Optimization and Optimal Control in Banach Spaces

Bernhard Schmitzer

January 22, 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 $C \stackrel{\text{def.}}{=} \bigcap_{i \in I} C_i$ 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* $\operatorname{conv} C$ 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.,

$$T \stackrel{\text{def.}}{=} \left\{ \sum_{i=1}^{k} \lambda_i\, x_i \middle| k \in \mathbb{N},\, x_1, \ldots, x_k \in C,\, \lambda_1, \ldots, \lambda_k > 0,\, \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

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

Proof. $\operatorname{conv} C \subset T$**.** 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 $\operatorname{conv} C \subset T$.
$\operatorname{conv} C \supset T$**.** Let S be convex and $S \supset C$. We will show that $S \supset T$ and thus $\operatorname{conv} C \supset T$, which with the previous step implies equality of the two sets.
We show $S \supset T$ by recursion. For some $k \in \mathbb{N}$, $x_1, \ldots, x_k \in C$, $\lambda_1, \ldots, \lambda_k > 0$, $\sum_{i=1}^{k} \lambda_i = 1$ let

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

When $k = 1$ clearly $s_k \in S$.
Otherwise, set $\tilde{\lambda}_i = \lambda_i/(1 - \lambda_k)$ for $i = 1, \ldots, k-1$. Then

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

We find that $s_k \in S$ if $s_{k-1} \in S$. Applying this argument recursively to s_{k-1} until we reach s_1, we have shown that $s_k \in S$. $\square$

Proposition 1.7 (Carathéodory)**.** Let $H = \mathbb{R}^n$. Every $x \in \operatorname{conv} C$ 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 λ_i becomes 0.

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

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

- Define $\tilde{\lambda}_i = \lambda_i - t^*\, \beta_i$ for $t^* = \frac{\lambda_{i^*}}{\beta_{i^*}}$ and $i^* = \operatorname{argmin}_{i=1,\ldots,k:\beta_i \neq 0} \frac{\lambda_i}{|\beta_i|}$.
- $\tilde{\lambda}_i \geq 0$: $\tilde{\lambda}_i = \lambda_i \cdot \big(1 - \underbrace{\tfrac{\lambda_{i^*}/\beta_{i^*}}{\lambda_i/\beta_i}}_{|\cdot| \leq 1}\big)$

- $\tilde{\lambda}_{i^*} = 0$
- $\sum_{i=1}^{k} \tilde{\lambda}_i = \underbrace{\sum_{i=1}^{k} \lambda_i}_{=1} - t^* \underbrace{\sum_{i=1}^{k} \beta_i}_{=0} = 1$
- $\sum_{i=1}^{k} \tilde{\lambda}_i \, x_i = \underbrace{\sum_{i=1}^{k} \lambda_i \, x_i}_{=x} - t^* \underbrace{\sum_{i=1}^{k} \beta_i \, x_i}_{=0} = x$

$\square$

1.2 Convex functions

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

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

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

---

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

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

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

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

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

(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*

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

is convex.

(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.,

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

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

$$\min\{f(x) + g(A\,x)|x \in \mathbb{R}^n\} \qquad \text{where} \qquad g = \iota_{\{y\}}\,.$$

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

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 $f \in \operatorname{Conv}(H)$ be proper and let $x_0 \in \operatorname{dom} f$. 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 $\alpha \overset{\text{def.}}{=} \|x - x_0\|/\rho \in (0, 1]$

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

$$f(x) \le (1-\alpha)\cdot f(x_0) + \alpha \cdot f(\tilde{x})$$
$$f(x) - f(x_0) \le \alpha \cdot (\eta - f(x_0)) = \|x - x_0\| \cdot \tfrac{\eta - f(x_0)}{\rho}$$

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:

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

  We combine to get:

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

- 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 $f(x_1) \geq f(x_0) - \frac{\rho}{4} \cdot \frac{\eta - f(x_0)}{\rho} \geq 2\,f(x_0) - \eta$. With arguments above get for every $x \in \overline{B(x_1,\rho/2)}$ that

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

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

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

□

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

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

- By assumption there is some $x_0 \in \operatorname{dom} f$, $\rho \in \mathbb{R}_{++}$ and $\eta < \infty$ such that $\sup f(\overline{B(x_0,\rho)}) \leq \eta$.
- For any $x \in \operatorname{int}\operatorname{dom} f$ there is some $y \in \operatorname{dom} f$ 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 $\overline{B(x,\alpha\cdot\rho)} \subset \operatorname{dom} f$ and $y \notin \overline{B(x,\alpha\cdot\rho)}$.
- Then, $\overline{B(x,\alpha\cdot\rho)} \subset \operatorname{conv}(\overline{B(x_0,\rho)} \cup \{y\})$.
- 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.

□

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

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 $\operatorname{int}\operatorname{dom} f$.

Proof.

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

□

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 $f : H \to \mathbb{R} \cup \{\infty\}$ 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

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

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 $f : H \to \mathbb{R} \cup \{\infty\}$. The following are equivalent:

(i) f is lower semicontinuous.

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

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

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

$$r = \lim_{k\to\infty} r_k \geq \liminf_{k\to\infty} f(y_k) \geq f(y) \qquad \Rightarrow \qquad (y, r) \in \operatorname{epi} f\,.$$

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

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

(iii) ⇒ (i). Assume (i) is false. Then there is a sequence $(y_k)_k$ in H converging to $y \in H$ such that $\rho \overset{\text{def.}}{=} \lim_{k\to\infty} f(y_k) < f(x)$. 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. □

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 $f : H \to \mathbb{R} \cup \{\infty\}$ be proper. The *subdifferential* of f is the set-valued operator

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

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 touch graph of function in x from below.

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

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

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 $f : H \to \mathbb{R} \cup \{\infty\}$ be proper. Then

$$\operatorname{argmin} f = \{x \in H \colon 0 \in \partial f(x)\}\,.$$

Proof. Let $x \in H$. Then

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

□

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

(i) $\partial f(x)$ is closed and convex.

(ii) If $x \in \operatorname{dom} \partial f$ then f is lower semicontinuous at x.

Proof. (i):

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

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

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

□

Definition 1.28 (Monotonicity)**.** A set-valued function $A : H \to 2^H$ 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:

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

$\square$

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 $f_i : H_i \to \mathbb{R} \cup \{\infty\}$ be proper and let $f : H \to \mathbb{R} \cup \{\infty\}$, $x = (x_i)_{i \in I} \mapsto \sum_{i \in I} f_i(x_i)$. Then $\partial f(x) = \bigotimes_{i \in I} \partial f_i(x_i)$.

Proof. $\partial f(x) \supset \bigotimes_{i \in I} \partial f_i(x_i)$**:** 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 $\partial f(0) = \overline{B(0,1)}$.

Sketch: Draw 'graph' of subdifferential.

- $H = \mathbb{R}$, $f(x) = |x|$ is a special case of the above.

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

- $H = \mathbb{R}^n$, $f(x) = \|x\|_1$. The L_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 $f, g : H \to \mathbb{R} \cup \{\infty\}$. 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 proof. 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 $f : H = \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. Let μ be a probability measure on H such that

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

and $\overline{x} \in \operatorname{dom} \partial f$. Then

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

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

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

□

Let us examine the subdifferential of differentiable functions.

Definition 1.35 (Gâteaux differentiability)**.** A function $f : H \to \mathbb{R} \cup \{\infty\}$ is *Gâteaux differentiable* in $x \in \operatorname{dom} f$ 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

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

Proposition 1.36. Let $f : H \to \mathbb{R} \cup \{\infty\}$ be proper and convex, let $x \in \operatorname{dom} f$. 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 $\phi : \mathbb{R}_++ \to \mathbb{R} \cup \{\infty\}$, $\alpha \mapsto \frac{f(x+\alpha \cdot y)-f(x)}{\alpha}$.
- ϕ 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.)

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

- 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 $\operatorname{cone} C$ 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 $\operatorname{cone} C \subset D$. Conversely, let $y \in D$. Then there are $x \in C$ and $\lambda \in \mathbb{R}_{++}$ such that $y = \lambda \cdot x$. So $x \in \operatorname{cone} C$, therefore $y \in \operatorname{cone} C$ and thus $D \subset \operatorname{cone} C$.

□

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. □

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

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

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

□

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

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

The *normal cone* to C at x is

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

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

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

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.

- $x \notin C$: $\partial \iota_C(x) = \emptyset = N_C x$.
- $x \in C$:

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

□

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

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

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:

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

- So by sending $k, l \to \infty$ 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:

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

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.

□

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 ε 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.

□

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

Comment: This condition is actually useful for computing projections.

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

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

Sketch: Draw normal cones attached to points in C.

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

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

Proof.

- By virtue of Corollary 1.50 (Characterization of projection with normal cone inclusion) we need to show

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

- **⇒:** 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.

- **⇐:** 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 $x-p \in N_K p$.

□

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: $K^\ominus = \{u \in H : \sup \langle K, u\rangle \leq 0\}$.
- 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 $K^{\ominus\ominus} \subset K$.

□

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) For all $x \in \operatorname{int} C$, $y \in \overline{C}$, $[x, y) \subset \operatorname{int} C$.

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

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

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

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 $x \in \operatorname{int} C$ 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:

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

- Therefore $z \in \operatorname{int} C$.
- **(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 $\lambda \cdot x + (1 - \lambda) \cdot y \subset \overline{C}$.
- **(iii):** Let $x, y \in \operatorname{int} C$. Then $y \in \overline{C}$. By (i) therefore $(x, y) \in \operatorname{int} C$.
- **(iv):** By definition $\operatorname{int} C \subset \operatorname{int} \overline{C}$. Show converse inclusion. Let $y \in \operatorname{int} \overline{C}$. Then there is $\varepsilon \in \mathbb{R}_{++}$ such that $B(y, \varepsilon) \subset \overline{C}$. Let $x \in \operatorname{int} C$, $x \neq y$. 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 $y \in \operatorname{int} C$.
- Similarly, it is clear that $\overline{\operatorname{int} C} \subset \overline{C}$. We show the converse inclusion. Let $x \in \operatorname{int} C$, $y \in \overline{C}$. For $\alpha \in (0, 1]$ let $y_\alpha = (1 - \alpha) \cdot y + \alpha \cdot x$. Then $y_\alpha \in \operatorname{int} C$ by (i) and thus $y = \lim_{\alpha \to 0} y_\alpha \in \overline{\operatorname{int} C}$.

∎

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

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 $\operatorname{int} C \neq \emptyset$ and $x \in C$. Then

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

Proof.

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

- One can show: if $D \subset H$ is convex with $\operatorname{int} D \neq \emptyset$ and $0 \in D$, then $[0 \in \operatorname{int} D] \Leftrightarrow [\overline{\operatorname{cone} D} = H]$.
- Sketch: assume $0 \in \operatorname{int} D$. Then $\overline{\operatorname{cone} D} = \operatorname{cone} D = H$ 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 $N_C x = \{u \in H : \sup \langle C - x, u \rangle \leq 0\}$. We can extend the supremum to $\operatorname{cone}(C - x)$ and we can then extend it to the closure $\overline{\operatorname{cone}(C - x)}$ without changing whether it will be ≤ 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 $\operatorname{spts} C$.

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

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

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

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

$\square$

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

$$\operatorname{bdry} C \subset \operatorname{spts} \overline{C} \qquad \text{and} \qquad C \cap \operatorname{bdry} C \subset \operatorname{spts} C\,.$$

Proof.
- If $C = H$ the result is clear. (Why?) So assume $C \neq H$.
- Let $x \in \operatorname{bdry} C \subset \overline{C}$. So $x \in \overline{C} \setminus \operatorname{int} C = \overline{C} \setminus \operatorname{int} \overline{C}$ (Prop. 1.54).
- Consequence of Prop. 1.56: $\exists u \in N_{\overline{C}} x \setminus \{0\}$.
- Consequence of Prop. 1.58: $x \in \operatorname{spts} \overline{C}$. Therefore $\operatorname{bdry} C \subset \operatorname{spts} \overline{C}$.
- Show $\operatorname{spts} C = C \cap \operatorname{spts} \overline{C}$: For this use $\sup \langle \overline{C}, u \rangle = \sup \langle C, u \rangle$ (why?).

- Let $x \in \operatorname{spts} C$: $\Rightarrow x \in C \subset \overline{C}$, $\exists\, u \neq 0$ s.t. $\sup \langle C, u\rangle \leq \langle x, u\rangle$. $\Rightarrow x \in C \cap \operatorname{spts} \overline{C}$.
- Let $x \in \operatorname{spts} \overline{C} \cap C$: $\Rightarrow x \in C$, $\exists\, u \neq 0$ s.t. $\sup \langle \overline{C}, u\rangle \leq \langle x, u\rangle$. $\Rightarrow x \in \operatorname{spts} C$.
- So: $C \cap \operatorname{bdry} C \subset C \cap \operatorname{spts} \overline{C} = \operatorname{spts} C$.

$\square$

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

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 $f : H \to \mathbb{R} \cup \{\infty\}$ be proper and convex and let $x \in \operatorname{dom} f$. If $x \in \operatorname{cont} f$ then $\partial f(x) \neq \emptyset$.

Proof.

- Since f is proper and convex, $\operatorname{epi} f$ is non-empty and convex.
- Since $x \in \operatorname{cont} f$, 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, $\operatorname{int} \operatorname{epi} f \neq \emptyset$ 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 $y_k \notin \operatorname{epi} f$ but $\lim_{k\to\infty} y_k = (x, f(x) \in \operatorname{epi} f$. Therefore $(x, f(x)) \in \operatorname{bdry} \operatorname{epi} f$.
- So by Proposition 1.59 there is some $(u, r) \in N_{\operatorname{epi} f}(x, f(x)) \setminus \{(0, 0)\}$.
- By definition of normal cone: For every $(v, s) \in \operatorname{epi} f$ 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 $y \in \operatorname{dom} f$ have $(y, f(y)) \in \operatorname{epi} f$ 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:

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

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

$\square$

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

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

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 $[x \notin \operatorname{dom} f] \Rightarrow [\partial f(x) = \emptyset]$.

□

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

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

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 $x \in \operatorname{spts} C$ 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 $u \in H$

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

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

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 $\operatorname{conv} C \subset D$. Since D is closed, also $\overline{\operatorname{conv} C} \subset D$.
- Now, let $x \in D$, set $p = P_{\overline{\operatorname{conv} C}} x$.
- Then $\langle x - p, y - p\rangle \leq 0$ for all $y \in \overline{\operatorname{conv} C}$ and thus $\sigma_{\overline{\operatorname{conv} C}}(x - p) = \sup \left\langle \overline{\operatorname{conv} C}, x - p \right\rangle = \langle p, x - p\rangle$.
- Moreover, $x \in D \subset A_{x-p}$. So $\langle x, x - p\rangle \leq \sigma_C(x - p)$.
- Since $C \subset \overline{\operatorname{conv} C}$ we get $\sigma_C \leq \sigma_{\overline{\operatorname{conv} C}}$.
- Now: $\|x - p\|^2 = \langle x, x - p\rangle - \langle p, x - p\rangle \leq \sigma_C(x - p) - \sigma_{\overline{\operatorname{conv} C}}(x - p) \leq 0$. Therefore $x = p \subset \overline{\operatorname{conv} C}$ and thus $D \subset \overline{\operatorname{conv} C}$.

□

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?

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

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

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

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

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

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

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

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

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

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

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

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

(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, $f_k : H_k \to [-\infty, \infty]$, let $H = \bigotimes_{k=1}^n H_k$, $f : H \to [-\infty, \infty]$, $((x_k)_k) \mapsto \sum_{k=1}^n f_k(x_k)$. Then $f^*((u_k)_k) = \sum_{k=1}^n f_k^*(u_k)$.

Proof. The proof is completely analogous to Example 1.69, (iv). □

A few simple 'transformation rules':

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

(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. □

Proposition 1.72 (Fenchel–Young inequality)**.** Let $f : H \to \mathbb{R} \cup \{\infty\}$ 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 $H \to [-\infty, \infty]$. Set $f : H \to [-\infty, \infty]$, $x \mapsto \sup_{i \in I} f_i(x)$. Then:

(i) $\operatorname{epi} f = \cap_{i \in I} \operatorname{epi} f_i$

(ii) If all f_i are lower semicontinuous, so is f.

(iii) If all f_i are convex, so is f.

Proof.

- **(i):** $[(x, r) \in \operatorname{epi} f] \Leftrightarrow [\mathbb{R} \ni r \geq f(x)] \Leftrightarrow [\mathbb{R} \ni r \geq f_i(x)$ for all $i \in I] \Leftrightarrow [(x, r) \in \operatorname{epi} f_i$ for all $i \in I] \Leftrightarrow [(x, r) \in \bigcap_{i \in I} \operatorname{epi} f_i]$.
- **(ii):** If all f_i are lower semicontinuous, all $\operatorname{epi} f_i$ are closed (Prop. 1.21). Then $\operatorname{epi} f = \bigcap_{i \in I} \operatorname{epi} f_i$ is closed, i.e. f is lower semicontinuous.
- **(iii):** If all f_i are convex, all $\operatorname{epi} f_i$ are convex (Prop. 1.9). Then $\operatorname{epi} f = \bigcap_{i \in I} \operatorname{epi} f_i$ is convex (Prop. 1.4), i.e. f is convex.

$\square$

Proposition 1.74 (Basic properties of conjugate)**.** Let $f : H \to [-\infty, \infty]$. 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: $f^*(u) = \sup_{x \in \operatorname{dom} f} \langle u, x \rangle - f(x)$.
- So conjugate is pointwise supremum over family of convex, lower semicontinuous functions: $(y \mapsto \langle y, x \rangle - f(x))_{x \in \operatorname{dom} f}$.
- 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 $f : H \to [-\infty, \infty]$.

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

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

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

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

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

Proof. This follows directly from Prop. 1.73. □

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

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

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

- For some $(y, \eta) \in D$, send $\eta \to \infty$ (which is still in D, why?). We deduce: $\zeta - \pi \leq 0$.
- Note that $(y, \eta) \in D \Rightarrow y \in \overline{\text{conv}\, \text{dom}\, f}$. (Details: any $(y, \eta) \in D = \overline{\text{conv}\, \text{epi}\, f}$ can be written as limit of sequence $(y_k, \eta_k)_k$ in $\text{conv}\, \text{epi}\, f$. Any (y_k, η_k) can be written as finite convex combination of some $(y_{k,i}, \eta_{k,i})_i$ in $\text{epi}\, f$. So all $y_{k,i} \in \text{dom}\, f$ and thus the convex combination $y_k \in \text{conv}\, \text{dom}\, f$ and therefore the limit $y \in \overline{\text{conv}\, \text{dom}\, f}$.)
- Also note: $\text{dom}\, F \subset \overline{\text{conv}\, \text{dom}\, f} = E$: 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 $\text{dom}\, F \subset E$.
- Assume $\zeta = \pi$. Then projection characterization yields: $\langle x - p, y - p \rangle \leq 0$ for all $y \in \overline{\text{conv}\, \text{dom}\, f}$. Since $[(x, \zeta) \in \text{epi}\, F] \Rightarrow [x \in \text{dom}\, F \subset \overline{\text{conv}\, \text{dom}\, f}]$ we may set $y = x$ 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 $\eta = f(y)$. Then from characterization pf projection get:

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

  Once more, set $y = x$ and use $\zeta \geq f(x)$ to get

$$\left[\zeta \geq \pi + \left\langle x - p, \tfrac{x-p}{\pi-\zeta} \right\rangle\right] \Leftarrow \left[-(\pi - \zeta)^2 \geq \|x - p\|^2\right] .$$

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

□

Now some basic properties of the biconjugate.

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

Proof.

- We find:

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

- 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.

□

We now prove the main result of this subsection.

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

Proof.

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

- Let $(v, r) \in H \times \mathbb{R}$ be the outward normal of a closed halfspace that contains $\operatorname{epi} F$. If $r > 0$ then $\operatorname{epi} F = \emptyset$ and then $f = +\infty = f^{**}$ and we are done.

- So assume that $\operatorname{epi} F \neq \emptyset$ and therefore $r \leq 0$ for all closed halfspaces that contain $\operatorname{epi} F$.

- Similarly, f^{**} is the pointwise supremum over all continuous affine lower bounds on f. Therefore, $\operatorname{epi} f^{**}$ is the intersection of all closed halfspaces that contain $\operatorname{epi} f^{**}$ and for which the outward normal (v, r) has $r < 0$.

- Therefore, $\operatorname{epi} F \subset \operatorname{epi} f^{**}$ 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 $(z, \zeta) \in \operatorname{epi} f^{**} \setminus \operatorname{epi} F$.

- Then there must be a closed halfspace in $H \times \mathbb{R}$ with horizontal outward normal (i.e. $r = 0$) that contains $\operatorname{epi} F$, but not (z, ζ). That is, there is some $(v, y) \in H^2$ such that $\langle x - y, v \rangle \leq 0$ for all $x \in \operatorname{dom} F$ but $\langle z - y, v \rangle > 0$.

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

- 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 $x \in \operatorname{dom} f \subset \operatorname{dom} F$ (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 $s \to \infty$ have $g_s(z) = g_0(z) + s \cdot \langle z - y, v \rangle \to \infty > \zeta \geq f^{**}(z)$.
- This is a contradiction, thus points like (z, ζ) cannot exist and $\operatorname{epi} f^{**} = \operatorname{epi} F$.

□

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

Corollary 1.80 (Fenchel–Moreau)**.** Let $f : H \to \mathbb{R} \cup \{\infty\}$ 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. $\Rightarrow$ $\operatorname{epi} f$ is convex, closed. $\Rightarrow$ it is intersection of all closed halfspaces that contain $\operatorname{epi} f$. 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 $f^{**} = \overline{\operatorname{conv} f}$ 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 $f^*(u) \geq \langle x_0, u \rangle - f(x_0)$.

□

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 ∂f or f^* onto the other. It gives the 'extreme cases' of the Fenchel–Young inequality.

Proposition 1.81. Let $f : H \to \mathbb{R} \cup \{\infty\}$ be convex, lsc. 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.

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

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

  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 $f : H \to \mathbb{R} \cup \{\infty\}$ 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 $f : H \to \mathbb{R} \cup \{\infty\}$. 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.
- **⇐:** ι_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 ι_C^* is positively 1-homogeneous.
- **⇒:** 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

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

- 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 have $f = f^{**} = \iota_C^*$.

□

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

Corollary 1.84. If $f : H \to \mathbb{R} \cup \{\infty\}$ 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 $x_a \rightharpoonup x$, if for all $u \in H$

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

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 $\langle u, x_k\rangle \to \langle u, x\rangle$ 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 $N \to \infty$) and $\langle u, x_k\rangle \to 0$ as $k \to \infty$. Therefore $x_k \rightharpoonup 0$. (But clearly not $x_k \to 0$.)

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:

$$[x_k \to x] \qquad \Leftrightarrow \qquad [x_k \rightharpoonup x \text{ and } \|x_k\| \to \|x\|]$$

Proof.

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

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

□

Remark 1.93. In the previous example we find indeed $\lim_{k\to\infty} \|x_k\| = 1 \neq \|0\|$. 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).

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: $[x_k \to x] \Rightarrow [x_k \rightharpoonup x]$. 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 $x \neq C$. 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.)

□

Corollary 1.97. For a convex function $f : H \to \mathbb{R} \cup \{\infty\}$ 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. □

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 $\|x_k\| \to 1$.

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 $x_k \rightharpoonup 0 \in B(0,1)$.

□

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 $f : H \to \mathbb{R} \cup \{\infty\}$ 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 bonded.

- 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.

□

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

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

Proposition 1.104. Let $f : H \to [-\infty, \infty]$. 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 $\|x_k\| \to \infty$ but $\limsup f(x_k) \leq r$.
- Assume $S_r(f)$ is bounded for every $r \in \mathbb{R}$. Let $(x_k)_k$ be an unbounded sequence with $\lim \|x_k\| \to \infty$. 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, $\liminf f(x_k) \geq s$. Since this holds for any $s \in \mathbb{R}$, have $\lim f(x_k) = \infty$.

□

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. □

1.7 Proximal operators

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

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

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 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 Prox_f *is well-defined.*

- Since f is convex and lsc $\Rightarrow$ 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

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

  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.

□

Characterization of proximal operator.

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

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

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:

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

- 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 $\partial_\alpha g(\alpha)|_{\alpha=0} \geq 0$. 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

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

□

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

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

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

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

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:

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

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 $x = p + \frac{\lambda \cdot p}{\|p\|}$ 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:

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

We summarize:

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

Interpretation: if $\|x\| > \lambda$ we move towards the origin with stepsize λ, 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 $x^{(0)} \in H$:

$$x^{(\ell+1)} \stackrel{\text{def.}}{=} x^{(\ell)} - \lambda \nabla f(x^{(\ell)})$$

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

$$y^{(\ell+1)} \stackrel{\text{def.}}{=} \mathrm{Prox}_{\lambda f}(y^{(\ell)})$$

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 $\operatorname{dom} f$.
- 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 $x^{(\ell)} \to 0$ 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 λ 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

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

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

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

(i) $[x \in \operatorname{argmin} f] \Leftrightarrow [x = \operatorname{Prox}_f(x)]$.

(ii) $[x \notin \operatorname{argmin} f] \Rightarrow [f(\operatorname{Prox}_f(x)) < f(x)]$.

(iii) Let $p = \operatorname{Prox}_f(x)$, $C = S_{f(p)}(f)$. Then $p = P_C x$.

(iv) The function $\lambda \mapsto \|x - \operatorname{Prox}_{\lambda f}(x)\|$ is increasing.

(v) The function $\lambda \mapsto f(\operatorname{Prox}_f(x))$ is decreasing.

Proof.

- **(i):** Assume $x \in \operatorname{argmin} f$. Then for all $p \in H$

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

Therefore, $x = \operatorname{Prox}_f(x)$.

- Conversely, assume $x = \operatorname{Prox}_f(x)$. $\Rightarrow [x \in x + \partial f(x)] \Rightarrow [0 \in \partial f(x)] \Rightarrow$ (Fermat's rule, Prop. 1.26) $[x \in \operatorname{argmin} f]$.

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

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

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

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

and therefore p' is a better candidate for $\operatorname{Prox}_f(x)$ 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 $p_i = \operatorname{Prox}_{\lambda_i f}(x)$ 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:

$$0 \le \left\langle \tfrac{u_2}{\lambda_2} - \tfrac{u_1}{\lambda_1}, p_2 - p_1 \right\rangle$$
$$0 \le \left\langle u_2 - \tfrac{\lambda_2}{\lambda_1} u_1, \Delta p \right\rangle = \left\langle \Delta u - \tfrac{\lambda_2 - \lambda_1}{\lambda_1} u_1, \Delta p \right\rangle = -\|\Delta p\|^2 - \tfrac{\lambda_2-\lambda_1}{\lambda_1} \langle u_1, \Delta p\rangle$$

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

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

- **(v):** Use notation from previous point. Assume $f(p_2) > f(p_1)$, let $C = S_{f(p_2)}(f)$. Then $p_1 \ne p_2$ 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 $f(p_2) \le f(p_1)$.

□

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 $f : H \to \mathbb{R} \cup \{\infty\}$ be convex, lsc and proper, $x \in H$. Then $\mathrm{Prox}_f(x) + \mathrm{Prox}_{f^*}(x) = x$.

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

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

□

Example 1.114 (Moreau decomposition for projections)**.** Let C be a closed subspace of H. Then ι_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 ι_C^* is the indicator of the orthogonal complement of C. Then $\mathrm{Prox}_{\iota_C} = P_C$ and $\mathrm{Prox}_{\iota_C^*} = P_{C^\perp}$ 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 $x^{(\ell+1)} = \mathrm{Prox}_f(x^{(\ell)})$ we now find that $x^{(\ell+1)} + \mathrm{Prox}_{f^*}(x^{(\ell)}) = x^{(\ell)}$, $\Rightarrow x^{(\ell+1)} = x^{(\ell)} - \mathrm{Prox}_{f^*}(x^{(\ell)})$, so Prox_{f^*} gives the 'implicit gradient steps' $\Delta x^{(\ell+1)}$.
Let $f(x) = \|x\|$. Then $f^* = \iota_{\overline{B(0,1)}}$ and

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

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 iterator: 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 $f : H \to \mathbb{R} \cup \{\infty\}$ 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 $\lim_{k\to\infty} \|x_k - y\| = \inf_k \|x_k - y\|$.
- **(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.

□

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.

□

Remark 1.120. One can in fact show a slightly stronger result: [$(x_k)_k$ converges weakly] $\Leftrightarrow$ [$(x_k)_k$ 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 $r \in \mathbb{R}$.
- Further, by weak convergence of the two extracted subsequences we find

$$\lim_k \langle x_{i_k}, x - y\rangle = \langle x, x - y\rangle, \qquad \lim_k \langle x_{j_k}, x - y\rangle = \langle y, x - y\rangle.$$

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

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

  and therefore the two cluster points must coincide.

□

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.

□

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 $\ell = 0, \ldots, N$ we obtain:

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

- **(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 $\lim_\ell f(x^{(\ell)}) = f(z)$ 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.

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 $\ell = 0, 1, \ldots$:

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

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

(i) $\operatorname{Prox}_{\gamma g}(x) \in \operatorname{argmin}(f + g)$.

(ii) $y^{(\ell)} - z^{(\ell)} \to 0$ strongly.

(iii) $x^{(\ell)} \rightharpoonup x$ weakly.

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

We start by going through an explicit example.

Example 1.126. Let $H = \mathbb{R}^2$. $f(x) = \frac{1}{2}\|x\|^2$, $C = \{x \in H \colon x_1 = 1\}$, $g(x) = \iota_C(x)$. Then $\operatorname{Prox}_{\gamma f}(x) = \frac{1}{1+\gamma}x$. Further $\operatorname{Prox}_{\gamma g}(x) = P_C x$. 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

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

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:

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

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

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

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

$$r = r \cdot (1 - \tfrac{\lambda}{1+\gamma}) + \tfrac{\lambda\,(1-\gamma)}{1+\gamma} \quad \Leftrightarrow \quad r = 1-\gamma$$

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

Remark 1.127 (Interpretation of Douglas–Rachford algorithm)**.** The goal is to find a minimizer $y = z$ and an offset vector Δx 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. $y = \operatorname{Prox}_{\gamma g}(y + \Delta x) = \operatorname{Prox}_{\gamma f}(y - \Delta x)$. Unless $\operatorname{argmin} f \cap \operatorname{argmin} g \neq \emptyset$ there can be no such point for $\Delta x = 0$, which is why mere alternating application of $\operatorname{Prox}_{\gamma f}$ and $\operatorname{Prox}_{\gamma g}$ 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 $x_2^{(\ell)}$-update depends on whether $x_2^{(\ell)}$ 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 $\operatorname{argmin} f$ contains more than one element, then Prox_f has more than one fixed-point (Prop. 1.112 (i)). So 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

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

(ii) *firmly nonexpansive* if

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

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

(i) Prox_f is firmly nonexpansive.

(ii) $\mathrm{id} - \operatorname{Prox}_f$ is firmly nonexpansive.

(iii) $2\operatorname{Prox}_f - \mathrm{id}$ is nonexpansive.

Proof.
- **(i)** Use monotonicity of subdifferential. Let $x, y \in H$, set $p = \operatorname{Prox}_f(x)$, $q = \operatorname{Prox}_f(y)$. 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:

$$[\langle (x-p) - (y-q), p - q\rangle \geq 0] \Leftrightarrow [\langle \Delta x - \Delta p, \Delta p\rangle \geq 0] \Leftrightarrow [2\|\Delta p\|^2 - 2\langle \Delta p, \Delta x\rangle \leq 0]$$
$$\Leftrightarrow [\|\Delta p\|^2 + \|\Delta p\|^2 - 2\langle \Delta p, \Delta x\rangle + \|\Delta x\|^2 \leq \|\Delta x\|^2] \Leftrightarrow [\|\Delta p\|^2 + \|\Delta p - \Delta x\|^2 \leq \|\Delta x\|^2]$$

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

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

□

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

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

Proof.

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

(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 $x_k \rightharpoonup x$, $x_k - T(x_k) \to 0$ and $(x_k)_k$ bounded, i.e. $\|x_k\| \leq C_1$ for some $C_1 < +\infty$. Further, since T is nonexpansive: $\|T(x_k)\| = \|T(x_k) - T(0) + T(0)\| \leq \|T(x_k) - T(0)\| + \|T(0)\| \leq C + \|T(0)\| \stackrel{\text{def.}}{=} C_2$. Then

$$\|x_k - T(x_k)\|^2 \to 0,$$
$$\limsup_k |\langle T(x) - T(x_k), x_k - T(x_k)\rangle| \leq \limsup_k (\|T(x)\| + C_2) \cdot \|x_k - T(x_k)\| = 0$$

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

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

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. $T(x) = x$.

□

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

$$\begin{aligned}
\operatorname{Fix} T &= \{x \in H \colon T(x) = x\} \text{ the set of fixed-points of } T. \\
\operatorname{zer} T &= \{x \in H \colon T(x) = 0\} \text{ the set of 'roots' of } T.
\end{aligned}$$

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

$$\operatorname{zer} T = \{x \in H \colon 0 \in T(x)\}.$$

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 $T : H \to H$ be nonexpansive, let $\operatorname{Fix} T \neq \emptyset$, $\lambda \in (0,1)$ 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 $\operatorname{Fix} T$.

(ii) $(T(x^{(\ell)}) - x^{(\ell)})_\ell$ converges strongly to 0.

(iii) $(x^{(\ell)})_\ell$ converges weakly to a point in $\operatorname{Fix} T$.

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 $y \in \operatorname{Fix} T$. 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 $\operatorname{Fix} T$.

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

- **(iii)** $(x^{(\ell)})_\ell$ is bounded due to Fejér monotonicity (Prop. 1.118). From (ii) we have $T(x^{(\ell)}) - x^{(\ell)} \to 0$. So by the previous result, Prop. 1.130, we obtain that any weak sequential cluster point of $(x^{(\ell)})_\ell$ is in $\operatorname{Fix} T$.

- It follows then from Cor. 1.122 that $(x^{(\ell)})_\ell$ converges weakly to a point in $\operatorname{Fix} T$.

□

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

$$R_f = 2\operatorname{Prox}_f - \mathrm{id}, \qquad R_g = 2\operatorname{Prox}_g - \mathrm{id}\,.$$

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

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:

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

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

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

□

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

Proof of Prop. 1.124.

- Let

$$R_f = 2\operatorname{Prox}_{\gamma f} - \mathrm{id}, \qquad R_g = 2\operatorname{Prox}_{\gamma g} - \mathrm{id}, \qquad T = R_f\, R_g.$$

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

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

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

- Apply Prop. 1.133 to γf and γg to obtain

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

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

- In view of **(i)** we note that for every $x \in \operatorname{Fix} T$ have therefore $\operatorname{Prox}_{\gamma g}(x) \in \operatorname{argmin}(f + g)$.

- 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 $z^{(\ell)} - y^{(\ell)} = \frac{1}{2}(T(x^{(\ell)}) - x^{(\ell)}) \to 0$ strongly.

- From Prop. 1.132 (iii) we get: there is some $x \in \operatorname{Fix} T$ such that $x^{(\ell)} \rightharpoonup x$ weakly. Since $\operatorname{Prox}_{\gamma g}(x) \in \operatorname{argmin}(f + g)$ (see above) this establishes **(i)** and **(iii)**.

□

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 $f, g : H \to \mathbb{R} \cup \{\infty\}$ 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, $0 \notin \operatorname{int} C$, $\operatorname{int} C \neq \emptyset$. 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 $\operatorname{int} D = \operatorname{int} C \neq \emptyset$ and in particular $0 \notin C \supset \operatorname{int} D$.
- 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 $A_u = \{x \in H : \langle u, x\rangle - r \geq 0\}$, $u \neq 0$, 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 $0 \notin \operatorname{int} D$. 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:

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

□

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

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

- A and B are convex. Since f is continuous in x_0 and $f(x_0) < \infty$, we have $\operatorname{int} A \neq \emptyset$.
- 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 $0 \notin \operatorname{int}(A - B)$. $A - B$ is convex. Also, $\emptyset \neq \operatorname{int} A - B \subset \operatorname{int}(A - B)$. 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 $\lambda \to \infty$ and $\mu \to -\infty$ (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.

□

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.

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

Proposition 1.138. 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

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

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

Remark 1.139. 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].

Proposition 1.140. 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:)

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

□

Remark 1.141. The strategy for the convergence proof of Prop. 1.138 is as follows: we show that the optimality condition of Prop. 1.140 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. 1.138. We then generalize the original proximal optimization algorithm (Prop. 1.116) to firmly nonexpansive operators.

Proposition 1.142. Let $A : H \to 2^H$ be monotone. Let $M : H \to H$ 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 $T : K \to H$ 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) $\operatorname{Fix} T = \operatorname{zer} A$.

(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)**

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

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

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

  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.

Proposition 1.143. Assume $T : H \to H$ is firmly nonexpansive and $\operatorname{Fix} T \neq \emptyset$. For $x^{(0)} \in H$ set

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

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

Proof.

- Let $z \in \operatorname{Fix} T$. 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 $\operatorname{Fix} T$.
- Further $\sum_{\ell=0}^N \|T(x^{(\ell)}) - x^{(\ell)}\|^2 \leq \|x^{(0)} - z\|^2$. Therefore $T(x^{(\ell)}) - x^{(\ell)} \to 0$ 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 $x \in \operatorname{Fix} T$.

$\square$

Proof of Proposition 1.138.

- Define set valued operator $A : H \times H \to 2^H \times 2^H$ as follows:

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

- By Prop. 1.140 the primal and dual optimizers are given precisely by $\operatorname{zer} A$, and therefore by assumption $\operatorname{zer} A \neq \emptyset$.
- 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 $H \times H \to 2^{H\times H}$.

Sketch: $2^H \times 2^H$ vs $2^{(H\times H)}$ for $H = \mathbb{R}$: product of two intervals vs general 2d sets.

- 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 Δx, Δy, Δa, Δb all pairwise differences. Then by monotonicity of ∂f and ∂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 $M : H^2 \to H^2$ as

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

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

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

- Now analyze operator T constructed from A and M via Prop. 1.142: From $(a, b) = T(x, y)$ we obtain the following inclusion condition:

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

  So the iterations of the algorithm, Prop. 1.138, 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. 1.142 (ii) have $\operatorname{Fix} T = \operatorname{zer} A \neq \emptyset$.
- 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. 1.142 (iii). So by Prop. 1.143 the sequence $(x^{(\ell)}, y^{(\ell)})_\ell$ converges weakly to some $(x, y) \in \operatorname{Fix} T = \operatorname{zer} A$, 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 1.144.
- Prop. 1.143 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. 1.142 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.

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 Γ-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.

Definition 2.2. Throughout this subsection V is a real vector space.

2.1 Foundations

Definition 2.3 (Norm and inner product).

- A map $\|\cdot\| : V \to \mathbb{R}_+$ on V is a *norm* if for all $x, y \in V$, $\lambda \in \mathbb{R}$ have
  (i) (positive definite) $\|x\| = 0 \Rightarrow x = 0$,
  (ii) (subadditive) $\|x + y\| \leq \|x\| + \|y\|$,
  (iii) (homogeneity) $\|\lambda \cdot x\| = |\lambda| \cdot \|x\|$.
- A norm induces a metric on V via $d(x, y) = \|x - y\|$.
- A map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ is an *inner product* on V if for all $x, y, z \in V$, $\lambda \in \mathbb{R}$ it satisfies
  (i) (symmetry) $\langle x, y\rangle = \langle y, x\rangle$,
  (ii) (linearity in first argument) $\langle \lambda \cdot x, y\rangle = \lambda \cdot \langle x, y\rangle$, $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$,
  (iii) (positive definite) $\langle x, x\rangle \geq 0$, $[\langle x, x\rangle = 0] \Rightarrow [x = 0]$.
- From the axioms for the inner product we quickly obtain the *Cauchy-Schwarz inequality*, $\langle x, y\rangle \leq \|x\| \cdot \|y\|$, and that an inner product induces a norm via $\|x\| = \sqrt{\langle x, x\rangle}$.
- A vector space V with a norm $\|\cdot\|$ is called a *normed space*. A vector space with an inner product $\langle \cdot, \cdot \rangle$ is called *pre-Hilbert space*.

Definition 2.4 (Convergence in metric). We say a sequence $(x_k)_k$ on a metric space (X, d) converges to some $x \in X$ if $d(x_k, x) \to 0$. x is called the *limit* of $(x_k)_k$ and is unique.

Definition 2.5 (Cauchy sequence and complete metric spaces). A sequence $(x_k)_k$ on a metric space (X, d) is a Cauchy sequence if for all $\varepsilon > 0$ there is some $N \in \mathbb{N}$ such that $d(x_i, x_j) \leq \varepsilon$ whenever $i, j \geq N$. A metric space is called *complete* if all Cauchy sequences converge.

Definition 2.6 (Banach and Hilbert spaces). A complete normed space is called *Banach space*. A complete pre-Hilbert space is called *Hilbert space*.

Definition 2.7.

- A finite set $\{x_1, \ldots, x_n\} \subset V$ 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 $\operatorname{span}\{x_1, \dots, x_n\} = \{\sum_{i=1}^n \alpha_i \cdot x_i \,|\, (\alpha_i)_{i=1}^n \in \mathbb{R}^n\}$.
- A basis of V is a linearly independent finite set $X \subset V$ such that $\operatorname{span} X = V$.
- The dimension of V is the cardinality of any basis. If no basis exists, the dimension is ∞.
- On an ∞-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

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

Remark 2.8 (Convexity in Banach spaces)**.** The notions of 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.

2.2 Reminders on topology

We recall a few basic facts about topologies.

Definition 2.9. Let X be a set. A set $T \subset 2^X$ of subsets of X is called a *topology* for X if

(i) $X, \emptyset \in T$,

(ii) For arbitrary subsets $S \subset T$ their union is also in T, i.e.

$$\bigcap_{s\in S} s \in T.$$

(iii) For finite subsets $S = \{s_1, \dots, s_n\} \subset T$ their intersection is also in T, i.e.

$$\bigcap_{i=1}^n s_i \in T.$$

Sets in T are called *open*. A set is *closed*, if its complement in X is open. The set X with a corresponding topology T is called *topological space*. If not required, we may drop the explicit reference to T.

Remark 2.10. The restriction that only finite intersections of open sets are open is necessary. Recall basic example from $X = \mathbb{R}$: for every $\varepsilon > 0$ the set $(-\varepsilon, \varepsilon)$ is open. But

$$\bigcap_{n=1}^\infty (-1/n, 1/n) = \{0\}$$

which is not open.

Proposition 2.11 (de Morgan's law)**.** Let $(A_i)_{i\in I}$ be a set of subsets of X where I is an arbitrary index set. Then

$$\bigcap_{i\in I} A_i = X \setminus \big(\bigcup_{i\in I} (X \setminus A_i)\big)\,.$$

Proof.

$$[x \in \bigcap_{i\in I} A_i] \Leftrightarrow [x \in A_i \,\forall\, i \in I] \Leftrightarrow [x \notin X \setminus A_i \,\forall\, i \in I]$$
$$\Leftrightarrow [x \notin \bigcup_{i\in I}(X \setminus A_i)] \Leftrightarrow [x \in X \setminus \big(\bigcup_{i\in I}(X \setminus A_i)\big)]$$

□

Corollary 2.12. With this we quickly find that arbitrary intersections of closed sets are closed and finite unions of closed sets are closed.

Definition 2.13. Let (X, T) be a topological space. The interior $\operatorname{int} A$ of a set $A \subset X$ is the union of all open sets contained in A:

$$\operatorname{int} A = \bigcup \{U \in T : U \subset A\}$$

Since this set is by construction open and contains all open sets contained in A it is also referred to as 'largest open set contained in A'. Similarly, the closure $\operatorname{cl} A$ of A is the intersection of all closed sets that contain A.

Definition 2.14 (Convergence of sequences)**.** A sequence $(x_k)_k$ in a topological space (X, T) is said to converge to a point $x \in X$ if for any any $t \in T$ with $x \in t$ there is some $N \in \mathbb{N}$ such that $x_k \in t$ for $k \geq N$.

Definition 2.15 (Continuous maps)**.** Let (X, T), (Y, S) be two topological spaces and let $f : X \to Y$. f is called *continuous* if $f^{-1}(s) \in T$ for all $s \in S$. (*'Preimages of open sets are open.'*)

Proposition 2.16. 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 2.17 (Induced topology)**.** Let $(f_i : X \to Y_i)_{i\in I}$ 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.

Comment: Transition from metric to metric topology often simply 'stated as fact'. Take some time to re-check.

Proposition 2.18 (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 ε 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.

Proof.

- Note that $B(x, \varepsilon) = d(x, \cdot)^{-1}((-\varepsilon, \varepsilon))$. Since $(-\varepsilon, \varepsilon)$ is open in $\mathbb{R}$, $B(x, \varepsilon)$ is therefore open in the metric topology by construction.

- The intersection of two open balls can be written as union of open balls: For $x_1, x_2 \in X$, $r_1, r_2 > 0$ and $x \in B(x_1, r_1) \cap B(x_2, r_2)$ set $\varepsilon_i = r_i - d(x_i, x)$. Let $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}/2$. Then for $z \in B(x, \varepsilon)$ have $d(z, x_i) \le d(z, x) + d(x, x_i) < \varepsilon + r_i - \varepsilon_i \le \varepsilon_i/2 + r_i - \varepsilon_i = r_i - \varepsilon_i/2$. So $z \in B(x_i, r_i)$ and $B(x, \varepsilon) \subset B(x_i, r_i)$.

- More generally, for any $y \in B(x_1, r_1) \cap B(x_2, r_2)$ denote by $\varepsilon_y > 0$ a positive radius such that $B(y, \varepsilon_y) \subset B(x_1, r_1) \cap B(x_2, r_2)$. Then

$$B(x_1, r_1) \cap B(x_2, r_2) = \bigcup_{y \in B(x_1, r_1) \cap B(x_2, r_2)} B(y, \varepsilon_y)$$

- Finite intersections of arbitrary unions of open balls can be written as unions of open balls. We start with the intersection of two unions. The rest follows by induction. Let

$$S = \left(\bigcup_{i \in I} A_i \right) \cap \left(\bigcup_{j \in J} B_j \right)$$

  where I and J are some index sets and $(A_i)_{i \in I}$, $(B_j)_{j \in J}$ are families of open balls. Then $x \in S \Leftrightarrow \exists\, i_x \in I$, $j_x \in J$ such that $x \in A_{i_x} \cap B_{j_x} \subset S$. So

$$S = \bigcup_{x \in S} (A_{i_x} \cap B_{j_x}) \, .$$

  By the previous point we can rewrite the intersection $A_{i_x} \cap B_{j_x}$ of two open balls as union of open balls, therefore we can rewrite S as union of open balls.

- Now set $T = \{\bigcup_{i \in I} B_i |$ index sets I, open balls $(B_i)_{i \in I}\}$ be the set of unions of open balls. We set by convention that $X, \emptyset \in T$. Then T is a topology (see previous point for the intersection property) and it contains all open balls. Therefore, the metric topology is a subset of T.

- The metric topology must contain at least all open balls, and therefore all arbitrary unions thereof. So T is contained in the metric topology. Therefore, the two coincide.

□

Corollary 2.19. 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 $x_i \in A_i$, $A_1 \cap A_2 = \emptyset$.

Proof. Since $x_1 \neq x_2$ have $d(x_1, x_2) > 0$. Set $A_i = B(x_i, d(x_1, x_2)/3)$. □

Corollary 2.20. Let $A \subset X$ be open. Then for every $x \in A$ there is some $\varepsilon > 0$ such that $B(x, \varepsilon) \subset A$.

Proof. By Prop. 2.18 A can be written as union of open balls. Therefore, we must have some $y \in A$, $\delta > 0$ such that $x \in B(y, \delta) \subset A$. Set $\varepsilon = (\delta - d(x, y))/2$. By triangle inequality $B(x, \varepsilon) \subset B(y, \delta) \subset A$. □

Corollary 2.21 (Convergence in metric topology)**.** Convergence in the metric (Def. 2.4) is equivalent to convergence in the metric topology (Prop. 2.18).

Proof.

- Assume $d(x_k, x) \to 0$. For any open set U containing x, by Cor. 2.20, there is some $\varepsilon > 0$ such that $B(x, \varepsilon) \subset U$ and consequentially some $N \in \mathbb{N}$ such that $d(x_k, x) < \varepsilon$ for $k > N$, i.e. $x_k \in U$. This implies convergence in metric topology.
- Assume convergence in metric topology. Then for any $\varepsilon > 0$ there is some $N \in \mathbb{N}$ such that $x_k \in B(x, \varepsilon)$ for $k > N$, i.e. $d(x_k, x) < \varepsilon$. This implies $d(x_k, x) \to 0$.

□

2.3 Examples

Let $\Omega \subset \mathbb{R}^n$ be open, bounded, non-empty. For a *'multiindex'* $a \in \{0, 1, 2, \ldots\}^n$ let $|a| = a_1 + \ldots + a_n$ and $D^a f \stackrel{\text{def.}}{=} \frac{\partial^{|a|}}{\partial_{x_1}^{a_1} \ldots \partial_{x_n}^{a_n}} f$.

Definition 2.22. 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

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

Remark 2.23. 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 Ω is bounded).

Proof.

- Start with $C^0(\overline{\Omega})$. $\|f\|_{C^0(\overline{\Omega})} = \max_{x \in \overline{\Omega}} |f(x)|$.
- $\|f\|_{C^0(\overline{\Omega})}$ is a norm: finiteness, positive definiteness and homogeneity are immediate. Check subadditivity: $\|f+g\|_{C^0(\overline{\Omega})} = \max_{x \in \overline{\Omega}} |f(x)+g(x)| \leq \max_{x,y \in \overline{\Omega}} |f(x)|+|g(y)| = \|f\|_{C^0(\overline{\Omega})} + \|g\|_{C^0(\overline{\Omega})}$.
- Now assume $(f_k)_k$ is Cauchy sequence in $C^0(\overline{\Omega})$. Then for any $\varepsilon > 0$ there is some N such that for $i, j > N$ one finds for all $x \in \overline{\Omega}$

  $$|f_i(x) - f_j(x)| \leq \|f_i - f_j\|_{C^0(\overline{\Omega})} < \varepsilon.$$

  So $(f_k(x))_k$ is a Cauchy sequence in $\mathbb{R}$ and thus, for every $x \in \overline{\Omega}$ there is a limit. We denote the limits by $f(x)$.
- $\|f_i - f\|_{C^0(\overline{\Omega})} \to 0$: For $\varepsilon > 0$ there is some N such that $|f_i(x) - f_j(x)| < \varepsilon/3$ for all $i, j > N$, $x \in \overline{\Omega}$.
- For all $x \in \overline{\Omega}$ there is some $j_x > N$ such that $|f_{j_x}(x) - f(x)| < \varepsilon/3$. Therefore

  $$|f_i(x) - f(x)| \leq |f_i(x) - f_{j_x}(x)| + |f_{j_x}(x) - f(x)| < 2\varepsilon/3$$

  for all $i > N$ and $x \in \overline{\Omega}$.
- $f \in C^0(\overline{\Omega})$: For $x \in \overline{\Omega}$ and $\varepsilon > 0$ choose $i \in \mathbb{N}$ such that $\|f_i - f\|_{C^0(\overline{\Omega})} < \varepsilon/3$. Then there exists some $\delta > 0$ such that $|f_i(x) - f_i(y)| < \varepsilon/3$ for $y \in B(x, \delta)$. Finally, for $y \in B(x, \delta)$ get

  $$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_i(x)| + |f_i(x) - f_i(y)| + |f_i(y) - f(y)| \\ &\leq 2\|f_i - f\|_{C^0(\overline{\Omega})} + |f_i(x) - f_i(y)| < \varepsilon \,. \end{aligned}$$

- Now $C^1(\overline{\Omega})$. $\|f\|_{C^1(\overline{\Omega})} = \max\left\{\|f\|_{C^0(\overline{\Omega})}, \|\nabla f\|_{C^0(\overline{\Omega},\mathbb{R}^n)}\right\}$ where the norm $\|\cdot\|_{C^0(\overline{\Omega},\mathbb{R}^n)}$ is the maximum of the $\|\cdot\|_{C^0(\overline{\Omega})}$ norm of each component. We show analogously that this is indeed a norm.

- Now let $(f_i)_i$ be a Cauchy sequence in $C^1(\overline{\Omega})$. Then $(f_i)_i$ is a Cauchy sequence in $C^0(\overline{\Omega})$ and $(\nabla f_i)_i$ is a Cauchy sequence in $C^0(\overline{\Omega},\mathbb{R}^n)$. Therefore, they have limits in $C^0(\overline{\Omega})$ and $C^0(\overline{\Omega},\mathbb{R}^n)$, which we denote by f and g.

- Now show: $\nabla f = g$: For $x, y \in \overline{\Omega}$, $i \in \mathbb{N}$ we get:

$$\|f_i(y) - f_i(x) - \langle y - x, \nabla f_i(x)\rangle\| = \left\|\int_0^1 \langle \nabla f_i(x_t) - \nabla f_i(x), y - x\rangle \,\mathrm{d}t\right\|$$

(Where $x_t = x + (y-x)\cdot t$. Note also: $\|\nabla f_i(z) - g(z)\| \leq \sqrt{n}\|\nabla f_i - g\|_{C^0(\overline{\Omega},\mathbb{R}^n)}$.)

$$\leq \left\|\int_0^1 \langle g(x_t) - g(x), y - x\rangle \,\mathrm{d}t\right\| + 2\sqrt{n}\|\nabla f_i - g\|_{C^0(\overline{\Omega},\mathbb{R}^n)} \cdot \|x - y\|$$

By sending $i \to \infty$ for $x \neq y$ we obtain

$$\left\|\frac{f(y) - f(x)}{\|y - x\|} - \frac{\langle y - x, g(x)\rangle}{\|y - x\|}\right\| \leq \left\|\int_0^1 \|g(x_t) - g(x)\|\mathrm{d}t\right\| \xrightarrow[y \to x]{} 0$$

- General $C^{k+1}(\overline{\Omega})$, $k \geq 1$, follow by recursion. Assume we have dealt with $C^k(\overline{\Omega})$. Then for a Cauchy sequence $(f_i)_i$ in $C^{k+1}(\overline{\Omega})$ each component of $(\nabla f_i)_i$ converges in $C^k(\overline{\Omega})$. With the above argument we show that the gradient of the limit f is the limit of the gradients $(\nabla f_i)_i$.

□

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

Proposition 2.24 (Relation between $C^k([0,1])$)**.** For some integers $k > 0$ 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 $f_i \in C^k([0,1])$.

$\square$

Definition 2.25. For an exponent $\beta \in (0,1]$ a function $f : \overline{\Omega} \to \mathbb{R}$ is *Hölder continuous* with exponent β if there is a constant $C < \infty$ such that for all $x, y \in \overline{\Omega}$

$$|f(x) - f(y)| \leq C \cdot |x - y|^\beta$$

The space of k times Hölder continuously differentiable functions on $\overline{\Omega}$ with exponent β is denoted by $C^{k,\beta}(\overline{\Omega})$. It is a Banach space when equipped with the norm

$$\|f\|_{C^{k,\beta}(\overline{\Omega})} = \|f\|_{C^k(\overline{\Omega})} + \max_{a:|a|=k} \sup_{\substack{x,y\in\overline{\Omega}:\\ x\neq y}} \frac{|D^a f(x) - D^a f(y)|}{|x-y|^\beta}$$

Proof.
- When all derivatives of f up to k-th order are Hölder continuous, $\|f\|_{C^{k,\beta}(\overline{\Omega})}$ is finite. Moreover, $\|\cdot\|_{C^{k,\beta}(\overline{\Omega})}$ is positive definite, homogeneous and subadditive and thus is a norm on $C^{k,\beta}(\overline{\Omega})$.
- Completeness for $k = 0$: Let $(f_i)_i$ be a Cauchy sequence in $C^{0,\beta}(\overline{\Omega})$. Then it is a Cauchy sequence in $C^0(\overline{\Omega})$ and thus its pointwise limit f exists and is in $C^0(\overline{\Omega})$. Then for any $x, y \in \overline{\Omega}$, $x \neq y$

$$\frac{|(f-f_i)(x) - (f-f_i)(y)|}{\|x-y\|^\beta} = \lim_{j\to\infty} \underbrace{\frac{|(f_j-f_i)(x) - (f_j-f_i)(y)|}{\|x-y\|^\beta}}_{\leq \|f_i - f_j\|_{C^{0,\beta}(\overline{\Omega})}}$$

  The right-hand-side is bounded, therefore $\|f\|_{C^{0,\beta}(\overline{\Omega})} \leq \|f - f_i\|_{C^{0,\beta}(\overline{\Omega})} + \|f_i\|_{C^{0,\beta}(\overline{\Omega})} < \infty$ and thus $f \in C^{0,k}(\overline{\Omega})$. Moreover, as $i \to \infty$ the right-hand-side goes to zero, therefore $\|f - f_i\|_{C^{0,k}(\overline{\Omega})} \to 0$.
- Extension to $C^{k,\beta}$, $k > 0$ as above.

$\square$

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

Definition 2.26. For $p \in [1, \infty]$ let $\ell^p = \{x = (x_1, x_2, \dots) \in \mathbb{R}^\mathbb{N} | \|x\|_{\ell^p} < \infty\}$ where

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

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

Proposition 2.27 (Hölder inequality for ℓ^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

Proof.
- For $p = 1$ or $q = 1$ the inequality is immediate. So assume $p, q \in (1, \infty)$.
- For $a, b \geq 0$, $\lambda \in (0, 1)$ one has $a^\lambda\, b^{1-\lambda} \leq \lambda \cdot a + (1-\lambda) \cdot b$ ('geometric average $\leq$ arithmetic average'), with equality only if $a = b$.
- The statement is trivial if $a = 0$ or $b = 0$ since then the left-hand-side is zero.
- So assume $a, b > 0$. Then both expressions are well defined for $\lambda \in [0, 1]$. We find equality for $\lambda = 0$ and $\lambda = 1$.
- Let $g(\lambda) = a^\lambda\, b^{1-\lambda} = \exp(\lambda \log(a) + (1-\lambda)\log(b))$. We find $\frac{\partial^k}{\partial \lambda^k} g(\lambda) = g(\lambda) \cdot (\log(a) - \log(b))^k$. So $\frac{\partial^2}{\partial \lambda^2} g(\lambda) \geq 0$ and therefore g is convex and so $g(\lambda) \leq g(0) \cdot (1-\lambda) + g(1) \cdot \lambda$.
- If $a \neq b$ then $\frac{\partial^2}{\partial \lambda^2} g(\lambda) > 0$ and thus the function is strictly convex. So equality can only happen if $a = b$.
- Now set
$$a = \left(\frac{|x_i|}{\|x\|_{\ell^p}} \right)^p, \qquad b = \left(\frac{|y_i|}{\|y\|_{\ell^q}} \right)^q, \qquad \lambda = \tfrac{1}{p}, \qquad 1 - \lambda = \tfrac{1}{q}\,.$$
- Then
$$\frac{|x_i|}{\|x\|_{\ell^p}} \frac{|y_i|}{\|y\|_{\ell^q}} \leq \tfrac{1}{p} \left(\frac{|x_i|}{\|x\|_{\ell^p}} \right)^p + \tfrac{1}{q} \left(\frac{|y_i|}{\|y\|_{\ell^q}} \right)^q$$
- Now sum both sides over i to get:
$$\sum_{i=1}^\infty \frac{|x_i\, y_i|}{\|x\|_{\ell^p} \|y\|_{\ell^q}} = \sum_{i=1}^\infty \frac{|x_i|}{\|x\|_{\ell^p}} \frac{|y_i|}{\|y\|_{\ell^q}} \leq \sum_{i=1}^\infty \left[\tfrac{1}{p} \left(\frac{|x_i|}{\|x\|_{\ell^p}} \right)^p + \tfrac{1}{q} \left(\frac{|y_i|}{\|y\|_{\ell^q}} \right)^q \right] = \tfrac{1}{p} + \tfrac{1}{q} = 1$$

$\square$

Proposition 2.28 (Minkowski inequality for ℓ^p)**.** For $p \in [1, \infty]$ find $x, y \in \ell^p \Rightarrow x + y \in \ell^p$ with $\|x+y\|_{\ell^p} \leq \|x\|_{\ell^p} + \|y\|_{\ell^q}$. For $p \in (1, \infty)$ there is equality if and only if $x = q \cdot y$ for some $q \geq 0$.

Proof.
- For $p = 1$, $p = \infty$ the inequality follows directly (for $p = 1$ from subadditivity of the function $s \mapsto |s|$; for $p = \infty$ as for the $C^0(\overline{\Omega})$ space).
- Inequality is also trivial if $x = 0$ or $y = 0$. So assume $p \in (1, \infty)$, $x, y \neq 0$. In the following let $q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$. In particular $(p-1) \cdot q = p$ and $\frac{1}{q} = \frac{1}{p}(p-1)$. Then:
$$\begin{aligned}
\|x+y\|_{\ell^p}^p &= \sum_{i=1}^\infty |x_i + y_i|^p \leq \sum_{i=1}^\infty |x_i + y_i|^{p-1} (|x_i| + |y_i|) \\
&\leq \sum_{i=1}^\infty |x_i + y_i|^{p-1} |x_i| + \sum_{i=1}^\infty |x_i + y_i|^{p-1} |y_i|
\end{aligned}$$
(using Hölder inequality)
$$\leq \left(\sum_{i=1}^\infty |x_i + y_i|^{(p-1)q} \right)^{1/q} (\|x\|_{\ell^p} + \|y\|_{\ell^p}) = \|x+y\|_{\ell^p}^{p-1} (\|x\|_{\ell^p} + \|y\|_{\ell^p})$$

- This implies the inequality. If $x = q \cdot y$ for some $q \geq 0$ we have equality. Conversely, for Hölders inequality to yield equality it is necessary that

$$\left(\frac{|x_i|}{\|x\|_{\ell^p}}\right)^p = \left(\frac{|x_i + y_i|^{p-1}}{\||x+y|^{p-1}\|_{\ell^q}}\right)^q = \left(\frac{|x_i + y_i|}{\|x+y\|_{\ell^p}}\right)^p$$

which requires existence of some q.

□

Proposition 2.29. For $p \in [1, \infty]$ the space ℓ^p equipped with $\|\cdot\|_{\ell^p}$ is a Banach space.

Proof.

- Finiteness, positive definiteness and homogeneity of $\|\cdot\|_{\ell^p}$ are immediate. Subadditivity follows from the Minkowski inequality. So ℓ^p is normed space.
- For $p = \infty$ the proof for completeness is analogous to $C^0(\overline{\Omega})$. So let $p < \infty$.
- Let $(x_k)_k$ be a Cauchy sequence in ℓ^p, where for each k $(x_{k,i})_i$ is a sequence in $\mathbb{R}$. Then $|x_{k,i} - x_{j,i}| \leq \|x_k - x_j\|_{\ell^p}$, so for each i $(x_{k,i})_k$ is a Cauchy sequence in $\mathbb{R}$. Denote the sequence of limits by $(z_i)_i$.
- Since $(x_k)_k$ is Cauchy, $\|x_k\|_{\ell^p} < M$ for some $M < \infty$. So for all $n \in \mathbb{N}$

$$\sum_{i=1}^{n} |x_{k,i}|^p \leq M^p \quad \Rightarrow \quad \sum_{i=1}^{n} |z_i|^p \leq M^p$$

and consequently as $n \to \infty$ find $\|z\|_{\ell^p} \leq M$, i.e. $z \in \ell^p$.

- For any $\varepsilon > 0$ $\exists$ N such that $\forall$ $m, n > N$, $k \in \mathbb{N}$ get

$$\sum_{i=1}^{k} |x_{m,i} - x_{n,i}|^p \leq \|x_m - x_n\|_{\ell^p}^p \leq \varepsilon$$

Now let $m \to \infty$, then $k \to \infty$ to get $\|z - x_n\|^p \leq \varepsilon$.

□

Remark 2.30. Other prominent examples that are also very common in applications are L^p spaces and Sobolev spaces.

2.4 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 2.31.

- Let $X = C^0([0, 1])$, equipped with the norm $\|\cdot\| = \|\cdot\|_{C^0([0,1])}$.
- Let $Y = \{f \in X : \int_0^1 f(x)\,\mathrm{d}x = 0\}$. 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 $f \neq Y$, 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 $\int_0^1 g \, \mathrm{d}x = 0$ we find that we need to find $h \in X$ with $\int_0^1 f \mathrm{d}x = \int_0^1 h \mathrm{d}x$ that has minimal norm.
- This implies that $\|h\| \geq |\int_0^1 f \mathrm{d}x|$ (otherwise $|\int_0^1 h \mathrm{d}x| \leq \int_0^1 \|h\| \mathrm{d}x < |\int_0^1 f \mathrm{d}x|$).
- Try constant function $h(x) = \int_0^1 f(y) \mathrm{d}y$. This satisfies integral constraint and we find $\|h\| = |\int_0^1 f(y) \mathrm{d}y|$. So h is optimal and $g(x) = f(x) - \int_0^1 f(y) \mathrm{d}y$.
- Note: h is unique. Assume, h were not constant. Then $\|h\| > |\int_0^1 h \mathrm{d}x|$.

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

Proposition 2.32. 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 $X = \{f \in C^0([0,1]) : f(0) = 0\}$ 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 $Y = \{g \in X : \int_0^1 g(x) \mathrm{d}x = 0\}$.
- Analogous to above: fix $f \in X \setminus Y$, rewrite problem. Let $g \in Y$, set $h = f - g$. Solve $s = \inf\{\|h\| | h \in X \colon \int_0^1 h \mathrm{d}x = \int_0^1 f \mathrm{d}x\}$. 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. $s \geq |\int_0^1 f \mathrm{d}x|$. Since feasible h cannot be constant, must have $\|h\| > |\int_0^1 f \mathrm{d}x|$ for all feasible h.
- Now show that infimum $s = |\int_0^1 f \mathrm{d}x|$ 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 $h_i(x) = (\int_0^1 f(y) \mathrm{d}y) \cdot (1 + 1/i) \cdot x^{1/i}$. $h_i(0) = 0$, $\int_0^1 h_i(x) \mathrm{d}x = \int_0^1 f(y) \mathrm{d}y$ and $\|h_i\| = |h_i(1)| = |\int_0^1 f(y) \mathrm{d}y| \cdot (1 + 1/i)$. 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 $X \subset C^0([0,1])$.

$\square$

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 2.33 (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 $\operatorname{dist}(x, Y) = \operatorname{dist}(x_\theta, Y) \leq \|x_\theta\| \leq \theta \cdot \operatorname{dist}(x, Y)$.

Proof.

- $\operatorname{dist}(x, Y) = \inf_{y \in Y} \|x - y\|$ is finite and non-negative.
- Let $y_\theta \in Y$ such that $\|x - y_\theta\| < \theta \cdot \operatorname{dist}(x, Y)$ (take from a minimizing sequence) and set $x_\theta = x - y_\theta$.
- $\operatorname{dist}(x_\theta, Y) = \inf_{y \in Y} \|x - y_\theta - y\| = \inf_{y \in Y} \|x - y\| = \operatorname{dist}(x, Y)$.

□

Remark 2.34. If Y is not closed, then may have $\operatorname{dist}(x, Y) = 0$ even when $x \notin Y$ and $\|x_\theta\| \to 0$ as $\theta \searrow 1$.

Corollary 2.35. If Y is closed and $Y \neq X$ then for any $\theta > 1$ there is some x_θ with $\|x_\theta\| = 1$ and $\operatorname{dist}(x_\theta, Y) \geq \frac{1}{\theta}$.

We have learned that an important ingredient for existence of minimizers is compactness. Now study (strong) compactness on Banach spaces.

Proposition 2.36. 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 $(y_1, \ldots, y_n)$:

$$A \subset \bigcup_{y \in A} B(y, \delta_y) \quad \Rightarrow \quad A \subset \bigcup_{i=1}^{n} B(y_i, \delta_{y_i}) \quad \Rightarrow \quad \mathbb{N} = \bigcup_{i=1}^{n} N_i$$

- 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 ε-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 $B(x, \varepsilon_0) \subset I_{i_x}$. 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 $B(x, \delta) \in U_i$ 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 ε_0 chose $x_1, \ldots, x_n \in A$ such that

$$A \subset \bigcup_{k=1}^{n} B(x_k, \varepsilon_0) \subset \bigcup_{k=1}^{n} U_{i_{x_k}}$$

□

Corollary 2.37. 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). □

Proposition 2.38. For a Banach space X we find: $[\overline{B(0,1)}$ compact$] \Leftrightarrow [X$ 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) $\overline{B(0,1)} \subset \bigcup_{i=1}^n B(y_i, \frac{1}{2})$.
- $Y = \text{span}\{y_1, \ldots, y_n\}$ is closed subspace. Assume $Y \neq X$.
- By Corollary 2.35 for $\theta > 1$ there is some x_θ with $\|x_\theta\| = 1$ and $\text{dist}(x_\theta, \{y_1, \ldots, y_n\}) \geq \text{dist}(x_\theta, Y) \geq \frac{1}{\theta}$.
- But since $\|x_\theta\| = 1 \Rightarrow x_\theta \in \overline{B(0,1)} \subset \bigcup_{i=1}^n B(y_i, \frac{1}{2}) \Rightarrow \text{dist}(x_\theta, \{y_1, \ldots, y_n\}) < \frac{1}{2}$. For $\theta < 2$ this is a contradiction.
- **$\Leftarrow$:** 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.

□

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 2.39 (Equicontinuity)**.** A family of functions $f_i : V \to \mathbb{R}$, $i \in I$ is equicontinuous if for any $x \in V$ and $\varepsilon > 0$ there is some $\delta > 0$ such that $|f_i(y) - f_i(x)| < \varepsilon$ for all y with $\|y - x\| < \delta$, $i \in I$.

Proposition 2.40 (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 2.41. 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: $|f(x) - f(y)| \leq |x - y|$.

□

Definition 2.42 (Separable metric space)**.** A topological space X is *separable* if it contains a countable, dense subset A.

Remark 2.43. 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 countable elements. On separable spaces many proofs are constructive and one can avoid the axiom of choice.

Example 2.44.
(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 2.45. 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 $X \subset \bigcup_{i \in 1}^{k_n} B(x_{n,i}, \frac{1}{n})$.
- The countable set $A = \bigcup_{n=1}^{\infty} \{x_{n,1}, \ldots, x_{n,k_n}\}$ 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 $i \in \{1, \ldots, n_k\}$.
- So we can generate a sequence in A that converges to x.

□

Proposition 2.46. An infinite-dimensional Banach space with a Schauder basis is separable.

Remark 2.47. 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 $A_n = \{\sum_{i=1}^n s_i \, x_i | (s_i)_i \in \mathbb{Q}^n\}$. 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 $\lim_{n \to \infty} \|x - \sum_{i=1}^n \alpha_i \, x_i\| \to 0$, 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 $z_n = \sum_{i=1}^n \alpha_i \, x_i$, $y_n = \sum_{i=1}^n \beta_i \, x_i \in A$.

$$\|x - y_n\| \le \|x - z_n\| + \|z_n + y_n\| < \tfrac{\varepsilon}{2} + \sum_{i=1}^{n} |\alpha_i - \beta_i| \cdot \|x_i\| < \tfrac{\varepsilon}{2} + \varepsilon \sum_{i=1}^{n} 2^{-i-1} < \varepsilon$$

□

Proposition 2.48. $C^0([0,1])$ is separable.

We use a small auxiliary Lemma for the proof that is often helpful when working on compact spaces.

Lemma 2.49. Let (X, d) be a compact metric space. A continuous function $f : X \to \mathbb{R}$ 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 $\{x_1, \ldots, x_n\}$. Let $\delta = \min\{\delta_{x_1}, \ldots, \delta_{x_n}\} > 0$.
- 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 2.48.
- By Lemma 2.49 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 $i \in \{0, \ldots, 2^n\}$.
- Define set of 'tent functions' of scale n for $i \in \{0, \ldots, 2^n\}$:

$$f_{n,i}(x) = \begin{cases} 0 & \text{if } |x - i \cdot 2^{-n}| \geq 2^{-n}, \\ 1 - |2^n x - i| & \text{else} \end{cases}$$

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. 2.46 does not depend on uniqueness of the decomposition. So separability of $C^0([0,1])$ follows.

$\square$

Remark 2.50. (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 subsets 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 2.51. The spaces ℓ^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 ℓ^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:

$$\|x - z_i\|_{\ell^p}^p = \sum_{j=1}^{\infty} |x_j - z_{i,j}|^p = \sum_{j=1}^{\infty} |x_j - x_j\, \delta_{j \leq i}|^p = \sum_{j=i+1}^{\infty} |x_j|^p = \|x\|_{\ell^p}^p - \sum_{j=1}^{i} |x_j|^p \to 0$$

as $i \to \infty$. So $z_i \to x$.

- This decomposition is unique. Let $(y_i)_i$ be another sequence such that $\lim_{i\to\infty} v_i \to x$ for $v_i = \sum_{j=1}^i y_i \cdot e_i$ 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 2.52. The space ℓ^∞ 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 $a_k = (a_{k,1}, a_{k,2}, \ldots) \in \ell^\infty$.
- 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 ℓ^∞.

□

Proposition 2.53. An infinite-dimensional Hilbert space H is separable if and only if it has a orthonormal Schauder basis.

Proof.

- **⇐:** If H has a orthonormal Schauder basis, separability follows from Prop. 2.46.
- **⇒:** 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 $a_k \notin \operatorname{span}\{x_j\}_{j=1}^i$. 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 ∞. Note that $i \leq k$ throughout the process and that $a_k \in \operatorname{span}\{x_j\}_{j=1}^k$ 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
$$\left\| z - \sum_{i=1}^{k_j} x_i \langle x_i, z\rangle \right\| \leq \|z - a_{k_j}\| + \left\| \sum_{i=1}^{k_j} x_i \left\langle x_i, a_{k_j} - z\right\rangle \right\|$$
$$\leq \|z - a_{k_j}\| + \left(\left\langle \sum_{i=1}^{k_j} x_i \left\langle x_i, a_{k_j} - z\right\rangle, \sum_{i=1}^{k_j} x_i \left\langle x_i, a_{k_j} - z\right\rangle \right\rangle \right)^{1/2}$$
(Bessel inequality, cf. Example 1.91)
$$\leq 2\|z - a_{k_j}\| \to 0 \quad \text{as } j \to \infty.$$
- By orthonormality of $(x_i)_i$ the coefficients $\langle x_i, z\rangle$ are unique.

□

Corollary 2.54. Every separable Hilbert space H is isomorphic to ℓ^2, i.e. there is a bijection $\phi : H \to \ell^2$ such that $\langle x, y\rangle_H = \langle \phi(x), \phi(y)\rangle_{\ell^2}$.

2.5 Linear transformations and topological dual

Definition 2.55 (Linear transformations and functionals)**.**

- For two normed spaces X, Y, $D \subset X$ a map $T : D \to Y$ is called *transformation* from X to Y with domain D.
- A transformation $T : D \to \mathbb{R}$ is called *functional.*
- $T : X \to Y$ is *linear* if $T(a\,x + b\,y) = a\,T(x) + b\,T(y)$ for all $x, y \in X$, $a, b \in \mathbb{R}$.
- The *operator norm* of a transformation $T : X \to Y$ 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 2.56. Let T be a linear transformation from X to Y.

(i) $[T$ continuous in $0] \Leftrightarrow [T$ continuous on $X]$

(ii) $[T$ bounded$] \Leftrightarrow [T$ continuous$]$

Proof.

- X and Y normed spaces $\Rightarrow$ "ε-δ-notion" of continuity is sufficient.
- **(i):** Let $x \in X$, $\varepsilon > 0$. $[\exists\, \delta > 0 : \|T(z)\|_Y < \varepsilon$ if $\|z\|_X < \delta] \Leftrightarrow [\exists\, \delta > 0 : \|T(x) - T(y)\|_Y = \|T(x - y)\|_Y < \varepsilon$ if $y \in B_X(x, \delta)]$
- **(ii):** $\Rightarrow$: $[T$ bounded$] \Rightarrow [\forall x \in X$: $\|T(x)\|_Y \leq \|T\| \cdot \|x\|_X \Rightarrow [T$ continuous in $0] \Leftrightarrow [T$ continuous$]$.
- $\Leftarrow$: let $\varepsilon > 0$. $[T$ continuous in $0] \Rightarrow [\exists\, \delta > 0$: $\|T(x)\|_Y \leq \varepsilon$ 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 $\|T\| < \frac{2\varepsilon}{\delta}$.

$\square$

Proposition 2.57. $L(X, Y)$ 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 $(a\,S + b\,T) : x \mapsto a\,S(x) + b\,T(x)$ is a linear transformation.
- $\|T\| \geq 0$ by definition. $[\|T\| = 0] \Leftrightarrow [T(x) = 0$ for all $x \in X] \Leftrightarrow [T = 0]$.
- $\|a \cdot T\| = |a| \cdot \|T\|$ by homogeneity of $\|\cdot\|_Y$.
- $\|S+T\| = \sup_{x \in X \setminus \{0\}} \frac{\|S(x)+T(x)\|_Y}{\|x\|_X} \leq \sup_{x \in X \setminus \{0\}} \frac{\|S(x)\|_Y}{\|x\|_X} + \sup_{y \in X \setminus \{0\}} \frac{\|S(y)\|_Y}{\|y\|_X} = \|S\| + \|T\|$.

- Since $\|\cdot\|$ is subadditive, whenever S and T are bounded, so is $S+T$. So $L(X,Y)$ is a vector space.

$\square$

Proposition 2.58. 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 $T : x \mapsto \lim_{n\to\infty} T_n(x)$.
- T is linear:

$$T(a\,x + b\,y) = \lim_{n\to\infty} T_n(a\,x + b\,y) = \lim_{n\to\infty} a\,T_n(x) + b\,T_n(y)$$

  (sum of Cauchy sequences is Cauchy)

$$= a\,T(x) + b\,T(y).$$

- T is bounded: let $x \in X \setminus \{0\}$, $n > N$.

$$\begin{aligned}\|T(x)\|_Y \leq \|T(x) - T_n(x)\|_Y + \|T_n(x)\|_Y &= \lim_{m\to\infty} \|T_m(x) - T_n(x)\|_Y + \|T_n(x)\|_Y \\ &\leq \varepsilon \cdot \|x\|_X + \|T_n\| \cdot \|x\|_X\end{aligned}$$

  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 2.59 (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 $X^* \times X \to \mathbb{R}$ 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 2.60. The duality pairing $X^* \times X \to \mathbb{R}$, $(t,x) \mapsto t(x)$ 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 $\delta = \min\{1, \frac{\varepsilon}{\|s\|+\|x\|_X+1}\}$. Then $\delta \leq \frac{\varepsilon}{\|s\|+\|x\|_X+1} \leq \frac{\varepsilon}{\|s\|+\|x\|_X+\delta}$.

- Now let $t \in B_{X^*}(s,\delta)$, $y \in B_X(x,\delta)$. Get

$$\begin{aligned}|s(x) - t(y)| \leq |s(x) - s(y)| + |s(y) - t(y)| \leq \|s\| \, \|x - y\|_X + \|s - t\| \, \|y\|_X \\ \leq \|s\| \, \delta + \delta \left(\|x\|_X + \delta\right) \leq \varepsilon\end{aligned}$$

□

Definition 2.61 (Weak topology)**.**
- The topology induced on X by the family of maps X^* (see Def. 2.17) 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 $x_n \rightharpoonup x$.
- $[x_n \rightharpoonup x] \Leftrightarrow [t(x_n) \to t(x)$ for all $t \in X^*]$.

Definition 2.62 (Weak$*$ topology)**.**
- For fixed $x \in X$ consider the map $f_x : X^* \to \mathbb{R}$, $t \mapsto \langle t, x\rangle$.
- The weak$*$ topology on X^* is the topology induced by the family of maps $\{f_x | x \in X\}$.
- Weak$*$ convergence is denoted by $t_n \overset{*}{\rightharpoonup} t$.
- $[t_n \overset{*}{\rightharpoonup} t] \Leftrightarrow [\langle t_n, x\rangle \to \langle t, x\rangle$ for all $x \in X]$.

Example 2.63.
- 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 2.64. 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 $x = \lim_{n\to\infty} x_n$ where $x_n = \sum_{i=1}^n \alpha_i \cdot z_i$.
- Let $t \in X^*$. By continuity of t find $t(x) = \lim_{n\to\infty} t(x_n) = \lim_{n\to\infty} \sum_{i=1}^n \alpha_i \cdot t(z_i)$.
- Can represent t by sequence $(\lambda_i = t(z_i))_i$: $t(x) = \lim_{n\to\infty} \sum_{i=1}^n \alpha_i \cdot \lambda_i$.
- 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 $t \neq \hat{t}$.

□

Remark 2.65. The above proposition does not specify which sequences $(\lambda_n)_n$ represent some $t \in X^*$. Doing this helps to fully characterize X^*.

Proposition 2.66. For $p \in [1,\infty)$ the dual space of ℓ^p is isomorphic (exists bijection that preserves norm / metric) with ℓ^q where $\frac{1}{p} + \frac{1}{q} = 1$.

Proof.
- First treat $p > 1$. By Propositions 2.51 (ℓ^p has Schauder basis $(e_i)_i$) and 2.64 $(\ell^p)^*$ can be identified with subset of real sequences.

- Let $y \in \ell^q$. Define $t(x) = \lim_{n\to\infty} \sum_{i=1}^{n} x_i\, y_i$. 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. 2.27), so operator norm of t equals ℓ^q norm of y. So can identify ℓ^q with subset of $(\ell^p)^*$.

- Assume $y \notin \ell^q$. So $\sum_{i=1}^{n} |y_i|^q$ is unbounded as $n \to \infty$.

- Let $\hat{y}_n = (\hat{y}_{n,1}, \hat{y}_{n,2}, \ldots)$ 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 ℓ^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 $\frac{1}{\|\hat{x}_n\|_{\ell^p}} \sum_{i=1}^{\infty} y_i\, \hat{x}_{n,i} = \frac{1}{\|\hat{x}_n\|_{\ell^p}} \sum_{i=1}^{\infty} \hat{y}_{n,i}\, \hat{x}_{n,i} = \|\hat{y}_n\|_{\ell^q} \to \infty$ as $n \to \infty$. So y does not represent a bounded linear functional on ℓ^p.

- So can identify dual of ℓ^p with ℓ^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 ℓ^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 $|\sum_{i=1}^{\infty} y_i\, \hat{x}_{k,i}| = |y_{n_k}| \to \infty$, therefore y does not represent bounded functional.

□

Remark 2.67. Since ℓ^∞ has no Schauder basis, cannot use this trick to study dual space of ℓ^∞. Will later see indirectly that it cannot be identified with ℓ^1.

Remark 2.68.
- 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 2.69. ℓ^1 is a (strict) subspace of ℓ^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 2.70.
(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 2.71. Any $x \in X$ defines a bounded linear functional on X^* via $t \mapsto \langle t, x \rangle_{X^* \times X}$, see Def. 2.59. 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 2.72. (i) Hilbert spaces are reflexive.

(ii) ℓ^p for $p \in (1, \infty)$ are reflexive.

(iii) Will soon see: ℓ^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 2.73 (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 2.74 (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 2.75. 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 2.76. [l^1 is not reflexive] $\Leftrightarrow$ [$(\ell^\infty)^* \neq \ell^1$]

Proof.

- The equivalence in the statement follows from $(\ell^1)^* = \ell^\infty$ (Prop. 2.66).
- By Theorem 2.75 it suffices to show that $\overline{B_{\ell^1}(0,1)}$ is not weakly compact.
- Consider sequence $(e_n)_n$ of canonical Schauder basis vectors lies in ℓ^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. 2.74 implies that it is not weakly compact.

□

2.6 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 2.77 (Hahn–Banach)**.** Let X be real vector space. Let $f : X \to \mathbb{R}$ be positively 1-homogeneous and sub-additive, i.e.

$$f(\alpha\, x) = \alpha\, f(x), \qquad f(x+y) \leq f(x) + f(y) \qquad \text{for all} \quad x, y \in X, \alpha \in \mathbb{R}_+.$$

Let $Y \subset X$ be a subspace and let $t : Y = \operatorname{dom}(t) \to \mathbb{R}$ 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 $T : X \to \mathbb{R}$ of t that is majorized by f on X, i.e.

$$T(x) = t(x) \quad \text{for} \quad x \in Y \quad \text{and} \quad T(x) \leq f(x) \quad \text{for} \quad x \in X.$$

The proof relies fundamentally on the axiom of choice, in form of Zorn's Lemma.

Lemma 2.78 (Zorn's Lemma)**.** Let S be a non-empty partially ordered set. Assume that every totally ordered subset of S has an upper bound in S. Then S has a maximal element.

Proof of Theorem 2.77.

- Let S be the set of extensions of t that are majorized by f, i.e.

$$S = \big\{s : \operatorname{dom}(s) \to \mathbb{R}, \operatorname{dom}(s) \text{ subspace of } X, \\ \operatorname{dom}(t) \subset \operatorname{dom}(s),\ s \text{ linear}, s(x) \leq f(x) \text{ for } x \in \operatorname{dom}(s)\big\}.$$

- $S \neq \emptyset$ since $t \in S$.
- Define partial ordering $\succeq$ on S via

  $$[a \succeq b] \quad \Leftrightarrow \quad [\operatorname{dom}(a) \supset \operatorname{dom}(b)] \wedge [a(x) = b(x)\, \forall\, x \in \operatorname{dom}(b)].$$

  Is indeed partial ordering: $[a \succeq a]$, $[a \succeq b] \wedge [b \succeq a] \Rightarrow a = b$, $[a \succeq b] \wedge [b \succeq c] \Rightarrow [a \succeq c]$.
- Let $C \subset S$ be totally ordered. Define s_C via:

  $$\operatorname{dom}(s_C) = \bigcup_{s \in C} \operatorname{dom}(s), \quad s_C(x) = s(x) \quad \text{if} \quad x \in \operatorname{dom}(s)$$

  Verify: $s_C(x)$ is well defined: for every $x \in \operatorname{dom}(s_C)$ there is some $s \in C$ such that $x \in \operatorname{dom}(s)$. Let $x \in \operatorname{dom}(s_1) \cap \operatorname{dom}(s_2)$, $s_1, s_2 \in C$. Then $[s_1 \succeq s_2]$ or $[s_2 \succeq s_1]$, so $s_1(x) = s_2(x) = s_C(x)$.
- It follows that s_C is linear, majorized by f and $s_C(x) = t(x)$ for $x \in Y$. So $s_C \in S$. Moreover, $s_C \geq s$ for all $s \in C$ since $\operatorname{dom}(s_C) \supset \operatorname{dom}(s)$.
- $\Rightarrow C$ has upper bound s_C in S. Zorn's Lemma: S has maximal element T.
- T is linear extension of t, majorized by f. Need to show $\operatorname{dom}(T) = X$. By contradiction.
- Assume $x_0 \in X \setminus \operatorname{dom}(T)$. Set $Z = \operatorname{dom}(T) \oplus \operatorname{span}\{x_0\}$. For any $z \in Z$ $\exists$ unique decomposition $z = x + \lambda \cdot x_0$, $x \in \operatorname{dom}(T)$, $\lambda \in \mathbb{R}$.

- Define $\hat{T} : Z \to \mathbb{R}$ via $\hat{T}(x + \lambda \cdot x_0) = T(x) + a \cdot \lambda$ for some $a \in \mathbb{R}$.
- $\forall\, x, y \in \mathrm{dom}(T)$ find (use linearity of T, majorization of by f, subadditivity of f):

$$\begin{aligned}
&T(x) - T(y) = T(x - y) \leq f(x - y) \leq f(x + x_0) + f(-y - x_0) \\
\Rightarrow &- T(y) - f(-y - x_0) \leq -T(x) + f(x + x_0)
\end{aligned}$$

  Choose $a \in \left[\sup_{y \in \mathrm{dom}(T)}(-T(y) - f(-y - x_0)), \inf_{x \in \mathrm{dom}(T)}(-T(x) + f(x + x_0))\right] \neq \emptyset$.

- Now let $z = x + \lambda \cdot x_0 \in Z$.
- Assume $\lambda = 0$: $\hat{T}(z) = T(x) \leq f(z)$.
- Assume $\lambda > 0$:

$$\begin{aligned}
\hat{T}(z) &= T(x) + a \cdot \lambda \leq T(x) + \lambda \cdot (-T(\xi) + f(\xi + x_0)) \quad \text{for all } \xi \in \mathrm{dom}(T) \\
&\leq f(x + \lambda\, x_0) = f(z) \quad \text{when setting } \xi = x/\lambda
\end{aligned}$$

- Assume $\lambda < 0$:

$$\begin{aligned}
\hat{T}(z) &= T(x) + a \cdot \lambda \leq T(x) + \lambda \cdot (-T(\xi) - f(-\xi - x_0)) \quad \text{for all } \xi \in \mathrm{dom}(T) \\
&\leq f(x + \lambda\, x_0) = f(z) \quad \text{when setting } \xi = x/\lambda
\end{aligned}$$

- So $\hat{T} \in S$, $\hat{T} \succeq T$, $\hat{T} \neq T$: contradiction! Therefore $\mathrm{dom}(T) = X$.

$\square$

Remark 2.79. In similar fashion can use Zorn's Lemma to prove existence of basis for vector spaces (possibly uncountable), existence of unbounded linear functionals, etcetera.
For instance can proof that ℓ^∞ has basis. Since ℓ^∞ is not separable, basis must be uncountable. Can use this to define bounded linear functionals on ℓ^∞ that have no correspondence in ℓ^1.

A few applications.

Proposition 2.80. Let Y be subspace of normed space X, $t \in L(Y, \mathbb{R})$. Then there exists some $T \in L(X, \mathbb{R})$, $T|_Y = t$ ($T|_Y$: restriction of T to Y), $\|T\| = \|t\|$.

Proof.
- Apply Hahn–Banach to t defined on Y with $f(x) = \|t\|\,\|x\|$.
- Get linear $T : X \to \mathbb{R}$ 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 2.81. Let X be normed space, $x \in X$. Then there exists some $T \in X^* \setminus \{0\}$ such that $T(x) = \|T\|\,\|x\|$.

Proof.
- Set subspace $Y = \mathrm{span}\{x\}$, $t : Y \to \mathbb{R}$, $\alpha \cdot x \mapsto \alpha \|x\|$. Note: $\|t\|_{Y^*} = 1$. $f(z) = \|z\|$.
- Apply Hahn–Banach, get $T : X \to \mathbb{R}$. $T(\alpha x) = t(\alpha x) = \alpha \|x\|$, $T(z) \leq \|z\| \Rightarrow T(x) = \|x\| = \|T\|\,\|x\|$.

$\square$

Remark 2.82.

- 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 2.83. 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 2.84. ${}^\perp[Y^\perp] = Y$ for any closed subspace $Y \subset X$.

Comment: Compare to polar cone, Def. 1.42 and Prop. 1.43.

Proof.

- ${}^\perp[Y^\perp] \supset Y$: $[x \in Y] \Rightarrow [t(x) = 0\ \forall\ t \in Y^\perp] \Rightarrow [y \in {}^\perp[Y^\perp]]$.
- ${}^\perp[Y^\perp] \subset Y$: For $y \notin Y$ define $t \in L(V = Y \oplus \operatorname{span}\{y\}, \mathbb{R})$ by $t(x + \lambda\, y) = \lambda$ for every $v = x + \lambda\, y \in V$ (the decomposition is unique). We get

$$\|t\| = \sup_{\substack{x\in Y, \lambda \in \mathbb{R}:\\ x+\lambda\, y \neq 0}} \frac{|t(x+\lambda\, y)|}{\|x + \lambda\, y\|} = \sup_{\substack{x \in Y\setminus\{0\},\\ \lambda\in\mathbb{R}\setminus\{0\}}} \frac{|t(x+\lambda\, y)|}{\|x+\lambda\, y\|} = \sup_{x\in Y\setminus\{0\}} \frac{1}{\|x+y\|} < \infty$$

since $\operatorname{dist}(y, Y) > 0$ (Y is closed).

- Use Hahn–Banach via Prop. 2.81 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 2.85 (Subdifferential)**.** Let X be a normed space. For a function $f : X \to \mathbb{R} \cup \{\infty\}$ the subdifferential of f at $x \in X$ is given by

$$\partial f(x) = \left\{ t \in X^* : f(y) \geq f(x) + \langle t, y - x\rangle_{X^*\times X} \text{ for all } y \in X \right\}.$$

Definition 2.86 (Fenchel–Legendre conjugates)**.** Let X be a normed space. For a proper function $f : X \to \mathbb{R} \cup \{\infty\}$ the Fenchel–Legendre conjugate $f^* : X^* \to \mathbb{R} \cup \{\infty\}$ is given by

$$f^*(t) = \sup_{x\in X} \langle t, x\rangle_{X^*\times X} - f(x).$$

The preconjugate of a proper function $g : X^* \to \mathbb{R} \cup \{\infty\}$ 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 (basic properties of conjugate), Prop. 1.74 (pointwise suprema over families of convex, lsc functions remain convex, lsc.), Prop. 1.72 (Fenchel–Young) and Prop. 1.81 ('extreme points' of Fenchel–Young) we can show:

Proposition 2.87. For a normed space X let $f : X \to \mathbb{R} \cup \{\infty\}$, $g : X^* \to \mathbb{R} \cup \{\infty\}$ 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 2.88. Since $X^{**} \not\simeq X$ 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 2.89 (Hahn–Banach: separation form)**.** Let X be a normed space, $C \subset X$ convex, $\operatorname{int} C \neq \emptyset$, $x \notin \operatorname{int} C$. 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 2.90 (Minkowski functional)**.** Let X be a normed space. Let $C \subset X$ be convex, $0 \in \operatorname{int} C$. The *Minkowski functional* of C is defined as

$$p_C : x \to \mathbb{R}, \qquad x \mapsto \inf\{r \in \mathbb{R}_+ : x \in r\, C\}.$$

p_C is nonnegative, positively 1-homogeneous, continuous and subadditive (this implies convexity).

Proof.
- p_C is indeed real valued: since $0 \in \operatorname{int} C$ $\exists\, \eta > 0$ 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: let $x \in r \cdot C$, $y \in s \cdot C$ (let $x = r\, a$, $y = s\, b$, $a, b \in C$). Then

$$x + y = r\, a + s\, b = (r + s)\left(\tfrac{r}{r+s} a + \tfrac{s}{r+s} b\right) \in (r + s) \cdot C$$

  So for any $x, y \in X$:

$$\begin{aligned} p_C(x) + p_C(y) &= \inf\left\{ r + s | r, s \in \mathbb{R}_+ : \underbrace{x \in r\, C, y \in s\, C}_{\Rightarrow x+y \in (r+s)\, C} \right\} \\ &\geq \inf\{r + s \in \mathbb{R}_+ : x + y \in (r + s)\, C\} = p_C(x + y) \end{aligned}$$

- continuity: for $x \in X$, $\varepsilon > 0$ set $\delta = \eta \cdot \varepsilon$. For $y \in B(x, \delta)$ find

$$\begin{aligned} &p_C(y) = p_C(x + (y - x)) \leq p_C(x) + p_C(y - x) \leq p_C(x) + \tfrac{\delta}{\eta} \leq p_C(x) + \varepsilon, \\ &p_C(x) = p_C(y + (x - y)) \leq p_C(y) + p_C(x - y) \\ &p_C(y) \geq p_C(x) - p_C(x - y) \geq p_C(x) - \varepsilon \end{aligned}$$

  So $p_C(B(x, \delta)) \in p_C(x) + [-\varepsilon, \varepsilon]$.

□

Proof of Theorem 2.89.

- W.l.o.g. assume $0 \in \operatorname{int} C$ (otherwise, simply translate).
- Define $t : \operatorname{span}\{x\} \to \mathbb{R}$, $t(\lambda x) = \lambda$.
- Set $f(z) = p_{\operatorname{int} C}(z)$. Since $x \notin \operatorname{int} C \Rightarrow f(x) \geq 1$. 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 $y \in \operatorname{int} C$ 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 2.91. Let X be a normed space. Let $f, g : X \to \mathbb{R} \cup \{\infty\}$ 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 2.92.

- Recall subspace projection problem on Banach space. Let Y be closed subspace of normed space X, $x \in X \setminus Y$.

$$\operatorname{dist}(x, Y) = \inf_{y \in X} (\|x - y\| + \iota_Y(y))$$

- Let $f(y) = \|x - y\|$, $g(y) = \iota_Y(y)$. Brief calculation yields:

$$f^*(t) = \langle t, x\rangle + \iota_{\overline{B(0,1)}}(t), \qquad g^*(t) = \iota_{Y^\perp}(t)$$

- By above duality we find:

$$\operatorname{dist}(x, Y) = \max\left\{\langle t, x\rangle \,|\, t \in \overline{B(0,1)} \cap Y^\perp\right\}$$

2.7 Γ-convergence

Remark 2.93 (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 2.94 (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 $s : \Omega \to \mathbb{R}$ 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 $|A| = \int_A \mathrm{d}x = c$ 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 $\mathrm{Per}(A)$.
- The energetically best configuration will be given by a minimizer to the following optimization problem for sets A:

$$\min \left\{ \int_A s(x)\mathrm{d}x + \mathrm{Per}(A) \,\middle|\, A \subset \Omega, |A| = c \right\}$$

- This is an optimization problem over sets. Could try to define sufficiently regular classes of sets for which Per is well defined, topology for sets, with notions of compactness etcetera.
- Alternative: try to approximate problem. Let $u : \Omega \to [0,1]$ 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

$$\int_\Omega \left(\frac{W(u(x))}{\varepsilon} + \varepsilon \, |\nabla u(x)|^2 \right) \mathrm{d}x.$$

- The full problem can then be approximated by

$$\min \left\{ \int_\Omega \left(s(x)\, u(x) + \frac{W(u(x))}{\varepsilon} + \varepsilon \, |\nabla u(x)|^2 \right) \mathrm{d}x \,\middle|\, u : \Omega \to [0,1], \int_\Omega u(x)\, \mathrm{d}x = c \right\}.$$

- Correspondences: s-term, volume constraint. As $\varepsilon \to 0$ 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 ε). For transition region: has 'width' $\mathcal{O}(\varepsilon)$, 'length' $L = \mathrm{Per}(A)$, 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 2.95 (Homogenization). • Stationary heat equation on domain Ω with spatially varying thermal conductivity $a : \Omega \to \mathbb{R}_+$ and source $f : \Omega \to \mathbb{R}$ 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

$$\min \left\{ \int_\Omega \left(\tfrac{a}{2} |\nabla u|^2 - u\, f \right) \mathrm{d}x \middle| u : \Omega \to \mathbb{R}, u(x) = 0 \text{ for } x \in \partial\Omega \right\}.$$

  (of course need to identify reasonable subspace for candidates u).

- Now assume a contains 'microscopic periodic' structure at scale ε. Solution will probably also contain microscopic structure.

Sketch: a_ε 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 ε. Corresponds to sending $\varepsilon \to 0$.

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

$$\min \left\{ \int_\Omega \left(\tfrac{1}{2} \langle \nabla u, A \nabla u \rangle - u\, f \right) \mathrm{d}x \middle| u : \Omega \to \mathbb{R}, u(x) = 0 \text{ for } x \in \partial\Omega \right\}.$$

with matrix A (can no contain spatial symmetry breaking).

- How do we formalize transition between minimization problems?

Example 2.96 (Dimension reduction). Consider again stationary heat equation on domain $\Omega \times [0, \varepsilon]$, i.e. a 'thin film'. As $\varepsilon \to 0$ 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 Ω.

Example 2.97 (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.

Remark 2.98. Throughout this subsection let X be a topological space. We will focus on sequential notion of Γ-convergence. As before, it is possible to introduce more general topological definitions.

Definition 2.99 ((Sequential) Γ-convergence)**.** A sequence $f_k : X \to \mathbb{R} \cup \{\infty\}$, $k \in \mathbb{N}$ is said to (sequentially) Γ-*converge* to $f : X \to \mathbb{R} \cup \{\infty\}$ if for all $x \in X$ we have

(i) (lim inf inequality) for every sequence $(x_k)_k$ converging to x

$$f(x) \leq \liminf_{k\to\infty} f_k(x_k);$$

(ii) (lim sup inequality) there exists a sequence $(x_k)_k$ converging to x such that

$$f(x) \geq \limsup_{k\to\infty} f_k(x_k).$$

The function f is called the Γ-*limit* of $(f_k)_k$ and we write $f = \Gamma\text{-}\lim_k f_k$.

Remark 2.100 (Recovery sequence)**.** Let $(x_k)_k$ satisfy the lim sup inequality. It must also satisfy the lim inf inequality and therefore

$$f(x) \geq \limsup_{k\to\infty} f_k(x_k) \geq \liminf_{k\to\infty} f_k(x_k) \geq f(x).$$

So $\lim_{k\to\infty} f_k(x_k) = f(x)$ and $(x_k)_k$ is referred to as *recovery sequence.*

Remark 2.101 (Motivation of definition)**.** The main motivation of Γ-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 $(f_k)_k$, $x_k \to x$, 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 2.102. 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 $u : \Omega \to [0, 1]$. Possible solution: reasonable extension of function space, to handle both limit and sequences.

(ii) dimension reduction: limit problem on functions $u : \Omega \to \mathbb{R}$, sequence on functions $u : \Omega \times [0, \varepsilon] \to \mathbb{R}$. Rescale sequence problems to functions $u : \Omega \times [0, 1]$ (remove ε from space), limit problem on functions $u : \Omega \to [0, 1] \to \mathbb{R}$, add constraint $u(x, s) = \text{const}$ for all $s \in [0, 1]$, for every fixed $x \in \Omega$.

Definition 2.103 ((Sequential) Γ-limits)**.** For a sequence of functions $(f_k)_k$, $f_k : X \to \mathbb{R} \cup \{\infty\}$ and some $x \in X$ the quantities

$$(\Gamma\text{-}\liminf_k f_k)(x) = \inf\{\liminf_k f_k(x_k) | (x_k)_k : x_k \to x\},$$
$$(\Gamma\text{-}\limsup_k f_k)(x) = \inf\{\limsup_k f_k(x_k) | (x_k)_k : x_k \to x\}$$

are called sequential Γ-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 Γ-limit of $(f_k)_k$ at x.

Proposition 2.104. If the sequence $(f_k)_k$ Γ-converges to f the Γ-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 2.105. This implies that Γ-limit f of sequence $(f_k)_k$ is unique (if it exists) since Γ-lower and upper limit do not depend on f. Can conversely show that if Γ-lower and upper limit coincide at all points x then their value determines the Γ-limit function f.

Proposition 2.106 (Stability under continuous perturbations)**.** Let $(f_k)_k$ be a sequence of functions with Γ-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 2.107.

- Set

$$g(t) = \begin{cases} +1 & \text{if } t = 1, \\ -1 & \text{if } t = -1, \\ 0 & \text{else.} \end{cases}$$

- Set $f_k(t) = g(k\,t)$ and
$$f(t) = \begin{cases} 0 & \text{if } t \neq 0, \\ -1 & \text{if } t = 0. \end{cases}$$

- Proof that f is Γ-limit of f: let $x \neq 0$, $x_k \to x$. $\exists\, N \in \mathbb{N}$ 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 lim inf and lim sup condition.

- Let $x = 0$. Since $f(x) = -1 \leq f_k(t)$ for all k and t, the lim inf condition is clear. As recovery sequence set $x_k = -1/k$. Then $f_k(x_k) = g(-1) = -1 = f(x)$.

- Note: $\lim_k \to \infty f_k(x) = 0$ for all x. So the Γ-limit does not necessarily coincide with the pointwise limit.

Example 2.108 (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 semicontinous, f is not the Γ-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 Γ-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 2.109. Let (X, d) be a metric space. If $f : X \to \mathbb{R} \cup \{\infty\}$ is a Γ-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 lim sup condition we obtain
$$f(x_k) = \lim_{j\to\infty} f_j(x_{k,j}).$$

- 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 $z_j \to x$.

- Now:
$$\liminf_k f(x_k) \geq \liminf_k f_{j_k}(x_{k,j_k}) - \tfrac{1}{k} = \liminf_k f_{j_k}(x_{k,j_k}) \geq \liminf_j f_j(z_j) \geq f(x)\,.$$

□

Now give simple prototypical results that show how sequence of minimizers of $(f_k)_k$ is related to minimizers of Γ-limit f.

Proposition 2.110. Let $(f_k)_k$ be a sequence of functions $X \to \mathbb{R} \cup \{\infty\}$ as well as $f : X \to \mathbb{R} \cup \{\infty\}$.

(i) If $(f_k)_k$ and f satisfy the lim inf inequality, Def. 2.99(i), and K is a sequentially compact set then

$$\inf_K f \leq \liminf_{k\to\infty} \inf_K f_k.$$

(ii) If $(f_k)_k$ and f satisfy the lim sup inequality, Def. 2.99(ii), and U is an open set then

$$\inf_U f \geq \limsup_{k\to\infty} \inf_U f_k.$$

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 $\tilde{x}_{k_j} \to \overline{x} \in K$.
- Set

$$x_k = \begin{cases} \tilde{x}_{k_j} & \text{if } k = k_j, \\ \overline{x} & \text{if } k \neq k_j \,\forall\, j. \end{cases}$$

  Then $x_k \to \overline{x}$.
- With the lim inf condition we get

$$\inf_K f \leq f(\overline{x}) \leq \liminf_k f_k(x_k) \leq \liminf_j f_{k_j}(\tilde{x}_{k_j}) = \lim_j f_{k_j}(\tilde{x}_{k_j}) = \liminf_k \inf_K f_k.$$

- **(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 2.111 (Variants of coerciveness).
- A function $f : X \to \mathbb{R} \cup \{\infty\}$ 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 2.112 (Convergence of minimizers)**.** Let $(f_k)_k$, $f_k : X \to \mathbb{R} \cup \{\infty\}$ be a sequence of equi-mildly sequentially coercive functions and let $f = \Gamma\text{-}\lim_k f_k$. Then

$$\min_X f = \lim_k \inf_X f_k.$$

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. 2.110(i) to K and (ii) to $U = X$ to obtain:

$$\inf_X f \leq \inf_K f \leq \liminf_k \inf_K f_k = \liminf_k \inf_X f_k \leq \limsup_k \inf_X f_k \leq \inf_X f$$

- 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 $\overline{x}$. 'Replace' all indices $k \neq k_j$ for all j by $\overline{x}$, as in proof of Prop. 2.110. Call this sequence $(\tilde{x}_k)_k$. Get via lim inf condition:

$$\inf_X f \leq f(x) \leq \liminf_k f_k(\tilde{x}_k) \leq \liminf_j f_{k_j}(x_{k_j}) = \lim_k f_k(x_k) = \lim_k \inf_X f_k = \inf_X f$$

  So 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 $\inf_X f = \inf_K f$.

□

2.8 Example: optimal transport

Comment: References:
Villani: Topics in Optimal Transportation, 2003,
Villani: Optimal Transport: Old and New, 2009

Remark 2.113. Throughout this section (Ω, d) is a compact metric space. Most result extend to non-compact spaces (with appropriate modifications) but the arguments become more technical.

2.8.1 Reminders on measure theory

Comment: Reference: Ambrosio, Fusco, Pallara: Functions of Bounded Variation and Free Discontinuity Problems, Chapters 1 & 2.

Definition 2.114 (σ-algebra)**.** A collection $\mathcal{E} \subset 2^X$ of subsets of a set X is called σ-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 2.115. Borel algebra: smallest σ algebra containing all open sets of a topological space.

Comment: Intersection of two σ-algebras is again σ-algebra. 'smallest' is well-defined.

Definition 2.116 (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 $A_n \in \mathcal{E} \Rightarrow \mu\left(\bigcup_{n=0}^{\infty} A_n\right) = \sum_{n=0}^{\infty} \mu(A_n)$

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 μ satisfies (i) and (ii) with absolute convergence.

Comment: Measures are vector space, measures are finite, positive measures may be infinite.

Example 2.117. 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 μ, function $f \in L^1(\mu)$, new measure $\nu = f \cdot \mu$. $\nu(A) \stackrel{\text{def.}}{=} \int_A f(x) \, \mathrm{d}\mu(x)$.

Definition 2.118 (Total variation)**.** For finite measure μ on $(X, \mathcal{E})$ the total variation $|\mu|$ of $A \in \mathcal{E}$ is

$$|\mu|(A) = \sup \left\{ \sum_{n=0}^{\infty} |\mu(A_n)| \middle| A_n \in \mathcal{E}, \text{ pairwise disjoint}, \bigcup_{n=0}^{\infty} A_n = A \right\} .$$

$|\mu|$ is finite, positive measure on $(X, \mathcal{E})$.

Definition 2.119 (Negligible sets)**.** A set $N \subset X$ is μ-negligible if $\exists\ A \in \mathcal{E}$ with $N \subset A$ and $\mu(A) = 0$. Two functions f, $g : X \to Y$ are identical 'μ-almost everywhere' when $\{x \in X | f(x) \neq g(x)\}$ is μ-negligible.

Example 2.120. Null sets are Lebesgue-negligible sets.

Definition 2.121 (Measurable functions, push-forward)**.** Let $(X, \mathcal{E})$, $(Y, \mathcal{F})$ be measurable spaces. A function $f : X \to Y$ is 'measurable' if $f^{-1}(A) \in \mathcal{E}$ for $A \in \mathcal{F}$.
For measure μ on $(X, \mathcal{E})$ the 'push-forward' of μ under f to $(Y, \mathcal{F})$, we write $f_\sharp\mu$, is defined by $f_\sharp\mu(A) = \mu(f^{-1}(A))$ for $A \in \mathcal{F}$.
Change of variables formula:

$$\int_X g(f(x))\, \mathrm{d}\mu(x) = \int_Y g(y)\, \mathrm{d}f_\sharp\mu(y)$$

Sketch: Varying densities.

Example 2.122 (Marginal)**.** Let $\mathrm{proj}_i : X \times X \to X$, $\mathrm{proj}_i(x_0, x_1) = x_i$. Marginals of measure γ on $X \times X$:

$$\mathrm{proj}_{0\,\sharp}\gamma(A) = \gamma(A \times X)\,, \qquad \mathrm{proj}_{1\,\sharp}\gamma(A) = \gamma(X \times A)\,.$$

Sketch: Discuss pre-images of proj_i.

Definition 2.123 (Absolute continuity, singularity)**.** Let μ be positive measure, ν measure on measurable space $(X, \mathcal{E})$. ν is 'absolutely continuous' w.r.t. μ, 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 μ, ν 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 μ, ν by $|\mu|$, $|\nu|$.

Definition 2.124 (σ-finite)**.** A positive measure μ is called σ-finite if $X = \bigcup_{n=0}^{\infty} A_n$ for sequence $A_n \in \mathcal{E}$ with $\mu(A_n) < +\infty$.

Example 2.125. Lebesgue measure is not finite but σ-finite.

Theorem 2.126 (Radon–Nikodym, Lebesgue decomposition [Ambrosio et al., Theorem 1.28])**.** Let μ be σ-finite positive measure. ν 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 μ-almost everywhere. It is called 'density of ν with respect to μ' and usually denoted by $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.
Lebesgue decomposition: there exist unique measures ν_a, ν_s such that

$$\nu = \nu_a + \nu_s, \qquad \nu_a \ll \mu, \qquad \nu_s \perp \mu\,.$$

Note: $\nu_a = f \cdot \mu$ for some $f \in L^1(\mu)$.

Corollary 2.127. A real-valued measure ν can be decomposed into $\nu = \nu_+ - \nu_-$ with ν_+, ν_- 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 $\nu_\pm(B) = |\nu(B \cap A_\pm)|$. □

Comment: f is only unique $|\nu|$-almost everywhere.

Definition 2.128 (Support of measure)**.** Let (Ω, d) be a compact metric space with its Borel σ-algebra and $\mu \in \mathcal{M}_+(\Omega)$. The support of μ, denoted $\operatorname{spt} \mu$ is the smallest closed set $A \subset \Omega$ such that $\mu(A) = \mu(\Omega)$. For $x \in \operatorname{spt} \mu$ one has $\mu(B_r(x)) > 0$ for any $r > 0$.

Definition 2.129 (Radon measures)**.** Let (Ω, d) be compact metric space, let $\mathcal{E}$ be Borel-σ-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 2.130 (Regularity [Ambrosio et al., Proposition 1.43])**.** For positive Radon measures on $(\Omega, \mathcal{E})$ one has for $A \in \mathcal{E}$

$$\mu(A) = \sup\{\mu(B) \,|\, B \in \mathcal{E},\, B \subset A,\, B \text{ compact}\} = \inf\{\mu(B) \,|\, B \in \mathcal{E},\, A \subset B,\, B \text{ open}\}\,.$$

Theorem 2.131 (Duality [Ambrosio et al., Theorem 1.54])**.** Let (Ω, d) be compact metric space. Let $C(\Omega)^m$ be space of continuous functions from Ω 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 $\|\mu\|_{\mathcal{M}} \stackrel{\text{def.}}{=} |\mu|(\Omega)$. Duality pairing for $\mu \in \mathcal{M}(\Omega)^m$, $f \in C(\Omega)^m$:

$$\mu(f) = \langle \mu, f \rangle_{\mathcal{M} \times C} = \int_\Omega f(x)\, \mathrm{d}\mu(x)$$

Comment: Notation: $C(\Omega) \equiv C^0(\Omega)$ from Section 2.3 (Examples for Banach spaces).

Corollary 2.132. Two measures μ, $\nu \in \mathcal{M}(\Omega)^m$ with $\mu(f) = \nu(f)$ for all $f \in C(\Omega)^m$ coincide.

Remark 2.133.

- By Theorem 2.73(i) (Banach–Alaoglu) $\overline{B_{\mathcal{M}}(0,1)}$ is weak$*$ compact.
- Analogous to Prop. 2.48 ($C([0,1])$ is separable) can show that $C(\Omega)$ is separable. So by Theorem 2.73(ii) ('sequential' Banach–Alaoglu) $\overline{B_{\mathcal{M}}(0,1)}$ is sequentially compact.

2.8.2 Monge formulation of optimal transport

Comment: Gaspard Monge: French mathematician and engineer, 18$^{\text{th}}$ century. Studied problem of optimal allocation of resources to minimize transport cost.

Sketch: Bakeries and cafes

Example 2.134 (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 $c : \Omega \times \Omega \to \mathbb{R}_+$. $c(x, y)$ gives cost of transporting 1 unit of bread from bakery at x to cafe at y.
- Describe transport by map $T : \Omega \to \Omega$. Bakery at x will deliver bread to cafe at $T(x)$. Consistency condition: $T_\sharp\mu = \nu$.

Comment: Each cafe receives precisely ordered amount of bread.

- Total cost of transport map

$$C_M(T) = \int_\Omega c(x, T(x))\,\mathrm{d}\mu(x)$$

Comment: For bakery at location x pay $c(x, T(x)) \cdot \mu(x)$. Sum (i.e. integrate) over all bakeries.

Definition 2.135. Monge optimal transport problem: find T that minimizes C_M.

Problems:

- Do maps T with $T_\sharp\mu = \nu$ 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.

2.8.3 Kantorovich formulation of optimal transport

Comment: Leonid Kantorovich: Russian mathematician, 20$^{\text{th}}$ 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 2.136 (Coupling / Transport Plan)**.** Let μ, $\nu \in \mathcal{P}(\Omega)$. Set of 'couplings' or 'transport plans' $\Pi(\mu, \nu)$ is given by

$$\Pi(\mu, \nu) = \left\{\pi \in \mathcal{P}(\Omega \times \Omega) \,\middle|\, \mathrm{proj}_{0\,\sharp}\pi = \mu,\ \mathrm{proj}_{1\,\sharp}\pi = \nu\right\}\ .$$

Example 2.137. $\Pi(\mu, \nu) \neq \emptyset$, 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 2.138. For compact metric space (Ω, d), μ, $\nu \in \mathcal{P}(\Omega)$, $c \in C(\Omega \times \Omega)$ the Kantorovich optimal transport problem is given by

$$\mathcal{C}(\mu, \nu) = \inf \left\{ \int_{\Omega\times\Omega} c(x, y)\,\mathrm{d}\pi(x, y) \,\middle|\, \pi \in \Pi(\mu, \nu) \right\}$$

Comment: Linear (continuous) objective, affine constraint set.

Comment: Language of measures covers finite dimensional and infinite dimensional case.

Proposition 2.139. Minimizers of Kantorovich problem (Def. 2.138) exist.

For proof use following result.

Proposition 2.140. The set $\Pi(\mu, \nu)$ is weak$*$ sequentially closed.

Proof.

- Let $(\pi_n)_n$ be sequence in $\Pi(\mu, \nu)$, with $\pi_n \overset{*}{\rightharpoonup} \pi \in \mathcal{M}(\Omega \times \Omega)$.
- Positivity: π is a positive measure. Otherwise find function $\phi \in C(\Omega \times \Omega)$, $\phi \geq 0$, with $\int_{\Omega\times\Omega} \phi \, \mathrm{d}\pi < 0$ (use Cor. 2.127 and Thm. 2.130 for construction) which contradicts weak$*$ convergence since $\int \phi \, \mathrm{d}\pi_n \geq 0 \; \forall \; n$.
- Unit mass: $\pi(\Omega \times \Omega) = \int_{\Omega\times\Omega} 1 \, \mathrm{d}\pi = \lim_{n\to\infty} \int_{\Omega\times\Omega} 1 \, \mathrm{d}\pi_n = \lim_{n\to\infty} \pi_n(\Omega \times \Omega) = 1$.
- Marginal constraint: For every $\phi \in C(\Omega)$

$$\int_\Omega \phi \, \mathrm{d}\mathrm{proj}_{0\,\sharp}\pi = \int_{\Omega\times\Omega} \phi \circ \mathrm{proj}_0 \, \mathrm{d}\pi$$

$$= \lim_{n\to\infty} \int_{\Omega\times\Omega} \phi \circ \mathrm{proj}_0 \, \mathrm{d}\pi_n = \lim_{n\to\infty} \int_\Omega \phi \, \mathrm{d}\mathrm{proj}_{0\,\sharp}\pi_n = \int_\Omega \phi \, \mathrm{d}\mu$$

  So $\mathrm{proj}_{0\,\sharp}\pi = \mu$. Analogous: $\mathrm{proj}_{1\,\sharp}\pi = \nu$.

$\square$

Proof of Proposition 2.139.

- Let π_n be minimizing sequence. Since $\pi_n \in \mathcal{P}(\Omega \times \Omega)$ have $\|\pi_n\|_\mathcal{M} = 1$. By Banach-Alaoglu (Thm. 2.73) $\exists$ converging subsequence. After extraction of subsequence have convergent minimizing sequence $\pi_n \overset{*}{\rightharpoonup} \pi$.
- By Prop. 2.140 $\pi \in \Pi(\mu, \nu)$.
- Since $c \in C(\Omega \times \Omega)$ and $\pi_n \overset{*}{\rightharpoonup} \pi$ have

$$\int_{\Omega\times\Omega} c \, \mathrm{d}\pi = \lim_{n\to\infty} \int_{\Omega\times\Omega} c \, \mathrm{d}\pi_n \,.$$

  Therefore, π is minimizer.

$\square$

Comment: For proof under more general conditions see for instance [Villani, 2009, Chapter 4].

Relation to Monge problem:

Proposition 2.141 (Kantorovich is a relaxation of the Monge problem)**.** Assume $T : \Omega \to \Omega$ is a feasible transport map for the Monge problem between μ and ν, Definition 2.135. In particular $T_\sharp\mu = \nu$.
Let

$$(\mathrm{id}, T) : \Omega \to \Omega \times \Omega, \qquad\qquad x \mapsto (x, T(x)) \,.$$

Then $\pi = (\mathrm{id}, T)_\sharp\mu \in \Pi(\mu, \nu)$ and

$$\int_{\Omega\times\Omega} c \, \mathrm{d}\pi = \int_\Omega c(x, T(x)) \, \mathrm{d}\mu(x) \,.$$

Proof.

- Clearly $\pi \in \mathcal{P}(\Omega \times \Omega)$: non-negative, unit mass.
- $\text{proj}_{0\sharp}\pi = \text{proj}_{0\sharp}(\text{id}, T)_\sharp \mu = (\text{proj}_0 \circ (\text{id}, T))_\sharp \mu = (\text{id})_\sharp \mu = \mu$.
- $\text{proj}_{1\sharp}\pi = \text{proj}_{1\sharp}(\text{id}, T)_\sharp \mu = (\text{proj}_1 \circ (\text{id}, T))_\sharp \mu = T_\sharp \mu = \nu$.
- Equality of cost:

$$\int_{\Omega\times\Omega} c(x, y)\, \mathrm{d}((\text{id}, T)_\sharp \pi)(x, y) = \int_\Omega (c \circ (\text{id}, T))(x)\, \mathrm{d}\mu(x) = \int_\Omega c\big(x, T(x)\big)\mathrm{d}\mu(x).$$

$\square$

Remark 2.142. Under suitable conditions (e.g. Ω compact subset of $\mathbb{R}^d$ with Euclidean distance, $c(x, y) = \|x-y\|^2$, $\partial\Omega$ Lebesgue-negligible, μ 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.

2.8.4 Duality

Now we study the corresponding dual problem.

Proposition 2.143. Given the setting of Definition 2.138 one finds

$$\mathcal{C}(\mu, \nu) = \sup\left\{\int_\Omega \alpha\, \mathrm{d}\mu + \int_\Omega \beta\, \mathrm{d}\nu \middle| \alpha,\, \beta \in C(\Omega), \alpha(x) + \beta(y) \leq c(x, y) \text{ for all } (x, y) \in \Omega^2\right\}$$

Proof.

- Problem of Prop. 2.143 can be written as

$$\mathcal{C}(\mu, \nu) = -\inf\left\{f(\alpha, \beta) + g(A(\alpha, \beta))\middle|(\alpha, \beta) \in C(\Omega)^2\right\}$$

  with

$$\begin{aligned}
&f : C(\Omega)^2 \to \mathbb{R}, && (\alpha, \beta) \mapsto -\int_\Omega \alpha\, \mathrm{d}\mu - \int_\Omega \beta\, \mathrm{d}\nu \\
&g : C(\Omega^2) \to \mathbb{R} \cup \{\infty\}, && \psi \mapsto \begin{cases} 0 & \text{if } \psi(x, y) \leq c(x, y) \text{ for all } (x, y) \in \Omega^2 \\ +\infty & \text{else.} \end{cases} \\
&A : C(\Omega)^2 \to C(\Omega^2), && [A(\alpha, \beta)](x, y) = \alpha(x) + \beta(y)\,.
\end{aligned}$$

- f, g are convex, lsc. A is bounded, linear.
- Let (α, β) be two constant, finite functions with $\alpha(x) + \beta(y) < \min\{c(x', y')|(x', y') \in \Omega^2\}$. 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. 2.91 to account for the linear transformation A)

$$\mathcal{C}(\mu, \nu) = \min\left\{f^*(-A^*\pi) + g^*(\pi)\middle|\pi \in \mathcal{M}(\Omega^2)\right\}.$$

- One obtains:

$$\begin{aligned}
f^*(-\rho,-\sigma) &= \sup\left\{-\int_\Omega \alpha\,\mathrm{d}\rho - \int_\Omega \beta\,\mathrm{d}\sigma + \int_\Omega \alpha\,\mathrm{d}\mu + \int_\Omega \beta\,\mathrm{d}\nu \middle| (\alpha,\beta)\in C(\Omega)^2\right\} \\
&= \begin{cases} 0 & \text{if } \rho=\mu,\ \sigma=\nu\,, \\ +\infty & \text{else.}\end{cases}
\end{aligned}$$

(Reasoning similar than for positivity of limit π in proof of Prop. 2.140.)

$$\begin{aligned}
g^*(\pi) &= \sup\left\{\int_{\Omega^2} \psi\,\mathrm{d}\pi \middle| \psi\in C(\Omega^2), \psi(x,y)\leq c(x,y) \text{ for all } (x,y)\in\Omega^2\right\} \\
&= \begin{cases} \int_{\Omega^2} c\,\mathrm{d}\pi & \text{if } \pi\in\mathcal{M}_+(\Omega^2), \\ +\infty & \text{else.}\end{cases}
\end{aligned}$$

- Adjoint of A:

$$\begin{aligned}
\langle A^*\pi,(\alpha,\beta)\rangle_{\mathcal{M}(\Omega)^2\times C(\Omega)^2} &= \langle \pi, A(\alpha,\beta)\rangle_{\mathcal{M}(\Omega\times\Omega)\times C(\Omega\times\Omega)} \\
&= \int_{\Omega\times\Omega} [\alpha(x)+\beta(y)]\,\mathrm{d}\pi(x,y) \\
&= \int_{\Omega\times\Omega} [\alpha\circ\mathrm{proj}_0 + \beta\circ\mathrm{proj}_1]\,\mathrm{d}\pi \\
&= \int_\Omega \alpha\,\mathrm{d}(\mathrm{proj}_{0\,\sharp}\pi) + \int_\Omega \beta\,\mathrm{d}(\mathrm{proj}_{1\,\sharp}\pi)
\end{aligned}$$

$\Rightarrow A^*\pi = (\mathrm{proj}_{0\,\sharp}\pi, \mathrm{proj}_{1\,\sharp}\pi)$.

- Summarize:

$$\begin{aligned}
f^*(-A^*\pi)+g^*(\pi) &= f^*(-\mathrm{proj}_{0\,\sharp}\pi, -\mathrm{proj}_{1\,\sharp}\pi) + g^*(\pi) \\
&= \begin{cases} \int_{\Omega^2} c\,\mathrm{d}\pi & \text{if } \pi\in\Pi(\mu,\nu), \\ +\infty & \text{else.}\end{cases}
\end{aligned}$$

□

Remark 2.144 (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 helpful for proving equivalence with Monge problem.
- Also: dual problem has interpretation in 'transport perspective': α and β can be interpreted as prices that both sides have to pay for transport. $\alpha(x)+\beta(y)\leq c(x,y) \Rightarrow$ 'price may never exceed cost'...

2.8.5 Discretization and Γ-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 Γ-convergence to original problem, as discretization is refined. Show that optimal plan can be extracted as cluster point of discrete optimal plans.

Definition 2.145 (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 $r_n \to 0$ as $n \to \infty$,
- $\Omega = \bigcup_{i=1}^n T_{n,i}$ and $T_{n,i} \cap T_{n,j} = \emptyset$ if $i \neq j$ ('collectively exhaustive and mutually exclusive').

Example 2.146. $\{T_{n,i}\}_{i=1}^n$ and $\{x_{n,i}\}_{i=1}^n$ and might be cells and centroids of subsequently finer triangulations or of Cartesian grids over Ω.

Definition 2.147 (Discretization of marginals). For given $\mu, \nu \in \mathcal{P}(\Omega)$ set

$$\hat{\mu}_n = (\hat{\mu}_{n,i})_{i=1}^n, \qquad \hat{\mu}_{n,i} = \mu(T_{n,i}), \qquad \mu_n = \sum_{i=1}^n \delta_{x_{n,i}} \cdot \hat{\mu}_{n,i},$$

$$\hat{\nu}_n = (\hat{\nu}_{n,i})_{i=1}^n, \qquad \hat{\nu}_{n,i} = \nu(T_{n,i}), \qquad \nu_n = \sum_{i=1}^n \delta_{x_{n,i}} \cdot \hat{\nu}_{n,i}.$$

Proposition 2.148. $\mu_n \in \mathcal{P}(\Omega)$ and $\mu_n \stackrel{*}{\rightharpoonup} \mu$ as $n \to \infty$. Analogously $\nu_n \in \mathcal{P}(\Omega)$ and $\nu_n \stackrel{*}{\rightharpoonup} \nu$.

Proof.

- Only proof for μ. Result for ν 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 $\phi \in C(\Omega)$. (Ω, d) compact $\Rightarrow \phi$ uniformly continuous (Lemma 2.49). $\Rightarrow \forall\, \varepsilon > 0\ \exists\, N \in \mathbb{N}$ such that for $n \geq N$ have $|\phi(x) - \phi(x_{n,i})| < \varepsilon$ for $x \in T_{n,i} \subset B(x_{n,i}, r_n)$.

$$\begin{aligned}\left|\int_\Omega \phi \,\mathrm{d}\mu - \int_\Omega \phi \,\mathrm{d}\mu_n\right| &\leq \sum_{i=1}^n \left|\int_{T_{n,i}} \phi \,\mathrm{d}\mu - \int_{T_{n,i}} \phi \,\mathrm{d}\mu_n\right| = \sum_{i=1}^n \left|\int_{T_{n,i}} \phi \,\mathrm{d}\mu - \phi(x_{i,n}) \cdot \hat{\mu}_{n,i}\right| \\ &\leq \sum_{i=1}^n \left|\int_{T_{n,i}} (\phi(x) - \phi(x_{i,n}))\, \mathrm{d}\mu(x)\right| \leq \sum_{i=1}^n \varepsilon\, \hat{\mu}_{n,i} = \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary find $\langle \mu_n, \phi \rangle_{\mathcal{M}\times C} \to \langle \mu, \phi \rangle_{\mathcal{M}\times C}$. This is true for all $\phi \in C(\Omega)$. □

Comment: In general $\mu_n \not\to \mu$ in the norm topology.

Proposition 2.149 (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 $[\pi \in \Pi(\mu_n,\nu_n)] \Leftrightarrow [\pi = \sum_{i,j=1}^n \delta_{(x_{n,i},x_{n,j})} \cdot \hat{\pi}_{i,j}$ for some $\hat{\pi} \in \hat{\Pi}_n(\hat{\mu}_n,\hat{\nu}_n)]$.

Proof.
- $\Rightarrow$: $[\pi \in \Pi(\mu_n,\nu_n)] \Rightarrow$ [for measurable $A \subset \Omega$: $\pi(A\times\Omega) = \mu_n(A) = \mu_n(A \cap \{x_{n,i}\}_{i=1}^n)$)] $\Rightarrow [\operatorname{spt}\pi \subset (\{x_{n,i}\}_{i=1}^n \times \Omega)]$. Likewise: $[\operatorname{spt}\pi \subset (\Omega \times \{x_{n,i}\}_{i=1}^n)] \Rightarrow [\operatorname{spt}\pi \subset (\{x_{n,i}\}_{i=1}^n \times \{x_{n,i}\}_{i=1}^n)]$.
- 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 π 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 2.150 (Discretized transport problem)**.**

$$\mathcal{C}(\mu_n,\nu_n) = \min\left\{ \sum_{i,j=1}^n c(x_{n,i},x_{n,j}) \cdot \hat{\pi}_{i,j} \,\middle|\, \hat{\pi} \in \hat{\Pi}_n(\hat{\mu}_n,\hat{\nu}_n) \right\}$$

Comment: This is a finite-dimensional problem. Discretization of marginals suffices such that problem becomes finite-dimensional and can be solved exactly. No discretization of derivatives etc. required.

Now we show Γ-convergence of discretized problems to original transport problem.

Proposition 2.151. Let

$$F : \mathcal{M}(\Omega\times\Omega) \to \mathbb{R}\cup\{\infty\}, \qquad \pi \mapsto \int_{\Omega\times\Omega} c \,\mathrm{d}\pi + \iota_{\Pi(\mu,\nu)}(\pi)$$

$$F_n : \mathcal{M}(\Omega\times\Omega) \to \mathbb{R}\cup\{\infty\}, \qquad \pi \mapsto \int_{\Omega\times\Omega} c \,\mathrm{d}\pi + \iota_{\Pi(\mu_n,\nu_n)}(\pi)$$

Then F_n Γ-converges sequentially to F in the weak$*$ topology as $n \to \infty$.

Proof.
- **(lim inf)**: Let $\pi_n \stackrel{*}{\rightharpoonup} \pi$. Since c is continuous $\Rightarrow \int_{\Omega\times\Omega} c\,\mathrm{d}\pi_n \to \int_{\Omega\times\Omega} c\,\mathrm{d}\pi$.

- 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 $\phi \in C(\Omega)$:

$$\int_\Omega \phi \, \mathrm{d}\mu \leftarrow \int_\Omega \phi \, \mathrm{d}\mu_n = \int_{\Omega\times\Omega} \phi(x) \, \mathrm{d}\pi_n(x, y) \to \int_{\Omega\times\Omega} \phi(x) \, \mathrm{d}\pi(x, y)$$

  So $\mathrm{proj}_{0\,\sharp}\pi = \mu$ and likewise $\mathrm{proj}_{1\,\sharp}\pi = \nu$.
- By weak$*$ convergence, π is also non-negative and has unit mass (as all π_n are). So $\pi \in \Pi(\mu, \nu)$ and thus

$$\liminf_n F_n(\pi_n) = \liminf_n \int_{\Omega\times\Omega} c \, \mathrm{d}\pi_n = \lim_n \int_{\Omega\times\Omega} c \, \mathrm{d}\pi_n = \int_{\Omega\times\Omega} c \, \mathrm{d}\pi = F(\pi).$$

- **(lim sup)**: Let π 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 $\sum_{j=1}^n \hat{\pi}_{n,i,j} = \hat{\nu}_{n,j}$. 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. 2.149).
- Analogous to $\mu_n \stackrel{*}{\rightharpoonup} \mu$ (Prop. 2.148) show that $\pi_n \stackrel{*}{\rightharpoonup} \pi$.
- So $\limsup_n F_n(\pi_n) = \limsup_n \int_{\Omega\times\Omega} c \, \mathrm{d}\pi_n = \int_{\Omega\times\Omega} c \, \mathrm{d}\pi = F(\pi)$.

$\square$

Proposition 2.152. 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 π 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 2.133). 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. 2.112.

$\square$

2.9 Example: binary image segmentation

2.9.1 Motivation and set formulation

Remark 2.153. Study prototypical and foundational problem in image analysis. Divide image region into fore- and background. For simplicity only consider one-dimensional problem, extension to higher dimensions is analogous.

Remark 2.154 (Original problem formulation)**.**

- Image domain $\Omega = [0, 1]$, throughout this subsection.
- Bounded function $a \in L^1(\Omega)$ indicates affinity of each point to be foreground (a negative) or background (a positive).

  a could be generated from color in photo: e.g., are we looking for a red object?

  Sketch: Starfish

- Set $S \subset \Omega$ describes foreground. First proposal for optimization problem:

$$\inf \left\{ \int_S a(x)\,\mathrm{d}x \,\middle|\, S \subset \Omega, S \text{ measurable} \right\}$$

- Without further assumptions best foreground would be obtained by thresholding of a: $[a(x) < 0] \Leftrightarrow [x \in S$, i.e. x in foreground]. Problem: if a contains noise / measurement errors. S obtained by thresholding may be very irregular: 'single pixels' missing within foreground, or single pixels far from object mistakenly identified as foreground.

  Sketch: Add noise to starfish

- Add assumption to model: S should be 'regular'. Most objects in real world have relatively smooth (piecewise smooth) boundary. 'Penalize irregular boundaries' by adding new term to minimization problem.
- Simplest model: add term that measures volume of boundary. In 1d: number of points in boundary. Express as $\#(\partial S)$, $\#(\cdot)$: counting measure. Minimize following energy:

$$E_{\text{set}}(S) = \int_S a(x)\,\mathrm{d}x + \lambda \cdot \#(\partial S)$$

  The first term is called *data term*: depends on observed (possibly noisy) image. Second term is called *regularizer*: mathematically model assumptions on 'true' / noiseless observation. $\lambda \geq 0$ is weight. 'Correct' choice of λ depends on 'how much we trust' observation a vs. model of regularizer.

Example 2.155 (Choice of λ)**.** For $\delta \in (0, \frac{1}{2})$ let

$$a(x) = \begin{cases} -1 & \text{if } x \in [\frac{1}{2} - \delta, \frac{1}{2} + \delta], \\ 1 & \text{else,} \end{cases} \qquad\qquad S = [\tfrac{1}{2} - \delta, \tfrac{1}{2} + \delta].$$

We find $\#(\partial S) = 2$ and so

$$E_{\text{set}}(S) = -2\delta + 2\lambda, \qquad E_{\text{set}}(\emptyset) = 0, \qquad E_{\text{set}}(\Omega) = 1 - 4\delta.$$

So for $\delta \in (0, \frac{1}{4})$, S is optimal if $\delta \geq \lambda$. Otherwise, $\emptyset$ is optimal. So λ specifies length scale to discriminate between noise and true image structure.

Simple observation on regularity of feasible candidates for E_{set}:

Corollary 2.156. $E_{\text{set}}(S) < \infty$ iff S is finite collection of intervals in Ω.

2.9.2 Functions of bounded variation and convex relaxation

Remark 2.157 (Choice of function space)**.**

- Minimizing E_{set} over sets is inconvenient: set of candidates has no linear structure, so no notion of convexity, many tools from analysis missing.
- Want to reformulate segmentation problem as convex optimization problem over vector space.
- Segmentation encoded by function $u : \Omega \to \{0, 1\}$, $[u(x) = 1] \Leftrightarrow [x \in S]$.
- What is suitable function space? Must contain discontinuous functions. For convexity, relax allowed values from $\{0, 1\}$ to $[0, 1]$. Need to reformulate E_{set} in terms of u. How to handle non-binary values of u?
- If we find optimal function u, which may be non-binary, can we extract optimal S for original problem?

To represent segmentation sets by functions, we use the following definition for convenience.

Definition 2.158 (Characteristic function)**.** For a measurable set $S \subset \Omega$ the *characteristic function* of S is given by

$$\chi_S : \Omega \to \{0, 1\}, \qquad\qquad x \mapsto \begin{cases} 1 & \text{if } x \in S, \\ 0 & \text{else.} \end{cases}$$

Now we need a suitable function space that contains characteristic functions of sufficiently regular sets. A good choice is called the *functions of bounded variation*, given in the next definition. A thorough introduction can be found in [Ambrosio, Fusco, Pallara: Functions of Bounded Variation and Free Discontinuity Problems, 2000].

Definition 2.159 (Functions of bounded variation)**.** A function $u \in L^1(\Omega)$ is called a function of *bounded variation* if its weak derivative can be expressed as a Radon measure $\mu \in \mathcal{M}(\operatorname{int}\Omega)$. More precisely, for every test function $\varphi \in C^1_0(\Omega) = \{\varphi \in C^1(\Omega) : \varphi(0) = \varphi(1) = 0\}$ we find

$$\int_\Omega \varphi'(x) \cdot u(x)\,\mathrm{d}x = -\int_\Omega \varphi(x)\,\mathrm{d}\mu(x).$$

Comment: Need to be a little careful about boundary conditions.

By a density argument it can be shown that this weak integration by parts formula holds for any Lipschitz test function (with appropriate boundary conditions). The space of functions of bounded variation is denoted by $\mathrm{BV}(\Omega)$. If $u \in \mathrm{BV}(\Omega)$ then the weak derivative is often denoted by Du. By duality between $C^0(\Omega)$ and $\mathcal{M}(\Omega)$ the weak derivative Du is unique (if it exists).

Example 2.160. (i) Let $u \in W^{1,1}(\Omega)$. Then u has weak derivative $Du \in L^1(\Omega)$ with

$$\int_\Omega \varphi'(x) \cdot u(x)\,\mathrm{d}x = -\int_\Omega \varphi(x) \cdot Du(x)\,\mathrm{d}x$$

for all $\varphi \in C^1_0(\Omega)$. Then $Du \cdot \mathcal{L}|_{\operatorname{int}\Omega} \in \mathcal{M}(\operatorname{int}\Omega)$.

(ii) Characteristic function of interval $[a, b]$, $0 < a < b < 1$: $u = \chi_{[a,b]}$. For integration of any test function $\varphi \in C_0^1(\Omega)$ find:

$$\int_\Omega \varphi'(x) \cdot u(x) \, \mathrm{d}x = \int_a^b \varphi'(x) \, \mathrm{d}x = \varphi(b) - \varphi(a) = \int_\Omega \varphi \, \mathrm{d}(\delta_b - \delta_a)$$

So $Du = \delta_a - \delta_b \in \mathcal{M}(\operatorname{int} \Omega)$.

Definition 2.161 (Total variation of functions)**.** For $u \in L^1(\Omega)$ its *total variation* is given by

$$\mathrm{TV}(u) = \sup \left\{ \int_\Omega \varphi'(x) \cdot u(x) \, \mathrm{d}x \middle| \varphi \in C_0^1(\Omega), \|\varphi\|_{C^0(\Omega)} \leq 1 \right\}$$

This definition should not be confused with Def. 2.118, the total variation norm for measures.

Proposition 2.162. For $u \in L^1(\Omega)$ have $[u \in \mathrm{BV}(\Omega)] \Leftrightarrow [\mathrm{TV}(u) < \infty]$. If $u \in \mathrm{BV}(\Omega)$ then $\mathrm{TV}(u) = \|Du\|_{\mathcal{M}}$.

Proof.

- **$\Rightarrow$:** Let $u \in \mathrm{BV}(\Omega)$. $\Rightarrow \exists \; Du \in \mathcal{M}(\operatorname{int} \Omega)$ such that $\forall \varphi \in C_0^1(\Omega)$ have $\int_\Omega \varphi' \cdot u \, \mathrm{d}x = -\int_\Omega \varphi \, \mathrm{d}Du$. So

$$\begin{aligned}
\mathrm{TV}(u) &= \sup \left\{ \int_\Omega \varphi' \cdot u \, \mathrm{d}x \middle| \varphi \in C_0^1(\Omega), \|\varphi\|_{C^0} \leq 1 \right\} \\
&= \sup \left\{ \int_\Omega \varphi \, \mathrm{d}Du \middle| \varphi \in C_0^1(\Omega), \|\varphi\|_{C^0} \leq 1 \right\} \\
&\leq \sup \left\{ \int_\Omega \varphi \, \mathrm{d}Du \middle| \varphi \in C^0(\Omega), \|\varphi\|_{C^0} \leq 1 \right\} = \|Du\|_{\mathcal{M}} < \infty.
\end{aligned}$$

- **$\Leftarrow$:** Let $\mathrm{TV}(u) < \infty$. Then by linearity of integral for any $\varphi \in C_0^1(\Omega)$ have

$$\left| \int_\Omega \varphi' \cdot u \, \mathrm{d}x \right| \leq \mathrm{TV}(u) \cdot \|\varphi\|_{C^0}.$$

  So $C^0(\Omega) \supset C_0^1(\Omega) \ni \varphi \mapsto \int_\Omega \varphi' \cdot u \, \mathrm{d}x$ is linear and continuous in $C^0(\Omega)$ norm and therefore can be represented as integration with respect to a measure $-Du \in \mathcal{M}(\Omega)$. Since $\varphi(0) = \varphi(1) = 0$ can confine $-Du \in \mathcal{M}(\operatorname{int} \Omega)$.

- (Some details required, since functional first only defined on $C_0^1(\Omega)$. Need to 'remove' mean slope for $\phi \in C^1(\Omega)$, need approximation argument for $\phi \in C^0(\Omega)$. But functional remains linear and bounded.)

- **$\mathrm{TV}(u) = \|Du\|_{\mathcal{M}}$:** Need to show that both suprema above yield same value. Have already seen above: $\mathrm{TV}(u) \leq \|Du\|_{\mathcal{M}}$. Now show converse relation.

- Have established that $C^1(\Omega)$ lies dense in $C^0(\Omega)$ in C^0-norm. So can confine second supremum for test functions to determine $\|Du\|_{\mathcal{M}}$ to $C^1(\Omega)$.

- Let now $\phi \in C^1(\Omega)$ and for $\varepsilon > 0$ let $\psi_\varepsilon \in C_0^1(\Omega)$ with $\psi_\varepsilon(x) = 1$ if $x \in [\varepsilon, 1 - \varepsilon]$ and $\psi_\varepsilon(x) \in [0, 1]$ for $x \in \Omega$. Then $\phi \cdot \psi_\varepsilon \in C_0^1(\Omega)$ and

$$\left| \int_\Omega \phi \, \mathrm{d}Du - \int_\Omega \phi \cdot \psi_\varepsilon \, \mathrm{d}Du \right| \leq \|\phi\|_{C^0} \cdot |Du|((0, \varepsilon) \cap (1 - \varepsilon, 1)).$$

By regularity of positive Radon measures (cf. Theorem 2.130) $|Du|((0,\varepsilon)\cap(1-\varepsilon,1)) \to 0$ as $\varepsilon \to 0$. So the integral of any $\phi \in C^1(\Omega)$ can be approximated arbitrarily well by some $\phi \cdot \psi_\varepsilon \in C_0^1(\Omega)$ and thus the two suprema coincide.

□

Example 2.163. Reconsider Example 2.160(ii), $u = \chi_{[a,b]}$ for $0 < a < b < 1$. Then $Du = \delta_a - \delta_b$ and therefore $\mathrm{TV}(u) = \|Du\|_{\mathcal{M}} = 2$. This equals $\#(\partial[a,b]) = \#(\{a,b\})$. A test function that achieves the supremum is any function $\phi \in C_0^1(\Omega)$ with $\phi(a) = 1$, $\phi(b) = -1$.

Sketch: $\delta_a - \delta_b$, suitable ϕ, comparison with integral of ϕ' over $[a,b]$

We will extend this example to more general sets and their characteristic functions. First, understand structure of $\mathrm{BV}(\Omega)$ a little better.

Proposition 2.164. Let u_1, $u_2 \in \mathrm{BV}(\Omega)$. $\mathrm{TV}(u_1 - u_2) = 0$ if and only if there is some $r \in \mathbb{R}$ such that $u_1(x) - u_2(x) = r$ x-almost everywhere.

Proof.

- One has $u_1 - u_2 \in \mathrm{BV}(\Omega)$ with $D(u_1 - u_2) = Du_1 - Du_2$.
- $[\mathrm{TV}(u_1 - u_2) = 0] \Leftrightarrow [\int_\Omega \varphi' \cdot (u_1 - u_2)\,\mathrm{d}x = 0$ for all $\varphi \in C_0^1(\Omega)]$
- **⇐:** Assume $u_1(x) - u_2(x) = r \in \mathbb{R}$ x-almost everywhere. Then for $\varphi \in C_0^1(\Omega)$ get $\int_\Omega \varphi' \cdot (u_1 - u_2)\,\mathrm{d}x = r \cdot (\varphi(1) - \varphi(0)) = 0$.
- **⇒:** Assume $\mathrm{TV}(u_1 - u_2) = 0$. Set $r = \int_\Omega (u_1 - u_2)\,\mathrm{d}x$. For some $\psi \in C^0(\Omega)$ set $\overline{\psi} = \int_\Omega \psi\,\mathrm{d}x$ and define $\phi : \Omega \to \mathbb{R}$ via $\phi(x) = \int_0^x \psi(t)\,\mathrm{d}t - \overline{\psi} \cdot x$. Clearly $\phi \in C_0^1(\Omega)$ and $\phi'(x) = \psi(x) - \overline{\psi}$. Then:

$$\int_\Omega \psi(x) \cdot (u_1(x) - u_2(x) - r)\,\mathrm{d}x = \int_\Omega \left(\psi(x) - \overline{\psi}\right) \cdot (u_1(x) - u_2(x))\,\mathrm{d}x$$
$$= \int_\Omega \phi'(x) \cdot (u_1(x) - u_2(x))\,\mathrm{d}x = 0$$

This is true for any $\psi \in C^0(\Omega)$. Therefore $u_1(x) - u_2(x) = r$ x-almost everywhere.

□

Corollary 2.165. TV defines a semi-norm on $\mathrm{BV}(\Omega)$. Two functions u_1, $u_2 \in \mathrm{BV}(\Omega)$ belong to the same equivalence class (w.r.t. the TV semi-norm) if $\exists\, r \in \mathbb{R}$ such that $u_1(x) - u_2(x) = r$ x-almost everywhere.
If one identifies all functions in $\mathrm{BV}(\Omega)$ that only differ on negligible sets, we can identify $\mathrm{BV}(\Omega)$ with $\mathbb{R} \times \mathcal{M}(\operatorname{int}\Omega)$ with the identification rule $\mathrm{BV}(\Omega) \ni u \sim (r, Du) \in \mathbb{R} \times \mathcal{M}(\operatorname{int}\Omega)$ where $u(x) = r + Du((0,x))$ and $\|(r, Du)\| \stackrel{\text{def.}}{=} |r| + \|Du\|_{\mathcal{M}}$ defines a norm on this space.

Proposition 2.166. The representative $u : x \mapsto r + Du((0,x))$ of each equivalence class is left-continuous.

Proof.

- Let $x \in (0,1]$.

$$\limsup_{\varepsilon \searrow 0} |u(x) - u(x-\varepsilon)| = \limsup_{\varepsilon \searrow 0} |Du([x-\varepsilon, x))|$$
$$\leq \limsup_{\varepsilon \searrow 0} |Du(\{x-\varepsilon\})| + \limsup_{\varepsilon \searrow 0} |Du|((x-\varepsilon, x))$$

The first limsup must be 0 since otherwise Du would not be finite. The second limsup must be 0 by regularity of Radon measures (cf. Thm. 2.130).

□

We generalize the above example for general indicator functions.

Proposition 2.167. Let $S \subset \Omega$ be measurable. If $\#(\partial S) < \infty$ then $\#(\partial S) = \mathrm{TV}(\chi_S)$. Conversely, if $\mathrm{TV}(\chi_S) < \infty$ then there is some measurable $\hat{S} \subset \Omega$ such that $(\hat{S} \setminus S) \cup (S \setminus \hat{S})$ is negligible and $\mathrm{TV}(\chi_S) = \mathrm{TV}(\chi_{\hat{S}}) = \#(\partial \hat{S})$.

This implies that $\mathrm{TV}(\chi_S)$ approximates $\#(\partial S)$ reasonably well. Possibly we need to modify S on a negligible set, e.g. adding missing 'isolated points'.

Proof.

- Let $\#(\partial S) < \infty$. Then there are some pairs $(a_i, b_i)_{i=1}^n \in \mathbb{R}^{n\times 2}$ with $a_i < b_i$, $b_i \leq a_{i+1}$, such that up to a negligible set one has $[x \in S] \Leftrightarrow [x \in (a_i, b_i)$ for some i in $1, \ldots, n]$, possibly $a_1 \leq 0$, $b_n \geq 1$ (cf. Corollary 2.156).
- Analogous to Example 2.160(ii) one finds $D\chi_S = \sum_{i=1}^n \delta_{a_i} - \delta_{b_i}$ (again, possibly dropping a_1, b_n).
- Then $\#(\partial S) = \mathrm{TV}(\chi_S) = \|D\chi_S\|_{\mathcal{M}} = \#(\operatorname{int}\Omega \cap \{a_1, b_1, \ldots, a_n, b_n\})$.
- Now assume $\mathrm{TV}(\chi_S) < \infty$. Then the weak derivative $D\chi_S$ lies in $\mathcal{M}(\operatorname{int}\Omega)$.
- Consider the left-continuous representative $u_l : x \mapsto r_l + D\chi_S((0, x))$. Since u_l is left-continuous and coincides with χ_S almost everywhere, we have hat $u_l(x) \in \{0, 1\}$ for all $x \in (0, 1]$. By regularity of Radon measures must have $r_l \in \{0, 1\}$ (otherwise $u_l(x) \notin \{0, 1\}$ on a non-negligible set). So $u_l = \chi_{S_l}$ for some measurable $S_l \subset \Omega$ and $\mathrm{TV}(S) = \mathrm{TV}(S_l)$.
- Analogously, consider $u_r : x \mapsto r_r + D\chi_S((0, x])$. u_r is right-continuous and the indicator function of some measurable $S_r \subset \Omega$. Since $u_l(x) = u_r(x)$ x-a.e., have $\partial S_l = \partial S_r$.
- By left-continuity of u_l and right-continuity of u_r have for every $x \in \partial S_l = \partial S_r$ that there is some $\varepsilon > 0$ such that $(x - \varepsilon, x)$ and $(x, x + \varepsilon)$ either lie completely in $S_l \cap S_r$ or in the complement.
- Then $D\chi_S|_{(x-\varepsilon, x+\varepsilon)} = \pm\delta_x$. If ∂S_l were not finite, neither were $D\chi_S$. So ∂S_l must be finite. Equality of $\#(\partial S_l)$ and $\mathrm{TV}(\chi_{S_l})$ then follows from the first part.

$\square$

Now we approximate the original functional E_{set} for sets in terms of functions of bounded variation.

Definition 2.168.

$$E_{\text{func}} : \mathrm{BV}(\Omega) \to \mathbb{R} \cup \{\infty\}, \qquad u \mapsto \begin{cases} \int_\Omega a(x) \cdot u(x)\,\mathrm{d}x + \lambda \cdot \mathrm{TV}(u) & \text{if } u(x) \in [0, 1]\ x\text{-a.e.,} \\ +\infty & \text{else.} \end{cases}$$

From Prop. 2.167 we find:

Corollary 2.169. Let $S \subset \Omega$ be measurable. If $E_{\text{func}}(S) < \infty$ ($\Leftrightarrow \#(\partial S) < \infty$) then $E_{\text{set}}(S) = E_{\text{func}}(\chi_S)$. Conversely, if $E_{\text{func}}(\chi_S) < \infty$ then there is some measurable $\hat{S} \subset \Omega$ such that $(\hat{S} \setminus S) \cup (S \setminus \hat{S})$ is negligible and $E_{\text{func}}(\chi_S) = E_{\text{func}}(\chi_{\hat{S}}) = E_{\text{set}}(\hat{S})$.

Proof.

- Clear with above Proposition and $\int_S a(x)\,\mathrm{d}x = \int_\Omega a(x) \cdot \chi_S(x)\,\mathrm{d}x$.

$\square$

Note that optimizing E_{func} is a convex optimization problem. We now establish that it has a solution.

Proposition 2.170. E_{func} has minimizers in $\mathrm{BV}(\Omega)$.

Proof.

- We express $E(u)$ for $u \in \mathrm{BV}(\Omega)$ via the left-continuous representative $u : x \mapsto r + Du((0,x))$, and in particular via the pair (r, Du) (cf. Cor. 2.165 and Prop. 2.166). Assume $u(x) \in [0,1]$ for all x. We find

$$
\begin{aligned}
E(u) &= \int_\Omega a(x) \cdot u(x)\,\mathrm{d}x + \lambda\ \mathrm{TV}(u) = \int_\Omega a(x) \cdot [r + Du((0,x))]\,\mathrm{d}x + \lambda \|Du\|_{\mathcal{M}} \\
&= \int_\Omega a\,r\,\mathrm{d}x + \int_\Omega a(x) \cdot \left(\int_\Omega \underbrace{\chi_{(0,x)}(t)}_{=\chi_{(t,1)}(x)}\,\mathrm{d}Du(t) \right) \mathrm{d}x + \lambda \|Du\|_{\mathcal{M}} \\
&= \int_\Omega \underbrace{\left(\int_\Omega a(x)\,\chi_{(t,1)}(x)\,\mathrm{d}x \right)}_{\stackrel{\text{def.}}{=} A(t)} \mathrm{d}Du(t) + \lambda \|Du\|_{\mathcal{M}} = A(0) \cdot r + \int_\Omega A\,\mathrm{d}Du + \lambda \|Du\|_{\mathcal{M}}
\end{aligned}
$$

- We have used Fubini (swap order of integration) for which we used $a \in L^1(\Omega)$ (provides finiteness of integral).
- The first term is continuous in r. The second term is linear and weak$*$ continuous (since $A(t) = \int_t^1 a(x)\,\mathrm{d}x$ is continuous), the third term is weak$*$ lower semicontinuous and coercive.
- Consider a minimizing sequence $(u_k)_k$ of E_{func} and the corresponding $(r_k, Du_k)_k$. Due to the constraint have $r_k \in [0,1]$ for all k and by the above coerciveness $\|Du_k\|_{\mathcal{M}}$ is bounded. Therefore (Banach–Alaoglu), there is a subsequence such that $Du_{k_j} \stackrel{*}{\rightharpoonup} Du$ and $r_{k_j} \to r$ for some (r, Du). It is easy to verify that the corresponding u satisfies $u(x) \in [0,1]$ for all $x \in \Omega$. By the above discussed continuity and lower-semicontinuity, (r, Du) (or u) is a minimizer of E_{func}.

$\square$

Remark 2.171 (Motivation).

- The next question is whether minimizers for the original functional E_{set} can be recovered from minimizers of E_{func}.
- While a minimizer u of E_{set} need not be binary, i.e. $u(x) \notin \{0,1\}$ for some $x \in \Omega$ (possibly all), we will show that by thresholding we can (almost surely) generate an optimal set for E_{set}, i.e. by setting $S \stackrel{\text{def.}}{=} \{x \in \Omega : u(x) > t\} = \Omega \setminus S_t(u)$ for almost every t in $[0,1]$.
- For this we need to express $E_{\text{func}}(u)$ in terms of the sublevel or superlevel sets of u. Data term is relatively easy, need a suitable result for the regularizer.

Theorem 2.172 (Coarea formula [Ambrosio et al., Theorem 3.40])**.** For $u \in L^1(\Omega)$ have

$$\mathrm{TV}(u) = \int_{-\infty}^{\infty} \mathrm{TV}(\chi_{S_t(u)})\,\mathrm{d}t.$$

We illustrate this result with a smooth example.

Example 2.173.

- Let $u \in C^1(\Omega)$. Then $Du = u' \cdot \mathcal{L}_{\text{int}\,\Omega}$ and so $\text{TV}(u) = \int_\Omega |u'(x)|\,\mathrm{d}x$.
- Let $(I_i)_{i=1}^n$ be the interiors of the connected components of Ω where $u'(x) \neq 0$. That is, on each I_i u' is non-zero and (by continuity of u') has constant sign. Let σ_i be the corresponding sign. Then

$$\text{TV}(u) = \sum_{i=1}^n \sigma_i \int_{I_i} u'(x)\mathrm{d}x = \sum_{i=1}^n \int_{a_i}^{b_i} \mathrm{d}t = \int_{-\infty}^{\infty} \sum_{i=1}^n \chi_{(a_i,b_i)}(t)\,\mathrm{d}t$$

  where $a_i < b_i$ are chosen such that $(a_i, b_i) = u(I_i)$.

- If $t \in (a_i, b_i) = u(I_i)$ then there exists (by continuity of u) a unique (by strict monotonicity of u on I_i) $x \in I_i$ such that $u(x) = t$ and thus $\#(\partial S_t(u) \cap I_i) = 1$. Conversely, if $t \notin (a_i, b_i)$ then $\#(\partial S_t(u) \cap I_i) = 0$. So $\chi_{(a_i,b_i)}(t) = \#(\partial S_t(u) \cap I_i)$.
- Let $i \in \{1, \ldots, n\}$, if $t \notin \{a_i, b_i\}$ then $\partial I_i \cap \partial S_t(u) = \emptyset$. Further, $\Omega = \bigcup_i \overline{I_i} = \bigcup_i (I_i \cup \partial I_i)$. Therefore, for t-a.e. have $\#(\partial S_t(u)) = \sum_i \#(\partial S_t(u) \cap I_i) = \sum_i \chi_{(a_i,b_i)}(t) < \infty$. Finally,

$$\text{TV}(u) = \int_{-\infty}^{\infty} \#(\partial S_t(u))\,\mathrm{d}t = \int_{-\infty}^{\infty} \text{TV}(\chi_{S_t(u)})\,\mathrm{d}t\,.$$

We can now show how to recover minimizers of E_{set} from minimizers of E_{func}. To avoid issues with 'spurious' discontinuities we pick the left-continuous minimizers of E_{func} (which can be constructed from any minimizer via integration, see Prop. 2.166).

Proposition 2.174. If u is a left-continuous minimizer of E_{func} then for almost every threshold $t \in [0, 1]$ the (superlevel) set $\Omega \setminus S_t(u)$ is a minimizer of E_{set}.

Proof.

- Show a formula similar to coarea formula (Theorem 2.172) for data term:

$$\int_\Omega a(x)\,u(x)\,\mathrm{d}x = \int_\Omega a(x) \left(\int_0^1 \chi_{(0,u(x))}(t)\,\mathrm{d}t \right) \mathrm{d}x = \int_\Omega \int_0^1 a(x) \chi_{[\Omega \setminus S_t(u)]}(x)\,\mathrm{d}t\mathrm{d}x$$
$$= \int_0^1 \left(\int_{\Omega \setminus S_t(u)} a(x)\,\mathrm{d}x \right) \mathrm{d}t$$

- Do a few manipulations with coarea formula. Use $u(x) \in [0, 1]$ (i.e. $\partial S_t(u) = \emptyset$ for $t \notin [0, 1]$) and $\text{TV}(\chi_S) = \text{TV}(1 - \chi_S) = \text{TV}(\chi_{\Omega \setminus S})$ (since TV does not change when one flips sign and adds constant function).

$$\text{TV}(u) = \int_0^1 \text{TV}(\chi_{[\Omega \setminus S_t(u)]})\,\mathrm{d}t$$

  Since u is optimal, have $\text{TV}(u) < \infty$ and therefore $\text{TV}(\chi_{[\Omega \setminus S_t(u)]}) < \infty$ t-almost everywhere. By left-continuity of u this implies that $\#(\partial[\Omega \setminus S_t(u)]) = \text{TV}(\chi_{[\Omega \setminus S_t(u)]})$.

- Together find:

$$E_{\text{func}}(u) = \int_\Omega a(x) \cdot u(x)\,\mathrm{d}x + \lambda\,\text{TV}(u) = \int_0^1 E_{\text{set}}(\Omega \setminus S_t(u))\,\mathrm{d}t$$

- This implies $E_{\text{set}}(\Omega \setminus S_t(u)) < \infty$ t-a.e. and so by Cor. 2.169 (and since u is optimal)

$$E_{\text{set}}(\Omega \setminus S_t(u)) = E_{\text{func}}(\chi_{[\Omega \setminus S_t(u)]}) \geq E_{\text{func}}(u)$$

  This now implies $E_{\text{set}}(\Omega \setminus S_t(u)) = E_{\text{func}}(u)$ t-almost everywhere.

- Let t be some threshold with equality. Assume now $E_{\text{set}}(S) < E_{\text{set}}(\Omega \setminus S_t(u)) = E_{\text{func}}(u) < \infty$. Then $E_{\text{func}}(\chi_S) = E_{\text{set}}(S) < E_{\text{func}}(u)$, which contradicts optimality of u. So $\Omega \setminus S_t(u)$ must be optimal for E_{set}.

□

2.9.3 Discretization and Γ-convergence

We introduce a finite number of equidistant points in Ω and discretize E_{func} by only considering functions that are constant between two neighbouring points.

Definition 2.175. For $n \in \mathbb{N}$ let $\Omega_n = \left\{\frac{i}{n+1} \middle| i \in \{1, \dots, n\}\right\}$ and

$$E_n(u) = \begin{cases} E_{\text{func}}(u) & \text{if } \operatorname{spt} Du \subset \Omega_n, \\ +\infty & \text{else.} \end{cases}$$

Set $E_\infty = E_{\text{func}}$.

Minimizing E_n corresponds to a finite-dimensional optimization problem.

Proposition 2.176 (Discrete functional)**.** If $E_n(u) < \infty$ then there is some $\hat{u} \in [0,1]^{n+1} \subset \mathbb{R}^{n+1}$ with $u(x) = \hat{u}_i$ for almost every $x \in (0,1)$ where $i \in \{1, \dots, n+1\}$ is determined by $x \in (\frac{i-1}{n+1}, \frac{i}{n+1}]$. Then $E_n(u) = \hat{E}_n(\hat{u})$ with

$$\hat{E}_n : \mathbb{R}^{n+1} \to \mathbb{R} \cup \{\infty\}, \qquad \hat{u} \mapsto \sum_{i=1}^{n+1} \hat{a}_{n,i} \cdot \hat{u}_i + \lambda \sum_{i=1}^{n} |\hat{u}_{i+1} - \hat{u}_i| + \sum_{i=1}^{n+1} \iota_{[0,1]}(\hat{u}_i)$$

where for $i \in \{1, \dots, n+1\}$ set

$$\hat{a}_{n,i} = \int_{\frac{i-1}{n+1}}^{\frac{i}{n+1}} a(x)\,\mathrm{d}x.$$

Conversely, if $\hat{E}_n(\hat{u}) < \infty$ then u constructed from $\hat{u}$ in the above way yields $E_n(u) = \hat{E}_n(\hat{u})$.

Proof.

- If $E_n(u) < \infty$ then $\operatorname{spt} Du \subset \Omega_n$, which implies $Du = \sum_{i=1}^n \delta_{\frac{i}{n+1}} \cdot \hat{Du}_i$ for some $\hat{Du} \in \mathbb{R}^n$.

- Pick now u to be the left-continuous representative of u, Prop. 2.166. For $x \in (\frac{i-1}{n+1}, \frac{i}{n+1}]$ where $i \in \{1, \dots, n+1\}$ get

$$u(x) = r + Du((0,x)) = r + \sum_{k=1}^{i-1} \hat{Du}_k$$

  which implies that u is constant on intervals $(\frac{i-1}{n+1}, \frac{i}{n+1})$ and $\hat{u}_1 = r$, $\hat{u}_i = r + \sum_{k=1}^{i-1} \hat{Du}_k$ for $i \in \{2, \dots, n+1\}$.

- $\Rightarrow \hat{Du}_i = \hat{u}_{i+1} - \hat{u}_i$ for $i \in \{1, \dots, n\}$. And $E_n(u) < \infty \Rightarrow u(x) \in [0,1]$ x-a.e. $\Rightarrow \hat{u} \in [0,1]^{n+1}$.
- Now:

$$\begin{aligned}E_n(u) &= \int_\Omega a(x) \cdot u(x)\,\mathrm{d}x + \lambda \|Du\| = \sum_{i=1}^{n+1} \int_{\frac{i-1}{n+1}}^{\frac{i}{n+1}} a(x)\mathrm{d}x \cdot \hat{u}_i + \lambda \sum_{i=1}^{n} |\hat{Du}_i| \\ &= \sum_{i=1}^{n+1} \hat{a}_i \cdot \hat{u}_i + \lambda \sum_{i=1}^{n} |\hat{u}_{i+1} - \hat{u}_i| + \sum_{i=1}^{n+1} \iota_{[0,1]}(\hat{u}_i) = \hat{E}_n(\hat{u})\end{aligned}$$

- Conversely, if $\hat{E}_n(\hat{u}) < \infty$ then define u as piecewise constant, as above. Then $u(x) \in [0,1]$ for all $x \in \Omega$. So we can ignore the $[0,1]$-constraint in E_{func} and therefore we can reverse the above computation to find $\hat{E}_n(\hat{u}) = E_n(u)$.

$\square$

Proposition 2.177 (Γ-convergence)**.** We represent functions $u \in \mathrm{BV}(\Omega)$ by pairs $(r, Du) \in \mathbb{R} \times \mathcal{M}(\operatorname{int}\Omega)$ (Cor. 2.165) and equip the space $\mathrm{BV}(\Omega)$ with the product topology of the standard topology on $\mathbb{R}$ and the weak$*$ topology on $\mathcal{M}(\Omega)$. For simplicity we denote convergence in this topology by $(u_n \equiv (r_n, Du_n)) \stackrel{*}{\rightharpoonup} (u \equiv (r, Du)) \Leftrightarrow [r_n \to r \wedge Du_n \stackrel{*}{\rightharpoonup} Du]$.
In this topology, E_n Γ-converges to E_∞. Further, $\lim_{n\to\infty} \min E_n = \min E_\infty$ and any sequence of minimizers of E_n has a convergent subsequence (in the topology described above) such that its limit is a minimizer of E_∞.

Proof.

- **liminf:** Let $u_n \stackrel{*}{\rightharpoonup} u$. W.l.o.g. we can focus on sequences where $E_n(u_n) < \infty$ (subsequences with $E_n(u_n) = \infty$ only contribute to the liminf with $+\infty$). Then $E_n(u_n) = E_\infty(u_n) = E_{\text{func}}(u_n)$. As shown in the proof of Prop. 2.170, E_{func} is sequentially lower-semicontinuous in the considered topology. So $\liminf_n E_n(u_n) = \liminf_n E_{\text{func}}(u_n) \geq E_{\text{func}}(u) = E_\infty(u)$.
- **limsup:** Let u be represented by (r, Du). For $n \in \mathbb{N}$ set $r_n = r$ and

$$Du_n = \sum_{i=1}^{n} \delta_{i/(n+1)} Du\left(\left(\tfrac{i-1}{n+1}, \tfrac{i}{n+1}\right]\right)$$

Sketch: Intervals of Ω, which areas get put where. Final interval is ignored.

Since $Du((\frac{n}{n+1}, 1)) \to 0$ as $n \to \infty$, analogous to Prop. 2.148 have $Du_n \stackrel{*}{\rightharpoonup} Du$. So $(r_n, Du_n) \stackrel{*}{\rightharpoonup} (r, Du)$.

- $\mathrm{TV}(u_n) = \|Du_n\|_{\mathcal{M}} = \sum_{i=1}^n |Du((\frac{i-1}{n+1}, \frac{i}{n+1}])| \leq |Du|(\Omega) = \|Du\|_{\mathcal{M}} = \mathrm{TV}(u)$. Moreover, by weak$*$ convergence have $\int_\Omega A\,\mathrm{d}Du_n \to \int_\Omega A\,\mathrm{d}Du$ and $A(0) \cdot r_n = A(0) \cdot r$ with A as defined in proof of Prop. 2.170. So

$$\limsup_{n\to\infty} E_n(u_n) = \limsup_{n\to\infty} A(0) \cdot r_n + \int_\Omega A\,\mathrm{d}Du_n + \lambda\,\mathrm{TV}(u_n) \leq E_\infty(u).$$

- Since the additional constraint $\operatorname{spt} Du \subset \Omega_n$ to change E_{func} into E_n is sequentially weak$*$ closed and non-empty, by Prop. 2.170 every E_n has a minimizer. So $\min E_n$ is well-defined.

- Due to the constraint $u(x) \in [0,1]$ for all x and the TV-term the functionals E_n are equi-mildly [weak$*$ + standard topology on r] coercive and sequences of minimizers are bounded. Convergence of optimal values and convergence of (subsequences) of minimizers then follow from Banach–Alaoglu and Prop. 2.112.

$\square$

2.9.4 Optimization algorithm

Functional $\hat{E}_n$ from Prop. 2.176 can be written as follows:

$$\hat{E}_n(\hat{u}) = \sum_{i=1}^{n+1} \left[\hat{a}_{n,i} \cdot \hat{u}_i + \iota_{[0,1]}(\hat{u}_i)\right] + \lambda \sum_{i=1}^{n} |\hat{u}_{i+1} - \hat{u}_i| = F(\hat{u}) + G(A\,\hat{u})$$

with

$$\begin{aligned}
&F : \mathbb{R}^{n+1} \to \mathbb{R} \cup \{\infty\}, && \hat{u} \mapsto \sum_{i=1}^{n+1} f_i(\hat{u}_i), \quad f_i(\hat{u}_i) = \hat{a}_{n,i} \cdot \hat{u}_i + \iota_{[0,1]}(\hat{u}_i), \\
&A : \mathbb{R}^{n+1} \to \mathbb{R}^n, && (A\,\hat{u})_i = \hat{u}_{i+1} - \hat{u}_i, \\
&G : \mathbb{R}^n \to \mathbb{R}, && \hat{v} \mapsto \sum_{i}^{n} g_i(\hat{v}_i), \quad g_i(\hat{v}_i) = |\hat{v}_i|.
\end{aligned}$$

Using $\hat{u} = 0 \in \mathbb{R}^{n+1}$ we find that $F(\hat{u}) = 0 < \infty$ and since G is finite and continuous, have $G(A\,\hat{u}) < \infty$ and G is continuous in $A\,\hat{u}$. So the Fenchel–Rockafellar theorem (cf. Prop. 1.135 and exercise sheet) implies that the corresponding dual problem has a solution. We can solve primal and dual problem with the extension of the primal-dual algorithm, Prop. 1.138 (see again exercise sheet):
Let $\tau, \sigma \in \mathbb{R}_{++}$, $\tau\sigma < \|A\|^{-2}$ and $(\hat{u}^{(0)}, \hat{w}^{(0)}) \in (X = \mathbb{R}^{n+1}) \times (Y = \mathbb{R}^n)$. Then set:

$$\begin{aligned}
\hat{u}^{(\ell+1)} &= \operatorname{Prox}_{\tau F}(\hat{u}^{(\ell)} - \tau A^* \hat{w}^{(\ell)}), \\
\hat{w}^{(\ell+1)} &= \operatorname{Prox}_{\sigma G^*}(\hat{w}^{(\ell)} + \sigma A(2\hat{u}^{(\ell+1)} - \hat{u}^{(\ell)})).
\end{aligned}$$

Then $\hat{u}^{(\ell)} \rightharpoonup \hat{u}$, $\hat{w}^{(\ell)} \rightharpoonup \hat{w}$ as $\ell \to \infty$ where $(\hat{u}, \hat{w})$ are a pair of primal and dual solutions.
Compute the proximal operators: let $\hat{p}$, $\hat{u} \in \mathbb{R}^{n+1}$. Then $\hat{p} = \operatorname{Prox}_{\tau F}(\hat{u}) \Leftrightarrow \hat{p}_i = \operatorname{Prox}_{\tau f_i}(\hat{u}_i) \Leftrightarrow \hat{u}_i - \hat{p}_i \in \tau \partial f_i(\hat{p}_i)$ (Prop. 1.108). We find:

$$\partial f_i(\hat{p}_i) = \begin{cases} (-\infty, \hat{a}_{n,i}] & \text{if } \hat{p}_i = 0, \\ \{\hat{a}_{n,i}\} & \text{if } \hat{p}_i \in (0,1), \\ [\hat{a}_{n,i}, \infty) & \text{if } \hat{p}_i = 1, \\ \emptyset & \text{else.} \end{cases} \qquad \hat{p}_i = \begin{cases} \hat{u}_i - \tau \cdot \hat{a}_{n,i} & \text{if } \hat{u}_i - \tau \cdot \hat{a}_{n,i} \in [0,1], \\ 0 & \text{if } \hat{u}_i - \tau \cdot \hat{a}_{n,i} < 0, \\ 1 & \text{if } \hat{u}_i - \tau \cdot \hat{a}_{n,i} > 1. \end{cases}$$

Now consider $\operatorname{Prox}_{\sigma G^*}$. We find $G^* = \sum_{i=1}^n g_i^*$. And $g_i^* = (|\cdot|)^* = \iota_{[-1,1]}$. Then, as above for $\hat{q}$, $\hat{w} \in \mathbb{R}^n$ have $\hat{q} = \operatorname{Prox}_{\sigma G^*}(\hat{w}) \Leftrightarrow \hat{q}_i = \operatorname{Prox}_{\sigma g_i^*}(\hat{w}_i) = P_{[-1,1]}\hat{w}_i$.
Both proximal steps are pointwise and very simple to evaluate. The non-smoothness of the problem is no problem for the algorithm.

2.10 Local optimality theory

Remark 2.178 (Motivation)**.** Not all problems can be solved with non-smooth convex analysis. Sometimes need to rely on more classical arguments: if a smooth function has a local minimum at some point its derivative vanishes in that point. On general Banach spaces need to be a little bit more careful about notion of derivative.

Definition 2.179 (Gâteaux and Fréchet differentials)**.**

- Let X, Y be normed vector spaces, $T : X \to Y$. The *Gâteaux differential* of T at $x \in X$ in direction $h \in X$ is

$$\delta T(x; h) = \lim_{\alpha \to 0} \frac{T(x + \alpha\, h) - T(x)}{a} = \frac{\mathrm{d}}{\mathrm{d}\alpha}\, T(x + \alpha\, h)|_{\alpha=0} \quad \text{if the limit exists.}$$

- T is *Gâteaux differentiable* at x if $\delta T(x; h)$ exists for all $h \in X$.
- If there is a $\delta T(x; \cdot) \in L(X, Y)$ with $\lim_{h \to 0} \frac{\|T(x+h) - T(x) - \delta T(x;h)\|}{\|h\|} = 0$ then T is *Fréchet differentiable* at x with *Fréchet differential* $\delta T(x, \cdot)$.

Proposition 2.180.

(i) The Fréchet differential is unique (if it exists).

(ii) Fréchet differentiable $\Rightarrow$ Gâteaux differentiable, and both differentials coincide.

(iii) Fréchet differentiable in x $\Rightarrow$ continuous in x.

The proof is a direct application of the definitions.

Example 2.181.

(i) $f \in C^1(\mathbb{R}^n)$ has Fréchet differential $\delta f(x; h) = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x) \cdot h_i = Df(x)\, h$. Indeed $\delta f(x; \cdot) \in L(\mathbb{R}^n, \mathbb{R})$ and by Taylor's theorem $f(x + h) = f(x) + \delta f(x; h) + o(h)$.

(ii) Let $g \in C^1(\mathbb{R}^2)$, define $f : C^0([0, 1]) \to \mathbb{R}$ by $f : x \mapsto \int_0^1 g\big(x(t), t\big)\, \mathrm{d}t$. The Gâteaux differential of f in direction $h \in C^0([0, 1])$ is

$$\delta f(x; h) = \tfrac{\mathrm{d}}{\mathrm{d}\alpha} \int_0^1 g(x(t) + \alpha\, h(t), t)\, \mathrm{d}t\bigg|_{\alpha=0} = \int_0^1 \frac{\partial g}{\partial x}(x(t), t)\, h(t)\, \mathrm{d}t.$$

f is even Fréchet differentiable: $\partial f(x; \cdot)$ is linear and

$$\begin{aligned} |f(x + h) - f(x) - \delta f(x; h)| &= \left| \int_0^1 \left[g(x + h, t) - g(x, t) - \frac{\partial g}{\partial x}(x, t)\, h \right] \mathrm{d}t \right| \\ &= \left| \int_0^1 \left[\frac{\partial g}{\partial x}(\overline{x}, t) - \frac{\partial g}{\partial x}(x, t) \right] \cdot h\, \mathrm{d}t \right| \end{aligned}$$

(use mean value theorem: $\overline{x}(t) \in [x(t), x(t) + h(t)]$)

$$\leq \|h\|_{C^0} \cdot \left\| \frac{\partial g}{\partial x}(\overline{x}, t) - \frac{\partial g}{\partial x}(x, t) \right\|_{C^0}$$

where the second term tends to 0 as $h \to 0$ in $C^0([0, 1])$: because $h(t) \to 0$ uniformly for all $t \in [0, 1]$, have $\overline{x}(t) \to x(t)$ uniformly, and by uniform continuity of $\frac{\partial g}{\partial x}$ (since $[0, 1]$ compact metric space, cf. Lemma 2.49).

(iii) Let $T \in L(X,Y)$, then $\delta T = T$ is its own Fréchet differential.

(iv) Let

$$f : \mathbb{R}^2 \to \mathbb{R}, \qquad (x_1, x_2) \mapsto \begin{cases} 1 & \text{if } x_2 \in [\frac{1}{2}x_1^2, x_1^2], \\ 0 & \text{else.} \end{cases}$$

Sketch: Draw $\mathbb{R}^2$, highlight 'parabola' where $f = 1$.

For all $(h_1, h_2) \in \mathbb{R}^2 \setminus (0,0)$ have $f(\alpha h) = 0$ for $\alpha > 0$ sufficiently small (why?). So the Gâteaux differential of f at 0 is $\delta f(0; h) = 0$. But f is not continuous in 0. So f cannot be Fréchet differentiable in 0 (Prop. 2.180(iii)).

Definition 2.182 (Fréchet derivative)**.** Let $T : X \to Y$, T Fréchet differentiable for all $x \in X$ with Fréchet differential $\delta T(x; \cdot)$. The map

$$T' : X \to L(X,Y), \qquad x \mapsto \delta T(x; \cdot)$$

is called Fréchet derivative of T.

Proposition 2.183.

(i) Taking derivative is linear: Let T_1, $T_2 : X \to Y$ Fréchet differentiable, α_1, $\alpha_2 \in \mathbb{R}$. Then $(\alpha_1 T_1 + \alpha_2 T_2)' = \alpha_1 T_1' + \alpha_2 T_2'$.

(ii) Chain rule: Let $T_1 : X \to Y$, $T_2 : Y \to Z$ Fréchet differentiable. Then $T = T_2 \circ T_1$ is Fréchet differentiable with $T'(x) \overset{\text{def.}}{=} T_2'(T_1(x)) \circ T_1'(x)$.

Proof.

- **(i):** Follows quickly from definition. $h \mapsto T_i(x+h) - T_i(x) - T_i'(x) = o(h)$. Then so is their linear combination.

- **(ii):** Show that $T'(x) = T_2'(T_1(x)) \circ T_1'(x)$ is Fréchet differential of T at x. Uniqueness was shown in Prop. 2.180(i). Need to show:

$$\lim_{h \to 0} \frac{1}{\|h\|_X} \big\| T_2(T_1(x+h) - T_2(T_1(x)) - T'(x) \big\|_Z = 0$$

- Have $T_2(T_1(x+h)) = T_2(T_1(x)) + T_2'(T_1(x))\,(T_1(x+h) - T_1(x)) + o(T_1(x+h) - T_1(x))$.

- $\lim_{h\to 0} \frac{\|T_1(x+h) - T_1(x)\|_Y}{\|h\|_X} \leq \|T_1'(x)\|_{L(X,Y)} < \infty$. So $o(T_1(x+h) - T_1(x)) = o(h)$. So:

$$\begin{aligned}
&\lim_{h \to 0} \frac{1}{\|h\|_X} \big\| T_2(T_1(x+h) - T_2(T_1(x)) - T'(x) \big\|_Z \\
&\qquad \leq \frac{1}{\|h\|_X} \big(o(h) + \|T_2'(T_1(x))\,(T_1(x+h) - T_1(x) - T_1'(x)h)\|_Z \big) \\
&\qquad\qquad \leq \frac{1}{\|h\|_X} \big(o(h) + \|T_2'(T_1(x))\|_{L(Y,Z)} \cdot o(h) \big) = 0
\end{aligned}$$

□

From the definition of the Fréchet differential, Def. 2.179, see that it can be used to locally approximate function.

Proposition 2.184 (Taylor-type formula)**.** Let $T : X \to Y$ be Fréchet differentiable on an open set $D \subset X$ with $[x, x+h] \subset D$. Then

$$\|T(x+h) - T(x)\| \leq \sup_{\alpha \in [0,1]} \|T'(x+\alpha\, h)h\|.$$

Proof. • Let $t \in Y^*$ be aligned with $T(x+h) - T(x)$ and $\|t\| = 1$ (so $\langle t, T(x+h) - T(x)\rangle_{Y^* \times Y} = \|t\|_{Y^*} \cdot \|T(x+h) - T(x)\|_Y$, see Def. 2.83. Existence of t provided by Hahn–Banach, see Prop. 2.81)

- Let $\varphi : [0,1] \to \mathbb{R}$, $\alpha \mapsto t(T(x+\alpha\, h))$. Find $\varphi'(\alpha) = t(T'(x+\alpha\, h)\, h)$. (Use chain rule, Prop. 2.183(ii) and $t' = t$, see Example 2.181(iii))
- mean value theorem: $\varphi(1) - \varphi(0) \leq \sup_{\alpha\in[0,1]} \varphi'(\alpha)$. Combine:

$$\begin{aligned}\|T(x+h) - T(x)\| = t(T(x+h) - T(x)) = |\varphi(1) - \varphi(0)|& \\ \leq \sup_{\alpha\in[0,1]} |t(T'(x+\alpha\, h)\, h| \leq \sup_{\alpha\in[0,1]} \|T'(x+\alpha\, h)\, h\|&\end{aligned}$$

□

Analogously can show:

Proposition 2.185. Let T be twice Fréchet differentiable on open $D \subset X$ with $[x, x+h] \subset D$. Then

$$\|T(x+h) - T(x) - T'(x)\, h\| \leq \tfrac{1}{2} \sup_{\alpha\in[0,1]} \|T''(x+\alpha\, h)(h)(h)\|.$$

The Gâteaux and Fréchet differentials can be used to characterize local minima of functions. Some examples for necessary conditions are given below.

Definition 2.186 (Local minimum)**.** Let $\Omega \subset X$, $f : \Omega \to \mathbb{R}$. $x_0 \in \Omega$ is a *(strict) local minimum* of f on Ω if $f(x_0) \leq f(x)$ (strict: $f(x_0) < f(x)$) for all x in a neighbourhood of x_0 (and a global minimum if the neighbourhood is Ω).

Proposition 2.187. Let $f : X \to \mathbb{R}$ be Gâteaux differentiable in x and have a local minimum at x. Then $\delta f(x; \cdot) = 0$.

Proof. • For all $h \in X$, $\alpha \mapsto f(x+\alpha\, h)$ must have a local minimum at $\alpha = 0$. $\Rightarrow \frac{\mathrm{d}}{\mathrm{d}\alpha} f(x + \alpha\, h)|_{\alpha=0} = 0$.

□

Proposition 2.188. Let $f : X \to \mathbb{R}$ be Gâteaux differentiable at $x_0 \in \Omega$ for $\Omega \subset X$ convex, and let f have a local minimum at x_0. Then $\delta f(x_0; x - x_0) \geq 0$ for all $x \in \Omega$.

Proof. • [Ω convex] $\Rightarrow$ [$x_0 + \alpha\,(x - x_0) \in \Omega$ for all $\alpha \in [0,1]$] $\Rightarrow$ [$\frac{\mathrm{d}}{\mathrm{d}\alpha} f(x_0 + \alpha(x - x_0))|_{\alpha=0} \geq 0$]

□

2.11 Euler–Lagrange equations

Proposition 2.189 (Euler–Lagrange equation)**.** Let $f \in C^1(\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R})$. For given $a, b \in \mathbb{R}$ set $X = \{x \in C^1([t_1, t_2]; \mathbb{R}^n) : x(t_1) = a,\, x(t_2) = b\}$. Define $J : X \to \mathbb{R}$ by

$$x \mapsto \int_{t_1}^{t_2} f\big(x(t), \dot{x}(t), t\big)\, \mathrm{d}t.$$

If $x \in X$ is a local minimizer of J then for $t \in [t_1, t_2]$

$$0 = \partial_x f(x, \dot{x}, t) - \tfrac{\mathrm{d}}{\mathrm{d}t} \partial_{\dot{x}} f(x, \dot{x}, t)$$

where the derivative $\frac{\mathrm{d}}{\mathrm{d}t}$ in the last term is well-defined. This is called the *Euler–Lagrange equation.*

The proof also uses the two Propositions below.

Proof.

- If $x \in X$ is a local minimizer then for $h \in C^1([t_1, t_2]; \mathbb{R}^n)$ with $h(t_1) = h(t_2) = 0$ need

$$0 = \delta J(x; h) = \int_{t_1}^{t_2} \Big[\partial_x f(x, \dot{x}, t)\, h + \partial_{\dot{x}} f(x, \dot{x}, t)\, \dot{h}\Big]\, \mathrm{d}t$$

(fow now assume that one can apply integration by parts on the second term, justify this below)

$$= \int_{t_1}^{t_2} \big[\partial_x f(x, \dot{x}, t) - \tfrac{\mathrm{d}}{\mathrm{d}t} \partial_{\dot{x}} f(x, \dot{x}, t)\big]\, h\, \mathrm{d}t$$

- Since this must hold for all allowed h, by Prop. 2.190 the first term in the integrand must be 0 almost everywhere, and then by continuity everywhere.

- Integration by parts: introduce $A : [t_1, t_2] \to \mathbb{R}$, $t \mapsto \int_{t_1}^{t} \partial_x f(x, \dot{x}, s)\, \mathrm{d}s$. Then $\frac{\mathrm{d}}{\mathrm{d}t} A(t) = \partial_x f(x, \dot{x}, t)$. Via integration by parts: $\int_{t_1}^{t_2} \partial_x f(x, \dot{x}, t)\, h\, \mathrm{d}t = -\int_{t_1}^{t_2} A(t)\, \dot{h}(t)\, \mathrm{d}t$. So from first line above find:

$$0 = \int_{t_1}^{t_2} [-A(t) + \partial_{\dot{x}} f(x, \dot{x}, t)]\, \dot{h}\, \mathrm{d}t$$

- So by Prop. 2.191 $-A(t) + \partial_{\dot{x}} f(x, \dot{x}, t) = \text{const}$ a.e., and then by continuity of both summands everywhere. Then differentiability of A implies differentiability of the other term and therefore, integration by parts is admissible.

□

Proposition 2.190 (Fundamental lemma of the calculus of variations)**.** Let $\Omega \subset \mathbb{R}^n$ open, $g \in L^1(\Omega)$. Then:

$$(1)\, \Big[\textstyle\int_\Omega h\, g\, \mathrm{d}x = 0\, \forall\, h \in C_0^\infty(\Omega)\Big] \Leftrightarrow (2)\, \Big[\textstyle\int_E g\, \mathrm{d}x = 0\, \forall\, \text{bounded measurable } E \subset\subset \Omega\Big]$$
$$\Leftrightarrow (3)[g = 0 \text{ almost everywhere}]$$

Proof.

- **(1) ⇒ (2):** Let $(\varphi_k)_k$ be a Dirac sequence in $C_0^\infty(\mathbb{R}^n)$. (This means $\varphi_k \cdot \mathcal{L} \overset{*}{\rightharpoonup} \delta_0$ as $k \to \infty$.) Set $h_k = \varphi_k * \chi_E$ (see Def. 2.158 for χ_E, $*$ denotes convolution)

- Then $0 = \int_\Omega h_k\, g \,\mathrm{d}x \to \int_\Omega \chi_E\, g \,\mathrm{d}x$ as $k \to \infty$, with the dominated convergence theorem (since $0 \leq h_k \leq 1$ and $h_k \to \chi_E$ a.e.).
- **(2) ⇒ (3):** For $\varepsilon > 0$ let $\varphi_\varepsilon = \frac{1}{|B(0,\varepsilon)|}\chi_{B(0,\varepsilon)}$. Then $g_\varepsilon(x) \overset{\text{def.}}{=} (\varphi_\varepsilon * g)(x) = \frac{1}{|B(x,\varepsilon)|}\int_{B(x,\varepsilon)} g \,\mathrm{d}x = 0$.
- $g_\varepsilon \to g$ in $L^1(\Omega)$ and thus also a.e. for a subsequence $\varepsilon \to 0$.
- **(3) ⇒ (1):** clear.

□

Comment: Have used similar arguments several times throughout Section 2.9 (Example: binary image segmentation), e.g. in Prop. 2.164.

Proposition 2.191. Let $\Omega \subset \mathbb{R}^n$ open and connected, $u \in L^1_{\mathrm{loc}}(\Omega)$. If $\int_\Omega u\, \partial_i h \,\mathrm{d}x = 0$ for all $h \in C_0^\infty(\Omega)$, $i = 1, \ldots, n$, then $u = \text{const}$ almost everywhere.

Proof.
- Let $B \subset\subset \Omega$, B open, $0 < \varepsilon < \mathrm{dist}(B, \partial\Omega)$, $\varphi_\varepsilon \in C_0^\infty(\mathbb{R}^n)$ a Dirac sequence (as $\varepsilon \to 0$) and $h \in C_0^\infty(B)$.

$$\int_\Omega \partial_i(u * \varphi_\varepsilon)\, h \,\mathrm{d}x = -\int_\Omega (u * \varphi_\varepsilon)\, \partial_i h \,\mathrm{d}x = -\int_\Omega u\, (\varphi_\varepsilon(-\cdot) * \partial_i h) \,\mathrm{d}x$$
$$= -\int_\Omega u\, \partial_i(\varphi_\varepsilon(-\cdot) * h) \,\mathrm{d}x = 0$$

- $\Rightarrow \nabla(u * \varphi_\varepsilon) = 0$ on B (see Proposition above) $\Rightarrow u * \varphi_\varepsilon = \text{const}$ on $B \Rightarrow u = \text{const}$ a.e. on $B \Rightarrow u = \text{const}$ a.e. on Ω.

□

Example 2.192 (Shortest curve)**.** Given (t_1, a), (t_2, b), what is the curve $x \in C^1([t_1, t_2])$ connecting the two points, i.e. $x(t_1) = a$, $x(t_2) = b$, with minimal arclength?

Sketch: Graph of x on $[t_1, t_2]$, arclength formula.

$$J(x) = \int_{t_1}^{t_2} \sqrt{1 + |\dot{x}|^2}\mathrm{d}t \quad \Rightarrow \quad 0 = \tfrac{\mathrm{d}}{\mathrm{d}t}\tfrac{\partial}{\partial \dot{x}}\sqrt{1 + |\dot{x}|^2} \quad \Rightarrow \quad \dot{x} = \text{const}$$

So the shortest path is a straight line.

Example 2.193 (Maximal utility)**.**
- $x(t)$: capital at time t, α: interest rate, $r(t)$: expenditure $\Rightarrow \dot{x}(t) = \alpha\, x(t) - r(t)$. $u(r(t))$: utility / "pleasure" derived from expenditure $r(t)$. T: lifetime, s initial capital. So maximize total utility:

$$J(x) = \int_0^T \exp(-\beta\, t)\, u(r(t)) \,\mathrm{d}t = \int_0^T \exp(-\beta\, t)\, u(\alpha\, x(t) - \dot{x}(t)) \,\mathrm{d}t$$

such that $x(0) = s$, $x(T) = 0$. (Weighting factor $\exp(-\beta\, t)$: future pleasure counts less.)

- Euler–Lagrange equation: $0 = \alpha \exp(-\beta\, t)\, u'(\alpha\, x - \dot{x}) + \frac{\mathrm{d}}{\mathrm{d}t} \exp(-\beta\, t) u'(\alpha\, x - \dot{x})$
- $\Rightarrow u'(r(t)) = u'(r(0)) \exp((\beta - \alpha)\, t)$
- e.g. $u(r) = 2\sqrt{r}$, $r(t) = r(0) \exp(2(\alpha - \beta)\, t) \Rightarrow x(t) = B \cdot \exp(\alpha\, t) + \frac{r(0)}{2\beta - \alpha} \exp(2(\alpha - \beta)\, t)$
- Now solve for $x(0) = s$, $x(T) = 0$ w.r.t. B, $r(0)$. E.g. for $\alpha > \beta > \alpha/2 \Rightarrow$ capital first grows, then decreases.

2.12 Inequality constraints

Throughout this section let X, Y be normed spaces.

Definition 2.194 (Positive cone)**.** A cone $P \subset X$ is *positive* if $P \cap (-P) = \{0\}$. A convex, positive cone induces a partial order on X via $[x \leq_P z] \Leftrightarrow [z - x \in P]$. We write $x <_P z$ if $z - x \in \operatorname{int} P$.

Proof that $\leq$ satisfies axioms of partial ordering.

- Reflexivity: $x \leq_P x$ on X: $[P$ positive$] \Rightarrow [0 \in P] \Rightarrow [x - x \in P] \Rightarrow x \leq_P x$.
- Transitivity: Let $x \leq_P y \leq_P z$. $\Rightarrow [y - x \in P \wedge z - y \in P] \Rightarrow$ (by convexity of P) $[\frac{1}{2}((z-y)-(y-x)) \in P] \Rightarrow$ (cone scaling property) $[z - x \in P] \Rightarrow [x \leq_P z]$.
- Antisymmetry: $[x \leq_P y \wedge y \leq_P x] \Rightarrow [x = y]$: By assumption $[y - x \in P] \wedge [x - y \in P]$. So $x - y \in P \cap (-P) = \{0\}$.

□

Proposition 2.195 (Dual cone)**.** For a positive, convex cone $P \subset X$ the *dual cone* is given by

$$P^* = \{t \in X^* : \langle t, x\rangle \geq 0 \text{ for all } x \in X\}.$$

If $\operatorname{int} P \neq \emptyset$ then P^* is a convex, positive cone.

Comment: Compare to Definition of polar cone, Def. 1.42. On Banach space must define 'bi-dual' cone asymmetrically, cf. Def. 2.86.

Proof.

- Clear that P^* is convex cone.
- Positivity: Assume $t \in P^* \cap -(P^*)$. $\Rightarrow \langle t, x\rangle = 0$ for all $x \in P$. Let $x_0 \in \operatorname{int} P$ with $\varepsilon > 0$ such that $B(x_0, \varepsilon) \subset P$. Then for any $x \in B(0, \varepsilon)$ have $x + x_0 \in P \Rightarrow 0 = \langle t, x + x_0\rangle = \langle t, x\rangle$ $\Rightarrow t = 0$.

□

Example 2.196. (i) $X = \mathbb{R}^n$, $P = [0, \infty)^n$. $P^* = P$. $y \leq_P x$ means $y \leq x$ component-wise.

(ii) $X = L^2([0,1])$, $P = \{u \in X | u(x) \geq 0 \text{ a.e.}\}$. $P^* = P$, $\operatorname{int} P = \emptyset$.

Definition 2.197. Let $G : X \to Y$, $P \subset Y$ be a convex, positive cone. G is *convex* if for all $\lambda \in [0,1]$, $x, y \in X$ have $G(\lambda\, x + (1-\lambda)\, y) \leq_P \lambda\, G(x) + (1-\lambda)\, G(y)$.

Proposition 2.198. Let $G : X \to Y$ be convex, $z \in Y$. Then $S = \{x \in X : G(x) \leq_P z\}$ is convex.

Proof.

- Let $x, y \in S$, $\lambda \in [0,1]$.
- $x \in S \Rightarrow G(x) \leq_P z$, $y \in S \Rightarrow G(y) \leq_P z$. Convexity of P: $\lambda\,(z - G(x)) + (1-\lambda)\,(z - G(y)) = z + (\lambda\, G(x) + (1-\lambda)\, G(y)) \in P$.
- $\Rightarrow$ (convexity of G) $G(\lambda\, x + (1-\lambda)\, y) \leq_P \lambda\, G(x) + (1-\lambda)\, G(y) \leq_P z$.

□

We can use positive cones to formulate general inequality constraints.

Definition 2.199. Let $f : X \to \mathbb{R} \cup \{\infty\}$ be convex, $G : X \to Y$ convex w.r.t. positive cone $P \subset Y$. We consider the following optimization problem:

$$\inf \{f(x)|x \in X : G(x) \leq_P 0\}$$

We will also consider the *perturbed* problem with modified inequality constraint where we demand $G(x) \leq_P z$ instead for some $z \in Y$. This has two reasons: 1) Analytical trick for understanding structure of minimization problem. 2) In some applications precise constraint may be unknown or depend on noisy measurement. Want to understand how sensitive optimal value and optimizer are w.r.t. small perturbations in constraint.

Sketch: Illustrate simple linear program: f linear on $\mathbb{R}^2$, $G(x) \leq_P 0$ describes polytope, where whole face is optimal. Small perturbation of polytope: only one vertex will remain optimal. $\Rightarrow$ possibly huge sensitivity of optimizer w.r.t. constraint.

Proposition 2.200 (Effective primal)**.** For the above minimization problem the effective primal functional is given by $w : Y \to \mathbb{R} \cup \{\infty\}$,

$$z \mapsto \inf \{f(x)|x \in X : G(x) \leq_P z\} .$$

Clearly, $w(0)$ yields the optimal value, sought-after above. Further w is convex and decreasing, i.e. $z_1 \leq_P z_2 \Rightarrow w(z_1) \geq w(z_2)$.

Proof.

- Convexity:

$$\begin{aligned}\lambda\, w(z_1) + (1-\lambda)\, w(z_2) = \inf\{&\lambda\, f(x_1) + (1-\lambda)\, f(x_2)|x_1, x_2 \in X : G(x_1) \leq_P z_1,\\ &G(x_2) \leq_P z_2\}\end{aligned}$$

  (using convexity of f, P and G:)

$$\begin{aligned}&\geq \inf\{f(\lambda\, x_1 + (1-\lambda)\, x_2)|x_1, x_2 \in X :\\ &\quad G(\lambda\, x_1 + (1-\lambda)\, x_2) \leq_P \lambda\, z_1 + (1-\lambda)\, z_2\}\\ &= \inf\{f(x)|x \in X : G(x) \leq_P \lambda\, z_1 + (1-\lambda)\, z_2\}\\ &= w(\lambda\, z_1 + (1-\lambda)\, z_2)\end{aligned}$$

- Decreasing: If $z_1 \leq_P z_2$ then feasible set for $w(z_1)$ is contained in feasible set for $w(z_2)$. So $w(z_2) \leq w(z_1)$.

$\square$

Proposition 2.201 (Karush–Kuhn–Tucker (KKT) conditions)**.** Consider setup of Def. 2.199. Denote the optimal value by μ. If $\exists\, x_0 \in X$ with $f(x_0) < \infty$ and $G(x_0) <_P 0$ (implies $\operatorname{int} P \neq \emptyset$) then $\exists\, \hat{t} \in Y^*$ with $\hat{t} \in P^*$ such that

$$\mu = \inf\{f(x) + \langle \hat{t}, G(x)\rangle : x \in X\} < \infty.$$

If the problem of Def. 2.199 has a solution $\hat{x} \in X$, then $\hat{x}$ minimizes also $\inf\{f(x)+\langle \hat{t}, G(x)\rangle : x \in X\}$ and $\langle \hat{t}, G(\hat{x})\rangle = 0$. $\hat{t}$ is also referred to as Lagrange multiplier.

Proof.

- Express original problem via effective primal functional, Prop. 2.200:

$$\begin{aligned}\mu &= \inf\{f(x)|x \in X : G(x) \leq_P 0\}\\ &= \inf\{f(x)|x \in X, z \in Y : G(x) \leq_P z \leq_P 0\}\\ &= \inf\{w(z) + \iota_{(-P)}(z)|z \in Y\}\end{aligned}$$

- P is convex and w is convex (Prop. 2.200). By assumption have $z_0 \stackrel{\text{def.}}{=} G(x_0) \in -\operatorname{int} P$. So $\iota_{(-P)}$ is finite and continuous at z_0 and $w(z_0) \leq f(x_0) < \infty$. So we can apply the duality result Prop. 2.91. Get:

$$\mu = \max\{-w^*(-t) - \iota^*_{(-P)}(t) | t \in Y^*\}$$

- $\iota^*_{(-P)}(t) = \sup\{\langle t, z\rangle \,| z \in -P\} = \iota_{P^*}(t)$. So $\mu = \max\{-w^*(-t) | t \in P^*\}$. In particular this problem has a solution $\hat{t}$.

- By Fenchel–Young (Prop. 1.72, but for Banach space, cf. Def. 2.86) get for all $z \in Y$: $w(z) + w^*(-\hat{t}) \geq \big\langle -\hat{t}, z\big\rangle \Rightarrow w(z) + \big\langle \hat{t}, z\big\rangle \geq \mu$ and equality for $z = 0$. So $\inf\{w(z) + \big\langle \hat{t}, z\big\rangle \,| z \in Y\} = \mu$. So

$$\mu = \inf\{f(x) + \big\langle \hat{t}, z\big\rangle \,| x \in X, z \in Y : G(x) \leq_P z\}$$

For all feasible candidates have $z - G(x) \in P$. Since $\hat{t} \in P^*$ have $\big\langle \hat{t}, z - G(x)\big\rangle \geq 0$ and therefore, for fixed x, optimal z is given by $G(x)$. So

$$\mu = \inf\{f(x) + \big\langle \hat{t}, G(x)\big\rangle \,| x \in X\}.$$

- Finally, assume primal problem has minimizer $\hat{x} \in X$ with $\mu = f(\hat{x})$ and $G(\hat{x}) \in -P$. So $\big\langle \hat{t}, G(\hat{x})\big\rangle \leq 0$. So $\mu = \inf\{f(x) + \big\langle \hat{t}, G(x)\big\rangle \,| x \in X\} \leq f(\hat{x}) + \big\langle \hat{t}, G(\hat{x})\big\rangle \leq \mu$, which implies that $\hat{x}$ solves the auxiliary problem and $\big\langle \hat{t}, G(\hat{x})\big\rangle = 0$.

□

Motivation behind this proposition is easier to understand in simple, finite-dimensional example.

Example 2.202. • Let $X = \mathbb{R}^m$, $Y = \mathbb{R}^n$. Let f, G be differentiable and $P = \mathbb{R}^n_+$. Then $G(x) = (G_i(x))_{i=1}^n \leq_P 0$ implies $G_i(x) \leq 0$ for $i \in \{1, \ldots, n\}$.

Sketch: Try to minimize f on sublevel sets of $(G_i(\cdot))_{i=1}^n$.

- Assume primal problem has solution $\hat{x}$. Recall that $Y^* = \mathbb{R}^n$. Then there is some $\hat{t} \in P^* = \mathbb{R}^n_+ \subset \mathbb{R}^n$ such that $\hat{x}$ is also the solution to the unconstrained problem $\inf\{f(x) + \big\langle \hat{t}, G(x)\big\rangle \,| x \in X\}$ and $\big\langle \hat{t}, G(\hat{x})\big\rangle = 0$.

- The first condition implies that $\nabla f(\hat{x}) + \sum_{i=1}^n \hat{t}_i \cdot \nabla G_i(\hat{x}) = 0$. By feasibility of $\hat{x}$ have $G_i(\hat{x}) \leq 0$. The second condition implies $G_i(\hat{x}) < 0 \Rightarrow \hat{t}_i = 0$.

- Recall that ∇G_i is orthogonal to level line of G_i. So the gradient of f at $\hat{x}$ is in span of normals of level lines of 'active' constraints of G. Since $\hat{t}$ has only positive entries, ∇f points 'inwards'. So $\hat{t}$ can be thought of 're-weighting' primal objective, such that new optimal point of unconstrained problem coincides with constrained optimal point.

Corollary 2.203 (Necessary optimality conditions)**.** The function $L : X \times Y^* \to \mathbb{R} \cup \{\infty\}$, $(x, t) \mapsto f(x) + \langle t, G(x)\rangle$ is called *Lagrangian* associated with the problem of Def. 2.199. If $\hat{x}$ is optimal for the primal problem $\exists\, \hat{t} \in Y^*$: $\hat{t} \in P^*$, $\big\langle \hat{t}, G(\hat{x})\big\rangle = 0$ such that L has saddle point at $(\hat{x}, \hat{t})$, i.e.

$$L(\hat{x}, t) \leq L(\hat{x}, \hat{t}) \leq L(x, \hat{t})$$

for all $x \in X$, $t \in P^*$.

Proof. • Existence of $\hat{t} \in P^*$ with $\langle \hat{t}, G(\hat{x}) \rangle = 0$ is given by Prop. 2.201. Since $G(\hat{x}) \in -P$ have $\forall\, t \in P^*$: $\langle t, G(\hat{x}) \rangle \leq 0 = \langle \hat{t}, G(\hat{x}) \rangle$. So $L(\hat{x}, t) \leq L(\hat{x}, \hat{t})$.

- The second inequality follows since $\hat{x}$ minimizes the auxiliary problem $\inf\{f(x) + \langle \hat{t}, G(x) \rangle : x \in X\}$ (Prop. 2.201).

□

Proposition 2.204 (Sufficient optimality conditions)**.** Let $f : X \to \mathbb{R} \cup \{\infty\}$, $G : X \to Y$, no convexity required this time. $P \subset Y$ positive cone.

(i) $[\exists\, \hat{t} \in P^* \subset Y^* \text{ s.t. } \hat{x} \in \operatorname{argmin}_{x \in X} L(x, \hat{t})] \Rightarrow [\hat{x} \in \operatorname{argmin}\{f(x) | x \in X : G(x) \leq_P G(\hat{x})\}]$.

(ii) $[\exists\, \hat{t} \in P^* \subset Y^* \text{ s.t. } (\hat{x}, \hat{t}) \text{ is saddle point of } L] \Rightarrow [\hat{x} \in \operatorname{argmin}\{f(x) | x \in X : G(x) \leq_P 0\}]$.

Proof. • **(i):** Assume $\exists\, x' \in X$ with $f(x') < f(\hat{x})$ and $G(x') \leq_P G(\hat{x})$. Then, since $\hat{t} \in P^*$ $\Rightarrow \langle \hat{t}, G(\hat{x}) - G(x') \in P \rangle \geq 0$. So $L(x', \hat{t}) = f(x') + \langle \hat{t}, G(x') \rangle < f(\hat{x}) + \langle \hat{t}, G(\hat{x}) \rangle = L(\hat{x}, \hat{t})$, which is a contradiction.

Sketch: $X = Y = \mathbb{R}$, f convex, decreasing $G = \mathrm{id}$, so constraint is $x \leq 0$. Tangent with slope $-\hat{t}$. Shifts effective minimum of f to tangent point.

- **(ii):** From saddle point get $\langle t, G(\hat{x}) \rangle \leq \langle \hat{t}, G(\hat{x}) \rangle$ for all $t \in P^*$. For $t' \in P^*$ have $t' + \hat{t} \in P^*$ since P^* is convex cone. Set $t = t' + \hat{t}$ to get $\langle t', G(\hat{x}) \rangle \leq 0$ for all $t' \in P^*$.

- Set $t = 0$ to get $0 \leq \langle \hat{t}, G(\hat{x}) \rangle \leq 0$.

- For $x \in X$ with $G(x) \leq_P 0$ have: $\langle \hat{t}, G(x) \rangle \leq 0$. Again, with saddle point property:

$$f(\hat{x}) = f(\hat{x}) + \langle \hat{t}, G(\hat{x}) \rangle \leq f(x) + \langle \hat{t}, G(x) \rangle \leq f(x)$$

Sketch: Second condition is stronger: also requires $\hat{t}$ to be optimal. This helps to get the right slope $\hat{t}$ such that the corresponding optimal $\hat{x}$ satisfies precisely $G(\hat{x}) \leq_P 0$.

□

As indicated: effective primal functional also helps to study the sensitivity of the optimal value w.r.t. small perturbations in the constraints. This can be encoded as first order approximation of w at 0.

Proposition 2.205 (Sensitivity w.r.t. constraints)**.** Consider the setting of Def. 2.199 and Prop. 2.201. Slightly modify problems: consider $\min\{f(x) | x \in X : G(x) \leq_P z_i\}$ for $z_1, z_2 \in Y$. Let x_i, t_i be the corresponding primal solutions and optimal Lagrange multipliers (dual solutions), $i = 1, 2$. Then

$$\langle t_2, z_2 - z_1 \rangle \leq f(x_1) - f(x_2) \leq \langle t_1, z_2 - z_1 \rangle .$$

This implies $t_i \in \partial w(z_i)$.

Proof. • Rewrite problems to bring into exact form of Def. 2.199: Set $G_i(x) = G(x) - z_i$. Then $G(x) \leq_P z_i \Leftrightarrow G_i(x) \leq_P 0$.