$ . As before, denote the converging subsequence also by  $(\pi_n)_n$ . One has:

$$W_p(\tilde{\mu}, \mu) \leq \lim_{n \rightarrow \infty} \int_{\Omega \times \Omega} d^p d\pi_n = \lim_{n \rightarrow \infty} W_p(\mu_n, \mu) = 0$$

Since  $W_p$  is a metric,  $\tilde{\mu} = \mu$ . Hence,  $\mu_n \xrightarrow{*} \mu$ .

- $\Leftarrow$ : assume  $\mu_n \xrightarrow{*} \mu$ . Let  $\pi_n$  be optimal plans for  $W_p(\mu_n, \mu)$ . Extract a converging subsequence, again denoted by  $(\pi_n)_n$ .
- Using duality one can show that any cluster point  $\pi$  of  $(\pi_n)_n$  is an optimal coupling for  $W_p(\mu, \mu)$  (e.g. [Villani, 2009, Theorem 5.20]). So:

$$0 = W_p(\mu, \mu) = \int_{\Omega \times \Omega} d^p d\pi = \lim_{n \rightarrow \infty} \int_{\Omega \times \Omega} d^p d\pi_n = \lim_{n \rightarrow \infty} W_p(\mu_n, \mu) \quad \square$$

In the following example we illustrate the qualitatively different behaviour of Wasserstein distances for  $p > 1$  and  $p = 1$ .

**Example 3.51** (‘Shifting a bookshelf’). Let  $n \in \mathbb{N}$ ,

$$\mu = \frac{1}{n} \sum_{i=0}^{n-1} \delta_i, \quad \nu = \frac{1}{n} \sum_{i=0}^{n-1} \delta_{i+1}.$$

---

**Sketch:**  $\mu$  and  $\nu$  and the optimal couplings.

What are the optimal couplings for  $W_p(\mu, \nu)$ ? For  $p > 1$  the unique optimal coupling is given by

$$\pi = \frac{1}{n} \sum_{i=0}^{n-1} \delta_{(i, i+1)}.$$

For  $p = 1$  any  $\pi \in \Pi(\mu, \nu)$  with  $y \geq x$   $\pi(x, y)$ -a.e. is optimal. Two extreme examples: same as for  $p > 1$ , or only move first mass from  $i = 0$  to  $i = n - 1$ . Any coupling that does not move mass to the left is optimal.

For  $p \in (0, 1)$   $W_p$  does not define a distance, but the behaviour is still interesting. The unique optimal is given by

$$\pi = \frac{1}{n} \left( \sum_{i=1}^{n-1} \delta_{(i, i)} + \delta_{(0, n)} \right).$$

That is, it is preferred to leave most masses in place and only move the first one from the beginning to the end.

This different behaviour is linked to the convexity / concavity of the function  $x \mapsto |x|^p$  depending on  $p$ : for  $p > 1$ , many small transports are cheaper than one large one. For  $p \in (0, 1)$  one larger transport is cheaper. For  $p = 1$  any combination of ‘monotonous’ transports has equal cost.

---

Comment: How to proof that these couplings are optimal? By guessing corresponding dual variables.

---

### 3.7 Displacement interpolation

An intriguing property of the Wasserstein space  $(\mathcal{P}(\Omega), W_p)$  is that it is a length space if  $(\Omega, d)$  is a length space.

**Definition 3.52** (Length space). A metric space  $(\Omega, d)$  is a length space if for every pair  $(x, y) \in \Omega$  there is a continuous map  $\gamma_{x,y} \in C([0, 1], \Omega)$  with

$$\gamma_{x,y}(0) = x, \quad \gamma_{x,y}(1) = y, \quad d(\gamma_{x,y}(s), \gamma_{x,y}(t)) = d(x, y) \cdot |s - t|$$

for  $s, t \in [0, 1]$ .

**Theorem 3.53.** If  $(\Omega, d)$  is a length space and there exists a measurable map  $(x, y) \mapsto \gamma_{x,y}$  that takes start and endpoint to a shortest path between them, then  $(\mathcal{P}(\Omega), W_p)$  is a length space.

---

Comment: Sufficient conditions for the measurability of  $(x, y) \mapsto \gamma_{x,y}$  can be found for instance in [Villani, 2009, Proposition 7.16].

---

*Proof.* • Let  $(\gamma_{x,y})_{(x,y) \in \Omega^2}$  be the family of maps for  $(\Omega, d)$  as given by Definition 3.52. For fixed  $s, t \in [0, 1]$  let

$$\begin{aligned} \Gamma_s &: \Omega \times \Omega \rightarrow \Omega, & (x, y) &\mapsto \gamma_{x,y}(s), \\ \Gamma_{s,t} &: \Omega \times \Omega \rightarrow \Omega \times \Omega, & (x, y) &\mapsto (\gamma_{x,y}(s), \gamma_{x,y}(t)). \end{aligned}$$

---

Comment: Between  $\gamma$  and  $\Gamma$  the roles of ‘index’ and ‘arguments’ of the functions are exchanged. This is formally helpful to use the push-forward of  $\Gamma$ .

---

- For given  $\mu, \nu \in \mathcal{P}(\Omega)$  let  $\pi$  be an optimal coupling for  $W_p(\mu, \nu)$ . Denote  $\rho_s = \Gamma_s \# \pi$ .

---

**Sketch:** Interpretation of  $\rho_s$ : put mass that is transported from  $x$  to  $y$  at  $\gamma_{x,y}(s)$ .

---

- Claim:  $s \mapsto \rho_s$  is a geodesic in  $(\mathcal{P}(\Omega), W_p)$  between  $\mu$  and  $\nu$ . A ‘length space map’ for  $(\mathcal{P}(\Omega), W_p)$  between  $\mu$  and  $\nu$ ,  $\gamma_{\mu,\nu} : [0, 1] \rightarrow \mathcal{P}(\Omega)$  is given by  $\gamma_{\mu,\nu}(s) = \rho_s$ . We will now show this.
- Measurability of  $\Gamma_s$ : By assumption  $S : (x, y) \mapsto \gamma_{x,y}$  is measurable. For fixed  $t \in [0, 1]$  the map  $e_t : C([0, 1], \Omega) \rightarrow \Omega$ ,  $\gamma \mapsto \gamma(t)$  is continuous and thus measurable. We find  $\Gamma_s = e_s \circ S$ . Similarly,  $\Gamma_{s,t} = (\Gamma_s, \Gamma_t) = (e_s, e_t) \circ S$  is measurable.
- Claim:  $\Gamma_{s,t} \# \pi \in \Pi(\rho_s, \rho_t)$ .

$$\text{proj}_0 \# \Gamma_{s,t} \# \pi = (\text{proj}_0 \circ \Gamma_{s,t}) \# \pi = \Gamma_s \# \pi = \rho_s$$

- Claim:  $W_p(\rho_s, \rho_t) = |s - t| \cdot W_p(\mu, \nu)$ .

$$\begin{aligned} W_p(\rho_s, \rho_t)^p &\leq \int_{\Omega \times \Omega} d(x, y)^p d(\Gamma_{s,t} \# \pi)(x, y) = \int_{\Omega \times \Omega} ((d \circ \Gamma_{s,t})(x, y))^p d\pi(x, y) \\ &= \int_{\Omega \times \Omega} (d(\gamma_{x,y}(s), \gamma_{x,y}(t)))^p d\pi(x, y) = |s - t|^p \int_{\Omega \times \Omega} (d(x, y))^p d\pi(x, y) \\ W_p(\rho_s, \rho_t) &\leq |s - t| \cdot W_p(\mu, \nu) \end{aligned}$$

So for  $0 \leq s \leq t \leq 1$  have

$$W_p(\mu, \rho_s) \leq s \cdot W_p(\mu, \nu), \quad W_p(\rho_s, \rho_t) \leq (t - s) \cdot W_p(\mu, \nu), \quad W_p(\rho_t, \nu) \leq (1 - t) \cdot W_p(\mu, \nu)$$

So

$$W_p(\mu, \rho_s) + W_p(\rho_s, \rho_t) + W_p(\rho_t, \nu) \leq W_p(\mu, \nu)$$

and by the triangle inequality

$$W_p(\mu, \rho_s) + W_p(\rho_s, \rho_t) + W_p(\rho_t, \nu) \geq W_p(\mu, \nu).$$

Hence we must have equality and in particular  $W_p(\rho_s, \rho_t) = |s - t| \cdot W_p(\mu, \nu)$ . □

### 3.8 Wasserstein-1 space

**Remark 3.54.** In this section we study in more detail the structure of the Wasserstein distance for  $p = 1$ . Short summary: the dual problem allows a particular simplification, only possible for  $p = 1$ , which then can be reformulated as min cost flow problem (also known as Beckmann's problem). This is particularly suitable for numerical optimization.

**Definition 3.55** (1-Lipschitz functions). A function  $\alpha : \Omega \rightarrow \mathbb{R}$  is called 1-Lipschitz if for any pair  $x, y \in \Omega$  one finds

$$|\alpha(x) - \alpha(y)| \leq d(x, y).$$

The set of 1-Lipschitz functions on  $\Omega$  is denoted by  $\text{Lip}(\Omega)$  and evidently  $\text{Lip}(\Omega) \subset C(\Omega)$ .

**Theorem 3.56** (Kantorovich–Rubinstein formula).

$$W_1(\mu, \nu) = \sup \left\{ \int_{\Omega} \alpha \, d(\mu - \nu) \mid \alpha \in \text{Lip}(\Omega) \right\} \quad (*)$$

For the proof use the following Lemma.

**Lemma 3.57.** For  $\psi \in C(\Omega)$  let

$$\psi^c(x) := \inf_{y \in \Omega} [d(x, y) - \psi(y)].$$

(i)  $\psi^c \in \text{Lip}(\Omega)$ .

(ii) If  $\psi \in \text{Lip}(\Omega)$  then  $\psi^c = -\psi$ .

*Proof.* • **(i):** Since  $\Omega$  is compact and  $d(x, \cdot)$  and  $\psi$  are continuous, the infimum in definition of  $\psi^c$  is actually a minimum.

• Let  $x, x' \in \Omega$  and let  $y \in \Omega$  such that  $\psi^c(x) = d(x, y) - \psi(y)$ . Then

$$\psi^c(x') \leq d(x', y) - \psi(y) \leq d(x', x) + d(x, y) - \psi(y) = d(x', x) + \psi^c(x)$$

Since this holds for all  $x, x' \in \Omega$ ,  $\psi^c$  is 1-Lipschitz.

• **(ii):**

$$\psi^c(x) = \inf_{y \in \Omega} [d(x, y) - \psi(y)] \leq -\psi(x) \quad (\text{try } y = x \text{ as candidate in infimum})$$

$$\psi^c(x) = \inf_{y \in \Omega} [d(x, y) - \underbrace{\psi(y)}_{\leq \psi(x) + d(x, y)}] \geq \inf_{y \in \Omega} [d(x, y) - \psi(x) + d(x, y)] = -\psi(x) \quad \square$$

**Remark 3.58** ( $c$ -transform). The transformation introduced in Lemma 3.57 is a special case of the so-called  $c$ -transform which can also be defined for more general transport costs. It is a crucial tool in the analysis of optimal transport plans.

*Proof of Theorem 3.56.* • By Proposition 3.31 (dual Kantorovich problem) we have

$$W_1(\mu, \nu) = \sup \left\{ \int_{\Omega} \alpha \, d\mu + \int_{\Omega} \beta \, d\nu \mid \alpha, \beta \in C(\Omega), \alpha(x) + \beta(y) \leq d(x, y) \text{ for all } (x, y) \in \Omega^2 \right\}. \quad (**)$$

- Let  $\alpha \in \text{Lip}(\Omega)$  which means that it is feasible for (\*).
- Set  $\beta = -\alpha = \alpha^c$  (Lemma 3.57). Then  $\alpha(x) + \beta(y) = \alpha(x) - \alpha(y) \leq d(x, y)$ , which means that  $(\alpha, -\alpha)$  is feasible for (\*\*) and has the same score as  $\alpha$  for (\*).
- We can do this construction for every  $\alpha$  that is feasible in (\*). Therefore,  $(*) \leq (**)$ .
- Conversely, let  $(\alpha, \beta)$  be feasible for (\*\*). Then for all  $(x, y) \in \Omega^2$  have  $\beta(y) \leq d(x, y) - \alpha(x)$  and thus

$$\beta(y) \leq \inf_{x' \in \Omega} [d(x', y) - \alpha(x')] = \alpha^c(y) \leq d(x, y) - \alpha(x).$$

So  $\beta \leq \alpha^c$  and  $(\alpha, \alpha^c)$  are still feasible for (\*\*) with a score that is not worse than the one of  $(\alpha, \beta)$ .

- Likewise, replace now  $\alpha$  by  $(\alpha^c)^c = -\alpha^c$  to get

$$\int_{\Omega} \alpha \, d\mu + \int_{\Omega} \beta \, d\nu \leq \int_{\Omega} \alpha \, d\mu + \int_{\Omega} \alpha^c \, d\nu \leq \int_{\Omega} (\alpha^c)^c \, d\mu + \int_{\Omega} \alpha^c \, d\nu = \int_{\Omega} (-\alpha^c) \, d(\mu - \nu)$$

- In particular,  $-\alpha^c$  is feasible for (\*) with a score not worse than the one of  $(\alpha, \beta)$  for (\*\*).
- Therefore,  $(*) \geq (**)$  and thus equality.  $\square$

---

**Sketch:** Interpretation with  $\mu, \nu$  being Dirac measures.

---

**Remark 3.59.** An advantage of the Kantorovich–Rubinstein formula is that  $\alpha \in \text{Lip}(\Omega)$  can often be turned into a local constraint (pointwise, or on small neighbourhoods) and thus, the large set of constraints of the original dual Kantorovich problem  $(\alpha(x) + \beta(y) \leq d(x, y))$  for all  $(x, y) \in \Omega^2$  can be replaced by a smaller number of constraints.

**Remark 3.60** (Additional assumptions). For the rest of this subsection assume that  $\Omega$  is compact and the closure of an open set of  $\mathbb{R}^d$ .  $d$  is the metric induced by shortest paths in  $\Omega$ , according to the Euclidean length of paths.

Then the Lipschitz constraint can be approximated by a local constraint on the gradient of the Kantorovich potential  $\alpha$ . Going then back to a primal formulation one finds a transport problem where movement of mass is described by a vector field  $\omega \in \mathcal{M}(\Omega)^2$  that describes how mass flows from  $\mu$  to  $\nu$ .

**Definition 3.61** (Weak divergence). A measure  $\omega \in \mathcal{M}(\Omega)^2$  is said to have weak divergence  $\rho \in \mathcal{M}(\Omega)$ , we write  $\text{div } \omega = \rho$ , if for every  $\phi \in C^1(\Omega)$  one has

$$\int_{\Omega} (\nabla \phi) \cdot d\omega + \int_{\Omega} \phi \, d\rho = 0.$$

---

Comment: If  $\omega$  is a smooth vector field can apply integration by parts on first term to get standard definition of divergence.

---

**Theorem 3.62** (Beckmann’s problem).

$$W_1(\mu, \nu) = \min \{ \|\omega\|_{\mathcal{M}(\Omega)^2} \mid \omega \in \mathcal{M}(\Omega)^2, \text{div } \omega = \nu - \mu \}$$

---

Comment: Interpretation:  $\omega$  is flow field that takes  $\mu$  to  $\nu$ . At each point,  $\omega$  gives orientation and amount of flow.

---

Comment: After discretization with  $|\Omega| = N$  discrete points, the standard Kantorovich dual problem needs  $\mathcal{O}(N)$  variables and  $\mathcal{O}(N^2)$  constraints. Beckmann's problem only requires  $\mathcal{O}(N)$  variables and constraints.

---

*Proof.* • Start with Kantorovich–Rubinstein formula, Theorem 3.56:

$$W_1(\mu, \nu) = \sup \left\{ \int_{\Omega} \alpha \, d(\mu - \nu) \mid \alpha \in \text{Lip}(\Omega) \right\}$$

- $C^1(\Omega) \cap \text{Lip}(\Omega)$  is dense in  $\text{Lip}(\Omega)$  for the sup-norm.
- 

Comment: The proof is intuitive but somewhat lengthy. It can be found, e.g., in [Schmitzer, Wirth: Dynamic Models of Wasserstein-1-Type Unbalanced Transport, 2017].

---

- Since  $\alpha \mapsto \int_{\Omega} \alpha \, d(\mu - \nu)$  is continuous, by density can restrict optimization of  $\alpha$  to  $C^1(\Omega) \cap \text{Lip}(\Omega)$ .

- Find

$$C^1(\Omega) \cap \text{Lip}(\Omega) = \{\alpha \in C^1(\Omega) \mid \|\nabla \alpha(x)\| \leq 1 \text{ for all } x \in \Omega\} = \{\alpha \in C^1(\Omega) \mid \|\nabla \alpha\|_{C(\Omega)^2} \leq 1\}$$

- Get:

$$W_1(\mu, \nu) = \sup \left\{ -f(\alpha) - g(\nabla \alpha) \mid \alpha \in C^1(\Omega) \right\}$$

with

$$\begin{aligned} f(\alpha) &= \int_{\Omega} \alpha \, d(\nu - \mu), \\ g(v) &= \begin{cases} 0 & \text{if } \|v\|_{C(\Omega)^2} \leq 1 \\ \infty & \text{else.} \end{cases} \end{aligned}$$

- Now apply Fenchel-Rockafellar.  $f, g$  are convex, lower semi-continuous ( $f$  is linear,  $g$  is indicator of closed unit ball of sup-norm).  $X = C^1(\Omega)$ ,  $Y = C(\Omega)^2$ ,  $A = \nabla$ . At  $\alpha : x \mapsto 0$  have  $f(\alpha) < \infty$  and  $g(\nabla \alpha) < \infty$  and  $g$  is continuous at  $\nabla \alpha$ .

- Dual problem:

$$W_1(\mu, \nu) = \min \left\{ f^*(-\nabla^* \omega) + g^*(\omega) \mid \omega \in \mathcal{M}(\Omega)^2 \right\}$$

- Compute conjugates:

$$\begin{aligned} f^*(-\nabla^* \omega) &= \sup \left\{ \langle \alpha, -\nabla^* \omega \rangle_{C^1(\Omega), C^1(\Omega)^*} - f(\alpha) \mid \alpha \in C^1(\Omega) \right\} \\ &= \sup \left\{ \langle -\nabla \alpha, \omega \rangle_{C(\Omega)^2, \mathcal{M}(\Omega)^2} + \int_{\Omega} \alpha \, d(\mu - \nu) \mid \alpha \in C^1(\Omega) \right\} \\ &= \begin{cases} 0 & \text{if } \text{div } \omega = \nu - \mu \\ +\infty & \text{else.} \end{cases} \\ g^*(\omega) &= \sup \left\{ \langle v, \omega \rangle_{C(\Omega)^2, \mathcal{M}(\Omega)^2} \mid v \in C(\Omega)^2, \|v\|_{C(\Omega)^2} \leq 1 \right\} \end{aligned}$$

So  $g^*(\omega)$  is operator-norm of  $\omega$  (interpreted as linear map on  $C(\Omega)^2$ ). By duality:

$$g^*(\omega) = \|\omega\|_{\mathcal{M}(\Omega)^2}$$

□

**Remark 3.63** (Min-cost flow on discrete graph). Assume  $\Omega$  is a finite, discrete vertex set of a graph with edge set  $E \subset \Omega \times \Omega$  and edge lengths  $\ell : E \rightarrow (0, \infty]$ . Let  $d$  be metric induced by shortest paths in  $(\Omega, E, \ell)$ . Then

$$[\alpha \in \text{Lip}(\Omega)] \Leftrightarrow [|\alpha(x) - \alpha(y)| \leq \ell(x, y) \text{ for all } (x, y) \in E].$$

In many applications  $|E| \ll |\Omega|^2$  and thus the right condition is easier to check than the left. Can define discrete gradient operator  $G : C(\Omega) \rightarrow C(E)$ ,  $(G\alpha)(x, y) = (\alpha(x) - \alpha(y))/\ell(x, y)$  and replace Lipschitz constraint by local constraint on gradient, as above. This leads to discrete min-cost flow problem on metric graph. See for instance [Bertsekas, Eckstein: Dual Coordinate Step Methods for Linear Network Flow Problems, 1988].