Optimization and Optimal Control in Banach Spaces

Bernhard Schmitzer

November 9, 2018

1 Convex non-smooth optimization with proximal operators

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

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

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

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

1.1 Convex sets

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

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

Sketch: Positive example with ellipsoid, counterexample with 'kidney'

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

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

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

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.,

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

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

Proof.

- **Part I,** $\operatorname{conv} C \subset T$**:** 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 $\operatorname{conv} C \subset T$.
- **Part II,** $\operatorname{conv} C \supset T$**:** 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 induction. By definition, any element in T can be represented as follows: 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 $s_k = x_1 \in C \subset S$.
- Assume, we have shown that all linear combinations up to $k-1$ elements in T are also contained in S.
- For $k > 1$ set $\tilde{\lambda}_i = \lambda_i / (1 - \lambda_k)$ for $i = 1, \ldots, k-1$. Then

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

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

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

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

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

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

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 $\{x \in H | f(x) < +\infty\}$. 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', $\operatorname{epi} f = \{(x,r) \in H \times \mathbb{R} | r \geq f(x)\}$.
- The *sublevel set* of f with respect to $r \in \mathbb{R}$ is $S_r(f) = \{x \in H | f(x) \leq r\}$.

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

---

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

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

(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*

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

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

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

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

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

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

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

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

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

Definition 1.12. A function $f : H \to \mathbb{R} \cup \{\infty\}$ 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 $\alpha := \|x - x_0\|/\rho \in (0, 1]$

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

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

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:

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

  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

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

Proposition 1.15 (On convexity and continuity II)**.** If any of the conditions of Proposition 1.14 hold for some $x_0 \in \operatorname{dom} f$, 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. $\square$

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

Proposition 1.17 (On convexity and continuity in finite dimensions)**.** If $f \in \operatorname{Conv}(H = \mathbb{R}^n)$ 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. $\square$

Comment: Why is interior necessary in Proposition above?

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

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

Definition 1.19 (Lower semi-continuity)**.** A function $f : H \to \mathbb{R} \cup \{\infty\}$ 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) $\Rightarrow$ (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) $\Rightarrow$ (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.

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

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

1.3 Subdifferential

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

Comment: Meaning of notation.

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

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

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

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

Definition 1.24. The *domain* $\operatorname{dom} A$ 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

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

Proof. Let $x \in H$. Then

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

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)**:

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

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

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

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:

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

□

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

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

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

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

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

Example 1.31.

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

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

Sketch: Draw 'graph' of subdifferential.

- $H = \mathbb{R}^n$, $f(x) = \|x\|_1$. The ℓ_1 norm is not induced by an inner product. Therefore the above does not apply. We can use Proposition 1.30:

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

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 $\phi : (0, \infty) \to \mathbb{R} \cup \{\infty\}$, $\alpha \mapsto \frac{f(x+\alpha \cdot y) - f(x)}{\alpha}$.

- 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.) ← VL2

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

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

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

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

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

1.4 Cones and support functions

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

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

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

Proposition 1.40. The intersection of a family $\{C_i\}_{i \in I}$ of cones is cone. The *conical hull* of a set $C \subset H$, denoted by $\operatorname{cone} C$ 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

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

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

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

Proof.
- Since C is a linear subspace, if $\langle x, y \rangle \neq 0$ for some $y \in H$, $x \in C$ then $\sup \langle C, y \rangle = \infty$.
- Therefore, $C^{\ominus} = \{y \in H \mid \langle x, y \rangle = 0 \text{ for all } x \in C\}$. □

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

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

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

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

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

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

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

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

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

Proof.

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

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

□

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

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

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

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

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

Proof.

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

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

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

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

- 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 $C = \{(x,y) \in \mathbb{R}^2 \mid |x|+|y| \leq 1\}$. We find:

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

Sketch: Draw normal cones attached to points in C.

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

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

Comment: Do computation in detail.

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

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

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

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

Proof.

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

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

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

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

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

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

Proof.

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

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

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

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

(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 $\overline{C} = \overline{\operatorname{int} C}$. ← VL3

Proof.

- **(i):** Assume $x \neq y$ (otherwise the result is trivial). Then for $z \in [x, y)$ there is some $\alpha \in (0, 1]$ such that $z = \alpha \cdot x + (1 - \alpha) \cdot y$.
- Since $x \in \operatorname{int} C$ 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 $N_C x = (C - x)^{\ominus} = \{u \in H : \sup \langle C - x, u\rangle \leq 0\}$. We can extend the supremum to $\operatorname{cone}(C - x)$ and we can then extend it to the closure $\overline{\operatorname{cone}(C - x)} = T_C x$ 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 $T_C x = H$. □

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:

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

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: $C \cap \operatorname{bdry} C \subset C \cap \operatorname{spts} \overline{C} = \operatorname{spts} C$. □

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 $r \neq 0$. □

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 $[x \notin \operatorname{dom} f] \Rightarrow [\partial f(x) = \emptyset]$. $\square$ ← VL4

Finally, we show that closed, convex sets can be expressed solely in terms of their supporting hyperplanes.
For notational convenience introduce 'support function'.

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

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

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 $D \subset \overline{\operatorname{conv} C}$. $\square$

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

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

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. □

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 $\overline{\operatorname{conv} f}$ is convex, lower continuous.

Proof. This follows directly from Prop. 1.73. □

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

Proof.

- The claim is trivial when $f = +\infty \Leftrightarrow \operatorname{epi} f = \emptyset$. So now assume f is proper: then both sets are non-empty.
- Set $F = \overline{\operatorname{conv} f}$ and $D = \overline{\operatorname{conv} \operatorname{epi} f}$.
- Since $F \leq f \Rightarrow \operatorname{epi} f \subset \operatorname{epi} F$. Since $\operatorname{epi} F$ is convex, have $\operatorname{conv} \operatorname{epi} f \subset \operatorname{epi} F$. Since $\operatorname{epi} F$ is also closed (why?), have $D = \overline{\operatorname{conv} \operatorname{epi} f} \subset \operatorname{epi} F$.
- Show converse inclusion. Let $(x, \zeta) \in \operatorname{epi} F \setminus D$. Since $D \neq emptyset$ is closed and convex, the projection onto D is well defined. Let $(p, \pi) = P_D(x, \zeta)$. Characterization of projection:

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

- For some $(y, \eta) \in D$, send $\eta \to \infty$ (which is still in D, why?). We deduce: $\zeta - \pi \leq 0$.
- Assume $\zeta = \pi$. Then $(*)$ implies $\langle x - p, y - p \rangle \leq 0$ for all $(y, \eta) \in D = \overline{\operatorname{conv} \operatorname{epi} f}$.
- Note: $[\exists\ \eta \in \mathbb{R}$ s.t. $(y, \eta) \in \operatorname{conv} \operatorname{epi} f] \Leftrightarrow [y \in \operatorname{conv} \operatorname{dom} f]$. So: $\langle x - p, y - p \rangle \leq 0$ for all $y \in \operatorname{conv} \operatorname{dom} f$ and therefore for all $y \in \overline{\operatorname{conv} \operatorname{dom} f}$.
- Also note: $\operatorname{dom} F \subset \overline{\operatorname{conv} \operatorname{dom} f} = E$: 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 $x \in \operatorname{dom} f \subset E$, have $g \leq f$. Since F is the convex lower semicontinuous envelope of f we must therefore have $g \leq F$ and therefore $\operatorname{dom} F \subset E$.
- So we can set $y = x$ in projection characterization and obtain: $\|x - p\|^2 \leq 0$. Therefore $x = p$ which contradicts $(x, \zeta) \notin D$.
- Now assume $\zeta < \pi$. Set $u = \frac{x-p}{\pi-\zeta}$ and let $y \in \operatorname{dom} f$, $\eta = f(y)$ (i.e. $(y, \eta) \in \operatorname{epi} f \subset D$). Then from characterization of projection get

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

  So f is lower bounded by affine function $g : y \mapsto \langle u, y - p \rangle + \pi$. Therefore, $g \leq F$.
- Since $(x, \zeta) \in \operatorname{epi} F$ get

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

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

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.

- If $f = \pm\infty$ then $f^* = \mp\infty \Rightarrow$ claim is true. Now assume f is proper.
- We find:

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

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

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

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

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 $\operatorname{epi} f$. Such a halfspace must exist: f has affine lower bound $\Rightarrow$ $\operatorname{epi} f$ contained in some (closed) halfspace. Halfspace is closed, convex $\Rightarrow$ $\overline{\operatorname{conv} \operatorname{epi} f}$ also contained in halfspace.
- 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, proper. Let $x, u \in H$. Then:

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

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

Proof.

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

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

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

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

  So $f(y) \geq \langle u, y - x\rangle + f(x)$ for all $y \in H$. $\Rightarrow u \in \partial f(x)$.
- For second equivalence, apply first equivalence to f^* and use that $f^{**} = f$. □

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

Definition 1.82. A function $f : H \to \mathbb{R} \cup \{\infty\}$ 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.

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

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

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

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

- 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, proper ($f(0) = 0$) have $f = f^{**} = \iota_C^*$. □

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

Corollary 1.84. If $f : H \to \mathbb{R} \cup \{\infty\}$ 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. □

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 $N \in \mathbb{N}$

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

So $\|u\|^2 \ge \sum_{k=1}^N \langle u, x_k\rangle^2$ 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). ← VL6

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

(i) C is weakly sequentially closed.

(ii) C is sequentially closed.

(iii) C is closed.

(iv) C is weakly closed.

Proof.

- **(i) $\Rightarrow$ (ii):** Let $(x_k)_k$ be a sequence in C that converges strongly to some $x \in H$. Prop. 1.92: $[x_k \to x] \Rightarrow [x_k \rightharpoonup x]$. 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.) $\square$

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. $\square$

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

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

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

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

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

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

Proposition 1.102. Let $f : H \to \mathbb{R} \cup \{\infty\}$ 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. $\square$

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

Definition 1.103 (Coerciveness)**.** A function $f : H \to [-\infty, \infty]$ 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 $\limsup_{k\to\infty} f(x_k) \leq r$.
- Assume $S_r(f)$ is bounded for every $r \in \mathbb{R}$. Let $(x_k)_k$ be an unbounded sequence with $\lim_{k\to\infty} \|x_k\| \to \infty$. 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, $\liminf_{k\to\infty} f(x_k) \geq s$. Since this holds for any $s \in \mathbb{R}$, have $\lim_{k\to\infty} f(x_k) = \infty$. $\square$

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

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

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

1.7 Proximal operators

Definition 1.106. Let $f : H \to \mathbb{R} \cup \{\infty\}$ convex, lsc and proper. Then the map $\mathrm{Prox}_f : H \to H$ 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, lsc, proper $\Rightarrow f^*$ is proper. Therefore f has a continuous affine lower bound, which we denote by $\tilde{f} : y \mapsto \langle u, y \rangle - r$.
- For fixed $x \in H$ let $g : y \mapsto \frac{1}{2}\|x - y\|^2$. By 'completing the square' we get

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

  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

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

Proof.

- The second equivalence is trivial. We prove the first.
- **$\Rightarrow$:** Assume $p = \mathrm{Prox}_f(x)$, 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 $\frac{\mathrm{d}}{\mathrm{d}\alpha} g(\alpha)|_{\alpha=0} \geq 0$. This implies:

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

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

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

□

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 $x^{(0)} \in H$:

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

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

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

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

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

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

The examples indicate that $\operatorname{Prox}_f(x)$ 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 $\lambda \mapsto f(\operatorname{Prox}_{\lambda f}(x))$ is decreasing. ← VL7

Proof.

- **(i):** $[x = \operatorname{Prox}_f(x)]$ (Characterization of Prox, Prop. 1.108) $\Leftrightarrow [x \in x + \partial f(x)] \Leftrightarrow [0 \in \partial f(x)] \Leftrightarrow$ (Fermat's rule, Prop. 1.26) $[x \in \operatorname{argmin} f]$.
- **(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:
$$0 \leq \left\langle \tfrac{u_2}{\lambda_2} - \tfrac{u_1}{\lambda_1}, p_2 - p_1 \right\rangle$$
$$0 \leq \left\langle u_2 - \tfrac{\lambda_2}{\lambda_1} u_1, \Delta p \right\rangle = \left\langle \Delta u - \tfrac{\lambda_2 - \lambda_1}{\lambda_1} u_1, \Delta p \right\rangle = -\|\Delta p\|^2 - \tfrac{\lambda_2 - \lambda_1}{\lambda_1} \langle u_1, \Delta p \rangle$$
We deduce $\langle u_1, \Delta p \rangle \leq 0$. Then
$$\begin{aligned} \|x - p_2\|^2 = \|x - p_1 - (p_2 - p_1)\|^2 &= \|x - p_1 - \Delta p\|^2 \\ &= \|x - p_1\|^2 - 2\langle u_1, \Delta p \rangle + \|\Delta p\|^2 \geq \|x - p_1\|^2 . \end{aligned}$$

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

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

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

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

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

$\square$

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

Proposition 1.116 (Proximal algorithm)**.** Let $f : H \to \mathbb{R} \cup \{\infty\}$ 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 $r \in \mathbb{R}$: $\lim_{k\to\infty} \langle x_k, x - y\rangle = r \in \mathbb{R}$.
- Further, by weak convergence of the two extracted subsequences we find
$$\lim_{k\to\infty} \langle x_{i_k}, x - y\rangle = \langle x, x - y\rangle\,, \qquad \lim_{k\to\infty} \langle x_{j_k}, x - y\rangle = \langle y, x - y\rangle\,.$$
- Both sequences are subsequences of the converging sequence $(\langle x_k, x - y\rangle)_k$. Therefore, their limits must therefore coincide and equal r:
$$\underbrace{\lim_{k\to\infty} \langle x_{i_k}, x - y\rangle}_{=\langle x, x-y\rangle} = \underbrace{\lim_{k\to\infty} \langle x_{j_k}, x - y\rangle}_{=\langle y, x-y\rangle} = \lim_{k\to\infty} \langle x_k, x - y\rangle = r$$
Then
$$\|x - y\|^2 = \langle x - y, x - y\rangle = r - r = 0$$
and therefore the two cluster points must coincide. □

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:

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

- 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:

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

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

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

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 $H = \mathbb{R}^2$. $f(x) = \frac{1}{2}\|x\|^2$, $C = \{x \in H \,|\, x_1 = 1\}$, $g(x) = \iota_C(x)$. 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 $x_1^{(\ell)}$-update depends on whether $x_1^{(\ell)}$ was too close to 0 or too close to 1.

Now we start proving Prop. 1.124. We need considerable auxiliary definitions and results.
First note, that if $\operatorname{argmin} f$ 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) $\operatorname{id} - \operatorname{Prox}_f$ is firmly nonexpansive.

(iii) $2 \operatorname{Prox}_f - \operatorname{id}$ 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:

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

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: $\|T(x_k)\| = \|T(x_k) - T(0) + T(0)\| \leq \|T(x_k) - T(0)\| + \|T(0)\| \leq C + \|T(0)\| := C_2$. 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. $T(x) = x$.

$\square$

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

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

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

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

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

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

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

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:

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

  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)**. □