SPLITTING METHODS FOR CONVEX OPTIMIZATION

Master ACSYON (second year)
First semester of the 2022/2023 school year

University of Limoges (France)

Teacher: Loïc Bourdin
loic.bourdin@unilim.fr

Contents

Splitting algorithms:

$$\min_{x\in\mathbb{R}^n} f(x) = ? \implies \boxed{\nabla f(x) = 0}, \text{ if } f \in C^1$$

$$\text{GDM} = \begin{cases} x_0 \in \mathbb{R}^n \\ \boxed{x_{k+1} = x_k - s_k \nabla f(x_k) \qquad s_k > 0} \end{cases}$$

Explicit discretizat°

What if f is nonsmooth?

$f \in \Gamma_0(\mathbb{R}^n)$ = set of extended real-valued function, convex, lower semi-continuous and proper. $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$

- f is proper iff $\underbrace{\{x \in \mathbb{R}^n : f(x) < +\infty\}}_{\text{effective domain}} \neq \emptyset$
- f is lsc iff $\text{epi}(f)$ is closed in $\mathbb{R}^n \times \mathbb{R}$.

$$\begin{cases} x_0 \in \mathbb{R}^n \\ \boxed{x_{k+1} \in x_k - s_k \partial f(x_{k+1})} \end{cases}$$

implicit discretizat°

$$\partial f(x) = \{p \in \mathbb{R}^n : \langle p, y-x \rangle \leq f(y) - f(x) \quad \forall y \in \mathbb{R}^n\}$$

$$x_{k+1} + s_k \partial f(x_{k+1}) \ni x_k$$

$x \in A \iff A \ni x$

$$\boxed{x \in F(y) \iff y \in F^{-1}(x)}$$

$$(Id + s_k \partial f)(x_{k+1}) \ni x_k$$

$$x_{k+1} \cancel{\in} \underbrace{(Id + s_k \partial f)^{-1}}_{\shortparallel}(x_k)$$

$$=$$

$$\boxed{x_{k+1} = \text{prox}_{s_k f}(x_k)}$$

Proximal point Algorithm (PPA)

$$x_{k+1} \in x_k - s_k \partial f(x_{k+1}) \iff x_k \in x_{k+1} + s_k \partial f(x_{k+1})$$

$$\iff x_k \in (\mathrm{Id} + s_k \partial f)(x_{k+1})$$

$$\iff x_{k+1} \boxed{\in} \underbrace{\boxed{(\mathrm{Id} + s_k \partial f)^{-1}}}_{\text{single-valued}}(x_k)$$

$$\iff \boxed{x_{k+1} = \mathrm{prox}_{s_k f}(x_k)}$$

- <u>single-valued operator</u>:

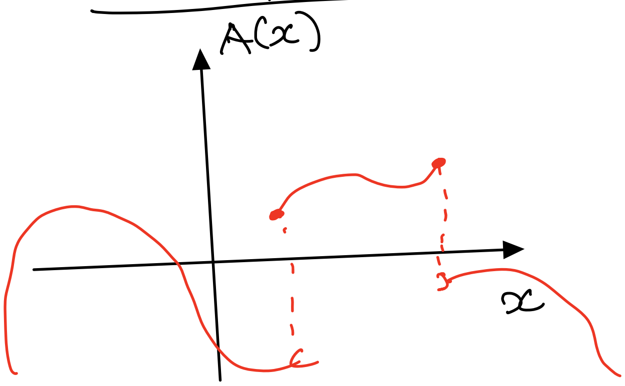
$$A : X \longrightarrow Y$$
$$x \longmapsto A(x) \in Y$$

- <u>set-valued operator</u>:

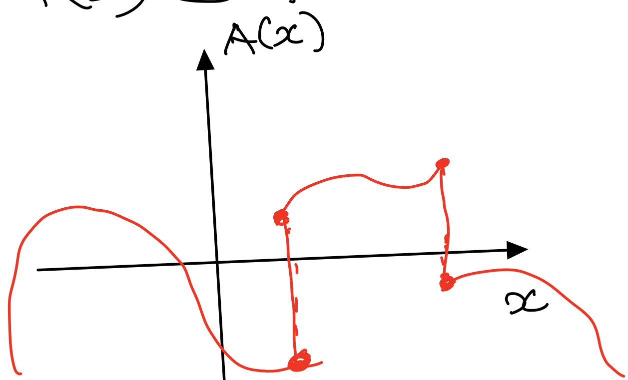
$$T : X \longrightarrow \mathcal{P}(Y)$$
$$x \longmapsto T(x) \subset Y$$

$$\boxed{T : X \rightrightarrows Y}$$

<u>multi-function</u>

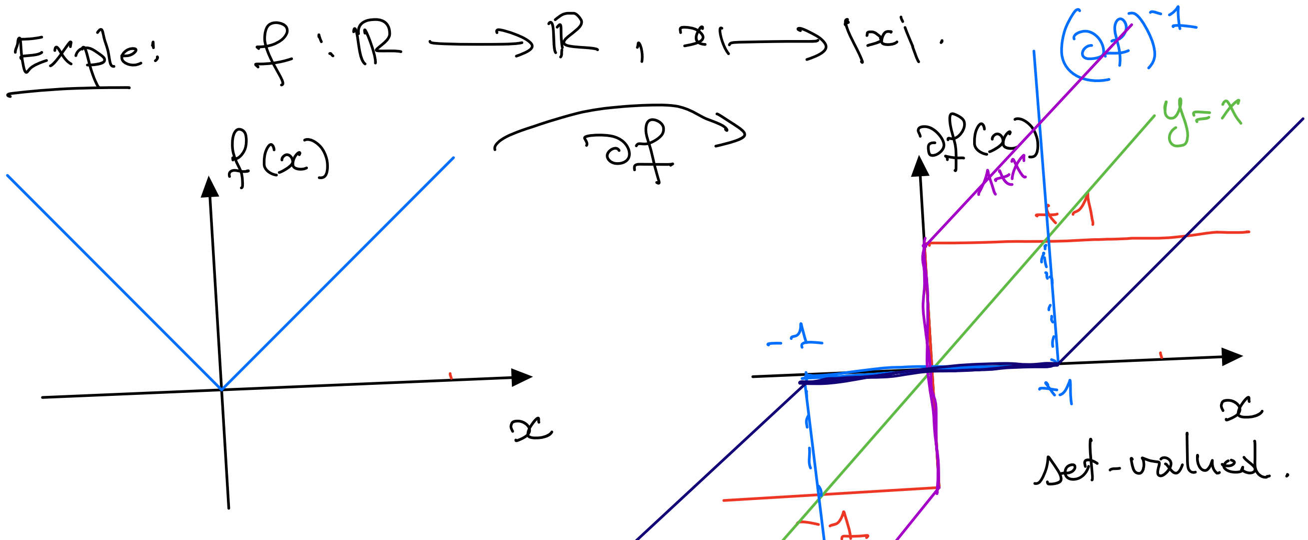


single-valued



set-valued.

Exple: $f : \mathbb{R} \longrightarrow \mathbb{R}$, $x \longmapsto |x|$.



$\forall f \in \Gamma_0(\mathbb{R}^n)$, $\forall s > 0$, we have:

$$\operatorname{prox}_{sf} = \left(Id + s\,\partial f\right)^{-1}$$

is single-valued.

$f = |\cdot|$ $\quad \operatorname{prox}_{sf}$ = threshold operator!
also

<u>LASSO</u>: $\displaystyle f(x) = \underbrace{\frac{1}{2}\|Ax-b\|_2^2}_{\text{smooth}} + \underbrace{\lambda\|x\|_1}_{\text{nonsmooth.}}$

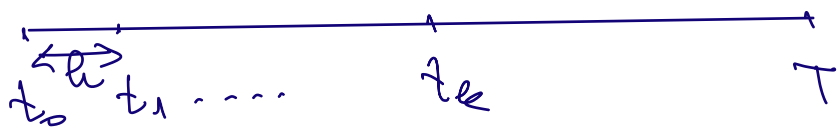
$f = g + h$ $\quad$ g is smooth, $h \in \Gamma_0(\mathbb{R}^n)$

$$\begin{cases} x_0 \\ x_{k+1} \in x_k - s_k \nabla g(x_k) - s_k\, \partial h(x_{k+1}) \end{cases}$$

$$x_{k+1} = \operatorname{prox}_{\boxed{s_k h}}\left(x_k - s_k \nabla g(x_k)\right)$$

Proximal gradient method

ODE: $\begin{cases} \dot{x}(t) = f(t, x(t)) & t \in [t_0, T]. \\ x(t_0) = x_0 \end{cases}$



$x(t_\ell) = x_\ell$. $\lim_{h \to 0} \dfrac{x(t+h) - x(t)}{h} = \dot{x}(t)$

$$\dot{x}(t_\ell) \simeq \frac{x(t_\ell + h) - x(t_\ell)}{h} = \frac{x_{\ell+1} - x_\ell}{h}$$

• Explicit Euler discretizat°:

$$\frac{x_{\ell+1} - x_\ell}{h} = f(t_\ell, x_\ell)$$

$$\begin{cases} x_0 \\ \boxed{x_{\ell+1} = x_\ell + h f(t_\ell, x_\ell)} \end{cases}$$

• Implicit Euler Method

$$\frac{x_{\ell+1} - x_\ell}{h} = f(t_\ell, x_{\ell+1})$$

$$\boxed{x_{\ell+1} = x_\ell + h f(t_\ell, x_{\ell+1})}$$

Chapter 1: Set-valued maps and fixed points of averaged operators

In what follows, $\mathcal{P}(\mathbb{R}^d)$ stands for the *set of all subsets* of $\mathbb{R}^d$. For example, one can write $\emptyset \in \mathcal{P}(\mathbb{R}^d)$ or $\mathbb{R}^d \in \mathcal{P}(\mathbb{R}^d)$. In the one-dimensional case $d = 1$, it holds that $\{-1, 4\} \in \mathcal{P}(\mathbb{R})$ or $[-3, 6] \in \mathcal{P}(\mathbb{R})$.

In order to solve a minimization problem of the form

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \, f(x),$$

where $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a given proper function, one way is to look for an element $x \in \text{dom}(f)$ such that $0_{\mathbb{R}^d} \in \partial f(x)$. For any $x \in \text{dom}(f)$, recall that $\partial f(x)$ is a subset of $\mathbb{R}^d$ that is possibly empty or possibly contains several different subgradients. Defining $\partial f(x) = \emptyset$ for all $x \in \mathbb{R}^d \backslash \text{dom}(f)$, one can see ∂f as a *set-valued map* of the form $\partial f : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^d)$. Indeed, for each $x \in \mathbb{R}^d$, we have $\partial f(x) \in \mathcal{P}(\mathbb{R}^d)$.

As a consequence, in this chapter, we will introduce and study the general notion of *set-valued map* $A : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^d)$ that we simply denote by $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. Note that a lot of mathematical problems (not only optimization problems) consist in solving an inclusion of the form $y \in A(x)$, where $A : \mathcal{X} \rightrightarrows \mathcal{Y}$ is a set-valued map between two sets $\mathcal{X}$ and $\mathcal{Y}$, where $y \in \mathcal{Y}$ is a given data and $x \in \mathcal{X}$ is the unknown to determine. In the case where $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, we can easily rewrite such an inclusion under the form $0_{\mathbb{R}^d} \in A(x)$. Since this inclusion is equivalent to $x \in x + A(x)$, one way to solve it is to look for a fixed point of $\text{Id} + A$ (in a sense to precise), or of its inverse $(\text{Id} + A)^{-1}$ (in a sense to precise). This is exactly the topic that we want to develop in this chapter.

$y \in A(x) \Longleftrightarrow (x,y) \in Gr(A$

1 Set-valued maps, monotonicity and resolvents

1.1 Set-valued maps (or multi-valued maps)

Definition 1.1 (Set-valued map)**.** A map $A : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^d)$ is called a *set-valued map* (or *multi-valued map)* and is simply denoted by $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. In that context, the *graph* of A is defined by

$$\mathrm{Gr}(A) := \{(x,y) \in \mathbb{R}^d \times \mathbb{R}^d \mid y \in A(x)\}.$$

Note that, for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$, we have $y \in A(x)$ if and only if $(x,y) \in \mathrm{Gr}(A)$.

Remark 1.1. In this lecture, we consider set-valued maps only of the form $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. Nevertheless, all notions developed here (and most of results) can be extended to set-valued maps of the form $A : \mathcal{X} \rightrightarrows \mathcal{Y}$ where $\mathcal{X}$, $\mathcal{Y}$ are general sets (possibly different, possibly finite, possibly finite- or infinite-dimensional linear spaces, etc.).

Remark 1.2. The graph $\mathrm{Gr}(A)$ of a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a subset of $\mathbb{R}^d \times \mathbb{R}^d$. Conversely, note that any subset $\mathcal{S}$ of $\mathbb{R}^d \times \mathbb{R}^d$ can be seen as the graph $\mathrm{Gr}(A)$ of the set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ defined by

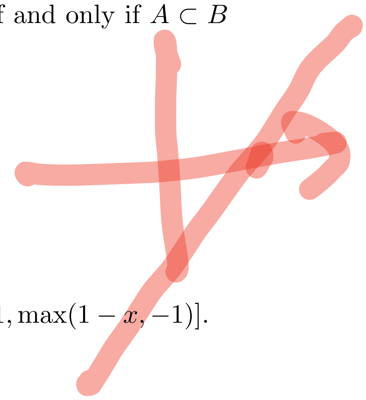
$$\forall x \in \mathbb{R}^d, \quad A(x) := \{y \in \mathbb{R}^d \mid (x,y) \in \mathcal{S}\}.$$

Therefore, a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is entirely determined by its graph, in the sense that

$$\forall x \in \mathbb{R}^d, \quad A(x) = \{y \in \mathbb{R}^d \mid (x,y) \in \mathrm{Gr}(A)\}.$$

Thus, two set-valued maps A, $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ are equal (denoted by $A = B$) if and only if $\mathrm{Gr}(A) = \mathrm{Gr}(B)$.

Remark 1.3. Let A, $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be two set-valued maps. If $A(x) \subset B(x)$ for all $x \in \mathbb{R}^d$, we write $A \subset B$. Note that $A \subset B$ if and only if $\mathrm{Gr}(A) \subset \mathrm{Gr}(B)$. With this notation, we easily obtain that $A = B$ if and only if $A \subset B$ and $B \subset A$.



Example 1.1. Consider the four set-valued maps on $\mathbb{R}$ ($d = 1$) defined by

$$\begin{array}{rccl} A_1 : & \mathbb{R} & \rightrightarrows & \mathbb{R} \\ & x & \longmapsto & A_1(x) := \emptyset \end{array} \quad \text{and} \quad \begin{array}{rccl} A_2 : & \mathbb{R} & \rightrightarrows & \mathbb{R} \\ & x & \longmapsto & A_2(x) := \{1-x\} \end{array}$$

$$\begin{array}{rccl} A_3 : & \mathbb{R} & \rightrightarrows & \mathbb{R} \\ & x & \longmapsto & A_3(x) := \begin{cases} \{1-x\} & \text{if} \quad x < 0 \\ [-1,1] & \text{if} \quad x = 0 \\ \{-1\} & \text{if} \quad x > 0 \end{cases} \end{array} \quad \text{and} \quad \begin{array}{rccl} A_4 : & \mathbb{R} & \rightrightarrows & \mathbb{R} \\ & x & \longmapsto & A_4(x) := [-1, \max(1-x,-1)]. \end{array}$$

Then the three inclusions $A_1 \subset A_2$ and $A_1 \subset A_3 \subset A_4$ hold true.

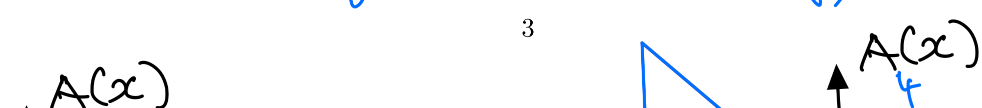
Definition 1.2 (Domain and range)**.** The *domain* and the *range* of a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ are respectively defined by

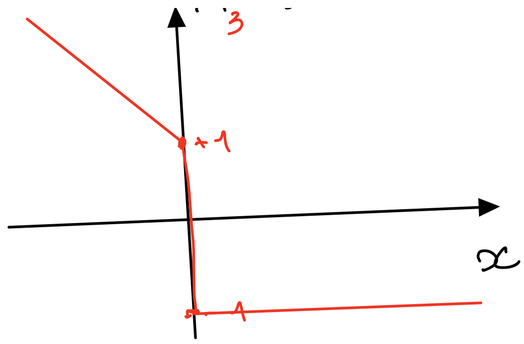
$$\mathrm{Dom}(A) := \{x \in \mathbb{R}^d \mid A(x) \neq \emptyset\} \qquad \text{and} \qquad \mathrm{Ran}(A) := \{y \in \mathbb{R}^d \mid \exists x \in \mathbb{R}^d,\ y \in A(x)\}.$$

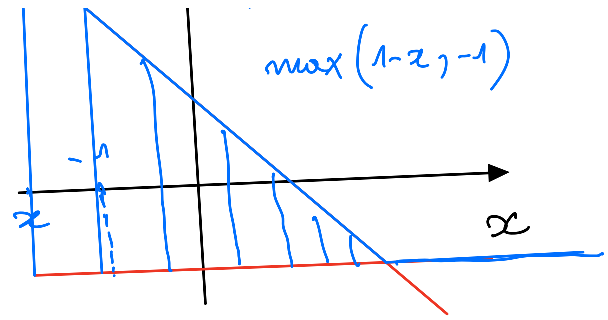
Definition 1.3 (Inverse map)**.** The *inverse map* $A^{-1} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ of a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is defined by

$$\forall y \in \mathbb{R}^d, \quad A^{-1}(y) := \{x \in \mathbb{R}^d \mid y \in A(x)\}.$$

$y \in A(x) \Longleftrightarrow x \in A^{-1}(y)$




3
+1
-1
x


max(1-x, -1)
-1
x
x

Remark 1.4. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. Note that

$$\forall x, y \in \mathbb{R}^d, \quad (x,y) \in \mathrm{Gr}(A) \iff y \in A(x) \iff x \in A^{-1}(y) \iff (y,x) \in \mathrm{Gr}(A^{-1}).$$

Furthermore, it holds that $\mathrm{Dom}(A^{-1}) = \mathrm{Ran}(A)$, $\mathrm{Ran}(A^{-1}) = \mathrm{Dom}(A)$ and $(A^{-1})^{-1} = A$.

Remark 1.5. Let A, $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be two set-valued maps such that $A \subset B$. Note that $A^{-1} \subset B^{-1}$ (and not $B^{-1} \subset A^{-1}$, be careful).

Definition 1.4 (Image and inverse image)**.** Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. The *image* of a subset E of $\mathbb{R}^d$ under A is defined by

$$A(E) := \{y \in \mathbb{R}^d \mid \exists x \in E,\ y \in A(x)\} = \bigcup_{x \in E} A(x).$$

The *inverse image* of a subset F of $\mathbb{R}^d$ under A is defined by $A^{-1}(F)$.

Remark 1.6. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. Note that $\mathrm{Dom}(A) = A^{-1}(\mathbb{R}^d)$ and $\mathrm{Ran}(A) = A(\mathbb{R}^d)$.

Definition 1.5 (Fixed point and zero)**.** The *set of all fixed points* and the *set of all zeros* of a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ are respectively defined by

$$\mathrm{Fix}(A) := \{x \in \mathbb{R}^d \mid x \in A(x)\} \quad \text{and} \quad \mathrm{Zer}(A) := \{x \in \mathbb{R}^d \mid 0_{\mathbb{R}^d} \in A(x)\}.$$

Remark 1.7. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. Note that $\mathrm{Fix}(A) = \mathrm{Fix}(A^{-1})$ and $\mathrm{Zer}(A) = A^{-1}(0_{\mathbb{R}^d})$.

Definition 1.6 (Operations on set-valued maps)**.** Let A, $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be two set-valued maps. We define the *union* $A \cup B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ by

$$\forall x \in \mathbb{R}^d, \quad (A \cup B)(x) := A(x) \cup B(x).$$

We define the *intersection* $A \cap B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ by

$$\forall x \in \mathbb{R}^d, \quad (A \cap B)(x) := A(x) \cap B(x).$$

We define the *composition* $A \circ B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ by

$$\forall x \in \mathbb{R}^d, \quad (A \circ B)(x) := A(B(x)).$$

Let Λ_1, Λ_2 be two subsets of $\mathbb{R}$. We define the *linear combination* $\Lambda_1 A + \Lambda_2 B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ by

$$\forall x \in \mathbb{R}^d, \quad (\Lambda_1 A + \Lambda_2 B)(x) := \Lambda_1 A(x) + \Lambda_2 B(x).$$

When $\Lambda_1 = \{\lambda_1\}$ and $\Lambda_2 = \{\lambda_2\}$ are both reduced to singletons, we simply write $\lambda_1 A + \lambda_2 B$.

Remark 1.8. Note that set-valued maps can be seen as a generalization of usual maps, as underlined in the next definition.

Definition 1.7 (Single-valued map)**.** Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. If $A(x)$ is a singleton for all $x \in \mathrm{Dom}(A)$, we say that A is a *single-valued map* and we denote by $A : \mathrm{Dom}(A) \to \mathbb{R}^d$.

Example 1.2. The *identity map* $\mathrm{Id} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ defined by

$$\forall x \in \mathbb{R}^d, \quad \mathrm{Id}(x) := \{x\},$$

is a single-valued map with $\mathrm{Dom}(\mathrm{Id}) = \mathbb{R}^d$. We simply denote by $\mathrm{Id} : \mathbb{R}^d \to \mathbb{R}^d$ and by $\mathrm{Id}(x) = x$ for all $x \in \mathbb{R}^d$.

1.2 Monotone and maximal monotone set-valued maps

Definition 1.8 (Monotone map)**.** A set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is said to be *monotone* if

$$\forall (x_1, y_1), (x_2, y_2) \in \mathrm{Gr}(A), \quad \langle y_2 - y_1, x_2 - x_1 \rangle \geqslant 0.$$

Example 1.3. The identity map $\mathrm{Id} : \mathbb{R}^d \to \mathbb{R}^d$ is monotone.

Remark 1.9. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. Two remarks:

(i) A is monotone if and only if A^{-1} is monotone.

(ii) If $d = 1$ and $A : \mathrm{Dom}(A) \to \mathbb{R}$ is singled-valued, then A is monotone if and only if A is nondecreasing.

Remark 1.10. If A, $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ are two monotone set-valued maps, then so is $\lambda_1 A + \lambda_2 B$ for all λ_1, $\lambda_2 \geqslant 0$.

Example 1.4. The two set-valued maps on $\mathbb{R}$ ($d = 1$) defined by

$$\begin{array}{rccl} A : & \mathbb{R} & \rightrightarrows & \mathbb{R} \\ & x & \longmapsto & A(x) := \begin{cases} \{x-1\} & \text{if} \quad x < 0 \\ \{-1, 1\} & \text{if} \quad x = 0 \\ \{x+1\} & \text{if} \quad x > 0 \end{cases} \end{array} \quad \text{and} \quad \begin{array}{rccl} B : & \mathbb{R} & \rightrightarrows & \mathbb{R} \\ & x & \longmapsto & B(x) := \begin{cases} \{x-1\} & \text{if} \quad x < 0 \\ [-1, 1] & \text{if} \quad x = 0 \\ \{x+1\} & \text{if} \quad x > 0, \end{cases} \end{array}$$

are both monotone with $A \subset B$.

Definition 1.9 (Maximal monotone map)**.** A monotone set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is said to be *maximal monotone* if the inclusion $A \subset B$, with another monotone set-valued map $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, implies that $B = A$.

In other words, a monotone set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is said to be *maximal monotone* if it is not strictly included into another monotone set-valued map

Remark 1.11. Thanks to Zorn's lemma, one can prove that any monotone set-valued map is included into a maximal monotone one.

Remark 1.12. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. Note that A is maximal monotone if and only if A^{-1} is maximal monotone.

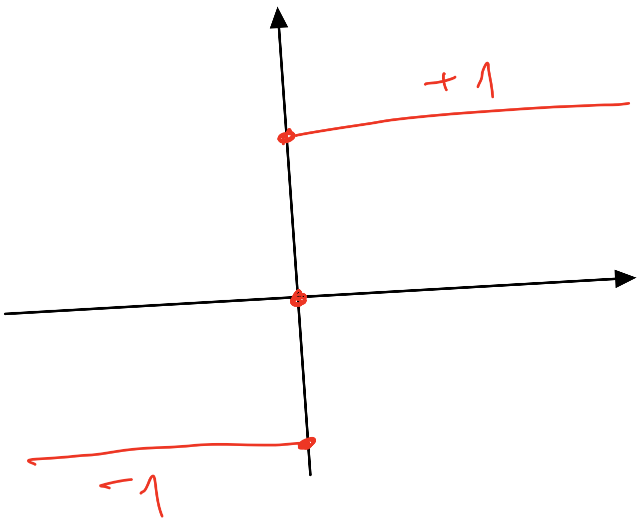
Remark 1.13. If a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximal monotone, then λA is maximal monotone for all $\lambda > 0$. What happens for $\lambda = 0$? (exercise)

Remark 1.14. Let A, $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be two maximal monotone set-valued maps. We know that $A + B$ is monotone. Is $A + B$ maximal monotone? (exercise)

Lemma 1.1 (Characterization)**.** A set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximal monotone if and only if

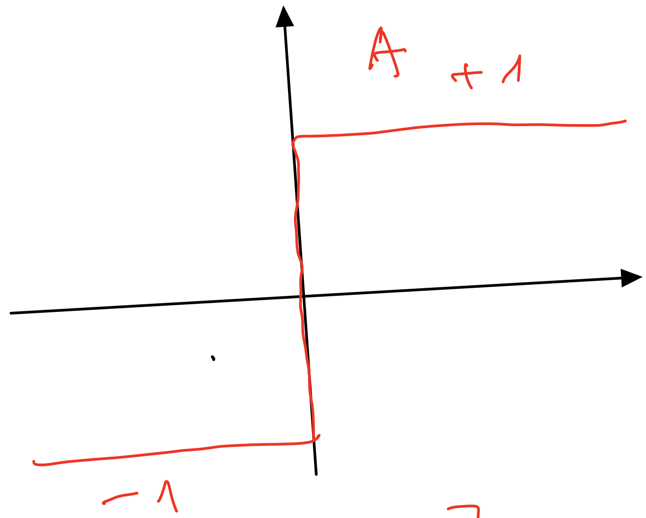
$$\forall (x_1, y_1) \in \mathbb{R}^d \times \mathbb{R}^d, \qquad \Big[\forall (x_2, y_2) \in \mathrm{Gr}(A), \quad \langle y_2 - y_1, x_2 - x_1 \rangle \geqslant 0 \Big] \quad \Longleftrightarrow \quad y_1 \in A(x_1).$$

Proof. Let us prove the necessary condition. Assume that A is maximal monotone and let $(x_1, y_1) \in \mathbb{R}^d \times \mathbb{R}^d$. If $y_1 \in A(x_1)$, then we have $(x_1, y_1) \in \mathrm{Gr}(A)$ and thus $\langle y_2 - y_1, x_2 - x_1 \rangle \geqslant 0$ for all $(x_2, y_2) \in \mathrm{Gr}(A)$ by monotonicity of A. Conversely, assume that $\langle y_2 - y_1, x_2 - x_1 \rangle \geqslant 0$ for all $(x_2, y_2) \in \mathrm{Gr}(A)$ and assume, by contradiction, that $y_1 \notin A(x_1)$. Then we introduce $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ by $\mathrm{Gr}(B) := \mathrm{Gr}(A) \cup \{(x_1, y_1)\}$. We conclude that B is monotone and contains strictly A, which is absurd since A is maximal monotone.



$A(0) = \{-1, 0, 1\}$

monotone



$A(0) = [-1, 1]$

monotone

$\exists B$ monotone s.t. $Gr(A) \subset Gr(B)$

Let us prove the sufficient condition. For this purpose, assume that the property is satisfied. The monotonicity of A is then trivial. By contradiction, assume that A is not maximal monotone, then it is strictly included in a monotone operator $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. Take $(x_1, y_1) \in \mathrm{Gr}(B)\backslash\mathrm{Gr}(A)$. From monotonicity of B, we obtain $\langle y_2 - y_1, x_2 - x_1 \rangle \geqslant 0$ for all $(x_2, y_2) \in \mathrm{Gr}(A) \subset \mathrm{Gr}(B)$. From the property, we obtain that $y_1 \in A(x_1)$ and thus $(x_1, y_1) \in \mathrm{Gr}(A)$, which is absurd. The proof is complete. □

Remark 1.15. In Lemma 1.1, note that the implication $\Leftarrow$ characterizes the monotonicity, while the reverse implication $\Rightarrow$ characterizes its maximality.

Proposition 1.1. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a maximal monotone set-valued map. Then $A(x)$ is a closed convex subset of $\mathbb{R}^d$ for all $x \in \mathbb{R}^d$. In that context, for all $x \in \mathrm{Dom}(A)$, we denote by

$$\mathfrak{m}(A(x)) := \mathrm{proj}_{A(x)}(0_{\mathbb{R}^d}),$$

the element of minimal norm of $A(x)$. Furthermore, $\mathrm{Gr}(A)$ is closed.

Proof. From Lemma 1.1, we have

$$A(x) := \bigcap_{(x',y') \in \mathrm{Gr}(A)} \{y \in \mathbb{R}^d \mid \langle y' - y, x' - x \rangle \geqslant 0\},$$

and thus $A(x)$ is a closed convex subset of $\mathbb{R}^d$ for all $x \in \mathbb{R}^d$. Thus, for all $x \in \mathrm{Dom}(A)$, $A(x)$ is a nonempty closed convex subset of $\mathbb{R}^d$ and thus $\mathfrak{m}(A(x))$ is well defined. Now consider a convergent sequence $(x_k, y_k)_{k\in\mathbb{N}} \subset \mathrm{Gr}(A)$ whose limit is denoted by (x, y). Let us prove that $(x, y) \in \mathrm{Gr}(A)$ thanks to Lemma 1.1. Let $(w, z) \in \mathrm{Gr}(A)$. Since $(x_k, y_k) \in \mathrm{Gr}(A)$, we have $\langle z - y_k, w - x_k \rangle \geqslant 0$ for all $k \in \mathbb{N}$, and thus $\langle z - y, w - x \rangle \geqslant 0$ by passing to the limit $k \to \infty$. The proof is complete. □

Remark 1.16. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a maximal monotone set-valued map. The closure of $\mathrm{Gr}(A)$ guarantees a kind of continuity property. Precisely, if two convergent sequences $(x_k)_{k\in\mathbb{N}}$ and $(y_k)_{k\in\mathbb{N}}$ of $\mathbb{R}^d$, with limits respectively denoted by x and y, are such that $y_k \in A(x_k)$ for all $k \in \mathbb{N}$, then $y \in A(x)$.

Proposition 1.2. If $A : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous monotone single-valued map, with $\mathrm{Dom}(A) = \mathbb{R}^d$, then A is maximal monotone.

Proof. Let us prove that A is maximal monotone by using Lemma 1.1 and, since A is assumed to be monotone, let us only prove the implication $\Rightarrow$ (see Remark 1.15). Let $(x_1, y_1) \in \mathbb{R}^d$ and let us assume that $\langle y_2 - y_1, x_2 - x_1 \rangle \geqslant 0$ for all $(x_2, y_2) \in \mathrm{Gr}(A)$, that is, $\langle A(x_2) - y_1, x_2 - x_1 \rangle \geqslant 0$ for all $x_2 \in \mathbb{R}^d$ (since $A : \mathbb{R}^d \to \mathbb{R}^d$ is assumed to be single-valued). The *Minty's trick* consists in considering $x_2 := x_1 + \lambda(w - x_1)$ for any $\lambda \in \mathbb{R}$ and any $w \in \mathbb{R}^d$. We deduce that $\lambda\langle A(x_1 + \lambda(w - x_1)) - y_1, w - x_1 \rangle \geqslant 0$ for all $w \in \mathbb{R}^d$. By dividing by $\lambda > 0$ and making tend $\lambda \to 0^+$, using the continuity of A, we obtain that $\langle A(x_1) - y_1, w - x_1 \rangle \geqslant 0$ for all $w \in \mathbb{R}^d$ and thus $y_1 = A(x_1)$. The proof is complete. □

Id + A is surjective !!

Theorem 1.1 (Minty's theorem)**.** A monotone set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximal monotone if and only if $\mathrm{Ran}(\mathrm{Id} + A) = \mathbb{R}^d$. $\forall y \in \mathbb{R}^d, \exists x \in \mathbb{R}^d$ s.t. $y \in x + A(x)$.

Proof. We only prove the sufficient condition (which is the condition used in most practical cases). Let us prove that A is maximal monotone by using Lemma 1.1 and, since A is assumed to be monotone, let us prove only the implication $\Rightarrow$ (see Remark 1.15). Let $(x_1, y_1) \in \mathbb{R}^d$ and let us assume that $\langle y_2 - y_1, x_2 - x_1 \rangle \geqslant 0$ for all $(x_2, y_2) \in \mathrm{Gr}(A)$. Let us prove that $y_1 \in A(x_1)$. Since $\mathrm{Ran}(\mathrm{Id} + A) = \mathbb{R}^d$, let us take $x_0 \in \mathbb{R}^d$ such that $x_1 + y_1 \in x_0 + A(x_0)$.

Now consider $y_0 \in A(x_0)$ such that $x_1 + y_1 = x_0 + y_0$. Taking $(x_2, y_2) = (x_0, y_0) \in \mathrm{Gr}(A)$, we obtain that $\|x_1 - x_0\|^2 = -\langle y_0 - y_1, x_0 - x_1 \rangle \leqslant 0$ and thus $x_0 = x_1$, then $y_0 = y_1$ and thus $y_1 = y_0 \in A(x_0) = A(x_1)$. The proof of the sufficient condition is complete. $\square$

1.3 Resolvents and Yosida approximations

Definition 1.10 (Resolvent)**.** The *resolvent* of a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is the set-valued map $J_A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ defined by

$$J_A := (\mathrm{Id} + A)^{-1},$$

with $\mathrm{Dom}(J_A) = \mathrm{Ran}(\mathrm{Id} + A)$ and $\mathrm{Ran}(J_A) = \mathrm{Dom}(\mathrm{Id} + A) = \mathrm{Dom}(A)$.

Remark 1.17. Note that any set-valued map $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ can be seen as the resolvent $B = J_A$ of another set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, just by considering $A := B^{-1} - \mathrm{Id}$.

Definition 1.11 (Yosida approximations)**.** The *Yosida approximations* of a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ are the set-valued maps $A_\lambda : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ defined by

$$A_\lambda := \frac{1}{\lambda}(\mathrm{Id} - J_{\lambda A}),$$

for all $\lambda > 0$. Note that $\mathrm{Dom}(A_\lambda) = \mathrm{Dom}(J_{\lambda A}) = \mathrm{Ran}(\mathrm{Id} + \lambda A)$ for all $\lambda > 0$.

Remark 1.18. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. Note that

$$\forall \lambda > 0, \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, \quad y \in A_\lambda(x) \Longleftrightarrow y \in A(x - \lambda y).$$

Lemma 1.2. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. It holds that $(A_\mu)_\lambda = A_{\mu+\lambda}$ for all λ, $\mu > 0$.

Proof. From Remark 1.18, we have

$$y \in (A_\mu)_\lambda(x) \Longleftrightarrow y \in A_\mu(x - \lambda y) \Longleftrightarrow y \in A(x - \lambda y - \mu y) \Longleftrightarrow y \in A(x - (\mu + \lambda)y) \Longleftrightarrow y \in A_{\mu+\lambda}(x),$$

for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ and all λ, $\mu > 0$. The proof is complete. $\square$

Lemma 1.3. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. Then

$$\forall \lambda > 0, \quad \forall x \in \mathrm{Dom}(A_\lambda) = \mathrm{Dom}(J_{\lambda A}), \quad A_\lambda(x) \subset A(J_{\lambda A}(x)).$$

Note that $\mathrm{Ran}(J_{\lambda A}) = \mathrm{Dom}(\mathrm{Id} + \lambda A) = \mathrm{Dom}(A)$ for all $\lambda > 0$.

Proof. Let $\lambda > 0$ and $x \in \mathrm{Dom}(A_\lambda)$. Let $y \in A_\lambda(x)$ and let us prove that $y \in A(J_{\lambda A}(x))$. Since $y \in A_\lambda(x) = \frac{1}{\lambda}(x - J_{\lambda A}(x))$, there exists $z \in J_{\lambda A}(x)$ such that $y = \frac{1}{\lambda}(x - z)$ and thus $x = z + \lambda y$. Since $z \in J_{\lambda A}(x) = (\mathrm{Id} + \lambda A)^{-1}(x)$, we have $x \in z + \lambda A(z)$, and thus there exists $w \in A(z)$ such that $x = z + \lambda w$. We deduce that $y = w \in A(z) \subset A(J_{\lambda A}(x))$. The proof is complete. $\square$

Proposition 1.3. If a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is monotone, then:

(i) $J_A : \mathrm{Dom}(J_A) \to \mathbb{R}^d$ is single-valued, 1-Lipschitz continuous and monotone;

(ii) $A_\lambda : \mathrm{Dom}(A_\lambda) \to \mathbb{R}^d$ is single-valued, $\frac{1}{\lambda}$-Lipschitz continuous and monotone, for all $\lambda > 0$.

Proof. Let us prove the first item. Let $(x_1, y_1), (x_2, y_2) \in \mathrm{Gr}(J_A)$. Then we have $(y_1, x_1 - y_1), (y_2, x_2 - y_2) \in \mathrm{Gr}(A)$. Since A is monotone, we have from the second item of Lemma A.2 that $\|y_2 - y_1\| \leqslant \|y_2 - y_1 + \delta(x_2 - y_2 - (x_1 - y_1))\|$ for all $\delta \geqslant 0$ and thus $\|y_2 - y_1\| \leqslant \|x_2 - x_1\|$ by taking $\delta = 1$. By taking $x_2 = x_1$, we obtain that $J_A : \mathrm{Dom}(J_A) \to \mathbb{R}^d$ is single-valued, and then the previous inequality shows that it is 1-Lipschitz continuous. The monotonicity of $J_A = (\mathrm{Id} + A)^{-1}$ is trivial since Id and A are monotone.

Let us prove the second item. Let $\lambda > 0$ and note that λA is monotone. From the previous item, we know that $J_{\lambda A} : \mathrm{Dom}(J_{\lambda A}) \to \mathbb{R}^d$ is single-valued and thus so is $A_\lambda : \mathrm{Dom}(A_\lambda) \to \mathbb{R}^d$, with $\mathrm{Dom}(A_\lambda) = \mathrm{Dom}(J_{\lambda A})$. From that we have

$$x = J_{\lambda A}(x) + \lambda A_\lambda(x),$$

for all $x \in \mathrm{Dom}(A_\lambda) = \mathrm{Dom}(J_{\lambda A})$.

1. We deduce that

$$\begin{aligned}\|x_2 - x_1\|^2 &= \|J_{\lambda A}(x_2) - J_{\lambda A}(x_1) + \lambda(A_\lambda(x_2) - A_\lambda(x_1))\|^2 \\ &= \|J_{\lambda A}(x_2) - J_{\lambda A}(x_1)\|^2 + 2\lambda\langle A_\lambda(x_2) - A_\lambda(x_1), J_{\lambda A}(x_2) - J_{\lambda A}(x_1)\rangle + \lambda^2\|A_\lambda(x_2) - A_\lambda(x_1)\|^2,\end{aligned}$$

for all x_1, $x_2 \in \mathrm{Dom}(A_\lambda) = \mathrm{Dom}(J_{\lambda A})$. The first term is clearly nonnegative. Since $A_\lambda(x_i) \in A(J_{\lambda A}(x_i))$ (from Lemma 1.3 and since A_λ is single-valued) and since A is monotone, the second term is also nonnegative. We deduce that

$$\|A_\lambda(x_2) - A_\lambda(x_1)\| \leqslant \frac{1}{\lambda}\|x_2 - x_1\|,$$

for all x_1, $x_2 \in \mathrm{Dom}(A_\lambda)$.

2. For similar reasons, we have

$$\langle A_\lambda(x_2) - A_\lambda(x_1), x_2 - x_1\rangle = \langle A_\lambda(x_2) - A_\lambda(x_1), J_{\lambda A}(x_2) - J_{\lambda A}(x_1)\rangle + \lambda\|A_\lambda(x_2) - A_\lambda(x_1)\|^2 \geqslant 0,$$

for all x_1, $x_2 \in \mathrm{Dom}(A_\lambda) = \mathrm{Dom}(J_{\lambda A})$.

The proof is complete. □

Theorem 1.2. If a set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximal monotone, then:

(i) $J_A : \mathbb{R}^d \to \mathbb{R}^d$ is single-valued, with $\mathrm{Dom}(J_A) = \mathbb{R}^d$, 1-Lipschitz continuous and maximal monotone;

(ii) $A_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ is single-valued, with $\mathrm{Dom}(A_\lambda) = \mathbb{R}^d$, $\frac{1}{\lambda}$-Lipschitz continuous and maximal monotone;

(iii) $0 \leqslant \|\mathfrak{m}(A(x)) - A_\lambda(x)\|^2 \leqslant \|\mathfrak{m}(A(x))\|^2 - \|A_\lambda(x)\|^2$ for all $x \in \mathrm{Dom}(A)$;

for all $\lambda > 0$. Furthermore

$$\forall x \in \mathrm{Dom}(A), \quad \lim_{\lambda \to 0^+} J_{\lambda A}(x) = x \quad \text{and} \quad \lim_{\lambda \to 0^+} A_\lambda(x) = \mathfrak{m}(A(x)).$$

Proof. (i) Since A is monotone, we know from Proposition 1.3 that its resolvent $J_A : \mathrm{Dom}(J_A) \to \mathbb{R}^d$ is single-valued, 1-Lipschitz continuous and monotone. Furthermore, from Minty's theorem, since A is maximal monotone, we have $\mathrm{Dom}(J_A) = \mathrm{Ran}(\mathrm{Id} + A) = \mathbb{R}^d$. Finally, since $J_A : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous monotone single-valued map, we deduce from Proposition 1.2 that it is maximal monotone.

(ii) Let $\lambda > 0$. Since λA is monotone, we know from Proposition 1.3 that $A_\lambda : \mathrm{Dom}(A_\lambda) \to \mathbb{R}^d$ is single-valued, $\frac{1}{\lambda}$-Lipschitz continuous and monotone. Furthermore, from Minty's theorem, since λA is maximal monotone, we have $\mathrm{Dom}(A_\lambda) = \mathrm{Dom}(J_{\lambda A}) = \mathrm{Ran}(\mathrm{Id} + \lambda A) = \mathbb{R}^d$. Finally, since $A_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous monotone single-valued map, we deduce from Proposition 1.2 that it is maximal monotone.

(iii) Let $\lambda > 0$ and $x \in \mathrm{Dom}(A)$. Recall that $\mathfrak{m}(A(x))$ is well-defined since A is assumed to be maximal monotone. We have

$$\begin{aligned}\|\mathfrak{m}(A(x)) - A_\lambda(x)\|^2 &= \|\mathfrak{m}(A(x))\|^2 + \|A_\lambda(x)\|^2 - 2\langle \mathfrak{m}(A(x)), A_\lambda(x)\rangle \\ &= \|m(A(x))\|^2 - \|A_\lambda(x)\|^2 - 2\langle \mathfrak{m}(A(x)) - A_\lambda(x), A_\lambda(x)\rangle \\ &= \|\mathfrak{m}(A(x))\|^2 - \|A_\lambda(x)\|^2 - \frac{2}{\lambda}\langle \mathfrak{m}(A(x)) - A_\lambda(x), x - J_{\lambda A}(x)\rangle.\end{aligned}$$

Since $\mathfrak{m}(A(x)) \in A(x)$ and $A_\lambda(x) \in A(J_{\lambda A}(x))$ (from Lemma 1.3) and since A is monotone, we obtain that the last term is nonnegative and thus $0 \leqslant \|\mathfrak{m}(A(x)) - A_\lambda(x)\|^2 \leqslant \|\mathfrak{m}(A(x))\|^2 - \|A_\lambda(x)\|^2$.

(iv) Let $x \in \mathrm{Dom}(A)$. From the previous item, we have $\|x - J_{\lambda A}(x)\| = \lambda\|A_\lambda(x)\| \leqslant \lambda\|\mathfrak{m}(A(x))\|$ for all $\lambda > 0$. We obtain that $J_{\lambda A}(x) \to x$ when $\lambda \to 0^+$.

(v) Let $x \in \mathrm{Dom}(A)$.

- In a first place, we prove that $\|A_\lambda(x)\|$ converges to some real number γ when $\lambda \to 0^+$. First, since $\|A_\lambda(x)\| \leqslant \|\mathfrak{m}(A(x))\|$ for all $\lambda > 0$, we know that the sequence $(\|A_\lambda(x)\|)_{\lambda>0}$ is bounded. Secondly, using Lemma 1.2 and the third item with $A = A_\mu$ which is single-valued and maximal monotone, we obtain that $\|A_{\mu+\lambda}(x)\| = \|(A_\mu)_\lambda(x)\| \leqslant \|\mathfrak{m}(A_\mu(x))\| = \|A_\mu(x)\|$ for all λ, $\mu > 0$. We deduce that the sequence $(\|A_\lambda(x)\|)_{\lambda>0}$ is monotone (precisely increasing when $\lambda \to 0^+$). This concludes this first step.

- Now let us prove that the sequence $(A_\lambda(x))_{\lambda>0}$ converges when $\lambda \to 0^+$, by proving that it is a Cauchy sequence. Using Lemma 1.2 and the third item with $A = A_\mu$ which is single-valued and maximal monotone, we obtain that $\|A_\mu(x) - A_{\mu+\lambda}(x)\|^2 = \|\mathfrak{m}(A_\mu(x)) - (A_\mu)_\lambda(x)\|^2 \leqslant \|\mathfrak{m}(A_\mu(x))\|^2 - \|(A_\mu)_\lambda(x)\|^2 = \|A_\mu(x)\|^2 - \|A_{\mu+\lambda}(x)\|^2$ which converges to $\gamma - \gamma = 0$ when $\lambda \to 0^+$ and $\mu \to 0^+$. Thus the sequence $(A_\lambda(x))_{\lambda>0}$ is a Cauchy sequence when $\lambda \to 0^+$, and thus converges when $\lambda \to 0^+$, denoting by ℓ its limit.

- Since $A_\lambda(x) \in A(J_{\lambda A}(x))$ for all $\lambda > 0$ (see Lemma 1.2), with $A_\lambda(x) \to \ell$ and $J_{\lambda A}(x) \to x$ when $\lambda \to 0^+$, and since $\mathrm{Gr}(A)$ is closed (from Proposition 1.1), we obtain that $\ell \in A(x)$.

- Finally, since $\|A_\lambda(x)\| \leqslant \|\mathfrak{m}(A(x))\|$ for all $\lambda > 0$, we obtain by passing to the limit $\lambda \to 0^+$ that $\|\ell\| \leqslant \|\mathfrak{m}(A(x))\|$. Since $\ell \in A(x)$ and by definition of $\mathfrak{m}(A(x))$, we obtain that $\ell = \mathfrak{m}(A(x))$.

Hence we have prove that $A_\lambda(x) \to \mathfrak{m}(A(x))$ when $\lambda \to 0^+$. The proof is complete. □

Remark 1.19. The above theorem shows that any maximal monotone set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ can be approximated, in a certain sense, by maximal monotone Lipschitz continuous single-valued maps $(A_\lambda)_{\lambda>0}$ defined on the whole space $\mathbb{R}^d$ which are called *Yosida approximations*.

2 *Fixed points of averaged operators*

In this lecture, the terminology *operators* will stand for single-valued maps of the form $T : D \to \mathbb{R}^d$ where $D = \mathrm{Dom}(T)$ is nonempty. Furthermore, for an operator $T : D \to \mathbb{R}^d$ with values in D, we will denote by $T : D \to D$.

Definition 2.1 (Contractive operator)**.** An operator $T : D \to \mathbb{R}^d$ is said to be *contractive* if it is L-Lipschitz continuous, that is, if

$$\forall (x_1, x_2) \in D \times D, \quad \|T(x_2) - T(x_1)\| \leqslant L\|x_2 - x_1\|,$$

with $0 \leqslant L < 1$.

Theorem 2.1 (Picard fixed point theorem)**.** If $T : D \to D$ is a contractive operator with D closed, then T admits a unique fixed point. Furthermore, the *fixed-point algorithm* given by

$$x_0 \in D \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := T(x_k),$$

converges to the unique fixed point of T.

Proof. This theorem is a very standard result. The proof can easily be found (on Internet or books of analysis). □

Note that $\mathrm{Zer}(A) = \mathrm{Fix}(J_A)$ for any set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ (exercise). In order to solve an inclusion of the form $0_{\mathbb{R}^d} \in A(x)$, a popular method in the literature is thus to determine a fixed point of the resolvent J_A. Therefore, the fixed-point algorithm $x_{k+1} \in J_A(x_k)$ is naturally considered in order to numerically approach a zero of A. When A is maximal monotone, we have seen that J_A enjoys several nice mathematical properties such as single-valuedness with domain $\mathrm{Dom}(J_A) = \mathbb{R}^d$ and 1-Lipschitz continuity. Unfortunately, even in that context, it is not guaranteed that J_A is a contractive operator and thus the fixed-point algorithm $x_{k+1} := J_A(x_k)$ may not converge. Therefore, the aim of this section is to provide alternative algorithms and convergence results with weaker assumptions than contraction.

2.1 *Preliminaries on Fejér monotone sequences*

Definition 2.2 (Fejér monotone sequence)**.** A sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^d$ is said to be *Fejér monotone* with respect to a nonempty subset E of $\mathbb{R}^d$ if

$$\forall x \in E, \quad \forall k \in \mathbb{N}, \quad \|x_{k+1} - x\| \leqslant \|x_k - x\|.$$

Remark 2.1. If a sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^d$ is Fejér monotone with respect to a nonempty subset E of $\mathbb{R}^d$, then the sequence $(\|x_k - x\|)_{k\in\mathbb{N}}$ converges for all $x \in E$, and the sequence $(x_k)_{k\in\mathbb{N}}$ is bounded.

Lemma 2.1. A sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^d$ converges if and only if it is bounded and admits at most one cluster point.

Proof. The necessary condition is trivial. Let us prove the sufficient condition. Assume that $(x_k)_{k\in\mathbb{N}}$ is bounded and admits at most one cluster point. From Bolzano-Weierstrass theorem, we know that $(x_k)_{k\in\mathbb{N}}$ admits exactly one cluster point denoted by ℓ. By contradiction, assume that $(x_k)_{k\in\mathbb{N}}$ does not converge to ℓ. Then, there exists $\varepsilon > 0$ and a subsequence $(x_{\psi(k)})_{k\in\mathbb{N}}$ such that $\|x_{\psi(k)} - \ell\| \geqslant \varepsilon$ for all $k \in \mathbb{N}$. Since the sequence $(x_{\psi(k)})_{k\in\mathbb{N}}$ is bounded (since $(x_k)_{k\in\mathbb{N}}$ is), it admits a subsequence which converges to some ℓ' with $\|\ell' - \ell\| \geqslant \varepsilon$. We deduce that $(x_k)_{k\in\mathbb{N}}$ admits two different cluster points ℓ and ℓ' which is absurd. □

In Optimizat°, usually we solve:

$$\begin{cases} \text{Find } x \in \mathbb{R}^d \text{ s.t.} \\ 0 \in A(x) \end{cases}$$

where $A: \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ maximal monotone

Think: $A = \partial f$, $f \in \Gamma_0(\mathbb{R}^d)$

$$\min_{x \in \mathbb{R}^d} f(x) = f(\bar{x})$$

$$\Updownarrow$$

$$\boxed{0 \in \partial f(\bar{x})}$$

$$0 \in A(x) \iff 0 \in \lambda A(x), \quad \lambda > 0$$

$$\iff x \in x + \lambda A(x), \quad \lambda > 0$$

$$\iff x \in (Id + \lambda A)(x)$$

$$\iff \underbrace{(Id + \lambda A)^{-1}}_{J_{\lambda A}}(x) \boxed{\ni} x$$

$$\boxed{0 \in A(x) \iff J_{\lambda A}(x) = x}$$

$$\mathrm{Zer}(A) = \mathrm{Fix}\left(J_{\lambda A}\right) \quad \forall \lambda > 0$$

$$\begin{cases} x_0 \in \mathbb{R}^d \\ x_{k+1} = J_{\lambda A}(x_k) \end{cases}$$

Fixed point Algorithm !

$\| J_{\lambda A}(x) - J_{\lambda A}(y) \| \leq \|x - y\|$ non expansive !

$$A = \partial f \longleftarrow J_{\lambda A} = \mathrm{prox}_{\lambda f}$$

$$\begin{cases} x_0 \\ x_{k+1} = \mathrm{prox}_{\lambda f}(x_k) \end{cases}$$

PPA Proximal Point Algorithm

Lemma 2.2. If a sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^d$ is Fejér monotone with respect to a nonempty subset E of $\mathbb{R}^d$ and such that every cluster point of $(x_k)_{k\in\mathbb{N}}$ belongs to E, then $(x_k)_{k\in\mathbb{N}}$ converges to some point of E.

Proof. From Lemma 2.1 and Remark 2.1, we only need to prove that $(x_k)_{k\in\mathbb{N}}$ admits at most one cluster point. Thus take x, y two cluster points of $(x_k)_{k\in\mathbb{N}}$ and let us prove that $x = y$. We write $x_{\psi_1(k)} \to x$ and $x_{\psi_2(k)} \to y$. From hypothesis we know that $x, y \in E$ and, from Remark 2.1, that $\|x_k - x\|$ and $\|x_k - y\|$ both converge when $k \to \infty$. We deduce from

$$\forall k \in \mathbb{N}, \quad 2\langle x_k, x - y\rangle = \|x_k - y\|^2 - \|x_k - x\|^2 + \|x\|^2 - \|y\|^2,$$

that $\langle x_k, x - y\rangle$ converges to some real $\gamma \in \mathbb{R}$. We deduce that $\langle x_{\psi_1(k)}, x - y\rangle$ converges to $\gamma = \langle x, x - y\rangle$, and similarly that $\langle x_{\psi_2(k)}, x - y\rangle$ converges to $\gamma = \langle y, x - y\rangle$. We deduce that $\|x - y\|^2 = 0$ and thus $x = y$. □

2.2 Nonexpansive operators and Krasnoselskii-Mann algorithm

Definition 2.3 (Nonexpansive operator)**.** An operator $T : D \to \mathbb{R}^d$ is said to be *nonexpansive* if it is 1-Lipschitz continuous, that is, if

$$\forall (x_1, x_2) \in D \times D, \quad \|T(x_2) - T(x_1)\| \leqslant \|x_2 - x_1\|.$$

Remark 2.2. The resolvent $J_A : \mathrm{Dom}(J_A) \to \mathbb{R}^d$ of a monotone set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is nonexpansive (see Proposition 1.3).

Theorem 2.2 (KM algorithm for nonexpansive operators)**.** Let $T : D \to D$ be a nonexpansive operator such that D is closed convex and $\mathrm{Fix}(T) \neq \emptyset$. Let $(\lambda_k)_{k\in\mathbb{N}} \subset [0,1]$ be a sequence such that $\sum_{k\in\mathbb{N}} \lambda_k(1-\lambda_k) = +\infty$. Then the *KM algorithm* given by

$$x_0 \in D \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := x_k + \lambda_k(T(x_k) - x_k),$$

converges to some point in $\mathrm{Fix}(T)$. Furthermore, the estimation

$$\forall k \in \mathbb{N}, \quad \|T(x_k) - x_k\|^2 \leqslant \frac{\|x_0 - x\|^2}{\sum_{j=0}^{k} \lambda_j(1-\lambda_j)},$$

$x_{k+1} = \cancel{x_k} - \lambda_k x_k$
$\cancel{- x_k}$

holds true for any point $x \in \mathrm{Fix}(T)$.

Proof. From convexity of D, one can see that the sequence $(x_k)_{k\in\mathbb{N}} \subset D$ is well defined. The proof is based on four steps:

1. Let us prove that $(\|x_k - T(x_k)\|)_{k\in\mathbb{N}}$ is decreasing and thus converges to some $\ell \geqslant 0$. It simply follows from

$$\begin{aligned}\|x_{k+1} - T(x_{k+1})\| &= \|(1-\lambda_k)(x_k - T(x_k)) + T(x_k) - T(x_{k+1})\| \\ &\leqslant (1-\lambda_k)\|x_k - T(x_k)\| + \|x_k - x_{k+1}\| = \|x_k - T(x_k)\|,\end{aligned}$$

   for all $k \in \mathbb{N}$.

2. Let us prove that $x_k - T(x_k) \to 0_{\mathbb{R}^d}$. To see this, consider some $x \in \mathrm{Fix}(T) \neq \emptyset$ and, from the first item of Lemma A.2 and nonexpansiveness of T, we get that

$$\begin{aligned}\|x_{k+1} - x\|^2 &= \|(1-\lambda_k)(x_k - x) + \lambda_k(T(x_k) - x)\|^2 \\ &= (1-\lambda_k)\|x_k - x\|^2 + \lambda_k\|T(x_k) - x\|^2 - \lambda_k(1-\lambda_k)\|T(x_k) - x_k\|^2 \\ &\leqslant \|x_k - x\|^2 - \lambda_k(1-\lambda_k)\|T(x_k) - x_k\|^2,\end{aligned}$$

$x_k = -\lambda x_{k-1}$

$x_{k-1} = -\lambda x_{k-2}$

$x_k = \lambda^2 x_{k-2}$

$\begin{cases} x_0 \\ x_{k+1} = -\lambda x_k \end{cases}$

$x_k = (-\lambda)^{k+1} x_0$

for all $k \in \mathbb{N}$. We obtain that

$$\lambda_k(1-\lambda_k)\|T(x_k)-x_k\|^2 \leqslant \|x_k-x\|^2 - \|x_{k+1}-x\|^2,$$

for all $k \in \mathbb{N}$. Therefore $\sum_{k\in\mathbb{N}} \lambda_k(1-\lambda_k)\|T(x_k)-x_k\|^2 \leqslant \|x_0-x\|^2$. Since $\sum_{k\in\mathbb{N}} \lambda_k(1-\lambda_k) = +\infty$ and $(\|x_k - T(x_k)\|)_{k\in\mathbb{N}}$ converges to some $\ell \geqslant 0$, we deduce that $\ell = 0$ and thus $x_k - T(x_k) \to 0_{\mathbb{R}^d}$.

3. Now let us prove that $(x_k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$. Let $x \in \mathrm{Fix}(T)$. Taking again the above computations, we get that

$$\|x_{k+1}-x\|^2 \leqslant \|x_k-x\|^2 - \lambda_k(1-\lambda_k)\|T(x_k)-x_k\|^2 \leqslant \|x_k-x\|^2,$$

for all $k \in \mathbb{N}$.

Since T is continuous and D is closed and $x_k - T(x_k) \to 0_{\mathbb{R}^d}$, one can easily conclude that every cluster point of $(x_k)_{k\in\mathbb{N}}$ belongs to $\mathrm{Fix}(T)$. Since moreover $(x_k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$, we deduce from Lemma 2.2 that $(x_k)_{k\in\mathbb{N}}$ converges to some point in $\mathrm{Fix}(T)$. To conclude the proof, we only need to prove the estimation. Consider some $x \in \mathrm{Fix}(T)$ and take the inequality from the above third item to obtain

$$\lambda_k(1-\lambda_k)\|T(x_k)-x_k\|^2 \leqslant \|x_k-x\|^2 - \|x_{k+1}-x\|^2,$$

for all $k \in \mathbb{N}$. Summing these inequalities over $j \in \{0,\dots,k\}$, and using the fact that the sequence $(\|x_k - T(x_k)\|)_{k\in\mathbb{N}}$ is decreasing, we obtain

$$\begin{aligned}\|T(x_k)-x_k\|^2 \sum_{j=0}^{k} \lambda_j(1-\lambda_j) &\leqslant \sum_{j=0}^{k} \lambda_j(1-\lambda_j)\|T(x_j)-x_j\|^2 \\ &\leqslant \sum_{j=0}^{k} \|x_j-x\|^2 - \|x_{j+1}-x\|^2 \leqslant \|x_0-x\|^2 - \|x_{k+1}-x\|^2 \leqslant \|x_0-x\|^2,\end{aligned}$$

for all $k \in \mathbb{N}$, which concludes the proof. □

Remark 2.3. In the framework of Theorem 2.2, the fixed point algorithm corresponds to $\lambda_k = 1$ for all $k \in \mathbb{N}$. Unfortunately, in that case, the condition $\sum_{k\in\mathbb{N}} \lambda_k(1-\lambda_k) = +\infty$ is not satisfied and thus the convergence is not guaranteed (take $T := -\mathrm{Id}$ and $x_0 \neq 0_{\mathbb{R}^d}$ for a counterexample). In the next sections, we will introduce stronger notions than nonexpansiveness allowing to obtain convergence results for the standard fixed-point algorithm.

2.3 *Firmly nonexpansive operators and averaged operators*

Definition 2.4 (Firmly nonexpansive and quasinonexpansive operators)**.** An operator $T : D \to \mathbb{R}^d$ is said to be:

(i) *firmly nonexpansive* if

$$\forall (x_1,x_2) \in D\times D, \quad \|T(x_2)-T(x_1)\|^2 + \|(\mathrm{Id}-T)(x_2)-(\mathrm{Id}-T)(x_1)\|^2 \leqslant \|x_2-x_1\|^2.$$

(ii) *quasinonexpansive* if

$$\forall (x_1,x_2) \in \mathrm{Fix}(T)\times D, \quad \|T(x_2)-x_1\| \leqslant \|x_2-x_1\|.$$

(iii) *strictly quasinonexpansive* if

$$\forall (x_1,x_2) \in \mathrm{Fix}(T)\times (D\backslash\mathrm{Fix}(T)), \quad \|T(x_2)-x_1\| < \|x_2-x_1\|.$$

$\|T(x_2)-T(x_1)\|^2 \leqslant \|T(x_2)-T(x_1)\|^2 + \underbrace{\|(\mathrm{Id}-T)(x_2)-(\mathrm{Id}-T)(x_1)\|^2}_{\geqslant 0} \leqslant \|x_2-x_1\|^2$

Remark 2.4. Note that

$$\text{firmly nonexpansive} \implies \text{nonexpansive} \implies \text{quasinonexpansive},$$

and that ?

$$\text{firmly nonexpansive} \implies \text{strictly quasinonexpansive} \implies \text{quasinonexpansive}.$$

Proposition 2.1 (Characterization)**.** Let $T : D \to \mathbb{R}^d$ be an operator. The following assertions are equivalent:

(i) T is firmly nonexpansive. $T \leftarrow \beta T$

(ii) $\mathrm{Id} - T$ is firmly nonexpansive.

(iii) $2T - \mathrm{Id}$ is nonexpansive.

(iv) $\|T(x_2) - T(x_1)\|^2 \leqslant \langle T(x_2) - T(x_1), x_2 - x_1\rangle$ for all $(x_1, x_2) \in D \times D$.

(v) $\langle T(x_2) - T(x_1), (\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\rangle \geqslant 0$ for all $(x_1, x_2) \in D \times D$.

(vi) $\|T(x_2) - T(x_1)\| \leqslant \|(1 - \delta)(T(x_2) - T(x_1)) + \delta(x_2 - x_1)\|$ for all $(x_1, x_2) \in D \times D$ and all $\delta \in [0, 1]$.

Proof. Note that the equivalence (i) $\Leftrightarrow$ (ii) is trivial. To prove (i) $\Leftrightarrow$ (iii), let us define $R := 2T - \mathrm{Id}$. From the first item of Lemma A.2, we have

$$\begin{aligned}\|R(x_2) - R(x_1)\|^2 &= \|2(T(x_2) - T(x_1)) + (1 - 2)(x_2 - x_1)\|^2 \\ &= 2\|T(x_2) - T(x_1)\|^2 + (1 - 2)\|x_2 - x_1\|^2 - 2(1 - 2)\|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2,\end{aligned}$$

and thus

$$\|R(x_2) - R(x_1)\|^2 - \|x_2 - x_1\|^2 = 2\Big(\|T(x_2) - T(x_1)\|^2 + \|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2 - \|x_2 - x_1\|^2\Big),$$

for all x_1, $x_2 \in D$. Thus the equivalence (i) $\Leftrightarrow$ (iii) follows. To prove that (i) $\Leftrightarrow$ (iv), note that

$$\|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2 = \|x_2 - x_1\|^2 + \|T(x_2) - T(x_1)\|^2 - 2\langle T(x_2) - T(x_1), x_2 - x_1\rangle,$$

and thus

$$\begin{aligned}&\|T(x_2) - T(x_1)\|^2 + \|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2 - \|x_2 - x_1\|^2 \\ &\qquad = 2\Big(\|T(x_2) - T(x_1)\|^2 - \langle T(x_2) - T(x_1), x_2 - x_1\rangle\Big),\end{aligned}$$

for all x_1, $x_2 \in D$. Thus the equivalence (i) $\Leftrightarrow$ (iv) follows. Note that the equivalence (iv) $\Leftrightarrow$ (v) is trivial. Finally, from the second item of Lemma A.2, the equivalence (v) $\Leftrightarrow$ (vi) is proved. □

Remark 2.5. From Proposition 2.1, note that, if $T : D \to \mathbb{R}^d$ is a firmly nonexpansive operator, then T is monotone.

Remark 2.6. Let $T : D \to \mathbb{R}^d$ be an operator. The operator $\mathrm{refl}_T : D \to \mathbb{R}^d$, defined by $\mathrm{refl}_T := 2T - \mathrm{Id}$, is usually called the *reflexion operator* of T. From Proposition 2.1, T is firmly nonexpansive if and only if refl_T is nonexpansive.

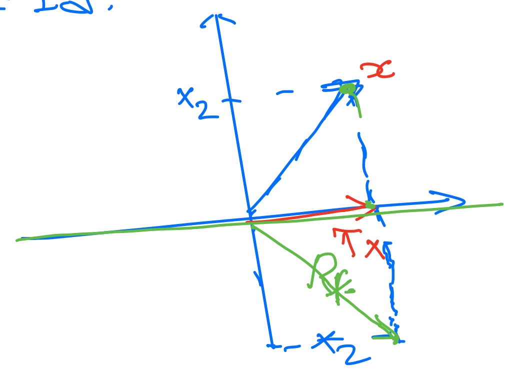
Theorem 2.3. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map with $J_A : \mathrm{Dom}(J_A) \to \mathbb{R}^d$ single-valued. Then the following properties are satisfied:

(i) A is monotone if and only if J_A is firmly nonexpansive.

(ii) A is maximal monotone if and only if J_A is firmly nonexpansive with $\mathrm{Dom}(J_A) = \mathbb{R}^d$.

  $J_A = (\mathrm{Id} + A)^{-1}$!!

$T = Id \longrightarrow R = 2T - Id = Id.$



$T = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$

$Tx = \begin{pmatrix} x_1 \\ 0 \end{pmatrix}$

$Rx = 2Tx - x = \begin{pmatrix} 2x_1 - x_1 \\ 0 - x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix}$

Reflexion operator !!

$(Id - T)\,x = x - Tx = \begin{pmatrix} x_1 - x_1 \\ x_2 - 0 \end{pmatrix} = \begin{pmatrix} 0 \\ x_2 \end{pmatrix}$

$T : D \longrightarrow D$

- T is monotone $\Leftrightarrow \langle Tx - Ty, x - y \rangle \geq 0, \ \forall x, y \in D$
- T is strongly monotone with modulus $\alpha > 0$ $\Leftrightarrow \langle Tx - Ty, x - y \rangle \geq \alpha \|x - y\|^2 \quad \forall x, y \in D$
- T is co-coercive with modulus $\alpha > 0$ $\Leftrightarrow \langle Tx - Ty, x - y \rangle \geq \alpha \|Tx - Ty\|^2 \quad \forall x, y \in D$

<u>Example:</u> $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ convex of class $\mathcal{C}^1$.

- f is convex $\Leftrightarrow \nabla f$ is monotone.
- f is strongly convex $\Leftrightarrow \nabla f$ is strongly monotone

T is k-Lipschitz : $\|Tx - Ty\| \leq k \|x - y\|$
$k > 0$.

Proof. In this proof, we denote by $D := \mathrm{Dom}(J_A)$ and $T := J_A : D \to \mathbb{R}^d$ which is single-valued. Let us prove (i). Assume that T is firmly nonexpansive. Take (x_1, y_1), $(x_2, y_2) \in \mathrm{Gr}(A)$. Then $x_i = T(x_i + y_i)$ and from firm nonexpansiveness we get that

$$\begin{aligned}\langle y_2 - y_1, x_2 - x_1\rangle &= \langle (x_2 + y_2) - (x_1 + y_1), x_2 - x_1\rangle - \|x_2 - x_1\|^2 \\ &= \langle (x_2 + y_2) - (x_1 + y_1), T(x_2 + y_2) - T(x_1 + y_1)\rangle - \|x_2 - x_1\|^2 \\ &\geqslant \|T(x_2 + y_2) - T(x_1 + y_1)\|^2 - \|x_2 - x_1\|^2 = 0,\end{aligned}$$

and thus A is monotone. Conversely assume that A is monotone. Take x_1, $x_2 \in D$ and define $z_1 := T(x_1)$, $z_2 := T(x_2)$. We obtain that $(z_i, x_i - z_i) \in \mathrm{Gr}(A)$ and thus

$$0 \leqslant \langle (x_2 - z_2) - (x_1 - z_1), z_2 - z_1\rangle = \langle x_2 - x_1, z_2 - z_1\rangle - \|z_2 - z_1\|^2$$

and thus

$$\langle x_2 - x_1, T(x_2) - T(x_1)\rangle \geqslant \|T(x_2) - T(x_1)\|^2,$$

and thus T is firmly nonexpansive from Proposition 2.1. Let us prove (ii). Assume that T is firmly nonexpansive and $D = \mathbb{R}^d$. We already know from the previous item that A is monotone. Since $\mathrm{Ran}(\mathrm{Id} + A) = \mathrm{Ran}(T^{-1}) = \mathrm{Dom}(T) = \mathbb{R}^d$, A is maximal monotone from Minty's theorem. Assume that A is maximal monotone. We already know from the previous item that T is firmly nonexpansive. Then $\mathrm{Dom}(T) = \mathrm{Ran}(T^{-1}) = \mathrm{Ran}(\mathrm{Id} + A) = \mathbb{R}^d$ from Minty's theorem. □

Definition 2.5 (Averaged operator)**.** An operator $T : D \to \mathbb{R}^d$ is said to be α-*averaged*, with $0 < \alpha < 1$, if there exists a nonexpansive operator $R : D \to \mathbb{R}^d$ such that $T = (1 - \alpha)\mathrm{Id} + \alpha R$.

Remark 2.7. An averaged operator $T : D \to \mathbb{R}^d$ is trivially nonexpansive. The converse is not true in general (take $T = -\mathrm{Id}$ for a counterexample).

Remark 2.8. An averaged operator corresponds to a strict convex combination of the identity map with another nonexpansive operator. In addition to nonexpansiveness, we will see that averaged operators enjoy several nice mathematical properties.

Proposition 2.2 (Characterization)**.** Let $T : D \to \mathbb{R}^d$ be an operator and $0 < \alpha < 1$. The following assertions are equivalent:

(i) T is α-averaged.

(ii) $(1 - \frac{1}{\alpha})\mathrm{Id} + \frac{1}{\alpha}T$ is nonexpansive.

(iii) $\|T(x_2) - T(x_1)\|^2 \leqslant \|x_2 - x_1\|^2 + (1 - \frac{1}{\alpha})\|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2$ for all $(x_1, x_2) \in D \times D$.

(iv) $\|T(x_2) - T(x_1)\|^2 + (1 - 2\alpha)\|x_2 - x_1\|^2 \leqslant 2(1 - \alpha)\langle T(x_2) - T(x_1), x_2 - x_1\rangle$ for all $(x_1, x_2) \in D \times D$.

In the above items, note that $\frac{1}{\alpha} > 1$ and $(1 - \frac{1}{\alpha}) < 0$.

Proof. In the whole proof, let us define $R := (1 - \frac{1}{\alpha})\mathrm{Id} + \frac{1}{\alpha}T$ and then we have $T = (1 - \alpha)\mathrm{Id} + \alpha R$. The equivalence (i) ⇔ (ii) follows. Now, from the first item of Lemma A.2, we have

$$\|R(x_2) - R(x_1)\|^2 = \frac{1}{\alpha}\|T(x_2) - T(x_1)\|^2 + \left(1 - \frac{1}{\alpha}\right)\|x_2 - x_1\|^2 - \frac{1}{\alpha}\left(1 - \frac{1}{\alpha}\right)\|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2,$$

$$T = -Id = (1-\alpha) Id + \alpha R \qquad 0 < \alpha < 1$$

$$\Updownarrow$$

$$\alpha R = -Id - Id + \alpha Id = (\alpha - 2) Id$$

$$\boxed{R = \frac{\alpha - 2}{\alpha} Id} \qquad \boxed{0 < \alpha < 1}$$

$$\|Rx - Ry\| = \frac{2-\alpha}{\alpha} \|\cancel{x-y}\| \overset{?}{\leq} \|\cancel{x-y}\|$$

$$2 - \alpha \leq \alpha \iff 2 \leq 2\alpha$$

$$\iff \alpha \geq 1 \quad \text{Contradicto!}$$

and thus

$$\alpha\Big(\|R(x_2) - R(x_1)\|^2 - \|x_2 - x_1\|^2\Big)$$
$$= \|T(x_2) - T(x_1)\|^2 - \|x_2 - x_1\|^2 - \left(1 - \frac{1}{\alpha}\right)\|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2,$$

for all x_1, $x_2 \in D$. The equivalence (ii) $\Leftrightarrow$ (iii) follows. For the last equivalence (iii) $\Leftrightarrow$ (iv), one has just to write

$$\|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2 = \|x_2 - x_1\|^2 + \|T(x_2) - T(x_1)\|^2 - 2\langle T(x_2) - T(x_1), x_2 - x_1\rangle,$$

for all x_1, $x_2 \in D$. $\square$

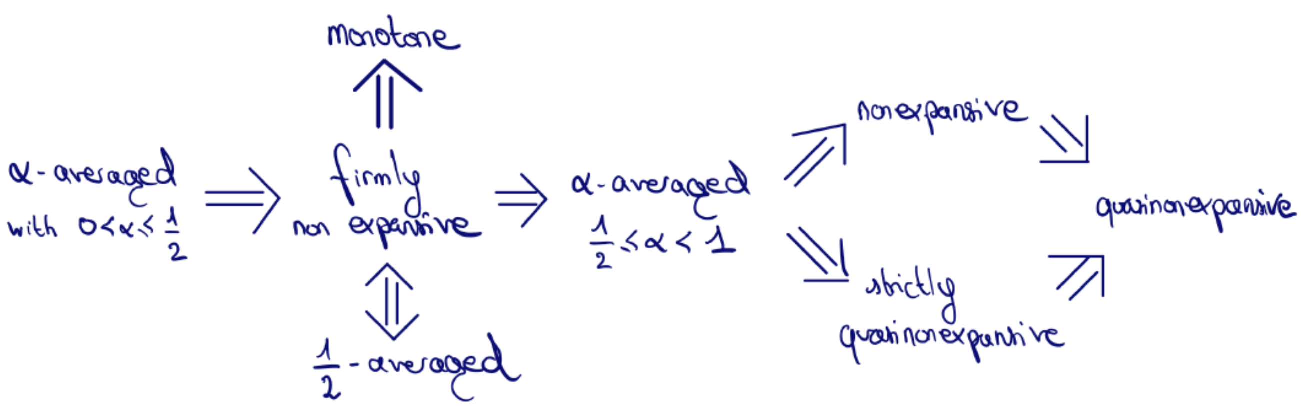
Remark 2.9. Several remarks:

(i) An operator $T : D \to \mathbb{R}^d$ is firmly nonexpansive if and only if it is $\frac{1}{2}$-averaged.

(ii) If an operator $T : D \to \mathbb{R}^d$ is α-averaged with $0 < \alpha < 1$, then it is α'-averaged for all $\alpha \leqslant \alpha' < 1$.

(iii) If an operator $T : D \to \mathbb{R}^d$ is α-averaged with $0 < \alpha \leqslant \frac{1}{2}$, then it is firmly nonexpansive.

(iv) If an operator $T : D \to \mathbb{R}^d$ is α-averaged with $0 < \alpha < 1$, then it is strictly quasinonexpansive.



Definition 2.6 (Cocoercive operator)**.** An operator $T : D \to \mathbb{R}^d$ is said to be $\boxed{\beta\text{-}cocoercive}$, with $\beta > 0$, if

$$\forall (x_1, x_2) \in D \times D, \quad \langle T(x_2) - T(x_1), x_2 - x_1\rangle \geqslant \beta\|T(x_2) - T(x_1)\|^2,$$

that is, if βT is firmly nonexpansive.

Remark 2.10. An operator $T : D \to \mathbb{R}^d$ is firmly nonexpansive if and only if it is 1-cocoercive.

Remark 2.11. An operator $T : D \to \mathbb{R}^d$ is β-cocoercive, with $\beta > 0$, if and only if βT is $\frac{1}{2}$-averaged.

Remark 2.12. Note that, if an operator $T : D \to \mathbb{R}^d$ is β-cocoercive, with $\beta > 0$, then it is β'-cocoercive for all $0 < \beta' \leqslant \beta$.

Lemma 2.3. An operator $T : D \to \mathbb{R}^d$ is α-averaged, with $0 < \alpha < 1$, if and only if $(1 - \lambda)\mathrm{Id} + \lambda T$ is nonexpansive for all $0 \leqslant \lambda \leqslant \frac{1}{\alpha}$.

Proof. The sufficient condition is trivial from Proposition 2.2 and by taking $\lambda = \frac{1}{\alpha}$. Now let us prove the necessary condition (which is trivial for $0 \leqslant \lambda \leqslant 1$ and $\lambda = \frac{1}{\alpha}$). Let $1 < \lambda < \frac{1}{\alpha}$ and take $\alpha' := \frac{1}{\lambda}$ which is such that $\alpha < \alpha' < 1$. From Remark 2.9, we know that T is α'-averaged and thus $(1-\lambda)\mathrm{Id} + \lambda T = (1-\frac{1}{\alpha'})\mathrm{Id} + \frac{1}{\alpha'}T$ is nonexpansive from Proposition 2.2. ☐

Proposition 2.3. Let $T : D \to \mathbb{R}^d$ be an operator. The following properties are satisfied:

1. Let $0 < \alpha < 1$ and $0 < \lambda < \frac{1}{\alpha}$. Then T is α-averaged if and only if $(1-\lambda)\mathrm{Id} + \lambda T$ is $\lambda\alpha$-averaged.
2. Let $0 < \lambda < 2$. Then T is firmly nonexpansive if and only if $(1-\lambda)\mathrm{Id} + \lambda T$ is $\frac{\lambda}{2}$-averaged.
3. If T is β-cocoercive, with $\beta > 0$, then $\mathrm{Id} - \lambda T$ is $\frac{\lambda}{2\beta}$-averaged for all $0 < \lambda < 2\beta$.

Proof. (1) Denote $S := (1-\lambda)\mathrm{Id} + \lambda T$ and $R := (1-\frac{1}{\alpha})\mathrm{Id} + \frac{1}{\alpha}T$. We get that

$$T = \left(1 - \frac{1}{\lambda}\right)\mathrm{Id} + \frac{1}{\lambda}S, \qquad T = (1-\alpha)\mathrm{Id} + \alpha R, \qquad S = (1-\lambda\alpha)\mathrm{Id} + \lambda\alpha R, \qquad R = \left(1 - \frac{1}{\lambda\alpha}\right)\mathrm{Id} + \frac{1}{\lambda\alpha}S.$$

Let us prove the equivalence. If T is α-averaged, then S and R are nonexpansive from Lemma 2.3. Since $0 < \lambda\alpha < 1$, we deduce that S is $\lambda\alpha$-averaged (by definition and from the third above equality). Conversely, if S is $\lambda\alpha$-averaged, then R is nonexpansive from Lemma 2.3 and thus T is α-averaged by definition. (2) Trivial from the previous item with $\alpha = \frac{1}{2}$. (3) Since T is β-cocoercive, βT is firmly nonexpansive. Thus $R := 2\beta T - \mathrm{Id}$ is nonexpansive (as well as $-R$). We obtain that $\mathrm{Id} - \lambda T = \mathrm{Id} - \lambda\frac{\mathrm{Id}+R}{2\beta} = (1 - \frac{\lambda}{2\beta})\mathrm{Id} + \frac{\lambda}{2\beta}(-R)$. Since $0 < \frac{\lambda}{2\beta} < 1$, we deduce that $\mathrm{Id} - \lambda T$ is $\frac{\lambda}{2\beta}$-averaged (by definition). ☐

Remark 2.13. The last item of Proposition 2.3 will play a central role in order to derive a convergence result for the *gradient descent method* introduced in Chapter 2 for solving smooth convex minimization problems (see the proof of Proposition 5.1).

2.4 *Operations on averaged operators and common fixed points*

Proposition 2.4 (Convex combination of averaged operators)**.** Let $p \in \mathbb{N}^*$, let $(T_i)_{i=1,\dots,p} : D \to \mathbb{R}^d$ be a finite family of α_i-averaged operators, with $0 < \alpha_i < 1$, for all $i \in \{1,\dots,p\}$, and let $(\lambda_i)_{i=1,\dots,p}$ be such that $\lambda_i \geqslant 0$ for all $i \in \{1,\dots,p\}$ with $\sum_{i=1}^p \lambda_i = 1$. Then the convex combination $\sum_{i=1}^p \lambda_i T_i$ is α-averaged with $\alpha := \max_{i=1}^p \alpha_i$.

Proof. We denote by $T := \sum_{i=1}^p \lambda_i T_i$ and let $x_1, x_2 \in D$. We know from Proposition 2.2 that

$$\|T_i(x_2) - T_i(x_1)\|^2 + \frac{1-\alpha_i}{\alpha_i}\|(\mathrm{Id} - T_i)(x_2) - (\mathrm{Id} - T_i)(x_1)\|^2 \leqslant \|x_2 - x_1\|^2,$$

for all $i \in \{1,\dots,p\}$. From convexity of $\|\cdot\|^2$ and since $\frac{1-\alpha}{\alpha} = \frac{1}{\alpha} - 1 = \min_{i=1}^p \frac{1}{\alpha_i} - 1 = \min_{i=1}^p \frac{1-\alpha_i}{\alpha_i}$, we get that

$$\begin{aligned}
&\|T(x_2) - T(x_1)\|^2 + \frac{1-\alpha}{\alpha}\|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2 \\
&\quad = \left\|\sum_{i=1}^p \lambda_i T_i(x_2) - \sum_{i=1}^p \lambda_i T_i(x_1)\right\|^2 + \frac{1-\alpha}{\alpha}\left\|\sum_{i=1}^p \lambda_i(\mathrm{Id} - T_i)(x_2) - \sum_{i=1}^p \lambda_i(\mathrm{Id} - T_i)(x_1)\right\|^2 \\
&\quad \leqslant \sum_{i=1}^p \lambda_i \left(\|T_i(x_2) - T_i(x_1)\|^2 + \frac{1-\alpha_i}{\alpha_i}\|(\mathrm{Id} - T_i)(x_2) - (\mathrm{Id} - T_i)(x_1)\|^2\right) \\
&\quad \leqslant \sum_{i=1}^p \lambda_i \|x_2 - x_1\|^2 = \|x_2 - x_1\|^2.
\end{aligned}$$

We deduce that T is α-averaged from Proposition 2.2. ☐

Remark 2.14. Let $p \in \mathbb{N}^*$ and $(T_i)_{i=1,\dots,p} : D \to \mathbb{R}^d$ be a finite family of firmly nonexpansive operators and let $(\lambda_i)_{i=1,\dots,p}$ be such that $\lambda_i \geqslant 0$ for all $i \in \{1,\dots,p\}$ with $\sum_{i=1}^p \lambda_i = 1$. Then the convex combination $\sum_{i=1}^p \lambda_i T_i$ is firmly nonexpansive.

Proposition 2.5 (Fixed points of a convex combination of quasinonexpansive operators)**.** Let $p \in \mathbb{N}^*$, let $(T_i)_{i=1,\dots,p} : D \to \mathbb{R}^d$ be a finite family of quasinonexpansive operators such that $\cap_{i=1}^p \mathrm{Fix}(T_i) \neq \emptyset$, and let $(\lambda_i)_{i=1,\dots,p}$ be such that $\lambda_i > 0$ for all $i \in \{1,\dots,p\}$ and $\sum_{i=1}^p \lambda_i = 1$. Then

$$\mathrm{Fix}\left(\sum_{i=1}^p \lambda_i T_i\right) = \bigcap_{i=1}^p \mathrm{Fix}(T_i).$$

$\mathrm{Fix}\left(\frac{1}{2}T_1 + \frac{1}{2}T_2\right) = \mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2)$

$\lambda_i > 0 \quad \sum \lambda_i = 1$

Proof. We denote by $T := \sum_{i=1}^p \lambda_i T_i$. The inclusion $\cap_{i=1}^p \mathrm{Fix}(T_i) \subset \mathrm{Fix}(T)$ is trivial (no need that $\lambda_i > 0$). To prove the reverse inclusion, let us fix some $x_0 \in \cap_{i=1}^p \mathrm{Fix}(T_i) \neq \emptyset$ and let $x \in \mathrm{Fix}(T)$. Recalling that $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$ and since each T_i is quasinonexpansive, we get that

$\mathrm{Fix}(\lambda_1 T_1 + \lambda_2 T_2) = \mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2)$

$\lambda_1, \lambda_2 \geqslant 0$

$\lambda_1 + \lambda_2 = 1$

$$0 = 2\langle T(x) - x, x - x_0\rangle = \sum_{i=1}^p \lambda_i \Big(2\langle T_i(x) - x, x - x_0\rangle\Big)$$
$$= \sum_{i=1}^p \lambda_i \Big(\|T_i(x) - x_0\|^2 - \|T_i(x) - x\|^2 - \|x - x_0\|^2\Big) \leqslant -\sum_{i=1}^p \lambda_i \|T_i(x) - x\|^2.$$

Since each $\lambda_i > 0$, we deduce that $\|T_i(x) - x\|^2 = 0$, and thus $x \in \mathrm{Fix}(T_i)$, for all $i \in \{1,\dots,p\}$. $\square$

Proposition 2.6 (Composition of averaged operators)**.** Let $p \in \mathbb{N}^*$ and $(T_i)_{i=1,\dots,p} : D \to D$ be a finite family of α_i-averaged operators, with $0 < \alpha_i < 1$, for all $i \in \{1,\dots,p\}$. Then $T_1 \circ \dots \circ T_p$ is κ-averaged with $\kappa := \frac{p}{p + \frac{1}{\alpha} - 1}$ where $\alpha := \max_{i=1}^p \alpha_i$.

Proof. We denote by $T := T_1 \circ \dots \circ T_p$ and let $x_1, x_2 \in D$. It holds that

$$\frac{1}{p}\|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2$$
$$= \frac{1}{p}\|(\mathrm{Id} - T_p)(x_2) - (\mathrm{Id} - T_p)(x_1) + (\mathrm{Id} - T_{p-1})(T_p(x_2)) - (\mathrm{Id} - T_{p-1})(T_p(x_1))$$
$$+ \dots + (\mathrm{Id} - T_1)(T_2 \circ \dots \circ T_p)(x_2) - (\mathrm{Id} - T_1)(T_2 \circ \dots \circ T_p)(x_1)\|^2,$$

and thus, from convexity of $\|\cdot\|^2$ and from Proposition 2.2, we get that

$$\frac{1}{p}\|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2$$
$$\leqslant \|(\mathrm{Id} - T_p)(x_2) - (\mathrm{Id} - T_p)(x_1)\|^2 + \|(\mathrm{Id} - T_{p-1})(T_p(x_2)) - (\mathrm{Id} - T_{p-1})(T_p(x_1))\|^2$$
$$+ \dots + \|(\mathrm{Id} - T_1)(T_2 \circ \dots \circ T_p)(x_2) - (\mathrm{Id} - T_1)(T_2 \circ \dots \circ T_p)(x_1)\|^2$$
$$\leqslant \frac{\alpha_p}{1 - \alpha_p}(\|x_2 - x_1\|^2 - \|T_p(x_2) - T_p(x_1)\|^2) + \frac{\alpha_{p-1}}{1 - \alpha_{p-1}}(\|T_p(x_2) - T_p(x_1)\|^2 - \|T_{p-1}(T_p(x_2)) - T_{p-1}(T_p(x_1))\|^2)$$
$$+ \dots + \frac{\alpha_1}{1 - \alpha_1}(\|(T_2 \circ \dots \circ T_p)(x_2) - (T_2 \circ \dots \circ T_p)(x_1)\|^2 - \|T(x_2) - T(x_1)\|^2)$$
$$\leqslant \gamma(\|x_2 - x_1\|^2 - \|T(x_2) - T(x_1)\|^2),$$

where $\gamma := \max_{i=1}^p \frac{\alpha_i}{1 - \alpha_i} > 0$. We deduce from Proposition 2.2 that T is κ-averaged with $1 - \frac{1}{\kappa} = -\frac{1}{p\gamma} < 0$ which gives $\kappa = \frac{p}{p - 1 + 1 + \frac{1}{\gamma}}$ which concludes the proof since $1 + \frac{1}{\gamma} = \frac{1}{\alpha}$ (exercise). $\square$

Lemma 2.4. Let $T_1, T_2 : D \to D$ be two quasinonexpansive operators such that $\mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2) \neq \emptyset$. Then:

(i) If T_1 or T_2 is strictly quasinonexpansive, then $\mathrm{Fix}(T_1 \circ T_2) = \mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2)$ and $T_1 \circ T_2$ is quasinonexpansive.

(ii) If T_1 and T_2 are both strictly quasinonexpansive, then $T_1 \circ T_2$ is strictly quasinonexpansive.

Proof. (i) It is trivial that $\mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2) \subset \mathrm{Fix}(T_1 \circ T_2)$. To prove the reverse inclusion, let us fix some $x_0 \in \mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2) \neq \emptyset$ and let $x \in \mathrm{Fix}(T_1 \circ T_2)$. There are three cases:

- If $T_2(x) \in \mathrm{Fix}(T_1)$, then $T_2(x) = T_1 \circ T_2(x) = x$ and thus $x \in \mathrm{Fix}(T_2)$ and finally $x \in \mathrm{Fix}(T_1)$.
- If $x \in \mathrm{Fix}(T_2)$, then $T_1(x) = T_1 \circ T_2(x) = x$ and thus $x \in \mathrm{Fix}(T_1)$.
- If $T_2(x) \notin \mathrm{Fix}(T_1)$ and $x \notin \mathrm{Fix}(T_2)$, then $\|x - x_0\| = \|T_1 \circ T_2(x) - x_0\| \leqslant \|T_2(x) - x_0\| \leqslant \|x - x_0\|$ by quasinonexpansiveness, with at least one strict inequality by strict quasinonexpansiveness of T_1 or of T_2, which is absurd.

Therefore we have proved that $\mathrm{Fix}(T_1 \circ T_2) = \mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2)$. Now let us prove that $T_1 \circ T_2$ is quasinonexpansive. Let $x_1 \in \mathrm{Fix}(T_1 \circ T_2) = \mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2)$ and $x_2 \in D$. We get that $\|T_1 \circ T_2(x_2) - x_1\| \leqslant \|T_2(x_2) - x_1\| \leqslant \|x_2 - x_1\|$ by quasinonexpansiveness. (ii) Let us prove that $T_1 \circ T_2$ is strictly quasinonexpansive. Let $x_1 \in \mathrm{Fix}(T_1 \circ T_2) = \mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2)$ and $x_2 \in D \backslash \mathrm{Fix}(T_1 \circ T_2)$. There are two cases:

- If $x_2 \notin \mathrm{Fix}(T_2)$, then $\|T_1 \circ T_2(x_2) - x_1\| \leqslant \|T_2(x_2) - x_1\| < \|x_2 - x_1\|$.
- If $x_2 \in \mathrm{Fix}(T_2)$, then $x_2 \notin \mathrm{Fix}(T_1)$ (otherwise $x_2 \in \mathrm{Fix}(T_1) \cap \mathrm{Fix}(T_2) = \mathrm{Fix}(T_1 \circ T_2)$) and then $\|T_1 \circ T_2(x_2) - x_1\| = \|T_1(x_2) - x_1\| < \|x_2 - x_1\|$.

The proof is complete. □

Proposition 2.7 (Fixed points of a composition of strictly quasinonexpansive operators)**.** Let $p \in \mathbb{N}^*$ and $(T_i)_{i=1,\ldots,p} : D \to D$ be a finite family of strictly quasinonexpansive operators such that $\cap_{i=1}^p \mathrm{Fix}(T_i) \neq \emptyset$. Then $T_1 \circ \ldots \circ T_p$ is strictly quasinonexpansive and

$$\mathrm{Fix}(T_1 \circ \ldots \circ T_p) = \bigcap_{i=1}^{p} \mathrm{Fix}(T_i).$$

Proof. We proceed by induction on p. The result is clear for $p = 1$ and $p = 2$ (from Lemma 2.4). Now assume that $p \geqslant 3$ and take a finite family $(T_i)_{i=1,\ldots,p} : D \to D$ of strictly quasinonexpansive operators such that $\cap_{i=1}^p \mathrm{Fix}(T_i) \neq \emptyset$. Since $\emptyset \neq \cap_{i=1}^p \mathrm{Fix}(T_i) \subset \cap_{i=1}^{p-1} \mathrm{Fix}(T_i)$, we obtain by induction hypothesis that $S := T_1 \circ \ldots \circ T_{p-1}$ is strictly quasinonexpansive with $\mathrm{Fix}(S) = \cap_{i=1}^{p-1} \mathrm{Fix}(T_i)$. Since T_p is stricly quasinonexpansive and $\mathrm{Fix}(S) \cap \mathrm{Fix}(T_p) = \cap_{i=1}^p \mathrm{Fix}(T_i) \neq \emptyset$, we obtain from Lemma 2.4 that $T_1 \circ \ldots \circ T_p = S \circ T_p$ is strictly quasinonexpansive with $\mathrm{Fix}(T_1 \circ \ldots \circ T_p) = \mathrm{Fix}(S \circ T_p) = \mathrm{Fix}(S) \cap \mathrm{Fix}(T_p) = \cap_{i=1}^p \mathrm{Fix}(T_i)$. The proof is complete. □

2.5 Fixed-point algorithms for averaged operators

Lemma 2.5. Let $T : D \to D$ be a quasinonexpansive operator with $\mathrm{Fix}(T) \neq \emptyset$. Consider the fixed-point algorithm given by

$$x_0 \in D \quad \text{and} \quad \forall k \in \mathbb{N}, \quad x_{k+1} := T(x_k).$$

Then the sequence $(x_k)_{k \in \mathbb{N}} \subset D$ is Fejér monotone with respect to $\mathrm{Fix}(T)$.

Proof. From quasinonexpansiveness, we have $\|x_{k+1} - x\| = \|T(x_k) - x\| \leqslant \|x_k - x\|$ for all $k \in \mathbb{N}$ and all $x \in \mathrm{Fix}(T)$. $\square$

Lemma 2.6. Let $T : D \to D$ be a continuous and quasinonexpansive operator such that D is closed and $\mathrm{Fix}(T) \neq \emptyset$. Consider the fixed-point algorithm given by

$$x_0 \in D \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := T(x_k).$$

$\lim_{k \infty} \|x_{k+1} - x_k\| = 0$

If $x_k - T(x_k) \to 0_{\mathbb{R}^d}$ when $k \to \infty$, then the sequence $(x_k)_{k\in\mathbb{N}}$ converges to some point in $\mathrm{Fix}(T)$.

Proof. From Lemma 2.5, the sequence $(x_k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$. Furthermore, since T is continuous and D is closed and $x_k - T(x_k) \to 0_{\mathbb{R}^d}$, one can easily see that every cluster point of $(x_k)_{k\in\mathbb{N}}$ belongs to $\mathrm{Fix}(T)$. From Lemma 2.2, the proof is complete. $\square$

Theorem 2.4 (Fixed-point algorithm for averaged operators)**.** Let $T : D \to D$ be an α-averaged operator, with $0 < \alpha < 1$, such that D is closed and $\mathrm{Fix}(T) \neq \emptyset$. Then, the fixed-point algorithm given by

$$x_0 \in D \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := T(x_k),$$

converges to some point in $\mathrm{Fix}(T)$.

Proof. Since T is α-averaged, we know that T is nonexpansive (and thus is continuous and quasinonexpansive). From Lemma 2.6, we only need to prove that $x_k - T(x_k) \to 0_{\mathbb{R}^d}$ when $k \to \infty$. Consider some $x \in \mathrm{Fix}(T) \neq \emptyset$. From Lemma 2.5, we know that $(x_k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$ and thus the sequence $(\|x_k - x\|)_{k\in\mathbb{N}}$ converges. Furthermore, since T is α-averaged, we have

$$\forall k \in \mathbb{N}, \quad \|x_{k+1} - x\|^2 = \|T(x_k) - T(x)\|^2 \leqslant \|x_k - x\|^2 + \left(1 - \frac{1}{\alpha}\right) \|(\mathrm{Id} - T)(x_k) - (\mathrm{Id} - T)(x)\|^2.$$

Since $(\mathrm{Id} - T)(x) = 0_{\mathbb{R}^d}$, we deduce that

$$\forall k \in \mathbb{N}, \quad \left(\frac{1}{\alpha} - 1\right) \|(\mathrm{Id} - T)(x_k)\|^2 \leqslant \|x_k - x\|^2 - \|x_{k+1} - x\|^2,$$

which converges to 0 and thus $x_k - T(x_k) \to 0_{\mathbb{R}^d}$ when $k \to \infty$. The proof is complete. $\square$

Corollary 2.1 (Fixed-point algorithm for compositions of averaged operators)**.** Let $p \in \mathbb{N}^*$ and $(T_i)_{i=1,\dots,p} : D \to D$ be a family of α_i-averaged operators, with $0 < \alpha_i < 1$ for all $i \in \{1, \dots, p\}$, such that D is closed and $\mathrm{Fix}(T) \neq \emptyset$ where $T = T_1 \circ \ldots \circ T_p$. Consider the fixed-point algorithm given by

$$x_0 \in D \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := T(x_k).$$

Then, there exists $y_1 \in \mathrm{Fix}(T_1 \circ \ldots \circ T_p)$, $y_2 \in \mathrm{Fix}(T_2 \circ \ldots \circ T_p \circ T_1)$, ..., $y_p \in \mathrm{Fix}(T_p \circ T_1 \circ \ldots \circ T_{p-1})$ such that

$$T_p(x_k) \to y_p = T_p(y_1),$$

$$T_{p-1} \circ T_p(x_k) \to y_{p-1} = T_{p-1}(y_p),$$

$$\ldots$$

$$T_2 \circ \ldots \circ T_p(x_k) \to y_2 = T_2(y_3),$$

$$T_1 \circ \ldots \circ T_p(x_k) \to y_1 = T_1(y_2),$$

when $k \to \infty$. In particular, we obtain that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to $y_1 \in \mathrm{Fix}(T)$.

Proof. Since each T_i is α_i-averaged, each T_i is nonexpansive and thus so is T (which is thus continuous and quasinonexpansive). Our aim now is to prove that $x_k - T(x_k) \to 0_{\mathbb{R}^d}$ when $k \to \infty$. Consider some $x \in \mathrm{Fix}(T) \neq \emptyset$. From Lemma 2.5, we know that $(x_k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$ and thus the sequence $(\|x_k - x\|)_{k\in\mathbb{N}}$ converges. Furthermore, since T_1 is α_1-averaged, we have

$$\begin{aligned}\forall k \in \mathbb{N}, \quad \|x_{k+1} - x\|^2 = \|T(x_k) - T(x)\|^2 &\leqslant \|(T_2 \circ \ldots \circ T_p)(x_k) - (T_2 \circ \ldots \circ T_p)(x)\|^2 \\ &+ \left(1 - \frac{1}{\alpha_1}\right) \|(\mathrm{Id} - T_1) \circ (T_2 \circ \ldots \circ T_p)(x_k) - (\mathrm{Id} - T_1) \circ (T_2 \circ \ldots \circ T_p)(x)\|^2.\end{aligned}$$

Since T_2 is α_2-averaged, we get that

$$\begin{aligned}\forall k \in \mathbb{N}, \quad \|x_{k+1} - x\|^2 &\leqslant \|(T_3 \circ \ldots \circ T_p)(x_k) - (T_3 \circ \ldots \circ T_p)(x)\|^2 \\ &+ \left(1 - \frac{1}{\alpha_2}\right) \|(\mathrm{Id} - T_2) \circ (T_3 \circ \ldots \circ T_p)(x_k) - (\mathrm{Id} - T_2) \circ (T_3 \circ \ldots \circ T_p)(x)\|^2 \\ &+ \left(1 - \frac{1}{\alpha_1}\right) \|(\mathrm{Id} - T_1) \circ (T_2 \circ \ldots \circ T_p)(x_k) - (\mathrm{Id} - T_1) \circ (T_2 \circ \ldots \circ T_p)(x)\|^2.\end{aligned}$$

By induction, we get that

$$\begin{aligned}\forall k \in \mathbb{N}, \quad \|x_{k+1} - x\|^2 &\leqslant \|x_k - x\|^2 + \left(1 - \frac{1}{\alpha_p}\right) \|(\mathrm{Id} - T_p)(x_k) - (\mathrm{Id} - T_p)(x)\|^2 \\ &\ldots \\ &+ \left(1 - \frac{1}{\alpha_2}\right) \|(\mathrm{Id} - T_2) \circ (T_3 \circ \ldots \circ T_p)(x_k) - (\mathrm{Id} - T_2) \circ (T_3 \circ \ldots \circ T_p)(x)\|^2 \\ &+ \left(1 - \frac{1}{\alpha_1}\right) \|(\mathrm{Id} - T_1) \circ (T_2 \circ \ldots \circ T_p)(x_k) - (\mathrm{Id} - T_1) \circ (T_2 \circ \ldots \circ T_p)(x)\|^2.\end{aligned}$$

Since each $1 - \frac{1}{\alpha_i}$ is negative and since $\|x_k - x\|^2 - \|x_{k+1} - x\|^2$ tends to zero, we deduce that

$$\begin{gathered}(\mathrm{Id} - T_1) \circ (T_2 \circ \ldots \circ T_p)(x_k) - (\mathrm{Id} - T_1) \circ (T_2 \circ \ldots \circ T_p)(x) \to 0_{\mathbb{R}^d}, \\ (\mathrm{Id} - T_2) \circ (T_3 \circ \ldots \circ T_p)(x_k) - (\mathrm{Id} - T_2) \circ (T_3 \circ \ldots \circ T_p)(x) \to 0_{\mathbb{R}^d}, \\ \ldots \\ (\mathrm{Id} - T_{p-1}) \circ T_p(x_k) - (\mathrm{Id} - T_{p-1}) \circ T_p(x) \to 0_{\mathbb{R}^d}, \\ (\mathrm{Id} - T_p)(x_k) - (\mathrm{Id} - T_p)(x) \to 0_{\mathbb{R}^d}.\end{gathered}$$

By developing and adding all these limits (try with $p = 3$ for an example), we only preserve the second and fourth terms of the first line, and the first and third terms of the last line, which gives

$$-T_1 \circ T_2 \circ \ldots \circ T_p(x_k) + T_1 \circ T_2 \circ \ldots \circ T_p(x) + x_k - x \to 0_{\mathbb{R}^d}.$$

Since $T_1 \circ T_2 \circ \ldots \circ T_p = T$ and $T(x) = x$, we get that $x_k - T(x_k) \to 0_{\mathbb{R}^d}$ when $k \to \infty$. From Lemma 2.6, we get that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to some point $y_1 \in \mathrm{Fix}(T) = \mathrm{Fix}(T_1 \circ \ldots \circ T_p)$. Then, defining $y_p := T_p(y_1)$, we easily get that $y_p \in \mathrm{Fix}(T_p \circ T_1 \circ \ldots \circ T_{p-1})$ and we deduce that $T_p(x_k) \to T_p(y_1) = y_p$. Similarly, we define $y_{p-1} := T_{p-1}(y_p), \ldots, y_2 := T_2(y_1)$ and obtain the convergence results. Finally, we obtain that $T_1 \circ \ldots \circ T_p(x_k) \to T_1(y_2) = T_1 \circ \ldots \circ T_p(y_1) = T(y_1) = y_1$. ☐

2.6 *Krasnoselskii-Mann algorithm for averaged operators with* $D = \mathbb{R}^d$

Proposition 2.8 (KM algorithm for averaged operators with $D = \mathbb{R}^d$)**.** Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be an α-averaged operator, with $0 < \alpha < 1$, such that $D = \mathbb{R}^d$ and $\mathrm{Fix}(T) \neq \emptyset$. Let $(\lambda_k)_{k\in\mathbb{N}} \subset [0, \frac{1}{\alpha}]$ be a sequence such that $\sum_{k\in\mathbb{N}} \lambda_k(\frac{1}{\alpha} - \lambda_k) = +\infty$. Then, the KM algorithm given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := \underbrace{\boxed{x_k + \lambda_k(T(x_k) - x_k),}}_{F(x_k)}$$

x_k

$\cancel{x^*} = \cancel{x^*} + \lambda T(x^*) - \lambda x^*$

$T(x^*) = x^*$

converges to some point in $\mathrm{Fix}(T)$.

Proof. Define the nonexpansive operator $R := (1 - \frac{1}{\alpha})\mathrm{Id} + \frac{1}{\alpha}T$. Note that $\mathrm{Fix}(R) = \mathrm{Fix}(T) \neq \emptyset$. Then the sequence $(x_k)_{k\in\mathbb{N}}$ satisfies

$$x_{k+1} = x_k + \mu_k(R(x_k) - x_k),$$

where $\mu_k := \alpha\lambda_k$ for all $k \in \mathbb{N}$. Note that $(\mu_k)_{k\in\mathbb{N}} \subset [0,1]$ satisfies $\sum_{k\in\mathbb{N}} \mu_k(1-\mu_k) = \alpha^2 \sum_{k\in\mathbb{N}} \lambda_k(\frac{1}{\alpha} - \lambda_k) = +\infty$. From Theorem 2.2, we know that $(x_k)_{k\in\mathbb{N}}$ converges to some point in $\mathrm{Fix}(R) = \mathrm{Fix}(T)$. □

Remark 2.15. Consider the framework of Proposition 2.8. Note that the KM algorithm can be seen as a relaxation of the fixed-point algorithm corresponding to the case where $\lambda_k = 1$ for all $k \in \mathbb{N}$ (this case being handling by Proposition 2.8, in contrary to Theorem 2.2 in which T is assumed to be nonexpansive only). However, Proposition 2.8 requires that $D = \mathbb{R}^d$ to guarantee that the KM algorithm is correctly defined.

Remark 2.16. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a firmly nonexpansive operator such that $D = \mathbb{R}^d$ and $\mathrm{Fix}(T) \neq \emptyset$. Let $(\lambda_k)_{k\in\mathbb{N}} \subset [0,2]$ be a sequence such that $\sum_{k\in\mathbb{N}} \lambda_k(2-\lambda_k) = +\infty$. Then the KM algorithm given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := x_k + \lambda_k(T(x_k) - x_k),$$

converges to some point in $\mathrm{Fix}(T)$.

Proposition 2.9 (KM algorithm for convex combinations of compositions of averaged operators with $D = \mathbb{R}^d$)**.** Let $p \in \mathbb{N}^*$ and $(T_i)_{i=1,\ldots,p} : \mathbb{R}^d \to \mathbb{R}^d$ be a finite family of α_i-averaged operators, with $0 < \alpha_i < 1$ and $D_i = \mathbb{R}^d$ for all $i \in \{1,\ldots,p\}$, such that $\cap_{i=1}^p \mathrm{Fix}(T_i) \neq \emptyset$. Let $q \in \mathbb{N}^*$ and, for all $j \in \{1,\ldots,q\}$, let $\mu_j > 0$ and

$$S_j := T_{\ell_j(1)} \circ \ldots \circ T_{\ell_j(p_j)},$$

where $p_j \in \mathbb{N}^*$ and where $\ell_j : \{1,\ldots,p_j\} \to \{1,\ldots,q\}$. Assume that $\sum_{j=1}^q \mu_j = 1$ and $\cup_{j=1}^q \ell_j(\{1,\ldots,p_j\}) = \{1,\ldots,p\}$ (that is, assume that each T_i is used at least one time in the definitions of $S_1,\ldots,S_q$). Now we define

$$\forall j = 1,\ldots,q, \quad \beta_j := \max_{i\in\ell_j(\{1,\ldots,p_j\})} \alpha_i \quad \text{and} \quad \kappa_j := \frac{p_j}{p_j + \frac{1}{\beta_j} - 1},$$

and $\gamma := \max_{j=1}^q \kappa_j$. Finally, let $(\lambda_k)_{k\in\mathbb{N}} \subset [0, \frac{1}{\gamma}]$ be a sequence such that $\sum_{k\in\mathbb{N}} \lambda_k(\frac{1}{\gamma} - \lambda_k) = +\infty$. Then, the KM algorithm given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := x_k + \lambda_k\left(\sum_{j=1}^q \mu_j S_j(x_k) - x_k\right),$$

converges to some point in $\cap_{i=1}^p \mathrm{Fix}(T_i)$.

Proof. Define $T := \sum_{j=1}^q \mu_j S_j$. From Proposition 2.8, we only need to prove that T is γ-averaged and that $\mathrm{Fix}(T) = \cap_{i=1}^p \mathrm{Fix}(T_i)$.

(i) From Proposition 2.6, we know that S_j is κ_j-averaged and, from Proposition 2.7, since $\emptyset \neq \cap_{i=1}^{p}\mathrm{Fix}(T_i) \subset \cap_{i\in I_j}\mathrm{Fix}(T_i)$, that $\mathrm{Fix}(S_j) = \cap_{i\in I_j}\mathrm{Fix}(T_i)$ for all $j \in \{1,\ldots,q\}$.

(ii) From the previous item and Proposition 2.4, we deduce that T is γ-averaged and, from Proposition 2.5, that $\mathrm{Fix}(T) = \cap_{j=1}^{q}\mathrm{Fix}(S_j) = \cap_{j=1}^{q} \cap_{i\in\ell_j(\{1,\ldots,p_j\})} \mathrm{Fix}(T_i) = \cap_{i=1}^{p}\mathrm{Fix}(T_i) \neq \emptyset$ since $\cup_{j=1}^{q}\ell_j(\{1,\ldots,p_j\}) = \{1,\ldots,p\}$.

The proof is complete. $\square$

3 Application to projections on nonempty closed convex sets

The best known examples of firmly nonexpansive operators are given by projection operators onto nonempty closed convex subsets of $\mathbb{R}^d$. Thanks to this nice mathematical property we will construct algorithms allowing to numerically solve *convex feasability problems*, that is, to numerically find a point $x \in \cap_{i=1}^{p} C_i$ for a given finite family $(C_i)_{i=1,\dots,p}$, with $p \in \mathbb{N}^*$, of nonempty closed convex sets.

3.1 Firm nonexpansiveness of projection operators

Definition 3.1 (Normal cone (set-valued map))**.** The *normal cone* $\mathrm{N}_C : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ associated with a convex subset C of $\mathbb{R}^d$ is defined by

$$\forall x \in \mathbb{R}^d, \quad \mathrm{N}_C(x) := \begin{cases} \{y \in \mathbb{R}^d \mid \forall x' \in C,\ \langle y, x' - x\rangle \leqslant 0\} & \text{if} \quad x \in C, \\ \emptyset & \text{if} \quad x \notin C. \end{cases}$$

Note that $\mathrm{Dom}(\mathrm{N}_C) = C$ and that N_C has closed convex cone values.

Lemma 3.1. Let C be a convex subset of $\mathbb{R}^d$. Then N_C is monotone.

Proof. Let (x_1, y_1), $(x_2, y_2) \in \mathrm{Gr}(N_C)$. In particular, we have x_1, $x_2 \in \mathrm{Dom}(\mathrm{N}_C) = C$ and thus

$$\langle y_1, x_2 - x_1\rangle \leqslant 0 \quad \text{and} \quad \langle y_2, x_1 - x_2\rangle \leqslant 0.$$

By adding these two inequalities we get that $\langle y_2 - y_1, x_2 - x_1\rangle \geqslant 0$ and thus N_C is monotone. □

Theorem 3.1. Let C be a nonempty closed convex subset of $\mathbb{R}^d$. Then:

(i) N_C is maximal monotone.

(ii) $J_{\mathrm{N}_C} = \mathrm{proj}_C$ which is single-valued, with $\mathrm{Dom}(\mathrm{proj}_C) = \mathbb{R}^d$, firmly nonexpansive and maximal monotone.

Proof. To use Minty's theorem, let us prove that $\mathrm{Ran}(\mathrm{Id} + \mathrm{N}_C) = \mathbb{R}^d$. Let $y \in \mathbb{R}^d$. We look for $x \in \mathbb{R}^d$ such that $y \in x + \mathrm{N}_C(x)$ which is equivalent to $x \in C$ and $\langle y - x, x' - x\rangle \leqslant 0$ for all $x' \in C$, which is equivalent to $x = \mathrm{proj}_C(y)$. We deduce that $\mathrm{N}_C : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximal monotone and that its resolvent is given by $J_{\mathrm{N}_C} = (\mathrm{Id} + \mathrm{N}_C)^{-1} = \mathrm{proj}_C$ which is thus, from Theorems 1.2 and 2.3, single-valued, with $\mathrm{Dom}(\mathrm{proj}_C) = \mathbb{R}^d$, firmly nonexpansive and maximal monotone. □

3.2 Some projection examples

Remark 3.1. If $C = C_1 \times \ldots \times C_d$, with each C_i being a nonempty closed convex subset of $\mathbb{R}$ (and thus nonempty closed interval of $\mathbb{R}$), then C is a nonempty closed convex subset of $\mathbb{R}^d$ and $\mathrm{proj}_C(x) = \mathrm{proj}_{C_1}(x_1) \times \ldots \times \mathrm{proj}_{C_d}(x_d)$ for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ (exercise).

Definition of the proximal operator:

Let $f \in \Gamma_0(\mathbb{R}^n)$: set of convex, lsc and proper functions.

The proximal operator of f is defined by:

$$\mathrm{prox}_f : \mathbb{R}^n \longrightarrow \mathbb{R}$$
$$x \longmapsto \mathrm{prox}_f(x) = (Id + \partial f)^{-1}(x)$$

Let $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a maximal monotone operator:

The resolvent of A is defined by:

$$J_A = (Id + A)^{-1}$$

$$J_{\lambda A} = (Id + \lambda A)^{-1}, \quad \lambda > 0$$

$$\boxed{\mathrm{prox}_f = J_{\partial f}}$$

$g(y) = \frac{1}{2}\|y-x\|^2$

We have:

$$y = \boxed{\mathrm{prox}_f(x)} \iff y = (Id + \partial f)^{-1}(x)$$
$$\iff x \in y + \partial f(y)$$
$$\iff 0 \in \underbrace{\boxed{y-x}}_{\nabla g(y)?} + \partial f(y)$$
$$\iff 0 \in \partial f(y) + \nabla_y \left(\frac{1}{2}\|y-x\|^2\right)$$
$$\iff 0 \in \partial_y \left(f + \frac{1}{2}\|\cdot - x\|^2\right)(y)$$

Propositio: $f \in \Gamma_0(\mathbb{R}^n)$, $g : \mathbb{R}^n \to \mathbb{R}$ convex and of class $\mathcal{C}^1$

then: $\partial(f+g) = \nabla g + \partial f$

$$\Leftrightarrow \quad y = \operatorname*{argmin}_{z\in\mathbb{R}^n} \left\{ f(z) + \frac{1}{2}\|z-x\|^2 \right\}$$

$$\boxed{\operatorname{prox}_f(x) = \operatorname*{argmin}_{z\in\mathbb{R}^n} \left\{ f(z) + \frac{1}{2}\|z-x\|^2 \right\}}$$

Let $x \in \mathbb{R}^n$ fixed.

The functio $z \longmapsto f(z) + \frac{1}{2}\|z-x\|^2$ is strongly convex functio. Then $\exists!\ \bar{z}(x,f) \in \mathbb{R}^n$ s.t.:

$$\min_{z\in\mathbb{R}^n} f(z) + \frac{1}{2}\|z-x\|^2 = f(\bar{z}) + \frac{1}{2}\|\bar{z}-x\|^2$$

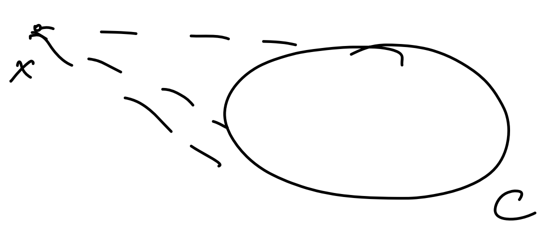
Notation: $\bar{z}(x,f) = \operatorname{prox}_f(x)$ the proximity / proximal operator!

Particular case: Let $C \subset \mathbb{R}^n$ be a closed cvx nonempty set.

Let $f = I_C$ the indicator function $I_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases}$ $\quad i_C,\ \delta_C,\ \psi_C$

$\operatorname{prox}_{I_C} \overset{?}{=}$

Let $x \in \mathbb{R}^n$ $\begin{matrix} \nearrow \\ \searrow \end{matrix}$



$$y = \mathrm{prox}_{I_C}(x) \iff y = \operatorname*{argmin}_{z\in\mathbb{R}^n} I_C(z) + \tfrac{1}{2}\|z-x\|$$

$$\iff y = \operatorname*{argmin}_{z\in C} \tfrac{1}{2}\|z-x\|^2 \; !$$

$$\iff \boxed{y = \mathrm{proj}_C(x)}$$

<u>other proof:</u> $\partial I_C = N_C$

$$\mathrm{prox}_{I_C}(x) = (Id + N_C)^{-1}(x)$$

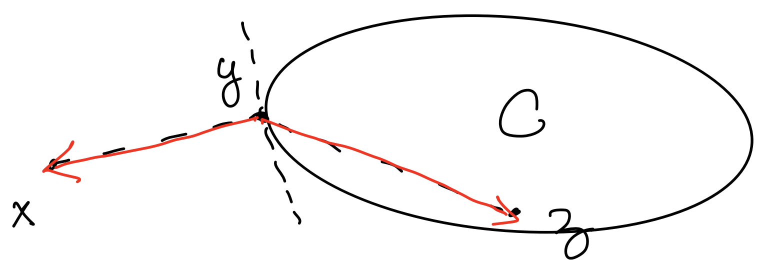
$$y = (Id + N_C)^{-1}(x) \iff x \in y + N_C(y)$$

$$\iff \overbrace{x-y}^{p} \in N_C(y)$$

$$\iff \langle \underbrace{x-y}_{p}, z-y \rangle \leq 0, \ \forall z \in C$$

$$\iff y = \mathrm{proj}_C(x).$$

$$p \in N_C(y) \iff \langle p, z-y \rangle \leq 0, \ \forall z \in C$$



$$y = P_C(x) \iff \langle x-y, z-y \rangle \leq 0, \ \forall z \in C$$

$$\boxed{(Id + N_C)^{-1} = \mathrm{Proj}_C}$$

Some prox-friendly functions

Example: (i) $f(x) = |x|$, $x \in \mathbb{R}$

Compute $\text{prox}_f(x)$?

(ii) $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

$x \longmapsto f(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i|$

$$\text{prox}_{\|\cdot\|_1} \overset{?}{=}$$

(iii) $\text{prox}_{\|\cdot\|_2} \overset{?}{=}$

$f : \mathbb{R}^n \longrightarrow \mathbb{R}$

$x \longmapsto f(x) = \|x\|_2 = \sqrt{\langle x, x \rangle}$

(iv) Prox of quadratic function

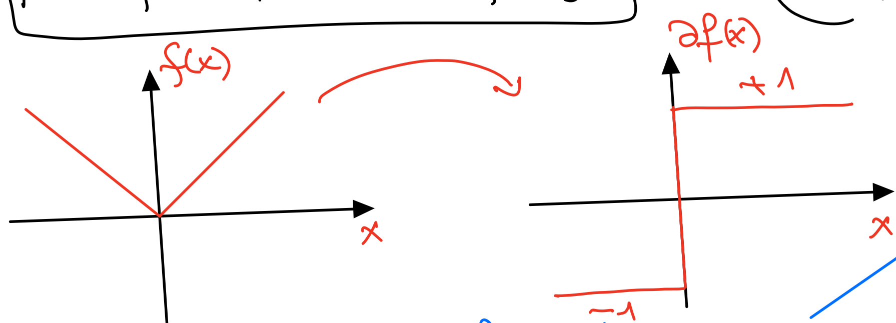
$f(x) = \frac{1}{2} x^T Q x + b^T x + c \qquad Q = Q^T \succ 0$

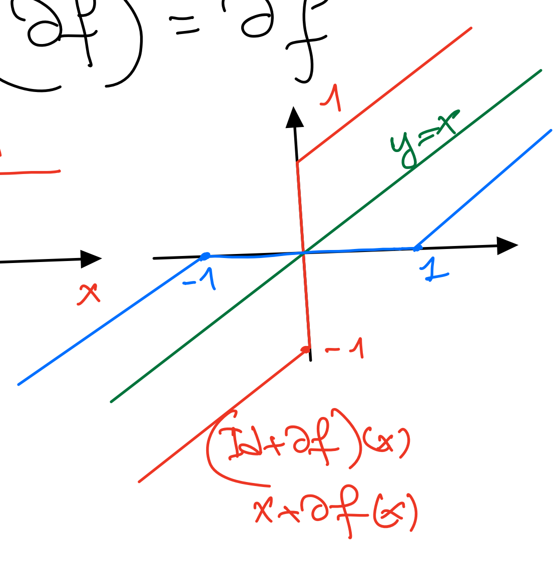
$y = \text{prox}_{|\cdot|}(x) \iff y = (\mathrm{Id} + \partial f)^{-1}(x)$

$\iff y + \partial f(y) \ni x \iff x - y \in \partial f(y)$

$f \in \Gamma_0(\mathbb{R}^n)$ $\iff y \in \partial f^*(x-y) \iff y \in \partial I_{[-1,1]}(x-y)$

$\boxed{p \in \partial f(x) \iff x \in \partial f^*(p)}$ $\qquad (\partial f)^{-1} = \partial f^*$





$$\text{prox}_{|\cdot|}(x) = \begin{cases} 1+x & \text{if } x \leq -1 \\ 0 & \text{if } |x| \leq 1 \\ x-1 & \text{if } x \geq 1 \end{cases}$$

Soft Thresholding Operator:

$$\mathcal{S}(x) = \begin{cases} x+1 & \text{if } x \leq -1 \\ 0 & \text{if } |x| \leq 1 \\ x-1 & \text{if } x \geq 1 \end{cases}$$

$$f(x) = |x| \Leftrightarrow f^* = I_{[-1,1]}$$

$$y \in N_{[-1,1]}(x-y)$$

$\lambda > 0$

$$\operatorname{prox}_{\lambda|\cdot|}(x) = (Id + \lambda\partial f)^{-1} = \operatorname{argmin}\left\{f(z) + \frac{1}{2\lambda}|z-x|^2\right\}$$

$$= \begin{cases} x+\lambda & \text{if } x \leq -\lambda \\ 0 & \text{if } |x| \leq \lambda \\ x-\lambda & \text{if } x \geq \lambda \end{cases} = \mathcal{S}_\lambda(x)$$

$x \in \mathbb{R}$
$\lambda > 0$

(ii) $\|x\|_1 = \sum_{i=1}^{n} |x_i|$

Extension: $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

$$x \longmapsto f(x) = \sum_{i=1}^{n} f_i(x_i) = f_1(x_1) + f_2(x_2) + \dots + f_n(x_n)$$

$f_i : \mathbb{R} \longrightarrow \mathbb{R}$

$z \longmapsto f_i(z)$

$\boxed{f_i \text{ convex}}$

Show that: $\operatorname{prox}_f(x) = \left(\operatorname{prox}_{f_1}(x_1), \operatorname{prox}_{f_2}(x_2), \dots, \operatorname{prox}_{f_n}(x_n)\right)$

Proof:

$$\partial f(x) = \begin{bmatrix} \partial f_1(x_1) \\ \partial f_2(x_2) \\ \vdots \\ \partial f_n(x_n) \end{bmatrix}$$

$\partial f_1(x_1) \subset \mathbb{R}$

$\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$

$x \longmapsto \partial f(x) \subset \mathbb{R}^n$

$$\partial f(x) = \left(\partial f_1(x_1), \partial f_2(x_2), \dots, \partial f_n(x_n)\right)$$

$$(\mathrm{Id} + \partial f)(x) = x + \partial f(x) = \left(x_1 + \partial f_1(x_1), x_2 + \partial f_2(x_2), \dots, x_n + \partial f_n(x_n)\right)$$

$$\underbrace{y \in (\mathrm{Id} + \partial f)(x)}_{\Updownarrow} \iff y_i \in x_i + \partial f_i(x_i) \quad \forall i = 1, \dots, n$$

$$x = \mathrm{prox}_f(y) \qquad \iff x_i = \mathrm{prox}_{f_i}(y_i)$$

$$x = (x_1, x_2, \dots, x_n) = \left(\mathrm{prox}_{f}(y_1), \mathrm{prox}_{f_2}(y_2), \dots, \mathrm{prox}_{f_n}(y_n)\right)$$

$$= \mathrm{prox}_f(y)$$

$$\forall y \in \mathbb{R}^n \qquad \mathrm{prox}_f(y) = \left(\mathrm{prox}_{f_1}(y_1), \mathrm{prox}_{f_2}(y_2), \dots, \mathrm{prox}_{f_n}(y_n)\right)$$

c/c: $\mathrm{prox}_{\|\cdot\|_1}(x) = \left(\mathcal{C}(x_1), \mathcal{C}(x_2), \dots, \mathcal{C}(x_n)\right)$.

LASSO: $\frac{1}{2}\|A\cancel{x} + b\|_2^2 + \lambda\|\cdot\|_1$

$$\mathrm{prox}_{\lambda\|\cdot\|_1}(x) = \left(\mathcal{C}_\lambda(x_1), \mathcal{C}_\lambda(x_2), \dots, \mathcal{C}_\lambda(x_n)\right). \qquad \underline{\underline{\lambda > 0}}$$

(iii) $$\mathrm{prox}_{\|\cdot\|_2}(x) = \begin{cases} \left(1 - \frac{1}{\|x\|_2}\right)x & \text{if } \|x\|_2 \geq 1 \\ 0 & \text{if } \|x\|_2 \leq 1 \end{cases}$$

$f = \|\cdot\|_2$

$$\operatorname{prox}_{\lambda\|\cdot\|_2}(x) = \begin{cases} \left(1 - \frac{\lambda}{\|x\|_2}\right)x & \text{if } \|x\|_2 \geq \lambda \\ 0 & \text{if } \|x\|_2 \leq \lambda \end{cases}$$

(iv) $f(x) = \frac{1}{2}x^T Q x + b^T x + c \qquad Q = Q^T \succ 0$

$$\operatorname{prox}_f(x) = (I_n + Q)^{-1}(x - b)$$

$$\operatorname{prox}_{\lambda f}(x) = (I_n + \lambda Q)^{-1}(x - \lambda b), \quad \underline{\underline{\lambda > 0}}$$

$$\partial f(x) = \begin{cases} \frac{x}{\|x\|_2} & \text{if } x \neq 0 \\ \overline{B}(0,1) & \text{if } x = 0 \end{cases}$$

$$y = \mathrm{prox}_f(x) \Leftrightarrow y + \partial f(y) \ni x$$
$$\Leftrightarrow x - y \in \partial f(y) \quad (*)$$

- if $(*)$ $\boxed{y = 0 \Leftrightarrow x - 0 \in \overline{B} \Leftrightarrow \|x\|_2 \leq 1}$
- if $y \neq 0$, $(*) \Leftrightarrow x - y = \frac{y}{\|y\|}$

$$\Leftrightarrow x = y\left(1 + \frac{1}{\|y\|}\right)$$

$\boxed{y \neq 0 \Leftrightarrow \|x\| > 1}$

$$\Leftrightarrow y = \frac{1}{1 + \frac{1}{\|y\|}} \cdot x$$

$$x = y\left(1 + \frac{1}{\|y\|}\right) \Rightarrow \|x\| = \|y\| \cdot \left(1 + \frac{1}{\|y\|}\right)$$
$$= \|y\| + 1$$
$$\Leftrightarrow \boxed{\|y\| = \|x\| - 1}$$

$$y = \frac{1}{1 + \frac{1}{\|x\| - 1}} \cdot x = \frac{\|x\| - 1}{\|x\| - 1 + 1} \cancel{x} = \frac{\|x\| - 1}{\|x\|} \cdot x$$

$$\boxed{y = \left(1 - \frac{1}{\|x\|}\right) \cdot x}$$

Proposition 3.1 (Some projection examples)**.** In the table below, we consider some $\ell = (\ell_i)_{i=1,\dots,d} \in (\mathbb{R} \cup \{-\infty\})^d$ and $u = (u_i)_{i=1,\dots,d} \in (\mathbb{R} \cup \{+\infty\})^d$ with $\ell_i \leqslant u_i$ for all $i \in \{1, \dots, d\}$. We also consider a matrix $M \in \mathbb{R}^{n \times d}$ with full range and $b \in \mathbb{R}^n$. Take also $a \in \mathbb{R}^d$, $\varepsilon > 0$, $h \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\}$ and $\alpha \in \mathbb{R}$.

Item	Comment	C	$\mathrm{proj}_C(x)$
(i)	Nonnegative orthant	$\mathbb{R}^d_+$	$(x_i^+)_{i=1,\dots,d}$
(ii)	Box	Box $[\ell, u] := [\ell_1, u_1] \times \dots \times [\ell_d, u_d]$	$(\min\{\max\{x_i, \ell_i\}, u_i\})_{i=1,\dots,d}$
(iii)	Affine set	$\{x \in \mathbb{R}^d \mid Mx = b\}$	$x - M^\top (MM^\top)^{-1}(Mx - b)$
(iv)	Ball (Euclidean norm)	$\overline{\mathrm{B}}(a, \varepsilon)$	$a + \dfrac{\varepsilon}{\max\{\lVert x - a \rVert, \varepsilon\}}(x - a)$
(v)	Half-space	$[\langle h, \cdot \rangle \leqslant \alpha]$	$x - \dfrac{(\langle h, x \rangle - \alpha)^+}{\lVert h \rVert^2} h$

Proof. (i)(ii) Trivial from Remark 3.1. Note that (i) is a particular case of (ii).

(iii) First, note that $MM^\top$ is invertible since M has full range. Indeed, if $MM^\top y = 0_{\mathbb{R}^n}$ for some $y \in \mathbb{R}^n$, then $\lVert M^\top y \rVert^2 = 0$ and thus $y \in \ker(M^\top) = \mathrm{im}(M)^\perp = \{0_{\mathbb{R}^n}\}$.[1] Now let us consider some $x \in \mathbb{R}^d$ and define $y := x - M^\top(MM^\top)^{-1}(Mx - b)$. Let us prove that $y = \mathrm{proj}_C(x)$. Firstly, note that $y \in C$ (since $My = b$). Since C is an affine set (with $C_{\mathrm{dir}} = \ker(M)$), we only need to prove that $x - y \in C_{\mathrm{dir}}^\perp = \ker(M)^\perp = \mathrm{im}(M^\top)$ which is clear.

(iv) Trivial. One has just to separate the two cases $x \in C$ and $x \notin C$. Make a drawing.

(v) One has to separate the two cases $x \in C$ (trivial) and $x \notin C$. For the second case, one has just to see that $\mathrm{proj}_C(x)$ corresponds to the projection of x onto the affine set $[\langle h, \cdot \rangle = \alpha]$ and use the item (iii). ☐

3.3 Algorithms for solving convex feasability problems

Proposition 3.2 (Parallel projection algorithm)**.** Let $p \in \mathbb{N}^*$ and $(C_i)_{i=1,\dots,p}$ be a finite family of nonempty closed convex sets such that $C := \cap_{i=1}^p C_i \neq \emptyset$. Consider $\mu_i > 0$ for all $i \in \{1, \dots, p\}$ such that $\sum_{i=1}^p \mu_i = 1$. Let $(\lambda_k)_{k \in \mathbb{N}} \subset [0, 2]$ be a sequence such that $\sum_{k \in \mathbb{N}} \lambda_k (2 - \lambda_k) = +\infty$. Then, the *parallel projection algorithm* given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := x_k + \lambda_k \left(\sum_{i=1}^p \mu_j \mathrm{proj}_{C_j}(x_k) - x_k \right),$$

converges to some point in C.

Proof. Since proj_{C_i} is firmly nonexpansive for all $i \in \{1, \dots, p\}$, note that $T := \sum_{i=1}^p \mu_i \mathrm{proj}_{C_i}$ is firmly nonexpansive from Proposition 2.4. From Proposition 2.5, since $\cap_{i=1}^p \mathrm{Fix}(\mathrm{proj}_{C_i}) = \cap_{i=1}^p C_i = C \neq \emptyset$, we know that $\mathrm{Fix}(T) = \cap_{i=1}^p \mathrm{Fix}(\mathrm{proj}_{C_i}) = \cap_{i=1}^p C_i = C \neq \emptyset$. Thus one has just to apply Proposition 2.8 with $\alpha = \frac{1}{2}$. ☐

[1] The fact that $\ker(M^\top) = \mathrm{im}(M)^\perp$ is simple to prove. We deduce that $\ker(M) = \mathrm{im}(M^\top)^\perp$. Passing to the orthogonal, we get that $\ker(M)^\perp = \mathrm{im}(M^\top)$.

Proposition 3.3 (String-averaged relaxed projection algorithm)**.** Let $p \in \mathbb{N}^*$ and $(C_i)_{i=1,\dots,p}$ be a finite family of nonempty closed convex sets such that $C := \cap_{i=1}^p C_i \neq \emptyset$. Consider

$$T_i := (1-\gamma_i)\mathrm{Id} + \gamma_i \mathrm{proj}_{C_i},$$

with $0 < \gamma_i < 2$ for all $i \in \{1,\dots,p\}$. Let $q \in \mathbb{N}^*$ and, for all $j = 1,\dots,q$, let $\mu_j > 0$ and

$$S_j := T_{\ell_j(1)} \circ \dots \circ T_{\ell_j(p_j)},$$

where $p_j \in \mathbb{N}^*$ and where $\ell_j : \{1,\dots,p_j\} \to \{1,\dots,p\}$. Assume that $\sum_{j=1}^q \mu_j = 1$ and $\cup_{j=1}^q \ell_j(\{1,\dots,p_j\}) = \{1,\dots,p\}$ (that is, assume that each T_i is used at least one time in the definitions of $S_1,\dots,S_q$). Then, the *string-averaged relaxed projection algorithm* given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := \sum_{j=1}^q \mu_j S_j(x_k),$$

converges to some point in C.

Proof. Since proj_{C_i} is firmly nonexpansive, we deduce from Proposition 2.3 that T_i is $\alpha_i := \frac{\gamma_i}{2}$-averaged for all $i \in \{1,\dots,p\}$. Furthermore, note that $\mathrm{Fix}(T_i) = C_i$ for all $i \in \{1,\dots,p\}$ and thus $\cap_{i=1}^p \mathrm{Fix}(T_i) = C \neq \emptyset$. Thus one has just to apply Proposition 2.9 with $\lambda_k := 1$ for all $k \in \mathbb{N}$. $\square$

Remark 3.2. As an application of both above algorithms, we have the following convergence result. Let $p \in \mathbb{N}^*$ and $(C_i)_{i=1,\dots,p}$ be a finite family of nonempty closed convex sets such that $C := \cap_{i=1}^p C_i \neq \emptyset$. Then, the algorithm

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := \frac{1}{p}\sum_{i=1}^p \mathrm{proj}_{C_i}(x_k),$$

converges to some point in C.

Proposition 3.4 (POCS algorithm)**.** Let $p \in \mathbb{N}^*$ and $(C_i)_{i=1,\dots,p}$ be a finite family of nonempty closed convex sets such that $\mathrm{Fix}(\mathrm{proj}_{C_1} \circ \dots \circ \mathrm{proj}_{C_p}) \neq \emptyset$. Consider the *Projection Onto Convex Sets (POCS) algorithm* given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} := \mathrm{proj}_{C_1} \circ \dots \circ \mathrm{proj}_{C_p}(x_k).$$

Then there exists $y_1 \in \mathrm{Fix}(\mathrm{proj}_{C_1} \circ \dots \circ \mathrm{proj}_{C_p})$, $y_2 \in \mathrm{Fix}(\mathrm{proj}_{C_2} \circ \dots \circ \mathrm{proj}_{C_p} \circ \mathrm{proj}_{C_1})$, …, $y_p \in \mathrm{Fix}(\mathrm{proj}_{C_p} \circ \mathrm{proj}_{C_1} \dots \circ \mathrm{proj}_{C_{p-1}})$ such that

$$\mathrm{proj}_{C_p}(x_k) \to y_p = \mathrm{proj}_{C_p}(y_1),$$
$$\mathrm{proj}_{C_{p-1}} \circ \mathrm{proj}_{C_p}(x_k) \to y_{p-1} = \mathrm{proj}_{C_{p-1}}(y_p),$$
$$\dots$$
$$\mathrm{proj}_{C_2} \circ \dots \circ \mathrm{proj}_{C_p}(x_k) \to y_2 = \mathrm{proj}_{C_2}(y_3),$$
$$\mathrm{proj}_{C_1} \circ \dots \circ \mathrm{proj}_{C_p}(x_k) \to y_1 = \mathrm{proj}_{C_1}(y_2).$$

In particular, we have:

(i) The sequence $(x_k)_{k\in\mathbb{N}}$ converges to y_1.

(ii) $(y_1,\dots,y_p) \in C_1 \times \dots \times C_p$.

(iii) If $C := \cap_{i=1}^p C_i \neq \emptyset$, then each $y_i \in C$.

Proof. The main result and the first item are a direct application of Corollary 2.1. The item (ii) is trivial. The point (iii) comes from the fact that each proj_{C_i} is firmly nonexpansive, and thus strictly quasinonexpansive. Therefore, from Proposition 2.7, if $\cap_{i=1}^{p}\mathrm{Fix}(\mathrm{proj}_{C_i}) = \cap_{i=1}^{p} C_i \neq \emptyset$, then $y_1 \in \mathrm{Fix}(\mathrm{proj}_{C_1} \circ \ldots \circ \mathrm{proj}_{C_p}) = \cap_{i=1}^{p}\mathrm{Fix}(\mathrm{proj}_{C_i}) = C$. A similar reasoning allows to conclude that each $y_i \in C$. $\square$

Remark 3.3. Consider the framework of Proposition 3.4. In the case where $C := \cap_{i=1}^{p} C_i \neq \emptyset$, note that the POCS algorithm is a particular case of the string-averaged relaxed projection algorithm (taking $\gamma_i = 1$ for all $i \in \{1, \ldots, p\}$, $q = 1$, $p_1 = p$, $S_1 = \mathrm{proj}_{C_1} \circ \ldots \circ \mathrm{proj}_{C_p}$ and $\mu_1 = 1$).

Chapter 2: Proximal operator and proximal algorithms

In this chapter, we start by introducing the well-known notion of *proximal operator* (also called *proximity operator* in the literature) introduced by Jean Jacques Moreau in 1965. This operator will be central in our work in order to set numerical algorithms dedicated to solve (nonsmooth) convex optimization problems. In this chapter, we will discuss several numerical algorithms, among the most known (such as the *forward-backward algorithm* or the *Douglas-Rachford algorithm*), but also others less known. It should be noted that the literature is full of many sophisticated numerical algorithms and it is not possible to make a state of the art on this field since it is very vast and in continuous evolution. Furthermore, it should be noted that the performance of the different algorithms depends strongly on the optimization problems studied.

In contrary to Chapter 1, for which exercises have been listed at the end of the chapter, here the exercises of Chapter 2 have been directly incorporated in the text.

4 Basic definitions and results

4.1 The proximal operator

Definition 4.1 (Argmin, subdifferential and proximal map)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function.

(i) The *set of all global minimizers* of f is defined by

$$\operatorname{Argmin}(f) := \{x \in \operatorname{dom}(f) \mid \forall x' \in \mathbb{R}^d,\ f(x) \leqslant f(x')\}.$$

(ii) The *subdifferential (set-valued) map* $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ of f is defined by

$$\forall x \in \mathbb{R}^d, \quad \partial f(x) := \begin{cases} \{y \in \mathbb{R}^d \mid \forall x' \in \mathbb{R}^d,\ \langle y, x' - x\rangle \leqslant f(x') - f(x)\} & \text{if } \ x \in \operatorname{dom}(f), \\ \emptyset & \text{if } \ x \notin \operatorname{dom}(f). \end{cases}$$

(iii) The *proximal (set-valued) map* $\operatorname{prox}_f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ of f is defined by

$$\operatorname{prox}_f := J_{\partial f} = (\operatorname{Id} + \partial f)^{-1}.$$

Note that $\operatorname{Ran}(\operatorname{prox}_f) = \operatorname{Dom}(\partial f) \subset \operatorname{dom}(f)$.

Remark 4.1. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function. Note that

$$\operatorname{Argmin}(f) = \operatorname{Zer}(\partial f) = \operatorname{Fix}(\operatorname{prox}_f).$$

Furthermore, note that $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is monotone (exercise), which implies that $\operatorname{prox}_f : \operatorname{Dom}(\operatorname{prox}_f) \to \mathbb{R}^d$ is singled-valued (and thus an operator), nonexpansive and monotone (see Proposition 1.3).

Lemma 4.1. If $f \in \Gamma_0(\mathbb{R}^d)$ (that is, if $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function), then

$$\operatorname{Argmin}\left(f + \frac{1}{2}\|\cdot - x\|^2\right),$$

$\leftarrow \min_y \left(f(y) + \frac{1}{2}\|y - x\|^2\right)$

is reduced to a singleton for any $x \in \mathbb{R}^d$.

Proof. Let $x \in \mathbb{R}^d$ and let us denote by $f_x := f + \frac{1}{2}\|\cdot - x\|^2$. Note that $f_x \in \Gamma_0(\mathbb{R}^d)$ and is strictly convex (and thus the uniqueness of a minimizer of f_x is trivial). Now take $(y_k)_{k\in\mathbb{N}}$ being a minimizing sequence of f_x. Since $f \in \Gamma_0(\mathbb{R}^d)$, we know that f admits an affine minorant (since f is convex and proper, see Theorem D.1), implying that there exists $(a, b) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$\forall y \in \mathbb{R}^d, \quad f_x(y) = f(y) + \frac{1}{2}\|y - x\|^2 \geqslant \frac{1}{2}\|y - x\|^2 + \langle a, y\rangle + b \geqslant \frac{1}{2}\|y\|^2 - \|a - x\|\|y\| + \frac{1}{2}\|x\|^2 + b.$$

We deduce that the sequence $(y_k)_{k\in\mathbb{N}}$ is bounded and thus, up to a subsequence, converges to some $y^* \in \mathbb{R}^d$. Since f_x is closed, we obtain that

$$\inf_{y\in\mathbb{R}^d} f_x(y) = \lim_{k\to+\infty} f_x(y_k) = \liminf_{k\to+\infty} f_x(y_k) \geqslant f_x(y^*),$$

which concludes the proof. $\square$

Theorem 4.1 (Proximal operator)**.** If $f \in \Gamma_0(\mathbb{R}^d)$, then:

(i) $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximal monotone.

(ii) $\operatorname{prox}_f : \mathbb{R}^d \to \mathbb{R}^d$ is single-valued, with $\operatorname{Dom}(\operatorname{prox}_f) = \mathbb{R}^d$, nonexpansive and maximal monotone.

(iii) $\operatorname{prox}_f : \mathbb{R}^d \to \mathbb{R}^d$ is firmly nonexpansive.

In that context, $\operatorname{prox}_f : \mathbb{R}^d \to \mathbb{R}^d$ is called the *proximal operator* of f.

Proof. From Theorems 1.2 and 2.3, we only need to prove the first item. To this aim, from Minty's theorem, we only need to prove that $\operatorname{Ran}(\operatorname{Id} + \partial f) = \mathbb{R}^d$. Let $x \in \mathbb{R}^d$ and denote by $f_x := f + \frac{1}{2}\| \cdot - x\|^2$ as in the proof of Lemma 4.1. Let us denote by y the unique minimizer of f_x given by Lemma 4.1. From basic subdifferential calculus rules (see Theorem E.5), we know that

$$0_{\mathbb{R}^d} \in \partial f_x(y) = \partial \left(f + \frac{1}{2}\| \cdot - x\|^2 \right)(y) = \partial f(y) + \nabla \left(\frac{1}{2}\| \cdot - x\|^2 \right)(x) = \partial f(y) + y - x,$$

and thus $\boxed{x \in (\operatorname{Id} + \partial f)(y)}$ which concludes the proof. $y = (\operatorname{Id} + \partial f)^{-1}(x) = \operatorname{prox}_f(x)$. □

Proposition 4.1 (Basic properties)**.** Let $f \in \Gamma_0(\mathbb{R}^d)$. Then, the following properties are satisfied:

(i) It holds that

$$\forall x \in \mathbb{R}^d, \quad \operatorname{prox}_f(x) = \operatorname{Argmin}\left(f + \frac{1}{2}\| \cdot - x\|^2 \right).$$

(ii) For any $x \in \mathbb{R}^d$, the element $y = \operatorname{prox}_f(x)$ is characterized by the property

$$\forall x' \in \mathbb{R}^d, \quad \langle x - y, x' - y \rangle \leqslant f(x') - f(y).$$

(iii) It holds that $(\partial f)^{-1} = \partial f^*$ and the *Moreau's decomposition formula*

$$\forall x \in \mathbb{R}^d, \quad x = \operatorname{prox}_f(x) + \operatorname{prox}_{f^*}(x),$$

holds true.

(iv) It holds that $\operatorname{prox}_{f^*}(x) \in \partial f(\operatorname{prox}_f(x))$ for all $x \in \mathbb{R}^d$.

(v) If $f = \iota_C$ where C is a nonempty closed convex subset of $\mathbb{R}^d$, then $\partial \iota_C = \mathrm{N}_C$ and $\operatorname{prox}_{\iota_C} = \operatorname{proj}_C$. Hence, the proximal operator can be seen as a generalization of the projection operator.

Proof. (i) Let $x \in \mathbb{R}^d$ and, as in the proof of Theorem 4.1, let us denote by $f_x := f + \frac{1}{2}\| \cdot - x\|^2$. Now, considering $y := \operatorname{prox}_f(x) = (\operatorname{Id} + \partial f)^{-1}(x)$, we get that $0_{\mathbb{R}^d} \in \partial f(y) + y - x = \partial f_x(y)$ and thus y is the unique minimizer of f_x (see Lemma 4.1). (ii) Let $x \in \mathbb{R}^d$. We have $y = \operatorname{prox}_f(x)$ if and only if $x - y \in \partial f(y)$ that is exactly $\langle x - y, x' - y \rangle \leqslant f(x') - f(y)$ for all $x' \in \mathbb{R}^d$. (iii) Since $f \in \Gamma_0(\mathbb{R}^d)$, we have $f^* \in \Gamma_0(\mathbb{R}^d)$ (see Remark F.1). Furthermore we have $(x, \ell) \in \operatorname{Gr}(\partial f)$ if and only if $x \in \operatorname{dom}(f)$ and $\ell \in \partial f(x)$ if and only if (see Proposition F.2) $\ell \in \operatorname{dom}(f^*)$ and $x \in \partial f^*(\ell)$ if and only if $(\ell, x) \in \operatorname{Gr}(\partial f^*)$. We deduce that $(\partial f)^{-1} = \partial f^*$. Now let us prove the Moreau's decomposition. Take $x \in \mathbb{R}^d$ and $y := \operatorname{prox}_f(y)$. Then $x - y \in \partial f(y)$, then $y \in \partial f^*(x-y)$, then $x \in (\operatorname{Id} + \partial f^*)(x-y)$ that is $x - y = \operatorname{prox}_{f^*}(x)$ which concludes the proof. (iv) Let $x \in \mathbb{R}^d$ and define $y := \operatorname{prox}_f(x)$. We have $\operatorname{prox}_{f^*}(x) = x - y \in \partial f(y)$. (v) Recall that $\partial \iota_C = \mathrm{N}_C$ (see Proposition E.2). Since C is nonempty closed convex, we know that N_C is maximal monotone and that $\operatorname{prox}_{\iota_C} = (\operatorname{Id} + \partial \iota_C)^{-1} = (\operatorname{Id} + \mathrm{N}_C)^{-1} = J_{\mathrm{N}_C} = \operatorname{proj}_C$ (see Theorem 3.1). □

Remark 4.2. Here we discuss some properties and simple examples in the unidimensional case $d = 1$.

$$\mathrm{prox}_{\lambda f}(x) = (\mathrm{Id} + \lambda \partial f)^{-1}(x) = y$$

$$\Updownarrow$$

$$x \in y + \lambda \partial f(y) \iff 0 \in y - x + \lambda \partial f(y), \ \lambda > 0$$

$$\iff 0 \in \nabla\left(\tfrac{1}{2}\|y-x\|^2\right) + \lambda \partial f(y)$$

$$\iff 0 \in \nabla\left(\tfrac{1}{2\lambda}\|y-x\|^2\right) + \partial f(y)$$

$$\iff 0 \in \partial\left(f(y) + \tfrac{1}{2\lambda}\|y-x\|^2\right)$$

$$\iff y = \operatorname{argmin}\left\{f(y) + \tfrac{1}{2\lambda}\|y-x\|^2\right\}$$

$$\underline{\underline{\lambda > 0}}$$

$$\lambda \partial f(y) \overset{?}{=} \partial f(\lambda y)$$

$$p \in \partial f(\lambda y) \iff \langle p, z - \lambda y\rangle \leq f(z) - f(\lambda y), \ \forall z \in \mathbb{R}^d$$

$$q \in \partial f(y) \iff \langle q, z' - y\rangle \leq f(z') - f(y) \ \forall z' \in \mathbb{R}^d$$

$$q \in \lambda \partial f(y) \iff \tfrac{q}{\lambda} \in \partial f(y)$$

$$\langle \tfrac{1}{\lambda} p, \tfrac{1}{\lambda} z - y\rangle \leq \tfrac{1}{\lambda} f(z) - \tfrac{1}{\lambda} f(\lambda y) \quad \forall z$$

$$\boxed{\mathrm{prox}_{\lambda f}(x) = \operatorname{argmin}\left\{f + \tfrac{1}{2\lambda}\|\cdot - x\|^2\right\}}$$

$$= (\mathrm{Id} + \lambda \partial f)^{-1}(x)$$

$$f'(x) = x$$

$$\boxed{f' = \mathrm{Id}}$$

$$\mathrm{prox}_f(x) = (\mathrm{Id} + f')^{-1}(x) = y$$

$$\iff y + y = x$$

$$\iff 2y = x \iff y = \tfrac{x}{2}$$

(i) If $f \in \Gamma_0(\mathbb{R})$, then $\mathrm{prox}_f : \mathbb{R} \to \mathbb{R}$ is nondecreasing (since monotone).

(ii) Consider the basic quadratic function $f \in \Gamma_0(\mathbb{R})$ defined by $f(x) := \frac{x^2}{2}$ for all $x \in \mathbb{R}$. Then, one can easily prove (exercise) that

$$\forall x \in \mathbb{R}, \quad \mathrm{prox}_f(x) = \frac{x}{2}.$$

(iii) One of the most known unidimensional example of proximal operator is given, for any $\lambda > 0$, by

$$\forall x \in \mathbb{R}, \quad \mathrm{prox}_{\lambda|\cdot|}(x) = \mathrm{sign}(x)(|x| - \lambda)^+ = \begin{cases} x + \lambda & \text{if} \quad x \leqslant -\lambda, \\ 0 & \text{if} \quad x \in [-\lambda, \lambda], \\ x - \lambda & \text{if} \quad x \geqslant \lambda. \end{cases}$$

To prove this (exercise), one has just to see that $\lambda|\cdot| = \sigma_{[-\lambda,\lambda]} = (\iota_{[-\lambda,\lambda]})^*$ and thus $\mathrm{prox}_{\lambda|\cdot|} = \mathrm{Id} - \mathrm{proj}_{[-\lambda,\lambda]}$. One can also make a simple drawing by remembering that $\partial f(x) = [f'_-(x), f'_+(x)]$ for all $x \in \mathbb{R}$ (see Proposition E.4).

(iv) In the unidimensional context, a function appears frequently in proximal calculus. For any $\omega_1 < \omega_2$, it is denoted by $\mathrm{soft}_{[\omega_1,\omega_2]} : \mathbb{R} \to \mathbb{R}$ and is defined by

$$\forall x \in \mathbb{R}, \quad \mathrm{soft}_{[\omega_1,\omega_2]}(x) := \begin{cases} x - \omega_1 & \text{if} \quad x \leqslant \omega_1, \\ 0 & \text{if} \quad x \in [\omega_1, \omega_2], \\ x - \omega_2 & \text{if} \quad x \geqslant \omega_2. \end{cases}$$

For example, for any $\lambda > 0$, it holds that

$$\forall x \in \mathbb{R}, \quad \mathrm{prox}_{\lambda|\cdot|} = \mathrm{soft}_{[-\lambda,\lambda]}.$$

Remark 4.3. Consider the quadratic function $f \in \Gamma_0(\mathbb{R}^d)$ defined by

$$\forall x \in \mathbb{R}^d, \quad f(x) := \frac{1}{2}\langle Mx, x\rangle + \langle a, x\rangle + b,$$

where $M \in \mathbb{R}^{d\times d}$ is a symmetric positive semidefinite matrix, $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$. In that context, one can easily prove (exercise) that

$$\forall x \in \mathbb{R}^d, \quad \mathrm{prox}_f(x) = (I_d + M)^{-1}(x - a).$$

Note that this formula generalizes the item (ii) of Remark 4.2.

4.2 The Moreau's envelope

Definition 4.2 (Moreau's envelope)**.** The *Moreau's envelope* $\mathcal{M}_f : \mathbb{R}^d \to \mathbb{R}$ associated with a function $f \in \Gamma_0(\mathbb{R}^d)$ is defined by

$$\mathcal{M}_f(x) := \min\left(f + \frac{1}{2}\|\cdot - x\|^2\right) = f(\mathrm{prox}_f(x)) + \frac{1}{2}\|\mathrm{prox}_f(x) - x\|^2 = f(\mathrm{prox}_f(x)) + \frac{1}{2}\|\mathrm{prox}_{f^*}(x)\|^2,$$

for all $x \in \mathbb{R}^d$.

Theorem 4.2. The Moreau's envelope $\mathcal{M}_f : \mathbb{R}^d \to \mathbb{R}$ associated with a function $f \in \Gamma_0(\mathbb{R}^d)$ is a differentiable convex function with

$$\nabla\mathcal{M}_f = \mathrm{prox}_{f^*}.$$

$$\min_{y\in\mathbb{R}^n} g(y) = g(\hat{y}) \qquad \hat{y} = \operatorname{argmin} g$$

$$\min_{y\in\mathbb{R}^n} \underbrace{\left(f(y) + \frac{1}{2}\|y-x\|^2\right)}_{g_x(y)} = \boxed{g_x\left(\operatorname{prox}_f(x)\right)} = M_f(x)$$

$$\boxed{\operatorname{prox}_f(x)} = \operatorname*{argmin}_{y\in\mathbb{R}^n} g_x(y)$$

$$= \boxed{f\left(\operatorname{prox}_f(x)\right) + \frac{1}{2}\|\operatorname{prox}_f(x) - x\|^2 = M_f(x)}$$

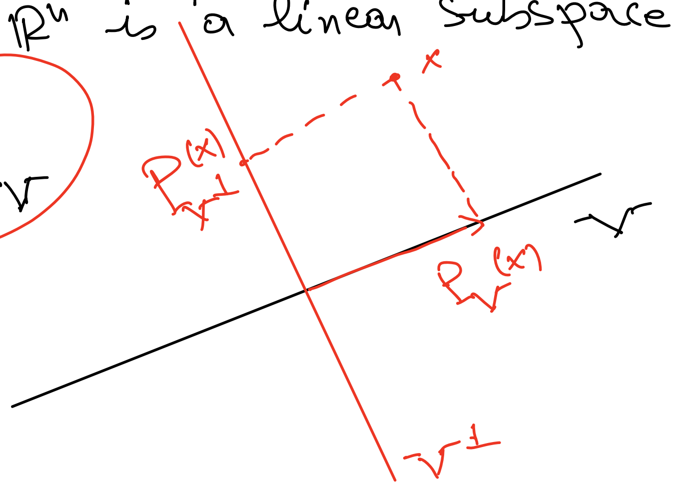
$$\boxed{\begin{array}{l} \forall f \in \Gamma_0(\mathbb{R}^n) \\ \qquad \operatorname{prox}_f + \operatorname{prox}_{f^*} = \mathrm{Id} \end{array}}$$

Moreau's decomposition.

Exples: (1) $f = I_{\mathcal{V}}$ $\quad \mathcal{V} \subset \mathbb{R}^n$ is a linear subspace

$$\operatorname{prox}_f = \operatorname{prox}_{I_{\mathcal{V}}} = \underbrace{\operatorname{proj}_{\mathcal{V}}}_{P_{\mathcal{V}}}$$

$$f^* = I_{\mathcal{V}^\perp}$$

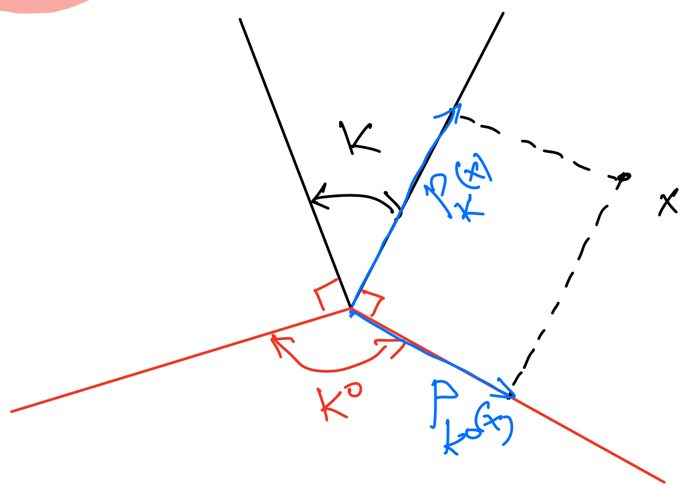


(2) $f = I_K$ where $K \subset \mathbb{R}^n$ is a closed convex cone.

$f^* = I_{K^\circ}$

K° = polar cone (negative)

$p \in K^\circ \iff \langle p, z \rangle \le 0, \ \forall z \in K$



$P_K(x) + P_{K^\circ}(x) = x$

$P_K(x) \perp P_{K^\circ}(x)$

Moreau's decompositio

$f \in \Gamma_0(\mathbb{R}^n)$

$$\mathrm{prox}_f + \mathrm{prox}_{f^*} = \mathrm{Id}$$

$$M_f + M_{f^*} = \frac{1}{2}\|\cdot\|^2$$

Proof. We first prove that $\mathcal{M}_f$ is differentiable. Let $x \in \mathbb{R}^d$ and define $y := \operatorname{prox}_{f^*}(x) = x - \operatorname{prox}_f(x)$. Our aim is to prove that

$$\lim_{h \to 0_{\mathbb{R}^d}} \frac{\mathcal{M}_f(x+h) - \mathcal{M}_f(x) - \langle y, h\rangle}{\|h\|} = 0.$$

To this aim, let us introduce the map

$$\begin{array}{rccl} \Phi: & \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ & h & \longmapsto & \Phi(h) := \mathcal{M}_f(x+h) - \mathcal{M}_f(x) - \langle y, h\rangle. \end{array}$$

Let $h \in \mathbb{R}^d$. Since

$$\mathcal{M}_f(x) = f(\operatorname{prox}_f(x)) + \frac{1}{2}\|\operatorname{prox}_f(x) - x\|^2 \quad \text{and} \quad \mathcal{M}_f(x+h) \leqslant f(\operatorname{prox}_f(x)) + \frac{1}{2}\|\operatorname{prox}_f(x) - (x+h)\|^2,$$

we obtain that

$$\Phi(h) \leqslant \frac{1}{2}(\|\operatorname{prox}_f(x) - (x+h)\|^2 - \|\operatorname{prox}_f(x) - x\|^2 - 2\langle y, h\rangle) = \frac{1}{2}(\|y+h\|^2 - \|y\|^2 - 2\langle y, h\rangle) = \frac{1}{2}\|h\|^2.$$

Furthermore, since $\operatorname{prox}_{f^*}(x) \in \partial f(\operatorname{prox}_f(x))$ (see Proposition 4.1), we have

$$\begin{aligned}
\mathcal{M}_f(x+h) - \mathcal{M}_f(x) &= f(\operatorname{prox}_f(x+h)) - f(\operatorname{prox}_f(x)) + \frac{1}{2}(\|\operatorname{prox}_{f^*}(x+h)\|^2 - \|\operatorname{prox}_{f^*}(x)\|^2) \\
&\geqslant \langle \operatorname{prox}_{f^*}(x), \operatorname{prox}_f(x+h) - \operatorname{prox}_f(x)\rangle + \frac{1}{2}(\|\operatorname{prox}_{f^*}(x+h)\|^2 - \|\operatorname{prox}_{f^*}(x)\|^2) \\
&= \frac{1}{2}(2\langle \operatorname{prox}_{f^*}(x), \operatorname{prox}_f(x+h) - \operatorname{prox}_f(x)\rangle + \|\operatorname{prox}_{f^*}(x+h)\|^2 - \|\operatorname{prox}_{f^*}(x)\|^2) \\
&= \frac{1}{2}(2\langle y, \operatorname{prox}_f(x+h) - (x+h) + x - \operatorname{prox}_f(x)\rangle + 2\langle y, h\rangle + \|\operatorname{prox}_{f^*}(x+h)\|^2 - \|y\|^2) \\
&= \frac{1}{2}(2\langle y, \operatorname{prox}_{f^*}(x) - \operatorname{prox}_{f^*}(x+h)\rangle + 2\langle y, h\rangle + \|\operatorname{prox}_{f^*}(x+h)\|^2 - \|y\|^2) \\
&= \frac{1}{2}(2\langle y, y - \operatorname{prox}_{f^*}(x+h)\rangle + 2\langle y, h\rangle + \|\operatorname{prox}_{f^*}(x+h)\|^2 - \|y\|^2) \\
&= \frac{1}{2}(2\langle y, h\rangle + \|\operatorname{prox}_{f^*}(x+h)\|^2 + \|y\|^2 - 2\langle y, \operatorname{prox}_{f^*}(x+h)\rangle) \\
&= \frac{1}{2}(\|\operatorname{prox}_{f^*}(x+h) - y\|^2 + 2\langle y, h\rangle).
\end{aligned}$$

We deduce that $\Phi(h) \geqslant 0$. This concludes the proof of differentiability of $\mathcal{M}_f$ with $\nabla\mathcal{M}_f = \operatorname{prox}_{f^*}$. The convexity of $\mathcal{M}_f$ is obtained from Theorem E.2 by noting that

$$\langle \nabla\mathcal{M}_f(x_2) - \nabla\mathcal{M}_f(x_1), x_2 - x_1\rangle = \langle \operatorname{prox}_{f^*}(x_2) - \operatorname{prox}_{f^*}(x_1), x_2 - x_1\rangle \geqslant 0,$$

for all x_1, $x_2 \in \mathbb{R}^d$, from monotocity of $\operatorname{prox}_{f^*}$. $\square$

Remark 4.4. Let $f \in \Gamma_0(\mathbb{R}^d)$. Then, it holds that $f \circ \operatorname{prox}_f \leqslant \mathcal{M}_f \leqslant f$ and $\operatorname{Argmin}(\mathcal{M}_f) = \operatorname{Argmin}(f)$ (exercise).

Proposition 4.2. The Moreau's envelope associated with $f = \iota_C \in \Gamma_0(\mathbb{R}^d)$, where C is a nonempty closed convex subset of $\mathbb{R}^d$, is given by $\mathcal{M}_f = \frac{1}{2}\mathrm{d}_C^2$. In particular, the map $\frac{1}{2}\mathrm{d}_C^2 : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable with

$$\nabla\left(\frac{1}{2}\mathrm{d}_C^2\right)(x) = x - \mathrm{proj}_C(x),$$

for all $x \in \mathbb{R}^d$.

Furthermore, the 1-Lipschitz continuous convex map $\mathrm{d}_C : \mathbb{R}^d \to \mathbb{R}$ is subdifferentiable at any $x \in \mathbb{R}^d = \mathrm{dom}(\mathrm{d}_C)$ with

$$\forall x \in \mathbb{R}^d, \quad \partial \mathrm{d}_C(x) = \begin{cases} \left\{\frac{1}{\mathrm{d}_C(x)}(x - \mathrm{proj}_C(x))\right\} & \text{if} \quad x \notin C, \\ \mathrm{N}_C[x] \cap \overline{\mathrm{B}}(0_{\mathbb{R}^d}, 1) & \text{if} \quad x \in C. \end{cases}$$

In particular, $\mathrm{d}_C : \mathbb{R}^d \to \mathbb{R}$ is differentiable at any $x \notin C$ with $\nabla \mathrm{d}_C(x) = \frac{1}{\mathrm{d}_C(x)}(x - \mathrm{proj}_C(x))$.

Proof. The first two sentences are trivial. Then, the fact that $\mathrm{d}_C : \mathbb{R}^d \to \mathbb{R}$ is 1-Lipschitz continuous and convex is recalled in Remark D.1. Let us prove the subdifferentiability. First, take $x \notin C$ and, since C is closed, we know that $\mathrm{d}_C(x) > 0$ (see Remark C.1) and thus $\mathrm{d}_C = \sqrt{\mathrm{d}_C^2}$ is differentiable at x with

$$\nabla \mathrm{d}_C(x) = \frac{1}{2\sqrt{\mathrm{d}_C(x)^2}} \nabla(\mathrm{d}_C^2)(x) = \frac{1}{\mathrm{d}_C(x)}(x - \mathrm{proj}_C(x)).$$

We conclude with Proposition E.1. Now take $x \in C$ and let us prove that $\partial \mathrm{d}_C(x) = \mathrm{N}_C[x] \cap \overline{\mathrm{B}}(0_{\mathbb{R}^d}, 1)$. Let $\ell \in \partial \mathrm{d}_C(x)$. Then $\langle \ell, y - x\rangle \leqslant \mathrm{d}_C(y) - \mathrm{d}_C(x) = \mathrm{d}_C(y) \leqslant \|y - x\|$ for all $y \in \mathbb{R}^d$. One can easily deduce that $\ell \in \mathrm{N}_C[x]$ (by taking all $y \in C$) and that $\ell \in \overline{\mathrm{B}}(0_{\mathbb{R}^d}, 1)$ (by taking $y = x + \ell$). Now let $\ell \in \mathrm{N}_C[x] \cap \overline{\mathrm{B}}(0_{\mathbb{R}^d}, 1)$. Then, it holds that $\langle \ell, y-x\rangle = \langle \ell, y - \mathrm{proj}_C(y)\rangle + \langle \ell, \mathrm{proj}_C(y) - x\rangle \leqslant \langle \ell, y - \mathrm{proj}_C(y)\rangle \leqslant \|\ell\|\|y - \mathrm{proj}_C(y)\| \leqslant \mathrm{d}_C(y) = \mathrm{d}_C(y) - \mathrm{d}_C(x)$ for all $y \in \mathbb{R}^d$. Thus we have $\ell \in \partial \mathrm{d}_C(x)$. □

Proposition 4.3. Let $f \in \Gamma_0(\mathbb{R}^d)$. Then, the so-called *Moreau's complement* of f given by

$$\frac{1}{2}\|\cdot\|^2 - \mathcal{M}_f,$$

is a differentiable convex function with $\nabla(\frac{1}{2}\|\cdot\|^2 - \mathcal{M}_f) = \mathrm{prox}_f$.

Proof. From Theorem 4.2, the differentiability of $\frac{1}{2}\|\cdot\|^2 - \mathcal{M}_f$ is trivial with $\nabla(\frac{1}{2}\|\cdot\|^2 - \mathcal{M}_f) = \mathrm{prox}_f$. The convexity of $\frac{1}{2}\|\cdot\|^2 - \mathcal{M}_f$ is obtained from monotonicity of prox_f and from Theorem E.2. □

Proposition 4.4. Let $f \in \Gamma_0(\mathbb{R}^d)$. Then, the so-called *Moreau's envelope decomposition* given by

$$\forall x \in \mathbb{R}^d, \quad \frac{1}{2}\|x\|^2 = \mathcal{M}_f(x) + \mathcal{M}_{f^*}(x),$$

holds true.

Proof. Let $x \in \mathbb{R}^d$ and let us define the proper convex map $\phi_x(y) := \frac{1}{2}\|y - x\|^2$ for all $y \in \mathbb{R}^d$. From conjugate calculus rules (see Propositions F.3 and F.4), one can easily derive that $\phi_x^*(\ell) = \frac{1}{2}\|\ell\|^2 + \langle \ell, x\rangle$ for all $\ell \in \mathbb{R}^d$ (exercise). Then, from the Fenchel-Rockafellar theorem, we have

$$\begin{aligned}\mathcal{M}_f(x) = \min_{y \in \mathbb{R}^d} f(y) + \phi_x(y) = \inf_{y \in \mathbb{R}^d} f(y) + \phi_x(y) = -\min_{\ell \in \mathbb{R}^d} f^*(\ell) + \phi_x^*(-\ell) = -\min_{\ell \in \mathbb{R}^d}\left(f^*(\ell) + \frac{1}{2}\|-\ell\|^2 - \langle \ell, x\rangle\right) \\ = -\min_{\ell \in \mathbb{R}^d}\left(f^*(\ell) + \frac{1}{2}\|x - \ell\|^2 - \frac{1}{2}\|x\|^2\right) = -M_{f^*}(x) + \frac{1}{2}\|x\|^2,\end{aligned}$$

which concludes the proof. □

Remark 4.5. Let $f \in \Gamma_0(\mathbb{R}^d)$. By differentiating the above Moreau's envelope decomposition, we recover the Moreau's decomposition.

4.3 Some proximal calculus rules and examples

Proposition 4.5 (Some proximal calculus rules and examples)**.** In the table below, we consider that C is a nonempty closed convex subset of $\mathbb{R}^d$ and that $\varphi \in \Gamma_0(\mathbb{R}^d)$ (or $\varphi \in \Gamma_0(\mathbb{R}^n)$ for some occasions).

Item	Comment	$f(x)$	$\mathrm{prox}_f(x)$
(i)	Translation/scaling	$\varphi(\lambda x + a)$ with $\lambda \neq 0$ and $a \in \mathbb{R}^d$	$\frac{1}{\lambda}(\mathrm{prox}_{\lambda^2\varphi}(\lambda x + a) - a)$
(ii)	Other scaling	$\lambda\varphi(\frac{x}{\lambda})$ with $\lambda > 0$	$\lambda \mathrm{prox}_{\varphi/\lambda}(\frac{x}{\lambda})$
(iii)	A quadratic perturbation	$\varphi(x) + \frac{\lambda}{2}\\|x\\|^2 + \langle a, x\rangle + b$ with $\lambda \geqslant 0$, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$	$\mathrm{prox}_{\varphi/(\lambda+1)}(\frac{x-a}{\lambda+1})$
(iv)	Conjugation	$\varphi^*(x)$	$x - \mathrm{prox}_\varphi(x)$
(v)	Moreau's envelope	$\mathcal{M}_\varphi$	$\frac{1}{2}(x + \mathrm{prox}_{2\varphi}(x))$
(vi)	Moreau's complement	$\frac{1}{2}\\|\cdot\\|^2 - \mathcal{M}_\varphi$	$x - \mathrm{prox}_{\varphi/2}(x/2)$
(vii)	Decomposition in orthonormal basis	$\sum_{i=1}^d \phi_i(\langle x, b_i\rangle)$ with $\phi_i \in \Gamma_0(\mathbb{R})$ and $(b_i)_{i=1,\dots,d}$ orthonormal basis	$\sum_{i=1}^d \mathrm{prox}_{\phi_i}(\langle x, b_i\rangle)\, b_i$
(viii)	Semi-orthogonal transformation	$\varphi(Mx)$ with $M \in \mathbb{R}^{n\times d}$ and $\mathrm{im}(M) \subset \mathrm{int}(\mathrm{dom}(\varphi))$ and $MM^\top = \lambda I_n$ with $\lambda > 0$	$x + \frac{1}{\lambda}M^\top(\mathrm{prox}_{\lambda\varphi}(Mx) - Mx)$
(ix)	A standard quadratic function	$\frac{\lambda}{2}\\|Mx - a\\|^2$ with $M \in \mathbb{R}^{n\times d}$, $a \in \mathbb{R}^n$, $\lambda > 0$	$(I_d + \lambda M^\top M)^{-1}(x + \lambda M^\top a)$
(x)	Indicator function	$\iota_C(x)$	$\mathrm{proj}_C(x)$
(xi)	Support function	$\sigma_C(x)$	$x - \mathrm{proj}_C(x)$
(xii)	Distance function	$\lambda \mathrm{d}_C(x)$ with $\lambda > 0$	$\begin{cases} x - \frac{\lambda}{\mathrm{d}_C(x)}(x - \mathrm{proj}_C(x)), & \text{if } \mathrm{d}_C(x) \geqslant \lambda \\ \mathrm{proj}_C(x), & \text{otherwise} \end{cases}$
(xiii)	Function of distance	$\phi(\mathrm{d}_C(x))$ with $\phi \in \Gamma_0(\mathbb{R})$ even and differentiable over $\mathbb{R}^*$	$\begin{cases} \mathrm{proj}_C(x), & \text{if } \mathrm{d}_C(x) \leqslant \phi'_+(0) \\ x + \left(\frac{\mathrm{prox}_\phi(\mathrm{d}_C(x))}{\mathrm{d}_C(x)} - 1\right)(x - \mathrm{proj}_C(x)), & \text{if } \mathrm{d}_C(x) > \phi'_+(0) \end{cases}$
(xiv)	Function of Euclidean norm	$\phi(\\|x\\|)$ with $\phi \in \Gamma_0(\mathbb{R})$ nondecreasing over $\mathrm{dom}(\phi) \subset \mathbb{R}_+$	$\begin{cases} \mathrm{prox}_\phi(\\|x\\|)\frac{x}{\\|x\\|} & \text{if } x \neq 0_{\mathbb{R}^d} \\ 0_{\mathbb{R}^d} & \text{if } x = 0_{\mathbb{R}^d} \end{cases}$

Proof. (i)(ii) Use the characterizing property of the proximal operator that is $y = \mathrm{prox}_f(x)$ if and only if $\langle x - y, x' - y\rangle \leqslant f(x') - f(y)$ for all $x' \in \mathbb{R}^d$.

(iii) Use the fact that $y = \mathrm{prox}_f(x)$ if and only if $x \in y + \partial f(y)$ and develop ∂f from subdifferential calculus rules.

(iv) Trivial.

(v) $y = \mathrm{prox}_f(x)$ if and only if $x \in y + \partial f(y) = y + \nabla M_\varphi(y) = y + \mathrm{prox}_{\varphi^*}(y) = 2y - \mathrm{prox}_\varphi(y)$ if and only if $y \in 2y - x + \partial\varphi(2y-x)$ if and only if $x - y \in \partial\varphi(2y-x)$ if and only if $2x - 2y \in \partial(2\varphi)(2y-x)$ if and only if $x \in 2y - x + \partial(2\varphi)(2y-x)$ if and only if $\mathrm{prox}_{2\varphi}(x) = 2y - x$ if and only if $y = \frac{1}{2}(x + \mathrm{prox}_{2\varphi}(x))$.

(vi) $y = \mathrm{prox}_f(x)$ if and only if $x \in y + \partial f(y) = y + \nabla f(y) = y + \mathrm{prox}_\varphi(y)$ if and only if $y \in x - y + \partial\varphi(x-y)$ if and only if $2y - x \in \partial\varphi(x-y)$ if and only if $y - \frac{x}{2} \in \partial(\frac{\varphi}{2})(x-y)$ if and only if $\frac{x}{2} \in x - y + \partial(\frac{\varphi}{2})(x-y)$ if and only if $x - y = \mathrm{prox}_{\varphi/2}(x/2)$. Another method consists in seeing that $f = \mathcal{M}_{\varphi^*}$ and thus $\mathrm{prox}_f(x) = \frac{1}{2}(x + \mathrm{prox}_{2\varphi^*}(x))$ for all $x \in \mathbb{R}^d$. Since $(2\varphi^*)^* = 2\varphi^{**}(\frac{\cdot}{2}) = 2\varphi(\frac{\cdot}{2})$, we obtain that $\mathrm{prox}_{2\varphi^*} = \mathrm{Id} - \mathrm{prox}_{2\varphi(\frac{\cdot}{2})} = \mathrm{Id} - 2\mathrm{prox}_{\varphi/2}(\frac{\cdot}{2})$. The last equality comes from item (ii).

(vii) Recall that, for any $t \in \mathbb{R}^d$, we have $t = \sum_{i=1}^d t_i b_i$ with $t_i := \langle t, b_i\rangle$ for all $i \in \{1,\dots,d\}$. Note that $\langle t, z\rangle = \sum_{i=1}^d t_i z_i$ for all $t,\, z \in \mathbb{R}^d$. Let $x \in \mathbb{R}^d$ and $y = \mathrm{prox}_f(x)$. It holds that

$$\langle x - y, x' - y\rangle \leqslant \sum_{i=1}^d \Big(\phi_i(\langle x', b_i\rangle) - \phi_i(\langle y, b_i\rangle)\Big),$$

for all $x' \in \mathbb{R}^d$. We develop and obtain

$$\sum_{i=1}^d (x_i - y_i)(x'_i - y_i) \leqslant \sum_{i=1}^d \Big(\phi_i(x'_i) - \phi_i(y_i)\Big),$$

for all $x' \in \mathbb{R}^d$. Taking $x' = x'_j b_j + \sum_{i=1, i\neq j}^d y_i b_i$, we obtain that

$$(x_j - y_j)(x'_j - y_j) \leqslant \phi_j(x'_j) - \phi_j(y_j),$$

for all $x'_j \in \mathbb{R}$ and all $j \in \{1,\dots,d\}$. We deduce that $y_j = \mathrm{prox}_{\phi_j}(x_j) = \mathrm{prox}_{\phi_j}(\langle x, b_j\rangle)$. We deduce that

$$y = \sum_{i=1}^d y_i b_i = \sum_{i=1}^d \mathrm{prox}_{\phi_i}(\langle x, b_i\rangle) b_i.$$

(viii) We start with some recalls. One can easily prove that $\ker(M) = \mathrm{im}(M^\top)^\perp$ (exercise) and thus $\ker(M)^\perp = \mathrm{im}(M^\top)$. Furthermore, when $MM^\top$ is invertible, one can deduce from Propositions 3.1 and C.1 that

$$\mathrm{proj}_{\ker(M)} = I_d - M^\top(MM^\top)^{-1}M \quad \text{and} \quad \mathrm{proj}_{\ker(M)^\perp} = I_d - \mathrm{proj}_{\ker(M)} = M^\top(MM^\top)^{-1}M.$$

In our context, we have

$$\mathrm{proj}_{\ker(M)} = I_d - \frac{1}{\lambda}M^\top M \quad \text{and} \quad \mathrm{proj}_{\ker(M)^\perp} = \frac{1}{\lambda}M^\top M.$$

Now let us start the proof. Let $x \in \mathbb{R}^d$ and $y := \mathrm{prox}_f(x)$. Since $My \in \mathrm{int}(\mathrm{dom}(\varphi))$, we get that $x \in y + \partial f(y) = y + \partial(\varphi \circ M)(y) = y + M^\top \partial\varphi(My)$ (see Proposition E.3) and thus $Mx \in My + \partial(\lambda\varphi)(My)$ and thus $My = \mathrm{prox}_{\lambda\varphi}(Mx)$. From the previous reasonning, we also know that $y - x \in \mathrm{im}(M^\top) = \ker(M)^\perp$ and thus

$$\begin{aligned} y = \mathrm{proj}_{\ker(M)}(y) + \mathrm{proj}_{\ker(M)^\perp}(y) &= \mathrm{proj}_{\ker(M)}(x) + \mathrm{proj}_{\ker(M)}(y-x) + \mathrm{proj}_{\ker(M)^\perp}(y) \\ &= \mathrm{proj}_{\ker(M)}(x) + \mathrm{proj}_{\ker(M)^\perp}(y) = x + \frac{1}{\lambda}M^\top(\mathrm{prox}_{\lambda\varphi}(Mx) - Mx). \end{aligned}$$

(ix) First, note that $I_d + \lambda M^\top M$ is invertible since $(I_d + \lambda M^\top M)x = 0_{\mathbb{R}^d}$ implies $\langle (I_d + \lambda M^\top M)x, x\rangle = 0 = \|x\|^2 + \lambda\|Mx\|^2$ and thus $x = 0_{\mathbb{R}^d}$. On the other hand, let $x \in \mathbb{R}^d$ and $y = \mathrm{prox}_f(x)$. We get that $x \in y + \partial f(y) = y + \nabla f(y) = y + \lambda M^\top(My - a)$ and thus $x + \lambda M^\top a = (I_d + \lambda M^\top M)y$ which concludes the proof. This example can also be solved from Remark 4.3.

(x) Trivial.

(xi) We have $\mathrm{prox}_{\sigma_C} = \mathrm{prox}_{(\iota_C)^*} = \mathrm{Id} - \mathrm{prox}_{\iota_C} = \mathrm{Id} - \mathrm{proj}_C$.

(xii) Let $x \in \mathbb{R}^d$ and $y := \mathrm{prox}_{\lambda \mathrm{d}_C}(x)$. Thus we have $x \in y + \lambda \partial \mathrm{d}_C(y)$. Let us prove some facts:

- If $y \in C$, then $x \in y + \lambda(\mathrm{N}_C[y] \cap \overline{\mathrm{B}}(0_{\mathbb{R}^d}, 1)) = y + \mathrm{N}_C[y] \cap \overline{\mathrm{B}}(0_{\mathbb{R}^d}, \lambda)$ thus $y = \mathrm{prox}_C(x)$ and $\mathrm{d}_C(x) = \|x - y\| \leqslant \lambda$.
- If $y \notin C$, then $x = y + \frac{\lambda}{\mathrm{d}_C(y)}(y - \mathrm{proj}_C(y))$ and thus

$$x - \mathrm{proj}_C(y) = \left(1 + \frac{\lambda}{\mathrm{d}_C(y)}\right)(y - \mathrm{proj}_C(y)) \in \mathrm{N}_C[\mathrm{proj}_C(y)],$$

  which leads to $\mathrm{proj}_C(y) = \mathrm{proj}_C(x)$. Using this last equality and passing to the norm in the previous equation (in middle text), we obtain that $\mathrm{d}_C(x) = \mathrm{d}_C(y) + \lambda > \lambda$. Finally, using this last equality with $\mathrm{proj}_C(y) = \mathrm{proj}_C(x)$ in the previous equation (in middle text), we obtain that $y = x - \frac{\lambda}{\mathrm{d}_C(x)}(x - \mathrm{proj}_C(x))$.

Now, if $\mathrm{d}_C(x) > \lambda$, then we cannot have $y \in C$ and thus $y \notin C$ and thus $y = x - \frac{\lambda}{\mathrm{d}_C(x)}(x - \mathrm{proj}_C(x))$. On the contrary, if $\mathrm{d}_C(x) \leqslant \lambda$, then we cannot have $y \notin C$ and thus $y \in C$ and thus $y = \mathrm{proj}_C(x)$.

(xiii) Since $\mathbb{R}^*$ is included in the convex set $\mathrm{dom}(\phi)$, we have $\mathrm{dom}(\phi) = \mathbb{R}$ and thus $0 \in \mathrm{int}(\mathrm{dom}(\phi))$ and thus ϕ is continuous and subdifferentiable at 0 with $\partial\phi(0) = [\phi'_-(0), \phi'_+(0)]$ (see Theorems E.1 and E.4 and Proposition E.4). Since ϕ even, one can easily see that $\partial\phi(0) = [-\beta, \beta]$ with $\beta := \phi'_+(0) = -\phi'_-(0) \geqslant 0$. Since $\partial\phi$ is monotone, we also know that $\phi'(\xi) \geqslant \beta$ for all $\xi > 0$. Also note that the assumptions on ϕ implies that ϕ is convex and nondecreasing over $\mathbb{R}_+$ and thus $f = \phi \circ \mathrm{d}_C \in \Gamma_0(\mathbb{R}^d)$ (convex in particular). Now let $x \in \mathbb{R}^d$ and $y := \mathrm{prox}_f(x)$. Two cases:

- If $y \in C$, let us prove that $y = \mathrm{proj}_C(x)$ and $\mathrm{d}_C(x) \leqslant \beta$. Indeed, for any $x' \in C$, we have $\langle x - y, x' - y\rangle \leqslant f(x') - f(y) = \phi(0) - \phi(0) = 0$ and thus $y = \mathrm{proj}_C(x)$. Since $y = \mathrm{proj}_C(x)$, one can easily prove that $\mathrm{proj}_C(y + \mu(x-y)) = y$ and thus $\langle x - y, y + \mu(x-y) - y\rangle \leqslant f(y + \mu(x-y)) - f(y) = \phi(\mathrm{d}_C(y + \mu(x-y))) - \phi(0) = \phi(\|y + \mu(x-y) - y\|) - \phi(0) = \phi(\mu\|x-y\|) - \phi(0)$ for all $\mu \geqslant 0$. We deduce that $\|x-y\|^2 \leqslant \phi'_+(0)\|x-y\| = \beta\|x-y\|$. We deduce that $\mathrm{d}_C(x) = \|x - y\| \leqslant \beta$.
- If $y \notin C$, let us prove that $y = x + (\frac{\mathrm{prox}_\phi(\mathrm{d}_C(x))}{\mathrm{d}_C(x)} - 1)(x - \mathrm{proj}_C(x))$ and $\mathrm{d}_C(x) > \beta$. Since $\mathrm{d}_C(y) > 0$ (since C is closed in particular, see Remark C.1), we know that f is differentiable at y and thus

$$x = y + f'(y) = y + \frac{\phi'(\mathrm{d}_C(y))}{\mathrm{d}_C(y)}(y - \mathrm{proj}_C(y)).$$

  Passing to the norm and since $\phi'(\mathrm{d}_C(y)) \geqslant \beta \geqslant 0$, we get that $\|x - y\| = \phi'(\mathrm{d}_C(y))$. Furthermore, from firm nonexpansiveness, we get that $\|\mathrm{proj}_C(x) - \mathrm{proj}_C(y)\|^2 \leqslant \langle x - y, \mathrm{proj}_C(x) - \mathrm{proj}_C(y)\rangle = \frac{\phi'(\mathrm{d}_C(y))}{\mathrm{d}_C(y)}\langle y - \mathrm{proj}_C(y), \mathrm{proj}_C(x) - \mathrm{proj}_C(y)\rangle \leqslant 0$ and thus $\mathrm{proj}_C(x) = \mathrm{proj}_C(y)$. Hence, we obtain that $x - y = \frac{\|x-y\|}{\mathrm{d}_C(y)}(y - x + x - \mathrm{proj}_C(x))$ and thus

$$\frac{\mathrm{d}_C(y) + \|x-y\|}{\mathrm{d}_C(y)}(x - y) = \frac{\|x-y\|}{\mathrm{d}_C(y)}(x - \mathrm{proj}_C(x)).$$

  Passing to the norm, we get that $\mathrm{d}_C(x) = \mathrm{d}_C(y) + \|x-y\| = \mathrm{d}_C(y) + \phi'(\mathrm{d}_C(y))$ and thus $\mathrm{d}_C(y) = \mathrm{prox}_\phi(\mathrm{d}_C(x))$ and thus $\|x - y\| = \mathrm{d}_C(x) - \mathrm{d}_C(y) = \mathrm{d}_C(x) - \mathrm{prox}_\phi(\mathrm{d}_C(x))$. Using the last equation (in the middle text), we get that $y = x + (\frac{\mathrm{prox}_\phi(\mathrm{d}_C(x))}{\mathrm{d}_C(x)} - 1)(x - \mathrm{proj}_C(x))$. On the other hand, since $\mathrm{prox}_\phi(\mathrm{d}_C(x)) = \mathrm{d}_C(y) > 0$, we deduce that $0 \leqslant \mathrm{d}_C(x) \notin \partial\phi(0) = [-\beta, \beta]$ and thus $\mathrm{d}_C(x) > \beta$.

Hence we conclude that $y \in C$ if and only if $\mathrm{d}_C(x) \leqslant \beta$. The proof easily follows.

(xiv) Since ϕ is nondecreasing, one can easily prove that $\partial\phi(y) \subset \mathbb{R}_+$ for all $y \in \mathbb{R}^d$ and thus $\mathrm{prox}_\phi(0) \leqslant 0$ (exercise). Furthermore, since ϕ is convex and nondecreasing over $\mathrm{dom}(\phi) \subset \mathbb{R}_+$, note that $f \in \Gamma_0(\mathbb{R}^d)$ (convex in particular).

- Let $y := \mathrm{prox}_f(0_{\mathbb{R}^d})$. Then y solves the problem
$$\underset{v\in\mathbb{R}^d}{\text{minimize}}\ f(v)+\frac{1}{2}\|v\|^2=\phi(\|v\|)+\frac{1}{2}\|v\|^2,$$
and thus $\|y\|$ solves the problem
$$\underset{\alpha\geqslant 0}{\text{minimize}}\ \phi(\alpha)+\frac{1}{2}|\alpha|^2,$$
and, since $\mathrm{dom}(\phi)\subset\mathbb{R}_+$, we get that $\|y\|$ solves the problem
$$\underset{\alpha\in\mathbb{R}}{\text{minimize}}\ \phi(\alpha)+\frac{1}{2}|\alpha|^2.$$
We deduce that $\|y\|=\mathrm{prox}_\phi(0)\leqslant 0$ and thus $y=0_{\mathbb{R}^d}$.

- Now let $x\neq 0_{\mathbb{R}^d}$ and define $y:=\mathrm{prox}_f(x)$. Denote by $\beta:=\mathrm{prox}_\phi(\|x\|)$ and $w:=\beta\frac{x}{\|x\|}$. Our aim is to prove that $y=w$. Thanks to Cauchy-Schwarz inequality and then to the definition of β (as minimizer), we have
$$\begin{aligned}
f(y)+\frac{1}{2}\|y-x\|^2 &= \phi(\|y\|)+\frac{1}{2}(\|y\|^2-2\langle x,y\rangle+\|y\|^2)\\
&\geqslant \phi(\|y\|)+\frac{1}{2}(\|y\|^2-2\|x\|\|y\|+\|y\|^2)\\
&= \phi(\|y\|)+\frac{1}{2}(\|y\|-\|x\|)^2\\
&\geqslant \phi(\beta)+\frac{1}{2}(\beta-\|x\|)^2\\
&= \phi(\beta)+\frac{1}{2}(\beta^2-2\beta\|x\|+\|x\|^2)\\
&= \phi(\|w\|)+\frac{1}{2}(\|w\|^2-2\langle w,x\rangle+\|x\|^2)\\
&= f(w)+\frac{1}{2}\|w-x\|^2,
\end{aligned}$$
and thus $y=w$ (since y is the unique minimizer of $f+\frac{1}{2}\|\cdot-x\|^2$).

The proof is complete. $\square$

Remark 4.6. Let $f\in\Gamma_0(\mathbb{R}^d)$ and $\lambda>0$. From item (ii) of Proposition 4.5 and Proposition F.4, the *extended Moreau's decomposition formula* given by
$$\forall x\in\mathbb{R}^d,\quad x=\mathrm{prox}_{\lambda f}(x)+\lambda\mathrm{prox}_{f^*/\lambda}\left(\frac{x}{\lambda}\right),$$
holds true (exercise). Some consequences:

(i) For C a nonempty closed convex subset of $\mathbb{R}^d$, we have
$$\forall x\in\mathbb{R}^d,\quad \mathrm{prox}_{\lambda\sigma_C}(x)=x-\lambda\mathrm{proj}_C\left(\frac{x}{\lambda}\right).$$

(ii) Now, from Proposition F.5, recall that, if $\mathcal{N}:\mathbb{R}^d\to\mathbb{R}$ stands for a general norm on $\mathbb{R}^d$, then $\mathcal{N}=\sigma_{\overline{\mathrm{B}}_{\mathcal{N}^D}(0_{\mathbb{R}^d},1)}$, where $\overline{\mathrm{B}}_{\mathcal{N}^D}(0_{\mathbb{R}^d},1)$ stands for the unit ball associated with the dual norm $\mathcal{N}^D$. Hence we obtain that
$$\forall x\in\mathbb{R}^d,\quad \mathrm{prox}_{\lambda\mathcal{N}}=\mathrm{prox}_{\lambda\sigma_{\overline{\mathrm{B}}_{\mathcal{N}^D}(0_{\mathbb{R}^d},1)}}=x-\lambda\mathrm{proj}_{\overline{\mathrm{B}}_{\mathcal{N}^D}(0_{\mathbb{R}^d},1)}\left(\frac{x}{\lambda}\right).$$
In particular, when $\mathcal{N}=\|\cdot\|$ is the Euclidean norm, we obtain from Propsotion 3.1 that
$$\forall x\in\mathbb{R}^d,\quad \mathrm{prox}_{\lambda\|\cdot\|}=x-\lambda\mathrm{proj}_{\overline{\mathrm{B}}(0_{\mathbb{R}^d},1)}\left(\frac{x}{\lambda}\right)=\left(1-\frac{\lambda}{\max\{\|x\|,\lambda\}}\right)x.$$

(iii) One can easily prove (exercise) that the max function defined by

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad f(x) := \max\{x_1, \ldots, x_d\},$$

satisfies $f = \sigma_{\Delta^{d-1}}$, where $\Delta^{d-1} := \{x \in \mathbb{R}^d_+ \mid \sum_{i=1}^d x_i = 1\} = \mathrm{conv}(\{e_1, \ldots, e_d\})$ stands for the unit simplex of $\mathbb{R}^d$ of dimension $d-1$. We deduce that

$$\forall x \in \mathbb{R}^d, \quad \mathrm{prox}_{\lambda f}(x) = x - \lambda \mathrm{proj}_{\Delta^{d-1}}\left(\frac{x}{\lambda}\right).$$

Thanks to *duality theory*, we will see in Chapter 3 how to compute easily $\mathrm{proj}_{\Delta^{d-1}}$ (see Remark 8.6).

Remark 4.7. Let $f \in \Gamma_0(\mathbb{R}^d)$ and $\lambda > 0$. With a similar proof than the one used for the classical Moreau's envelope decomposition, one can obtain (exercise) the *extended Moreau's envelope decomposition* given by

$$\forall x \in \mathbb{R}^d, \quad \frac{1}{2}\|x\|^2 = \mathcal{M}_{\lambda f}(x) + \lambda^2 \mathcal{M}_{f^*/\lambda}\left(\frac{x}{\lambda}\right).$$

By differentiating it, one can recover (exercise) the above extended Moreau's decomposition formula.

Remark 4.8. Let C be a nonempty closed convex subset of $\mathbb{R}^d$. From item (v) of Proposition 4.5 with $\varphi := \iota_C$, one can easily obtain (exercise) that

$$\forall x \in \mathbb{R}^d, \quad \mathrm{prox}_{\mathrm{d}_C^2/2}(x) = \frac{1}{2}(x + \mathrm{proj}_C(x)).$$

One can also recover this result by using item (xiii) of Proposition 4.5 with $\phi(z) := \frac{z^2}{2}$ for all $z \in \mathbb{R}$ (exercise).

Remark 4.9. Consider item (vii) of Proposition 4.5 with $(b_i)_{i=1,\ldots,d} = (e_i)_{i=1,\ldots,d}$ the canonical basis of $\mathbb{R}^d$. If $f \in \Gamma_0(\mathbb{R}^d)$ is a *separable function* of the form

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad f(x) = \sum_{i=1}^d f_i(x_i),$$

with $f_i \in \Gamma(\mathbb{R})$ for all $i \in \{1, \ldots, d\}$, then we have (exercise)

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad \mathrm{prox}_f(x) = (\mathrm{prox}_{f_1}(x_1), \ldots, \mathrm{prox}_{f_d}(x_d)).$$

In particular, if $f = \lambda \| \cdot \|_1$ for some $\lambda > 0$, then

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad \mathrm{prox}_{\lambda\|\cdot\|_1}(x) = \Big(\mathrm{soft}_{[-\lambda,\lambda]}(x_i)\Big)_{i=1,\ldots,d}.$$

In particular, we recover (exercise) from item (ii) of Remark 4.6 the trivial fact that

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad \mathrm{proj}_{\overline{\mathrm{B}}_{\|\cdot\|_\infty}(0_{\mathbb{R}^d},1)}(x) = \Big(x_i - \mathrm{soft}_{[-1,1]}(x_i)\Big)_{i=1,\ldots,d}.$$

Remark 4.10. Let $f \in \Gamma_0(\mathbb{R}^d)$ be defined by

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad f(x) := \varphi(x_1 + \ldots + x_d),$$

where $\varphi \in \Gamma_0(\mathbb{R})$ with $\mathrm{dom}(\varphi) = \mathbb{R}$. From item (viii) of Proposition 4.5, we get that (exercise)

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad \mathrm{prox}_f(x) = \left(x_i + \frac{1}{d}\Big(\mathrm{prox}_{d\varphi}(x_1 + \ldots + x_d) - (x_1 + \ldots + x_d)\Big)\right)_{i=1,\ldots,d}.$$

Remark 4.11. Let $f \in \Gamma_0(\mathbb{R}^d)$ be defined by

$$\forall x \in \mathbb{R}^d, \quad f(x) := |\langle h, x \rangle|,$$

for some $h \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\}$. From item (viii) of Proposition 4.5, we get that (exercise)

$$\forall x \in \mathbb{R}^d, \quad \mathrm{prox}_f(x) = x + \frac{1}{\|h\|^2}\Big(\mathrm{soft}_{[-\|h\|^2,\|h\|^2]}(\langle h, x \rangle) - \langle h, x \rangle\Big)h.$$

Remark 4.12. It holds that

$$\forall x \in \mathbb{R}^d, \quad \mathrm{prox}_{\lambda\|\cdot\|} = \begin{cases} \left(1 - \dfrac{\lambda}{\|x\|}\right) x & \text{if} \quad \|x\| \geqslant \lambda, \\ 0_{\mathbb{R}^d} & \text{if} \quad \|x\| \leqslant \lambda. \end{cases}$$

This result can be obtained in many manners (exercises). For example, one can consider item (ii) of Remark 4.6. One can also consider item (xii) of Proposition 4.5 with $C = \{0_{\mathbb{R}^d}\}$. One can also find this result from item (xiii) of Proposition 4.5 (taking $\phi := \lambda|\cdot|$) or from item (xiv) of Proposition 4.5 (taking $\phi(z) := +\infty$ if $z < 0$ and $\phi(z) := \lambda z$ if $z \geqslant 0$). For this last strategy, one should use item (iv) of the next Proposition 4.6.

Remark 4.13. For the needs of the next Proposition 4.6, a few remarks are required:

(i) Let $f \in \Gamma_0(\mathbb{R}^d)$. Note that, for any $x \in \mathbb{R}^d$, we have $\mathrm{prox}_f(x) = 0_{\mathbb{R}^d}$ if and only if $x \in \partial f(0_{\mathbb{R}^d})$.

(ii) From the scaling property, note that, if $f \in \Gamma_0(\mathbb{R}^d)$ is even, then $\mathrm{prox}_f : \mathbb{R}^d \to \mathbb{R}^d$ is odd (exercise). Thus, in that case, we have $\mathrm{prox}_f(0_{\mathbb{R}^d}) = 0_{\mathbb{R}^d}$ and, when $d = 1$, since $\mathrm{prox}_f : \mathbb{R} \to \mathbb{R}$ is nondecreasing (since monotone), we know that prox_f is nonpositive over $\mathbb{R}_-$ and nonnegative over $\mathbb{R}_+$.

(iii) Let $f \in \Gamma_0(\mathbb{R}^d)$ and recall that $\mathrm{Ran}(\mathrm{prox}_f) = \mathrm{Dom}(\partial f) \subset \mathrm{dom}(f)$. In the case where $d = 1$, note that, if $\mathrm{dom}(f) \subset \mathbb{R}_+$, then $\mathrm{prox}_f : \mathbb{R} \to \mathbb{R}$ has nonnegative values.

(iv) Only for the last item of the next Proposition 4.6, the following result is required. Let $f : \mathbb{R} \to \mathbb{R}$ be a function such that f is convex/nonincreasing over $\mathbb{R}_-$ and convex/nondecreasing over $\mathbb{R}_+$. Then f is convex over $\mathbb{R}$ (exercise). To see this, one should consider $x_1 < 0 < x_2$ and $z := (1-\lambda)x_1 + \lambda x_2$ for some $\lambda \in [0, 1]$. Considering that $x_1 \leqslant z \leqslant 0$ (the other case is similar), we can write $z = (1-\mu)x_1 + \mu 0 = (1-\mu)x_1$ for some $\mu \in [0, 1]$. We easily obtain that $\mu = \lambda(1 - \frac{x_2}{x_1}) > \lambda$. Finally we have $f((1-\lambda)x_1 + \lambda x_2) = f(z) \leqslant (1-\mu)f(x_1) + \mu f(0) = f(x_1) + \mu(f(0) - f(x_1)) \leqslant f(x_1) + \lambda(f(0) - f(x_1)) \leqslant f(x_1) + \lambda(f(x_2) - f(x_1)) = (1-\lambda)f(x_1) + \lambda f(x_2)$ since f is convex over $\mathbb{R}_-$, $f(0) \leqslant f(x_1)$ and $f(0) \leqslant f(x_2)$.

Proposition 4.6 (Examples in the case $d = 1$)**.** In the table below, we consider $d = 1$, $\omega_1 \leqslant \omega_2$ (assuming that $\omega_1 < 0 < \omega_2$ if necessary), $\lambda \geqslant 0$, $\mu \geqslant 0$, $\omega \geqslant 0$, $\gamma \in \mathbb{R}$, $p > 1$ and $\tau > 0$.

Item	$f(x)$	$\operatorname{prox}_f(x)$
(i)	λx^2	$\frac{1}{1+2\lambda}x$
(ii)	$\sigma_{[\omega_1,\omega_2]}(x) = \begin{cases} \omega_1 x & \text{if } x \leqslant 0 \\ \omega_2 x & \text{if } x \geqslant 0 \end{cases}$	$\operatorname{soft}_{[\omega_1,\omega_2]} = \begin{cases} x-\omega_1 & \text{if } x \leqslant \omega_1 \\ 0 & \text{if } x \in [\omega_1,\omega_2] \\ x-\omega_2 & \text{if } x \geqslant \omega_2 \end{cases}$
(iii)	$\phi(x) + \sigma_{[\omega_1,\omega_2]}(x)$ with $\phi \in \Gamma_0(\mathbb{R})$ differentiable at 0 with $\phi'(0) = 0$	$\operatorname{prox}_\phi(\operatorname{soft}_{[\omega_1,\omega_2]}(x))$
(iv)	$\begin{cases} \gamma x & \text{if } x \geqslant 0 \\ +\infty & \text{if } x < 0 \end{cases}$	$(x-\gamma)^+ = \begin{cases} 0 & \text{if } x \leqslant \gamma \\ x-\gamma & \text{if } x \geqslant \gamma \end{cases}$
(v)	$\begin{cases} \gamma x & \text{if } 0 \leqslant x \leqslant \mu \\ +\infty & \text{otherwise} \end{cases}$	$\min\{(x-\gamma)^+, \mu\}$
(vi)	$\begin{cases} \lambda \lvert x\rvert & \text{if } -\mu \leqslant x \leqslant \mu \\ +\infty & \text{otherwise} \end{cases}$	$\min\{(\lvert x\rvert-\lambda)^+, \mu\}\operatorname{sign}(x)$
(vii)	$(\lvert x\rvert - \lambda)^+$	$\begin{cases} x & \text{if } \lvert x\rvert \leqslant \lambda \\ \lambda\operatorname{sign}(x) & \text{if } \lambda \leqslant \lvert x\rvert \leqslant \lambda+1 \\ x - \operatorname{sign}(x) & \text{if } \lambda+1 \leqslant \lvert x\rvert \end{cases}$
(viii)	$\begin{cases} 2x^2 & \text{if } \lvert x\rvert \leqslant \frac{\lambda}{2} \\ 2\lambda\lvert x\rvert - \frac{\lambda^2}{2} & \text{if } \lvert x\rvert \geqslant \frac{\lambda}{2} \end{cases}$	$\begin{cases} \frac{x}{5} & \text{if } \lvert x\rvert \leqslant \frac{5\lambda}{2} \\ x - 2\lambda\operatorname{sign}(x) & \text{if } \lvert x\rvert \geqslant \frac{5\lambda}{2} \end{cases}$
(ix)	$\lambda\lvert x\rvert^p$	$\rho(x)\operatorname{sign}(x)$ where $\rho(x) \geqslant 0$ is the unique solution to $\rho(x) + \lambda p \rho(x)^{p-1} = \lvert x\rvert$
(x)	$\lambda\lvert x\rvert^p + \mu\lvert x\rvert^2 + \omega\lvert x\rvert$	$\operatorname{prox}_{\lambda\lvert\cdot\rvert^p/(2\mu+1)}\left(\operatorname{soft}_{[-\frac{\omega}{2\mu+1},\frac{\omega}{2\mu+1}]}\left(\frac{x}{2\mu+1}\right)\right)$
(xi)	$\lvert x\rvert - \ln(1+\lvert x\rvert)$	$\frac{\operatorname{sign}(x)}{2}(\lvert x\rvert - 2 + \sqrt{\lvert x\rvert^2+4})$
(xii)	$\begin{cases} -\tau\ln(x) & \text{if } x > 0 \\ +\infty & \text{if } x \leqslant 0 \end{cases}$	$\frac{1}{2}(x + \sqrt{x^2+4\tau})$
(xiii)	$\begin{cases} \ln(-\omega_1) - \ln(x-\omega_1) & \text{if } \omega_1 < x \leqslant 0 \\ \ln(\omega_2) - \ln(\omega_2 - x) & \text{if } 0 \leqslant x < \omega_2 \\ +\infty & \text{otherwise} \end{cases}$	$\begin{cases} \frac{1}{2}(x+\omega_1+\sqrt{(x-\omega_1)^2+4}) & \text{if } x \leqslant \frac{1}{\omega_1} \\ \frac{1}{2}(x+\omega_2-\sqrt{(x-\omega_2)^2+4}) & \text{if } x \geqslant \frac{1}{\omega_2} \\ 0 & \text{otherwise} \end{cases}$

Proof. (i) Trivial (one can also make a drawing).

(ii) One has just to see that $\operatorname{prox}_{\sigma_{[\omega_1,\omega_2]}} = \operatorname{prox}_{(\iota_{[\omega_1,\omega_2]})^*} = \operatorname{Id} - \operatorname{proj}_{[\omega_1,\omega_2]}$. Or simply make a drawing.

(iii) Let $x \in \mathbb{R}$ and $y := \operatorname{prox}_f(x)$. Since $\operatorname{int}(\operatorname{dom}(\sigma_{[\omega_1,\omega_2]})) \cap \operatorname{dom}(\phi) = \operatorname{dom}(\phi) \neq \emptyset$, we obtain that $x \in y + \partial f(y) = y + \partial(\phi + \sigma_{[\omega_1,\omega_2]})(y) = y + \partial\phi(y) + \partial\sigma_{[\omega_1,\omega_2]}(y)$. Hence there exists $z \in \partial\phi(y)$ such that $x \in y + z + \partial\sigma_{[\omega_1,\omega_2]}(y)$. Since $z \in \partial\phi(y)$ and 0 is a minimizer of ϕ, we obtain that $z(0-y) \leqslant \phi(0) - \phi(y) \leqslant 0$ and thus $zy \geqslant 0$. Three cases:

- If $y > 0$, then $z \geqslant 0$ and we obtain that $\partial\sigma_{[\omega_1,\omega_2]}(y) = \omega_2 = \partial\sigma_{[\omega_1,\omega_2]}(y+z)$. We deduce that $x \in y+z+\partial\sigma_{[\omega_1,\omega_2]}(y+z)$ and thus $y+z = \operatorname{prox}_{\sigma_{[\omega_1,\omega_2]}}(x) = \operatorname{soft}_{[\omega_1,\omega_2]}(x)$. Thus $\operatorname{soft}_{[\omega_1,\omega_2]}(x) = y+z \in y + \partial\phi(y)$ and $y = \operatorname{prox}_\phi(\operatorname{soft}_{[\omega_1,\omega_2]}(x))$.

- A similar reasoning holds for $y < 0$.

- If $y = 0$, then $\partial\phi(y) = \partial\phi(0) = \phi'(0) = 0$ and thus $x \in y + \partial\sigma_{[\omega_1,\omega_2]}(y) = [\omega_1, \omega_2]$. We have obtained that $\mathrm{prox}_\phi(\mathrm{soft}_{[\omega_1,\omega_2]}(x)) = \mathrm{prox}_\phi(0) = 0 = y$ which concludes the proof.

(iv) A simple drawing can be done, but we are going to prove this result mathematically (with two different proofs). Let $x \in \mathbb{R}$ and $y = \mathrm{prox}_f(x)$. Then we have $x - y \in \partial f(y)$ with $y \in \mathrm{Dom}(\partial f) \subset \mathrm{dom}(f) = \mathbb{R}_+$ and thus $y \geqslant 0$. Two cases:

- If $y = 0$, then $x \in \partial f(0) =]-\infty, \gamma]$ and thus $x \leqslant \gamma$.
- If $y > 0$, then $x - y \in \partial f(y) = f'(y) = \gamma$ and thus $x = y + \gamma > \gamma$ and $y = x - \gamma$.

From the second bullet, we deduce that, if $x \leqslant \gamma$, then $y \leqslant 0$ and thus $y = 0$ (since $y \geqslant 0$, established before bullets) and thus $y = (x-\gamma)^+$. From the first bullet, we deduce that, if $x > \gamma$, then $y \neq 0$ and thus $y > 0$ (since $y \geqslant 0$, established before bullets) and thus $y = x - \gamma = (x-\gamma)^+$. The proof is complete. Note that another proof is possible by noting that $f = \iota_{\mathbb{R}_+} + \gamma x$ is a linear perturbation (see item (iii) of Proposition 4.5) and thus

$$\forall x \in \mathbb{R}, \quad \mathrm{prox}_f(x) = \mathrm{proj}_{\mathbb{R}_+}(x - \gamma) = (x-\gamma)^+.$$

(v) Idem, several methods are possible (including a simple drawing). For example, note that $f = \iota_{[0,\mu]} + \gamma x$ is a linear perturbation (see item (iii) of Proposition 4.5). Thus we obtain that

$$\forall x \in \mathbb{R}, \quad \mathrm{prox}_f(x) = \mathrm{proj}_{[0,\mu]}(x - \gamma) = \min\{(x-\gamma)^+, \mu\}.$$

(vi) Idem, several methods are possible (including a simple drawing). Let us deal with the case where $\mu > 0$ (the case $\mu = 0$ is trivial). Since f is even, we know that prox_f is odd, satisfying $\mathrm{prox}_f(0) = 0$ and being nonnegative over $\mathbb{R}_+$ (see Remark 4.13). So, let $x > 0$ and $y := \mathrm{prox}_f(x) \geqslant 0$. We have $x - y \in \partial f(y)$ and thus $y \in \mathrm{Dom}(\partial f) \subset \mathrm{dom}(f) = [-\mu, \mu]$. Then $0 \leqslant y \leqslant \mu$ that we decompose in three cases:

- If $y = 0$, then $x \in \partial f(0) = [-\lambda, \lambda]$ and thus $0 < x \leqslant \lambda$.
- If $0 < y < \mu$, then $x - y = f'(y) = \lambda$ and thus $x = y + \lambda \in]\lambda, \lambda + \mu[$ and $y = x - \lambda$.
- If $y = \mu$, then $x - \mu \in \partial f(\mu) = [\lambda, +\infty[$ and thus $x \geqslant \lambda + \mu > \lambda$.

From these three points, we deduce that:

- if $0 < x \leqslant \lambda$, then necessarily we have $y = 0$ and thus $y = \min\{(x-\lambda)^+, \mu)$.
- if $\lambda < x < \lambda + \mu$, then necessarily we have $y = x - \lambda$ and thus $y = \min\{(x-\lambda)^+, \mu)$.
- if $x \geqslant \lambda + \mu$, then necessarily we have $y = \mu$ and thus $y = \min\{(x-\lambda)^+, \mu)$.

Thus we have proved that $\mathrm{prox}_f(0) = 0$ and $\mathrm{prox}_f(x) = \min\{(x-\lambda)^+, \mu\}$ for all $x > 0$. Therefore it is true that $\mathrm{prox}_f(x) = \min\{(|x|-\lambda)^+, \mu\}\mathrm{sign}(x)$ for all $x \geqslant 0$. To conclude, one has just to invoke that prox_f is odd and thus $\mathrm{prox}_f(x) = \mathrm{sign}(x)\mathrm{prox}_f(-x) = \mathrm{sign}(x)\min\{(-x-\lambda)^+, \mu\} = \min\{(|x|-\lambda)^+, \mu\}\mathrm{sign}(x)$ for all $x < 0$.

(vii) Idem, several methods are possible (including a simple drawing). Here we only deal with the case $\lambda > 0$ (the case $\lambda = 0$ has already been treated in item (iv)). Since f is even, we know that prox_f is odd, satisfying $\mathrm{prox}_f(0) = 0$ and being nonnegative over $\mathbb{R}_+$ (see Remark 4.13). So, let $x > 0$ and $y := \mathrm{prox}_f(x) \geqslant 0$. Then we have $x - y \in \partial f(y)$ and three cases:

- If $0 \leqslant y < \lambda$, then $x - y = f'(y) = 0$ and thus $x = y \in [0, \lambda[$.
- If $y > \lambda$, then $x - y = f'(y) = 1$ and thus $x = y + 1 > \lambda + 1$ and $y = x - 1$.
- If $y = \lambda$, then $x - \lambda \in \partial f(\lambda) = [0, 1]$ and thus $x \in [\lambda, \lambda + 1]$.

From these three points, we deduce that:

- if $0 < x < \lambda$, then necessarily we have $y = x$.
- if $x \in [\lambda, \lambda+1]$, then necessarily $y = \lambda$.
- if $x > \lambda + 1$, then necessarily $y = x - 1$.

Thus, since moreover $\mathrm{prox}_f(0) = 0$, we have proved that

$$\forall x \geqslant 0, \quad \mathrm{prox}_f(x) = \begin{cases} x & \text{if} \quad x \leqslant \lambda, \\ \lambda & \text{if} \quad \lambda \leqslant x \leqslant \lambda+1, \\ x-1 & \text{if} \quad \lambda+1 \leqslant x, \end{cases} = \begin{cases} x & \text{if} \quad |x| \leqslant \lambda, \\ \lambda \mathrm{sign}(x) & \text{if} \quad \lambda \leqslant |x| \leqslant \lambda+1, \\ x - \mathrm{sign}(x) & \text{if} \quad \lambda+1 \leqslant |x|. \end{cases}$$

To conclude, one has just to invoke that prox_f is odd and thus

$$\forall x < 0, \quad \mathrm{prox}_f(x) = \mathrm{sign}(x)\mathrm{prox}_f(-x) = \mathrm{sign}(x) \begin{cases} -x & \text{if} \quad -x \leqslant \lambda, \\ \lambda & \text{if} \quad \lambda \leqslant -x \leqslant \lambda+1, \\ -x-1 & \text{if} \quad \lambda+1 \leqslant -x, \end{cases}$$

which gives exactly

$$\forall x < 0, \quad \mathrm{prox}_f(x) = \begin{cases} x & \text{if} \quad |x| \leqslant \lambda, \\ \lambda \mathrm{sign}(x) & \text{if} \quad \lambda \leqslant |x| \leqslant \lambda+1, \\ x - \mathrm{sign}(x) & \text{if} \quad \lambda+1 \leqslant |x|. \end{cases}$$

(viii) We only deal with the case $\lambda > 0$ (since the case $\lambda = 0$ gives the null function whose proximal operator is the identity map). First of all, let us note that f is indeed convex since it is differentiable over $\mathbb{R}$ (even at $-\frac{\lambda}{2}$ and $\frac{\lambda}{2}$) and its derivative is clearly monotone, given by

$$\forall x \in \mathbb{R}, \quad f'(x) = \begin{cases} -2\lambda & \text{if} \quad x \leqslant -\dfrac{\lambda}{2}, \\ 4x & \text{if} \quad -\frac{\lambda}{2} \leqslant x \leqslant \dfrac{\lambda}{2}, \\ 2\lambda & \text{if} \quad x \geqslant \dfrac{\lambda}{2}. \end{cases}$$

To prove this result, a drawing can be done but we are going to prove it mathematically. Since f is even, we know that prox_f is odd, satisfying $\mathrm{prox}_f(0) = 0$ and being nonnegative over $\mathbb{R}_+$ (see Remark 4.13). So, let $x > 0$ and $y := \mathrm{prox}_f(x) \geqslant 0$. Then we have $x - y \in \partial f(y)$ and two cases:

- If $0 \leqslant y \leqslant \frac{\lambda}{2}$, then $x - y = f'(y) = 4y$ and thus $x = 5y \in [0, \frac{5\lambda}{2}]$ and $y = \frac{x}{5}$.
- If $y > \frac{\lambda}{2}$, then $x - y = f'(y) = 2\lambda$ and thus $x = y + 2\lambda > \frac{5\lambda}{2}$ and $y = x - 2\lambda$.

From these two points, we deduce that:

- if $0 < x \leqslant \frac{5\lambda}{2}$, then necessarily we have $y = \frac{x}{5}$.
- if $x > \frac{5\lambda}{2}$, then necessarily $y = x - 2\lambda$.

Thus, since moreover $\mathrm{prox}_f(0) = 0$, we have proved that

$$\forall x \geqslant 0, \quad \mathrm{prox}_f(x) = \begin{cases} \frac{x}{5} & \text{if} \quad x \leqslant \frac{5\lambda}{2}, \\ x - 2\lambda & \text{if} \quad x > \frac{5\lambda}{2}, \end{cases} = \begin{cases} \frac{x}{5} & \text{if} \quad |x| \leqslant \frac{5\lambda}{2}, \\ x - 2\lambda \mathrm{sign}(x) & \text{if} \quad |x| \geqslant \frac{5\lambda}{2}, \end{cases}$$

To conclude, one has just to invoke that prox_f is odd and thus

$$\forall x < 0, \; \mathrm{prox}_f(x) = \mathrm{sign}(x)\mathrm{prox}_f(-x) = \mathrm{sign}(x) \begin{cases} -\frac{x}{5} & \text{if} \quad -x \leqslant \frac{5\lambda}{2}, \\ -x - 2\lambda & \text{if} \quad -x > \frac{5\lambda}{2}, \end{cases} = \begin{cases} \frac{x}{5} & \text{if} \quad |x| \leqslant \frac{5\lambda}{2}, \\ x - 2\lambda \mathrm{sign}(x) & \text{if} \quad |x| \geqslant \frac{5\lambda}{2}. \end{cases}$$

(ix) We only deal with the case $\lambda > 0$ (since the case $\lambda = 0$ is trivial). Note that the case $p = 1$ is not true! Let us prove the result for $p > 1$. A drawing can be done but we are going to prove this result mathematically. Since f is even, we know that prox_f is odd, $\mathrm{prox}_f(0) = 0$ and prox_f is nonnegative over $\mathbb{R}_+$ (see Remark 4.13).

- Let $x > 0$ and $y := \mathrm{prox}_f(x) \geqslant 0$. Note that f is differentiable at y (even if $y = 0$, since $p > 1$) with $f'(y) = \lambda p y^{p-1}$. We deduce that $x = y + f'(y) = y + \lambda p y^{p-1}$ and thus $y \geqslant 0$ is a solution to $y + \lambda p y^{p-1} = x$. Furthermore, it is the unique solution since the map $y \mapsto y + \lambda p y^{p-1}$ is (strictly) increasing. We have obtained that $\mathrm{prox}_f(x) = \rho(x) = \rho(x)\mathrm{sign}(x)$.
- Let $x < 0$. We get that $\mathrm{prox}_f(x) = -\mathrm{prox}_f(-x) = -y$ where $y \geqslant 0$ is the unique solution to $y + \lambda p y^{p-1} = -x = |x|$. We have obtained that $\mathrm{prox}_f(x) = -\rho(x) = \rho(x)\mathrm{sign}(x)$.

The proof is complete.

(x) We have $f = \phi + \mu|\cdot|^2$ is a quadratic perturbation of $\phi = \lambda|\cdot|^p + \sigma_{[-\omega,\omega]}$. From item (iii) of Proposition 4.5 and item (iii) of the present Proposition 4.6 (the differentiability assumption is satisfied since $p > 1$), we obtain that

$$\mathrm{prox}_f(x) = \mathrm{prox}_{\phi/(2\mu+1)}\left(\frac{x}{2\mu+1}\right) = \mathrm{prox}_{\varphi}\left(\frac{x}{2\mu+1}\right) = \mathrm{prox}_{\lambda|\cdot|^p/(2\mu+1)}\left(\mathrm{soft}_{[-\frac{\omega}{2\mu+1},\frac{\omega}{2\mu+1}]}\left(\frac{x}{2\mu+1}\right)\right),$$

where $\varphi = \frac{\phi}{2\mu+1} = \frac{\lambda}{2\mu+1}|\cdot|^p + \sigma_{[-\frac{\omega}{2\mu+1},\frac{\omega}{2\mu+1}]}$ since $\frac{1}{2\mu+1}\sigma_{[-\omega,\omega]} = \sigma_{[-\frac{\omega}{2\mu+1},\frac{\omega}{2\mu+1}]}$ (exercise).

(xi) First of all, note that f is differentiable over $\mathbb{R}$ (even at 0) with $f'(x) = \frac{x}{1+|x|}$ for all $x \in \mathbb{R}$, which is monotone (since continuous over $\mathbb{R}$, and monotone over $\mathbb{R}_-$ and over $\mathbb{R}_+$ by computing the second-order derivative, exercise) and thus f is indeed convex. Since f is even, we know that prox_f is odd, $\mathrm{prox}_f(0) = 0$ and prox_f is nonnegative over $\mathbb{R}_+$ (see Remark 4.13).

- Let $x > 0$ and $y := \mathrm{prox}_f(x) \geqslant 0$. We deduce that $x = y + f'(y) = \frac{y(y+2)}{y+1}$ and thus $y^2 - (x-2)y - x = 0$. By solving, and taking into account that $y \geqslant 0$, we obtain that $y = \frac{x-2+\sqrt{x^2+4}}{2}$.
- Let $x < 0$. We get that $\mathrm{prox}_f(x) = -\mathrm{prox}_f(-x) = \mathrm{sign}(x)\frac{(-x)-2+\sqrt{(-x)^2+4}}{2}$.

The proof is complete.

(xii) Note that, if $\tau = 0$, then f is not closed. In the sequel, consider that $\tau > 0$. Since $\mathrm{dom}(f) = \mathbb{R}_+^*$, we know that $\mathrm{prox}_f : \mathbb{R} \to \mathbb{R}$ has positive values (see Remark 4.13). Let $x \in \mathbb{R}$ and $y := \mathrm{prox}_f(x) > 0$. We have $x = y + f'(y) = y - \frac{\tau}{y}$ and thus $y^2 - xy - \tau = 0$ which gives $y = \frac{1}{2}(x + \sqrt{x^2 + 4\tau})$ (taking into account the fact that $y > 0$).

(xiii) The fact that f is convex comes from item (iv) of Remark 4.13. Note that $\partial f(0) = [f'_-(0), f'_+(0)] = [\frac{1}{\omega_1}, \frac{1}{\omega_2}]$ and thus $\mathrm{prox}_f(x) = 0$ if and only $x \in [\frac{1}{\omega_1}, \frac{1}{\omega_2}]$. Now let $x > \frac{1}{\omega_2}$ and take $y := \mathrm{prox}_f(x) > 0$ (since prox_f is nondecreasing). We also know that $y < \omega_2$ since $\mathrm{Ran}(\mathrm{prox}_f) = \mathrm{Dom}(\partial f) \subset \mathrm{dom}(f) =]\omega_1, \omega_2[$. We have $x = y + f'(y) = y + \frac{1}{\omega_2 - y}$ which gives $y = \frac{1}{2}(x + \omega_2 - \sqrt{(x-\omega_2)^2 + 4})$ (taking into account the fact that $y \in]0, \omega_2[$). The proof is similar for $x < \frac{1}{\omega_1}$. □

Remark 4.14. Consider the framework of item (ix) of Proposition 4.6 with $p = 3$.

(i) We get that $\mathrm{prox}_f(x) = \frac{\mathrm{sign}(x)}{6\lambda}(\sqrt{1 + 12\lambda|x|} - 1)$ for all $x \in \mathbb{R}$.

(ii) Our aim in this item is to determine prox_g when $g = \mu|\cdot|^{2/3}$ for some $\mu > 0$. We have $f^*(y) = (\lambda|\cdot|^3)^*(y) = (3\lambda\frac{1}{3}|\cdot|^3)^*(y) = 3\lambda(\frac{1}{3}|\cdot|^3)^*(\frac{y}{3\lambda}) = 3\lambda(\frac{2}{3}|\cdot|^{3/2})(\frac{y}{3\lambda}) = \frac{1}{\sqrt{3\lambda}}\frac{2}{3}|y|^{3/2}$ for all $y \in \mathbb{R}$. Solving $\frac{1}{\sqrt{3\lambda}}\frac{2}{3} = \mu$, we obtain $\lambda = \frac{4}{27\mu^2}$ and $f^* = g$. Therefore

$$\forall x \in \mathbb{R}, \quad \mathrm{prox}_g(x) = x - \mathrm{prox}_f(x) = x - \frac{\mathrm{sign}(x)}{6\lambda}(\sqrt{1 + 12\lambda|x|} - 1)$$
$$= x - \frac{9\mu^2\mathrm{sign}(x)}{8}\left(\sqrt{1 + \frac{16}{9\mu^2}|x|} - 1\right).$$

Remark 4.15. Consider the function $f \in \Gamma_0(\mathbb{R})$ defined by

$$\forall x \in \mathbb{R}, \quad f(x) := \begin{cases} \lambda x^3 & \text{if} \quad x \geqslant 0, \\ +\infty & \text{if} \quad x < 0, \end{cases}$$

where $\lambda > 0$. Using similar techniques than in the proof of Proposition 4.6, one can prove (exercise) that

$$\forall x \in \mathbb{R}, \quad \operatorname{prox}_f(x) = \frac{1}{6\lambda}(\sqrt{1 + 12\lambda x^+} - 1).$$

Remark 4.16. Consider the separable function $f \in \Gamma_0(\mathbb{R}^d)$ defined by

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad f(x) := \begin{cases} -\lambda \sum_{i=1}^d \ln(x_i) & \text{if} \quad x \in (\mathbb{R}_+^*)^d, \\ +\infty & \text{otherwise}, \end{cases}$$

for some $\lambda > 0$. From Remark 4.9 and item (xii) of Proposition 4.6, we get that

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad \operatorname{prox}_f(x) = \frac{1}{2}\left(x_i + \sqrt{x_i^2 + 4\lambda}\right)_{i=1,\ldots,d}.$$

Remark 4.17. Consider the separable function $f \in \Gamma_0(\mathbb{R}^d)$ defined by

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad f(x) := \begin{cases} \sum_{i=1}^d \lambda_i |x_i| & \text{if} \quad \forall i = 1, \ldots, d, \ |x_i| \leqslant \mu_i, \\ +\infty & \text{otherwise}, \end{cases}$$

for some λ_i, $\mu_i \geqslant 0$ for all $i \in \{1, \ldots, d\}$, which represents a *weigthed one-norm*. From Remark 4.9 and item (vi) of Proposition 4.6, we get that

$$\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad \operatorname{prox}_f(x) = \Big(\min\{(|x_i| - \lambda_i)^+, \mu_i\}\operatorname{sign}(x_i)\Big)_{i=1,\ldots,d}.$$

Remark 4.18. Take $\lambda > 0$. From item (xiv) of Proposition 4.5 and Remark 4.15, one can prove (exercise) that

$$\forall x \in \mathbb{R}^d, \quad \operatorname{prox}_{\lambda\|\cdot\|^3}(x) = \frac{2}{1 + \sqrt{1 + 12\lambda\|x\|}}x.$$

$A: \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is α- cocoercive iff:

$\langle Ax - Ay, x - y \rangle \geqslant \alpha \|Ax - Ay\|^2.$

5 The gradient descent method and the proximal point algorithm

In this section, we introduce two numerical algorithms dedicated to solve convex minimization problems of the form

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\; f(x),$$

where $f \in \Gamma_0(\mathbb{R}^d)$ with $\operatorname{Argmin}(f) \neq \emptyset$.

5.1 Baillon-Haddad theorem and gradient descent method when f is smooth

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function. Then, for any $\lambda > 0$, we have

$$\nabla f(x) = 0 \iff \lambda \nabla f(x) = 0 \iff \boxed{x - \lambda \nabla f(x) = x} \quad \forall \lambda > 0$$

$$\operatorname{Argmin}(f) = \operatorname{Fix}(\mathrm{Id} - \lambda \nabla f).$$

In that context, the *gradient descent method* consists in the fixed-point algorithm given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = x_k - \lambda \nabla f(x_k),$$

for some $\lambda > 0$. The rest of this section is dedicated to establish a convergence result for the gradient descent method.

Lemma 5.1 (Baillon-Haddad theorem)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function and $L > 0$. Then $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz continuous if and only if $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is $\frac{1}{L}$-cocoercive.

Proof. Let us prove the necessary condition in several steps:

(i) Let us prove that
$$\forall x, y \in \mathbb{R}^d, \quad f(y) - f(x) \leqslant \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2.$$
Let x, $y \in \mathbb{R}^d$ and consider $g : [0,1] \to \mathbb{R}$ defned by $g(t) := f(x + t(y - x))$ for all $t \in [0,1]$. Then g is differentiable over $[0,1]$ with $g'(t) = \langle \nabla f(x + t(y - x)), y - x\rangle$ for all $t \in [0,1]$ and thus g' is Lipschitz continuous with the constant $L\|y - x\|^2$. In particular, g is of class C^1. We obtain that
$$\begin{aligned} f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t)\,dt = g'(0) + \int_0^1 g'(t) - g'(0)\,dt \\ \leqslant g'(0) + L\|y - x\|^2 \int_0^1 t\,dt = \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \end{aligned}$$

(ii) For all $x \in \mathbb{R}^d$, we introduce
$$\begin{array}{rccl} h_x : & \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ & z & \longmapsto & h_x(z) := f(z) - f(x) - \langle \nabla f(x), z - x\rangle, \end{array}$$
which is clearly a convex differentiable function with $\nabla h_x(z) = \nabla f(z) - \nabla f(x)$ for all $z \in \mathbb{R}^d$. Note that:

- $\nabla h_x(x) = 0_{\mathbb{R}^d}$ and therefore, from convexity of h_x, we deduce that x is a (global) minimizer of h_x for all $x \in \mathbb{R}^d$, that is
$$\forall x, z \in \mathbb{R}^d, \quad 0 = h_x(x) \leqslant h_x(z).$$
- ∇h_x is also L-Lipschitz continuous for all $x \in \mathbb{R}^d$, and we obtain from the above item (i) that
$$\forall x, y, z \in \mathbb{R}^d, \quad h_x(z) - h_x(y) \leqslant \langle \nabla h_x(y), z - y\rangle + \frac{L}{2}\|z - y\|^2.$$

(iii) Now, let us prove that

$$\forall x, y \in \mathbb{R}^d, \quad f(y) - f(x) \geqslant \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2.$$

Let x, $y \in \mathbb{R}^d$ and consider some $v \in \mathbb{R}^d$ such that $\langle \nabla h_x(y), v \rangle = \|\nabla h_x(y)\|$ (existence in both cases $\nabla h_x(y) \neq 0_{\mathbb{R}^d}$ and $\nabla h_x(y) = 0_{\mathbb{R}^d}$). Taking $z := y - \frac{\|\nabla h_x(y)\|}{L} v$ in the previous item (ii), we obtain that

$$\begin{aligned} 0 = h_x(x) \leqslant h_x(z) \leqslant h_x(y) + \langle \nabla h_x(y), z - y \rangle + \frac{L}{2} \|z - y\|^2 \\ = f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \frac{1}{2L} \|\nabla h_x(y)\|^2 \\ = f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2. \end{aligned}$$

(iv) Now, let us prove that $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is $\frac{1}{L}$-cocoercive. Let x, $y \in \mathbb{R}^d$. From the previous item (iii), we have

$$f(y) - f(x) \geqslant \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2,$$

and (by inverting the roles of x and y)

$$f(x) - f(y) \geqslant \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2.$$

By adding these two inequalities, we obtain that

$$\langle \nabla f(y) - \nabla f(x), x - y \rangle \geqslant \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2,$$

which concludes the proof of the necessary condition.

Finally the sufficient condition is trivial from the Cauchy-Schwarz inequality. □

Remark 5.1. A nonexpansive operator $T : D \to \mathbb{R}^d$ is not firmly nonexpansive in general (take $T = -\mathrm{Id}$ for a counterexample). Nevertheless, from the Baillon-Haddad theorem, note that, if $T = \nabla f$ for some convex differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, then T is nonexpansive if and only if it is firmly nonexpansive.

Proposition 5.1 (Gradient descent method)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function with $\mathrm{Argmin}(f) \neq \emptyset$. Assume that $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz continuous for some $L > 0$. Then, for any $0 < \lambda < \frac{2}{L}$, the *descent gradient method*

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = x_k - \lambda \nabla f(x_k),$$

converges to some point in $\mathrm{Argmin}(f)$.

Proof. From the Baillon-Haddad theorem, we know that $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is $\frac{1}{L}$-cocoercive. From Proposition 2.3, we deduce that $T := \mathrm{Id} - \lambda \nabla f$ is $\frac{\lambda L}{2}$-averaged for all $0 < \lambda < \frac{2}{L}$. To conclude, one has just to invoke the fixed-point algorithm for averaged operators. □

Remark 5.2 (Relaxed gradient descent method)**.** Consider the framework of Proposition 5.1 (and its proof) and let $0 < \lambda < \frac{2}{L}$. Instead of considering the fixed-point algorithm for averaged operators, since the operator $T := \mathrm{Id} - \lambda \nabla f$ is $\frac{\lambda L}{2}$-averaged, one can also consider the KM algorithm for averaged operators given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = x_k + \mu_k (T(x_k) - x_k),$$

with any sequence $(\mu_k)_{k \in \mathbb{N}} \subset [0, \frac{2}{\lambda L}]$ such that $\sum_{k \in \mathbb{N}} \mu_k (\frac{2}{\lambda L} - \mu_k) = +\infty$. Defining $\gamma_k := \lambda \mu_k$ for all $k \in \mathbb{N}$, the *relaxed gradient descent method* given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = x_k - \gamma_k \nabla f(x_k),$$

for any sequence $(\gamma_k)_{k \in \mathbb{N}} \subset [0, \frac{2}{L}]$ such that $\sum_{k \in \mathbb{N}} \gamma_k (\frac{2}{L} - \gamma_k) = +\infty$, converges to some point in $\mathrm{Argmin}(f)$. Note that the gradient descent method corresponds to the particular case $\gamma_k := \lambda$ for all $k \in \mathbb{N}$.

5.2 The proximal point algorithm and comments

Let $f \in \Gamma_0(\mathbb{R}^d)$. Then, for any $\lambda > 0$, we have

$$\operatorname{Argmin}(f) = \operatorname{Fix}(\operatorname{prox}_{\lambda f}).$$

In that context, the *proximal point algorithm* consists in the fixed-point algorithm given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = \operatorname{prox}_{\lambda f}(x_k),$$

for some $\lambda > 0$.

Proposition 5.2 (Proximal point algorithm)**.** Let $f \in \Gamma_0(\mathbb{R}^d)$ with $\operatorname{Argmin}(f) \neq \emptyset$. Then, for any $\lambda > 0$, the proximal point algorithm

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = \operatorname{prox}_{\lambda f}(x_k),$$

converges to some point in $\operatorname{Argmin}(f)$.

Proof. Trivial since $\operatorname{prox}_{\lambda f} : \mathbb{R}^d \to \mathbb{R}^d$ is firmly nonexpansive. □

Remark 5.3 (First interpretation)**.** For $f \in \Gamma_0(\mathbb{R}^d)$ and $\lambda > 0$, note that

$$\forall x \in \mathbb{R}^d, \quad \operatorname{prox}_{\lambda f}(x) = \operatorname{Argmin}\left(\lambda f + \frac{1}{2}\|\cdot - x\|^2\right) = \operatorname{Argmin}\left(f + \frac{1}{2\lambda}\|\cdot - x\|^2\right).$$

Hence, for any $x \in \mathbb{R}^d$, the term $\operatorname{prox}_{\lambda f}(x)$ can be seen as a compromise between minimizing f and being near to x. The parameter λ can be interpreted as a relative weight between these two terms. Note that the larger the parameter λ is, then the closer to the minimum of f the term $\operatorname{prox}_{\lambda f}(x)$ is. Conversely, the smaller the parameter λ is, then the closer to x the term $\operatorname{prox}_{\lambda f}(x)$ is. In particular, recall that $\operatorname{prox}_{\lambda f}(x)$ tends to x when $\lambda \to 0^+$ (see Theorem 1.2).

Remark 5.4 (Second interpretation)**.** For $f \in \Gamma_0(\mathbb{R}^d)$ and $\lambda > 0$, note that

$$\forall x \in \mathbb{R}^d, \quad \operatorname{prox}_{\lambda f}(x) = x - \operatorname{prox}_{(\lambda f)^*}(x) = x - \nabla \mathcal{M}_{\lambda f}(x) = x - \lambda \nabla \widetilde{\mathcal{M}}_{\lambda f}(x),$$

where $\widetilde{\mathcal{M}}_{\lambda f} := \frac{1}{\lambda}\mathcal{M}_{\lambda f}$. As a consequence, the proximal point algorithm can be seen as a gradient descent method applied to the convex differentiable map $\widetilde{\mathcal{M}}_{\lambda f} : \mathbb{R}^d \to \mathbb{R}$ which satisfies

$$\operatorname{Argmin}(\widetilde{\mathcal{M}}_{\lambda f}) = \operatorname{Argmin}\left(\frac{1}{\lambda}\mathcal{M}_{\lambda f}\right) = \operatorname{Argmin}(\mathcal{M}_{\lambda f}) = \operatorname{Argmin}(\lambda f) = \operatorname{Argmin}(f).$$

We refer to Remark 4.4 for the third equality.

Remark 5.5 (Third interpretation)**.** Let us consider the smooth case where $f : \mathbb{R}^d \to \mathbb{R}$ is a convex differentiable function with $\operatorname{Argmin}(f) \neq \emptyset$. In that context

$$x \in \operatorname{Argmin}(f) \quad \Longleftrightarrow \quad x \text{ is an equilibrium of the differential equation } \underbrace{\dot{y}(t) = -\nabla f(y(t))}_{\text{Gradient flow}},\ t \geqslant 0.$$

Therefore (and roughly speaking):

(i) If a point $x \in \operatorname{Argmin}(f)$ is an attractive equilibrium of the differential equation and the *forward Euler algorithm* (also called *explicit Euler algorithm*) given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \frac{x_{k+1} - x_k}{\varepsilon} = -\nabla f(x_k),$$

for some $\varepsilon > 0$, provides a good approximation of the unique solution $y(\cdot)$ to the differential equation over $\mathbb{R}_+$ and starting at $y(0) = x_0$, one can expect to get the convergence of the sequence $(x_k)_{k\in\mathbb{N}}$ to the point x. In that context the above algorithm can be rewritten as the gradient descent method

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = x_k - \varepsilon \nabla f(x_k).$$

This is the reason why each step of a gradient descent method is usually called a *forward step* in the literature.

(ii) If a point $x \in \operatorname{Argmin}(f)$ is an attractive equilibrium of the differential equation and the *backward Euler algorithm* (also called *implicit Euler algorithm*) given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \frac{x_{k+1} - x_k}{\varepsilon} = -\nabla f(x_{k+1}),$$

for some $\varepsilon > 0$, provides a good approximation of the unique solution $y(\cdot)$ to the differential equation over $\mathbb{R}_+$ and starting at $y(0) = x_0$, one can expect to get the convergence of the sequence $(x_k)_{k\in\mathbb{N}}$ to the point x. In that context the above algorithm can be rewritten as the proximal point algorithm

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = \operatorname{prox}_{\varepsilon f}(x_k).$$

This is the reason why each step of a proximal point algorithm is usually called a *backward step* in the literature.

Remark 5.6. Of course, for some $f \in \Gamma_0(\mathbb{R}^d)$ with $\operatorname{Argmin}(f) \neq \emptyset$, one of the worst drawbacks of the proximal point algorithm in order to solve the convex minimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f(x),$$

is that it requires to know explicitly the proximal operator $\operatorname{prox}_{\lambda f} : \mathbb{R}^d \to \mathbb{R}^d$, for some $\lambda > 0$, or at least to know how to numerically compute it in a easy way. Fortunately, the literature is very developed on this topic and many examples (such as those seen in Propositions 4.5 and 4.6 and the following remarks) have already been treated.

Remark 5.7 (Relaxed proximal point algorithm)**.** Consider the framework of Proposition 5.2 and let $\lambda > 0$. Since the operator $\operatorname{prox}_{\lambda f} : \mathbb{R}^d \to \mathbb{R}^d$ is firmly nonexpansive (that is, $\frac{1}{2}$-averaged), one can also consider the KM algorihtm for averaged operators. Therefore, the *relaxed proximal point algorithm*

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = x_k + \mu_k(\operatorname{prox}_{\lambda f}(x_k) - x_k),$$

for any sequence $(\mu_k)_{k\in\mathbb{N}} \subset [0,2]$ such that $\sum_{k\in\mathbb{N}} \mu_k(2-\mu_k) = +\infty$, converges to some point in $\operatorname{Argmin}(f)$. Note that the proximal point algorithm corresponds to the particular case $\mu_k := 1$ for all $k \in \mathbb{N}$.

LASSO: $\min_x \boxed{\underbrace{\frac{1}{2}\|Ax-b\|^2}_{g(x)} + \underbrace{\lambda\|x\|_1}_{f(x)}}$

smooth nonsmooth

Gradient descent operator:

$G_{\lambda,f}(x) = x - \lambda\nabla f(x) \leftarrow \boxed{G_{\lambda f} = \mathrm{Id} - \lambda\nabla f}$

$\operatorname{prox}_{\lambda f} = (\mathrm{Id} + \lambda\partial f)^{-1}$

6 Splitting methods for convex optimization

In the literature, a *proximal algorithm* is any algorithm that involves the proximal operator of a given function, such as the proximal point algorithm seen in the previous section. In practice, it is very frequent that a convex minimization problem is written as

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f(x)+g(x),$$

where f, $g\in\Gamma_0(\mathbb{R}^d)$. Unfortunately, even if the proximal operators prox_f and prox_g are separately known (or easy to numerically compute, separately), it is not true that prox_{f+g} is easy to compute in general. In particular, one has to note that

$$\mathrm{prox}_{f+g}\neq \mathrm{prox}_f\circ\mathrm{prox}_g,$$

in general (take $f=g=\frac{1}{2}\|\cdot\|^2$ for a counterexample). There exist some sufficient conditions on f and g ensuring that $\mathrm{prox}_{f+g}=\mathrm{prox}_f\circ\mathrm{prox}_g$ (such as $\partial g\subset\partial g\circ\mathrm{prox}_f$ for example, exercise) but, unfortunately, they are not satisfied in practice. The aim of this section is to propose proximal algorithms that handle convex minimization problems of the above form, by using informations on f and g (separately) and without using the proximal operator prox_{f+g}. These methods are well-known in the literature as *splitting methods*.

6.1 When g is smooth: proximal gradient algorithm (also called forward-backward algorithm)

Lemma 6.1. Let $f\in\Gamma_0(\mathbb{R}^d)$ and $g:\mathbb{R}^d\to\mathbb{R}$ be a convex differentiable function. Then, for any $\lambda>0$, we have

$$\mathrm{Argmin}(f+g)=\mathrm{Fix}\Big(\mathrm{prox}_{\lambda f}\circ(\mathrm{Id}-\lambda\nabla g)\Big).$$

Proof. In this proof, we will use Theorem E.5 at several occasions. Note that $\mathrm{dom}(f)\cap\mathrm{int}(\mathrm{dom}(g))\neq\emptyset$ and let $\lambda>0$. Let us prove the necessary condition. If $x\in\mathrm{Argmin}(f+g)$, then $x\in\mathrm{Argmin}(\lambda f+\lambda g)$ and thus the sum $\lambda f+\lambda g$ is subdifferentiable at x with $0_{\mathbb{R}^d}\in\partial(\lambda f+\lambda g)(x)$. Since $\mathrm{dom}(\lambda f)\cap\mathrm{int}(\mathrm{dom}(\lambda g))=\mathrm{dom}(f)\cap\mathrm{int}(\mathrm{dom}(g))\neq\emptyset$, we know that λf and λg are subdifferentiable at x with $\partial(\lambda f+\lambda g)(x)=\partial(\lambda f)(x)+\partial(\lambda g)(x)$. We deduce that $0_{\mathbb{R}^d}\in\partial(\lambda f)(x)+\partial(\lambda g)(x)=\partial(\lambda f)+\lambda\nabla g(x)$. We deduce that $x-\lambda\nabla g(x)\in x+\partial(\lambda f)(x)$ and thus $x=\mathrm{prox}_{\lambda f}(x-\lambda\nabla g(x))$.

Conversely, from $x=\mathrm{prox}_{\lambda f}(x-\lambda\nabla g(x))$, one can find similarly that $0_{\mathbb{R}^d}\in\partial(\lambda f)+\lambda\nabla g(x)=\partial(\lambda f)(x)+\partial(\lambda g)(x)\subset\partial(\lambda f+\lambda g)(x)$ and thus $x\in\mathrm{Argmin}(\lambda f+\lambda g)=\mathrm{Argmin}(f+g)$. $\square$

Let $f\in\Gamma_0(\mathbb{R}^d)$ and $g:\mathbb{R}^d\to\mathbb{R}$ be a convex differentiable function. According to Lemma 6.1, the *proximal gradient algorithm* (also called *forward-backward algorithm*) consists in the fixed-point algorithm given by

$$x_0\in\mathbb{R}^d\qquad\text{and}\qquad\forall k\in\mathbb{N},\quad x_{k+1}=\mathrm{prox}_{\lambda f}(x_k-\lambda\nabla g(x_k)),$$

for some $\lambda>0$.

Proposition 6.1 (Proximal gradient algorithm (forward-backward algorithm))**.** Let $f\in\Gamma_0(\mathbb{R}^d)$ and $g:\mathbb{R}^d\to\mathbb{R}$ be a convex differentiable function with $\mathrm{Argmin}(f+g)\neq\emptyset$. Assume that $\nabla g:\mathbb{R}^d\to\mathbb{R}^d$ is L-Lipschitz continuous for some $L>0$. Then, for any $0<\lambda<\frac{2}{L}$, the *proximal gradient algorithm* (also called *forward-backward algorithm*)

$$x_0\in\mathbb{R}^d\qquad\text{and}\qquad\forall k\in\mathbb{N},\quad x_{k+1}=\mathrm{prox}_{\lambda f}(x_k-\lambda\nabla g(x_k)).$$

converges to some point in $\mathrm{Argmin}(f+g)$.

$$\min_x f(x) + g(x)$$

$$0 \in \partial (f+g)(x) \iff 0 \in \nabla g(x) + \partial f(x)$$

$$\iff 0 \in \lambda \nabla g(x) + \lambda \partial f(x), \quad \lambda > 0$$

$$\iff x \in \lambda \nabla g(x) + x + \lambda \partial f(x)$$

$$\iff x - \lambda \nabla g(x) \in (Id + \lambda \partial f)(x)$$

$$\iff \left[\mathrm{prox}_{\lambda f} \circ \underbrace{(Id - \lambda \nabla g)}_{G_{\lambda f}}\right](x) = x$$

$$\mathrm{Argmin}\,(f+g) = \mathrm{Fix}\left(\mathrm{prox}_{\lambda f} \circ G_{\lambda f}\right) \quad \underline{\forall \lambda > 0}$$

$$\begin{cases} x_0 \\ x_{k+1} = \mathrm{prox}_{\lambda f}\left(x_k - \lambda \nabla g(x_k)\right) \end{cases}$$

$\begin{cases} \text{proximal gradient method} \\ \text{Forward-Backward method} \end{cases}$

<u>Applicat°</u> to LASSO

$$f(x) = \|\cdot\|_1$$
$$g(x) = \frac{1}{2}\|Ax - b\|^2$$

$$\nabla g(x) = A^T A x - A^T b$$

$$\mathrm{prox}_{\lambda f} = \mathrm{prox}_{\lambda \|\cdot\|_1} =: \text{soft thresholding operator!} = S_\lambda(\cdot)$$

$\boxed{\text{ISTA}} \longleftarrow$

$$S_\lambda(x) = \left(S_\lambda(x_1), S_\lambda(x_2), \dots, S_\lambda(x_n)\right)$$

$$S_\lambda(x_i) = \left(|x_i| - \lambda\right)^+ \mathrm{sg}(x_i) \quad i = 1, \dots, n.$$

ISTA ← Iterative Shrinkage-Thresholding Algorithm

Proof. Let $0 < \lambda < \frac{2}{L}$ and $T := T_1 \circ T_2$ with $T_1 := \mathrm{prox}_{\lambda f}$ that is firmly nonexpansive (and thus $\frac{1}{2}$-averaged) and $T_2 := \mathrm{Id} - \lambda \nabla g$ that is $\frac{\lambda L}{2}$-averaged (as in the proof of Proposition 5.1). Furthermore, from Lemma 6.1, we have $\mathrm{Fix}(T) = \mathrm{Argmin}(f+g) \neq \emptyset$. From the fixed-point algorithm for composition of averaged operators (see Corollary 2.1), we obtain that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to some point in $\mathrm{Fix}(T) = \mathrm{Argmin}(f+g)$. $\square$

Remark 6.1. Consider the framework of Proposition 6.1. Note that the numerical implementation of the proximal gradient algorithm requires (only) the knowledge of $\mathrm{prox}_{\lambda f}$ and ∇g (or at least that these two operators are easy to compute numerically). Also note that:

- when $f \equiv 0$, we recover the gradient descent method.
- when $g \equiv 0$, we recover the proximal point algorithm.
- when $f := \iota_C$ for some nonempty closed convex set C of $\mathbb{R}^d$, we recover the classical *projected gradient algorithm* given by
$$x_0 \in \mathbb{R}^d \quad \text{and} \quad \forall k \in \mathbb{N}, \quad x_{k+1} = \mathrm{proj}_C(x_k - \lambda \nabla g(x_k)),$$
which allows to solve the constrained convex minimization problem
$$\underset{x\in C}{\text{minimize}}\; g(x).$$

The forward-backward algorithm has been widely used in the 2000's in the case where $f := \|\cdot\|_1$ is the one-norm of $\mathbb{R}^d$. In that context, it was called *Iterative Shrinkage Thresholding Algorithm (ISTA)*. There exists now an accelerated version of this algorithm called *Fast Iterative Shrinkage Thresholding Algorithm (FISTA)* but this is beyond the scope of this lecture.

Remark 6.2 (Relaxed proximal gradient algorithm)**.** Consider the framework of Proposition 6.1 (and its proof) and let $0 < \lambda < \frac{2}{L}$. By composition (see Proposition 2.6), we do know that $T := T_1 \circ T_2$ is κ-averaged with $\kappa := \frac{1}{\frac{1}{2}+\min(1,\frac{1}{\lambda L})} \in \,]0,1[$. Therefore, one can also consider the KM algorihtm for averaged operators. Hence the *relaxed proximal gradient algorithm*
$$x_0 \in \mathbb{R}^d \quad \text{and} \quad \forall k \in \mathbb{N}, \quad x_{k+1} = x_k + \mu_k\Big(\mathrm{prox}_{\lambda f}(x_k - \lambda \nabla g(x_k)) - x_k\Big),$$
for any sequence $(\mu_k)_{k\in\mathbb{N}} \subset [0, \frac{1}{\kappa}]$ such that $\sum_{k\in\mathbb{N}} \mu_k(\frac{1}{\kappa} - \mu_k) = +\infty$, converges to some point in $\mathrm{Argmin}(f+g)$. Note that the proximal gradient algorithm corresponds to the particular case $\mu_k := 1$ for all $k \in \mathbb{N}$.

Remark 6.3 (Backward-backward algorithm)**.** For $f, g \in \Gamma_0(\mathbb{R}^d)$, we have $\mathrm{Argmin}(f + \mathcal{M}_g) = \mathrm{Fix}(\mathrm{prox}_f \circ \mathrm{prox}_g)$ (exercise). As a consequence, if $\mathrm{Argmin}(f+\mathcal{M}_g) \neq \emptyset$, then the *backward-backward algorithm*
$$x_0 \in \mathbb{R}^d \quad \text{and} \quad \forall k \in \mathbb{N}, \quad x_{k+1} = \mathrm{prox}_f(\mathrm{prox}_g(x_k)),$$
converges to some point in $\mathrm{Argmin}(f+\mathcal{M}_g)$ (exercise), and not to a point in $\mathrm{Argmin}(f+g)$ in general.

6.2 Douglas-Rachford algorithm and Peaceman-Rachford algorithm

Lemma 6.2. Let $f, g \in \Gamma_0(\mathbb{R}^d)$ with $\mathrm{dom}(f) \cap \mathrm{int}(\mathrm{dom}(g)) \neq \emptyset$. Then, for any $\lambda > 0$ and any $\rho \in \mathbb{R}^*$, we have
$$x \in \mathrm{Argmin}(f+g) \iff \exists y \in \mathbb{R}^d,\; x = \mathrm{prox}_{\lambda f}(y) \;\text{ and }\; y = y + \rho\Big(\mathrm{prox}_{\lambda g}(\mathrm{refl}_{\lambda f}(y)) - \mathrm{prox}_{\lambda f}(y)\Big),$$
where $\mathrm{refl}_{\lambda f} : \mathbb{R}^d \to \mathbb{R}^d$ stands for the nonexpansive reflexion operator $\mathrm{refl}_{\lambda f} := \mathrm{refl}_{\mathrm{prox}_{\lambda f}} = 2\mathrm{prox}_{\lambda f} - \mathrm{Id}$ (for the ease of notation).

Continue with Thibault Liard!

Proof. In this proof, we will use Theorem E.5 at several occasions. Let us prove the necessary condition. If $x \in \operatorname{Argmin}(f+g)$, then $x \in \operatorname{Argmin}(\lambda f+\lambda g)$ and thus the sum $\lambda f+\lambda g$ is subdifferentiable at x with $0_{\mathbb{R}^d} \in \partial(\lambda f+\lambda g)(x)$. Since $\operatorname{dom}(\lambda f) \cap \operatorname{int}(\operatorname{dom}(\lambda g)) = \operatorname{dom}(f) \cap \operatorname{int}(\operatorname{dom}(g)) \neq \emptyset$, we know that λf and λg are subdifferentiable at x with $\partial(\lambda f+\lambda g)(x) = \partial(\lambda f)(x)+\partial(\lambda g)(x)$. We deduce that $0_{\mathbb{R}^d} \in \partial(\lambda f)(x)+\partial(\lambda g)(x)$ and thus there exists $z \in \mathbb{R}^d$ such that $z \in \partial(\lambda f)(x)$ and $-z \in \partial(\lambda g)(x)$. Defining $y := x + z$, we find $y - x \in \partial(\lambda f)(x)$, and thus $x = \operatorname{prox}_{\lambda f}(y)$, and that $x - y \in \partial(\lambda g)(x)$, and thus $x = \operatorname{prox}_{\lambda g}(2x - y)$. Therefore, replacing x in the last equality, we get $\operatorname{prox}_{\lambda f}(y) = \operatorname{prox}_{\lambda g}(\operatorname{refl}_{\lambda f}(y))$ and thus $y = y + \rho(\operatorname{prox}_{\lambda g}(\operatorname{refl}_{\lambda f}(y)) - \operatorname{prox}_{\lambda f}(y))$.

Conversely, from $x = \operatorname{prox}_{\lambda f}(y)$ and $y = y + \rho(\operatorname{prox}_{\lambda g}(\operatorname{refl}_{\lambda f}(y)) - \operatorname{prox}_{\lambda f}(y))$, one can find similarly that $y - x \in \partial(\lambda f)(x)$ and $x - y \in \partial(\lambda g)(x)$. We deduce that $0_{\mathbb{R}^d} \in \partial(\lambda f)(x) + \partial(\lambda g)(x) \subset \partial(\lambda f + \lambda g)(x)$ and thus $x \in \operatorname{Argmin}(\lambda f + \lambda g) = \operatorname{Argmin}(f + g)$. □

Remark 6.4. From the proof of Lemma 6.2, one can easily see that the assumption $\operatorname{dom}(f) \cap \operatorname{int}(\operatorname{dom}(g)) \neq \emptyset$ can be replaced by $\operatorname{int}(\operatorname{dom}(f)) \cap \operatorname{dom}(g) \neq \emptyset$ (without inversing the role of f and g). The same remark holds for all statements hereafter.

Remark 6.5. Consider the framework of Lemma 6.2 and let $\lambda > 0$. Two particular well-known cases:

(i) Taking $\rho = 2$, Lemma 6.2 leads to

$$\operatorname{Argmin}(f + g) = \operatorname{prox}_{\lambda f}\Big(\operatorname{Fix}(\operatorname{refl}_{\lambda g} \circ \operatorname{refl}_{\lambda f})\Big).$$

The nonexpansive operator $\operatorname{refl}_{\lambda g} \circ \operatorname{refl}_{\lambda f}$ is known as the *Peaceman-Rachford operator*.

(ii) Taking $\rho = 1$, Lemma 6.2 leads to

$$\operatorname{Argmin}(f + g) = \operatorname{prox}_{\lambda f}\left(\operatorname{Fix}\left(\frac{1}{2}(\operatorname{Id} + \operatorname{refl}_{\lambda g} \circ \operatorname{refl}_{\lambda f})\right)\right).$$

The firmly nonexpansive operator $\frac{1}{2}(\operatorname{Id} + \operatorname{refl}_{\lambda g} \circ \operatorname{refl}_{\lambda f})$ is known as the *Douglas-Rachford operator*.

Proposition 6.2 (Douglas-Rachford algorithm)**.** Let $f, g \in \Gamma_0(\mathbb{R}^d)$ with $\operatorname{dom}(f) \cap \operatorname{int}(\operatorname{dom}(g)) \neq \emptyset$ and $\operatorname{Argmin}(f + g) \neq \emptyset$. Then, for any $\lambda > 0$, the *Douglas-Rachford algorithm*

$$y_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad y_{k+1} = \frac{y_k + \operatorname{refl}_{\lambda g}(\operatorname{refl}_{\lambda f}(y_k))}{2}.$$

converges to some point $y \in \operatorname{Fix}(\frac{1}{2}(\operatorname{Id} + \operatorname{refl}_{\lambda g} \circ \operatorname{refl}_{\lambda f}))$ and $x := \operatorname{prox}_{\lambda f}(y) \in \operatorname{Argmin}(f + g)$.

Proof. Since $\operatorname{Argmin}(f + g) \neq \emptyset$, we deduce from Lemma 6.2 that $\operatorname{Fix}(T) \neq \emptyset$ where $T := \frac{1}{2}(\operatorname{Id} + \operatorname{refl}_{\lambda g} \circ \operatorname{refl}_{\lambda f})$. The Douglas-Rachford algorithm corresponds to the fixed-point algorithm for the $\frac{1}{2}$-averaged operator T. From Theorem 2.4, the algorithm converges to some point $y \in \operatorname{Fix}(T)$ and $x := \operatorname{prox}_{\lambda f}(y) \in \operatorname{Argmin}(f + g)$ from Lemma 6.2. □

Remark 6.6 (Many different presentations of the Douglas-Rachford algorithm)**.** Consider the framework of Proposition 6.2 and let $\lambda > 0$. The Douglas-Rachford algorithm appears in many forms in the literature:

- Developing the reflexion operator, the algorithm

$$y_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad y_{k+1} = y_k + \operatorname{prox}_{\lambda g}\Big(2\operatorname{prox}_{\lambda f}(y_k) - y_k\Big) - \operatorname{prox}_{\lambda f}(y_k),$$

converges to some point y and $x := \operatorname{prox}_{\lambda f}(y) \in \operatorname{Argmin}(f + g)$.

- Introducing the variable $x_{k+1} := \mathrm{prox}_{\lambda f}(y_k)$, the algorithm

$$y_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \begin{cases} x_{k+1} = \mathrm{prox}_{\lambda f}(y_k), \\ y_{k+1} = y_k + \mathrm{prox}_{\lambda g}(2x_{k+1} - y_k) - x_{k+1}, \end{cases}$$

provides a sequence $(x_k)_{k\in\mathbb{N}^*}$ converging to some point in $\mathrm{Argmin}(f+g)$.

- Inverting the updates, we get the algorithm

$$y_0,\, z_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \begin{cases} y_{k+1} = y_k + \mathrm{prox}_{\lambda g}(2z_k - y_k) - z_k. \\ z_{k+1} = \mathrm{prox}_{\lambda f}(y_{k+1}). \end{cases}$$

In that case, y_1 plays the role of y_0 and z_k plays the role of x_k for all $k \geqslant 1$. We get that $(z_k)_{k\in\mathbb{N}}$ converges to some point in $\mathrm{Argmin}(f+g)$.

- Introducing the variable $u_{k+1} := \mathrm{prox}_{\lambda g}(2z_k - y_k)$ for all $k \in \mathbb{N}$, we obtain

$$y_0,\, z_0 \in \mathbb{R}^d, \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \begin{cases} u_{k+1} = \mathrm{prox}_{\lambda g}(2z_k - y_k), \\ y_{k+1} = y_k + u_{k+1} - z_k, \\ z_{k+1} = \mathrm{prox}_{\lambda f}(y_k + u_{k+1} - z_k). \end{cases}$$

In that context too, the sequence $(z_k)_{k\in\mathbb{N}}$ converges to some point in $\mathrm{Argmin}(f+g)$.

- Using the bijective change of variable $(v, w) := (z, z - y)$, we obtain

$$v_0,\, w_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \begin{cases} u_{k+1} = \mathrm{prox}_{\lambda g}(v_k + w_k), \\ v_{k+1} = \mathrm{prox}_{\lambda f}(u_{k+1} - w_k), \\ w_{k+1} = w_k + v_{k+1} - u_{k+1}. \end{cases}$$

With this algorithm, the sequence $(v_k)_{k\in\mathbb{N}}$ converges to some point in $\mathrm{Argmin}(f+g)$.

- Introducing the change of variable $t = -w$, we obtain

$$v_0,\, t_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \begin{cases} u_{k+1} = \mathrm{prox}_{\lambda g}(v_k - t_k), \\ v_{k+1} = \mathrm{prox}_{\lambda f}(u_{k+1} + t_k), \\ t_{k+1} = t_k + u_{k+1} - v_{k+1}. \end{cases}$$

With this algorithm, the sequence $(v_k)_{k\in\mathbb{N}}$ converges to some point in $\mathrm{Argmin}(f+g)$.

As one can see, there is a lot of possibilities to write the Douglas-Rachford algorithm. Sometimes, it can be tricky to recognize it in the literature.

Proposition 6.3 ((Relaxed) Peaceman-Rachford algorithm)**.** Let $f,\, g \in \Gamma_0(\mathbb{R}^d)$ with $\mathrm{dom}(f) \cap \mathrm{int}(\mathrm{dom}(g)) \neq \emptyset$ and $\mathrm{Argmin}(f+g) \neq \emptyset$. Then, for any $\lambda > 0$, the *(relaxed) Peaceman-Rachford algorithm*

$$\forall k \in \mathbb{N}, \quad y_{k+1} = y_k + \mu_k(\mathrm{refl}_{\lambda g} \circ \mathrm{refl}_{\lambda f}(y_k) - y_k),$$

for any sequence $(\mu_k)_{k\in\mathbb{N}} \subset [0,1]$ such that $\sum_{k\in\mathbb{N}} \mu_k(1-\mu_k) = +\infty$, converges to some point $y \in \mathrm{Fix}(\mathrm{refl}_{\lambda g} \circ \mathrm{refl}_{\lambda f})$ and $x := \mathrm{prox}_{\lambda f}(y) \in \mathrm{Argmin}(f+g)$.

Proof. Since $\mathrm{Argmin}(f+g) \neq \emptyset$, we deduce from Lemma 6.2 that $\mathrm{Fix}(T) \neq \emptyset$ where $T := \mathrm{refl}_{\lambda g} \circ \mathrm{refl}_{\lambda f}$ is the nonexpansive Peaceman-Rachford operator. The (relaxed) Peaceman-Rachford algorithm corresponds to the KM algorithm for T. From Theorem 2.2, the algorithm converges to some point $y \in \mathrm{Fix}(T)$ and thus $x := \mathrm{prox}_{\lambda f}(y) \in \mathrm{Argmin}(f+g)$ from Lemma 6.2. $\square$

Remark 6.7 (Douglas-Rachford algorithm is a particular case of the (relaxed) Peaceman-Rachford algorithm)**.** Consider the framework of Propositions 6.2 and 6.3 and let $\lambda > 0$. Note that the Douglas-Rachford algorithm is a particular case of the (relaxed) Peaceman-Rachford algorithm by taking $\mu_k := \frac{1}{2}$ for all $k \in \mathbb{N}$.

Remark 6.8 (Relaxed Douglas-Rachford algorithm gives exactly (relaxed) Peaceman-Rachford algorithm)**.** Consider the framework of Propositions 6.2 and 6.3 and let $\lambda > 0$. Since the Douglas-Rachfor operator $\frac{1}{2}(\mathrm{Id} + \mathrm{refl}_{\lambda g} \circ \mathrm{refl}_{\lambda f})$ is $\frac{1}{2}$-averaged, one could be interested in applying the KM algorithm for averaged operators. However this leads exactly to the (relaxed) Peaceman-Rachford algorithm (exercise).

Remark 6.9. Consider the framework of Propositions 6.2 and 6.3 and let $\lambda > 0$. Note that the numerical implementation of the Douglas-Rachford algorithm or of the (relaxed) Peaceman-Rachford algorithm, and also the final evaluation $x := \mathrm{prox}_{\lambda f}(y)$, require (only) the knowledge of $\mathrm{prox}_{\lambda f}$ and of $\mathrm{prox}_{\lambda g}$ (or at least that these operators are both easy to compute numerically).

Remark 6.10 (Composite Douglas-Rachford algorithm)**.** In this remark, we consider the composite convex minimization problem given by

$$\underset{x \in \mathbb{R}^d}{\mathrm{minimize}}\, f(x) + g(Mx),$$

where $f \in \Gamma_0(\mathbb{R}^d)$, $g \in \Gamma_0(\mathbb{R}^n)$ and $M \in \mathbb{R}^{n \times d}$ is a given matrix. Assume that there exists $x_0 \in \mathrm{int}(\mathrm{dom}(f))$ such that $Mx_0 \in \mathrm{int}(\mathrm{dom}(g))$ and consider some $\lambda > 0$. If g is not differentiable and M has no particular form or property, then, since one cannot a priori express $\mathrm{prox}_{\lambda(g \circ M)}$ easily, one cannot apply directly the Douglas-Rachford algorithm, nor the forward-backward algorithm (even if f is differentiable). One idea is to rewrite the problem as follows

$$\underset{(x,y) \in \mathbb{R}^{d+n}}{\mathrm{minimize}}\, h(x,y) + \iota_C(x,y),$$

where $h \in \Gamma_0(\mathbb{R}^{d+n})$ is defined by $h(x,y) := f(x) + g(y)$ for all $(x,y) \in \mathbb{R}^{d+n}$ and $C := \{(x,y) \in \mathbb{R}^{d+n} \mid Mx = y\}$ is an affine subset of $\mathbb{R}^{d+n}$. Then, one can apply a Douglas-Rachford algorithm on this new problem. Indeed:

- Since h is a separable function, we have $\mathrm{prox}_{\lambda h}(x,y) = (\mathrm{prox}_{\lambda f}(x), \mathrm{prox}_{\lambda g}(y))$ for all $(x,y) \in \mathbb{R}^{d+n}$.
- Furthermore, proj_C can be easily expressed since C is an affine subset of $\mathbb{R}^{d+n}$ (see Proposition 3.1, note that the full range property is trivially satisfied).
- It holds that $\mathrm{int}(\mathrm{dom}(h)) \cap \mathrm{dom}(\iota_C) = \mathrm{int}(\mathrm{dom}(h)) \cap C \neq \emptyset$ since it contains (x_0, Mx_0).

The resulting algorithm

$$z_0 = (z_0^d, z_0^n) \in \mathbb{R}^{d+n} \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \begin{cases} x_{k+1} = \mathrm{prox}_{\lambda f}(z_k^d), \\ y_{k+1} = \mathrm{prox}_{\lambda g}(z_k^n), \\ z_{k+1} = z_k + \mathrm{proj}_C(2(x_{k+1}, y_{k+1}) - z_k) - (x_{k+1}, y_{k+1}), \end{cases}$$

is called the *composite Douglas-Rachford algorithm* and provides a sequence $(x_k, y_k)_{k \in \mathbb{N}^*}$ converging to some point in $\mathrm{Argmin}(h + \iota_C)$ and thus the sequence $(x_k)_{k \in \mathbb{N}^*}$ converges to some point in $\mathrm{Argmin}(f + g \circ M)$.

6.3 Sum of three convex functions: the example of the Davis-Yin algorithm

Lemma 6.3. Let f, $g \in \Gamma_0(\mathbb{R}^d)$, with $\mathrm{dom}(f) \cap \mathrm{int}(\mathrm{dom}(g)) \neq \emptyset$, and $h : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function. Then, for any $\lambda > 0$, we have

$$\mathrm{Argmin}(f + g + h) = \mathrm{prox}_{\lambda f}(\mathrm{Fix}(\mathcal{T}_{f,g,h})),$$

where

$$\mathcal{T}_{f,g,h} := \mathrm{Id} - \mathrm{prox}_{\lambda f} + \mathrm{prox}_{\lambda g} \circ \Big(2\mathrm{prox}_{\lambda f} - \mathrm{Id} - \lambda \nabla h \circ \mathrm{prox}_{\lambda f}\Big).$$

Proof. In this proof, we will use Theorem E.5 at several occasions. Let us prove the necessary condition. If $x \in \operatorname{Argmin}(f+g+h)$, then $x \in \operatorname{Argmin}(\lambda f + \lambda g + \lambda h)$ and thus the sum $\lambda f + \lambda g + \lambda h$ is subdifferentiable at x with $0_{\mathbb{R}^d} \in \partial(\lambda f + \lambda g + \lambda h)(x)$. Since $\operatorname{dom}(\lambda f + \lambda g) \cap \operatorname{int}(\operatorname{dom}(\lambda h)) = \operatorname{dom}(f+g) \cap \operatorname{int}(\operatorname{dom}(h)) = \operatorname{dom}(f+g) = \operatorname{dom}(f) \cap \operatorname{dom}(g) \neq \emptyset$, we know that $\lambda f + \lambda g$ and λh are subdifferentiable at x with $\partial(\lambda f + \lambda g + \lambda h)(x) = \partial(\lambda f + \lambda g)(x) + \lambda \nabla h(x)$. Since $\operatorname{dom}(\lambda f) \cap \operatorname{int}(\operatorname{dom}(\lambda g)) = \operatorname{dom}(f) \cap \operatorname{int}(\operatorname{dom}(g)) \neq \emptyset$, we obtain that λf and λg are subdifferentiable at x with $\partial(\lambda f + \lambda g)(x) = \partial(\lambda f)(x) + \partial(\lambda g)(x)$. We finally obtain that

$$0_{\mathbb{R}^d} \in \partial(\lambda f)(x) + \partial(\lambda g)(x) + \lambda \nabla h(x).$$

Take $y \in \partial(\lambda f)(x)$ and $z \in \partial(\lambda g)(x)$ such that $y+z+\lambda\nabla h(x) = 0_{\mathbb{R}^d}$. Define $r := x+y$ which satisfies $x = \operatorname{prox}_{\lambda f}(r)$. Therefore, our aim is to prove that $r \in \operatorname{Fix}(\mathcal{T}_{f,g,h})$. Note that

$$2\operatorname{prox}_{\lambda f}(r) - r - \lambda\nabla h(\operatorname{prox}_{\lambda f}(r)) = 2x - (x+y) - \lambda \nabla h(x) = x - y - \lambda\nabla h(x) = x + z \in x + \partial(\lambda g)(x),$$

and thus

$$x = \operatorname{prox}_{\lambda g}\Big(2\operatorname{prox}_{\lambda f}(r) - r - \lambda\nabla h(\operatorname{prox}_{\lambda f}(r))\Big).$$

We deduce that $\mathcal{T}_{f,g,h}(r) = r - \operatorname{prox}_{\lambda f}(r) + x = r - x + x = r$ which concludes the necessary condition.

Conversely, assume that $x = \operatorname{prox}_{\lambda f}(r)$ with $r \in \operatorname{Fix}(\mathcal{T}_{f,g,h})$. Since

$$r = \mathcal{T}_{f,g,h}(r) = r - \operatorname{prox}_{\lambda f}(r) + \operatorname{prox}_{\lambda g}(w),$$

where $w := 2\operatorname{prox}_{\lambda f}(r) - r - \lambda\nabla h(\operatorname{prox}_{\lambda f}(r))$, we deduce that $x = \operatorname{prox}_{\lambda f}(r) = \operatorname{prox}_{\lambda g}(w)$. We obtain that $w \in x + \partial(\lambda g)(x)$ and $r - x \in \partial(\lambda f)(x)$. Since moreover $w = 2x - r - \lambda\nabla h(x) \in x + \partial(\lambda g)(x)$, we obtain that $0_{\mathbb{R}^d} \in \partial(\lambda f)(x) + \partial(\lambda g) + \lambda \nabla h(x) \subset \partial(\lambda f + \lambda g + \lambda h)(x)$ and thus $x \in \operatorname{Argmin}(\lambda f + \lambda g + \lambda h) = \operatorname{Argmin}(f+g+h)$. $\square$

Lemma 6.4. Let U, $V : \mathbb{R}^d \to \mathbb{R}^d$ be firmly nonexpansive operators and $S : \mathbb{R}^d \to \mathbb{R}^d$ be an operator. Denote by $T := U + (V \circ S)$ and $W := \mathrm{Id} - (2U + S)$. Then, it holds that

$$\forall x_1, x_2 \in \mathbb{R}^d, \quad \|T(x_2) - T(x_1)\|^2 \leqslant \|x_2 - x_1\|^2 - \|(\mathrm{Id} - T)(x_2) - \mathrm{Id} - T)(x_1)\|^2 \\ - 2\langle (V\circ S)(x_2) - (V \circ S)(x_1), W(x_2) - W(x_1)\rangle.$$

Proof. Let x_1, $x_2 \in \mathbb{R}^d$. From firm nonexpansiveness of U and V (for obtaining the inequality below), since $2U+S = \mathrm{Id} - W$ and $T = U + (V\circ S)$, it holds that

$$\begin{aligned}
&\|T(x_2) - T(x_1)\|^2 \\
&= \|U(x_2) - U(x_1)\|^2 + \|(V\circ S)(x_2) - (V\circ S)(x_1)\|^2 + 2\langle (V\circ S)(x_2) - (V\circ S)(x_1), U(x_2) - U(x_1)\rangle \\
&\leqslant \langle U(x_2) - U(x_1), x_2 - x_1\rangle + \langle (V\circ S)(x_2) - (V\circ S)(x_1), S(x_2) - S(x_1)\rangle + 2\langle (V\circ S)(x_2) - (V\circ S)(x_1), U(x_2) - U(x_1)\rangle \\
&= \langle U(x_2) - U(x_1), x_2 - x_1\rangle + \langle (V\circ S)(x_2) - (V\circ S)(x_1), (x_2 - x_1) - (W(x_2) - W(x_1))\rangle \\
&= \langle T(x_2) - T(x_1), x_2 - x_1\rangle - \langle (V\circ S)(x_2) - (V\circ S)(x_1), W(x_2) - W(x_1)\rangle.
\end{aligned}$$

To conclude, one has just to see that

$$\langle T(x_2) - T(x_1), x_2 - x_1\rangle = \frac{1}{2}\Big(\|T(x_2) - T(x_1)\|^2 + \|x_2 - x_1\|^2 - \|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2\Big),$$

and thus

$$\begin{aligned}
&\|T(x_2) - T(x_1)\|^2 - \langle T(x_2) - T(x_1), x_2 - x_1\rangle \\
&= \frac{1}{2}\Big(\|T(x_2) - T(x_1)\|^2 - \|x_2 - x_1\|^2 + \|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2\Big),
\end{aligned}$$

which concludes the proof. $\square$

Lemma 6.5. Let f, $g \in \Gamma_0(\mathbb{R}^d)$, with $\mathrm{dom}(f) \cap \mathrm{int}(\mathrm{dom}(g)) \neq \emptyset$, and $h : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function. Assume that $\nabla h : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz continuous for some $L > 0$. Then, for any $0 < \lambda < \frac{2}{L}$, the operator

$$\mathcal{T}_{f,g,h} := \mathrm{Id} - \mathrm{prox}_{\lambda f} + \mathrm{prox}_{\lambda g} \circ \Big(2\mathrm{prox}_{\lambda f} - \mathrm{Id} - \lambda \nabla h \circ \mathrm{prox}_{\lambda f}\Big),$$

is α-averaged with $\alpha := \frac{2}{4-\lambda L} \in]\frac{1}{2}, 1[$.

Proof. Let $0 < \lambda < \frac{2}{L}$ and let x_1, $x_2 \in \mathbb{R}^d$. To use Lemma 6.4, let us denote by

$$\begin{aligned} U &:= \mathrm{Id} - \mathrm{prox}_{\lambda f}, \\ V &:= \mathrm{prox}_{\lambda g}, \\ S &:= 2\mathrm{prox}_{\lambda f} - \mathrm{Id} - \lambda \nabla h \circ \mathrm{prox}_{\lambda f}, \\ W &:= \lambda \nabla h \circ \mathrm{prox}_{\lambda f}. \end{aligned}$$

Note that U and V are firmly nonexpansive, that $W = \mathrm{Id} - (2U + S)$ and $\mathcal{T}_{f,g,h} = T := U + V \circ S$. From Lemma 6.4 it holds that

$$\|T(x_2) - T(x_1)\|^2 \leqslant \|x_2 - x_1\|^2 - \|(\mathrm{Id} - T)(x_2) - \mathrm{Id} - T)(x_1)\|^2 - 2\langle (V \circ S)(x_2) - (V \circ S)(x_1), W(x_2) - W(x_1)\rangle.$$

Let us compute the last term. Since $-V \circ S = U - T = \mathrm{Id} - \mathrm{prox}_{\lambda f} - T$ and $W = \lambda \nabla h \circ \mathrm{prox}_{\lambda f}$, we get that

$$\begin{aligned} &- 2\langle (V \circ S)(x_2) - (V \circ S)(x_1), W(x_2) - W(x_1)\rangle \\ &\qquad = 2\langle (\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1), \lambda \nabla h(\mathrm{prox}_{\lambda f}(x_2)) - \lambda \nabla h(\mathrm{prox}_{\lambda f}(x_1))\rangle \\ &\qquad\qquad - 2\langle \mathrm{prox}_{\lambda f}(x_2) - \mathrm{prox}_{\lambda f}(x_1), \lambda \nabla h(\mathrm{prox}_{\lambda f}(x_2)) - \lambda \nabla h(\mathrm{prox}_{\lambda f}(x_1))\rangle. \end{aligned}$$

Using Young inequality (see Lemma A.1) on the first term and using the $\frac{1}{L}$-cocoercivity of ∇h (from Baillon-Haddad theorem and since ∇h is L-Lipschitz) on the second term, we obtain that

$$\begin{aligned} &- 2\langle (V \circ S)(x_2) - (V \circ S)(x_1), W(x_2) - W(x_1)\rangle \\ &\qquad \leqslant \delta \|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2 + \frac{\lambda^2}{\delta} \|\nabla h(\mathrm{prox}_{\lambda f}(x_2)) - \nabla h(\mathrm{prox}_{\lambda f}(x_1))\|^2 \\ &\qquad\qquad - \frac{2\lambda}{L} \|\nabla h(\mathrm{prox}_{\lambda f}(x_2)) - \nabla h(\mathrm{prox}_{\lambda f}(x_1))\|^2 \\ &\qquad = \delta \|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2 + \left(\frac{\lambda^2}{\delta} - \frac{2\lambda}{L}\right) \|\nabla h(\mathrm{prox}_{\lambda f}(x_2)) - \nabla h(\mathrm{prox}_{\lambda f}(x_1))\|^2, \end{aligned}$$

for any $\delta > 0$. Taking $\delta := \frac{\lambda L}{2} > 0$ (so that $\frac{\lambda^2}{\delta} - \frac{2\lambda}{L} = 0$), we deduce that

$$\begin{aligned} \|T(x_2) - T(x_1)\|^2 &\leqslant \|x_2 - x_1\|^2 - \|(\mathrm{Id} - T)(x_2) - \mathrm{Id} - T)(x_1)\|^2 + \frac{\lambda L}{2} \|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2 \\ &= \|x_2 - x_1\|^2 + \left(\frac{\lambda L}{2} - 1\right) \|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2 \\ &= \|x_2 - x_1\|^2 + \left(1 - \frac{1}{\alpha}\right) \|(\mathrm{Id} - T)(x_2) - (\mathrm{Id} - T)(x_1)\|^2, \end{aligned}$$

with $\alpha := \frac{2}{4-\lambda L} \in]\frac{1}{2}, 1[$, which concludes the proof from Proposition 2.2. □

Proposition 6.4 (Davis-Yin algorithm)**.** Let $f,\ g \in \Gamma_0(\mathbb{R}^d)$, with $\mathrm{dom}(f) \cap \mathrm{int}(\mathrm{dom}(g)) \neq \emptyset$, and $h : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function. Assume that $\mathrm{Argmin}(f+g+h) \neq \emptyset$ and that $\nabla h : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz continuous for some $L > 0$. Then, for any $0 < \lambda < \frac{2}{L}$, the *Davis-Yin algorithm* given by

$$y_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad y_{k+1} = \mathcal{T}_{f,g,h}(y_k),$$

where

$$\mathcal{T}_{f,g,h} := \mathrm{Id} - \mathrm{prox}_{\lambda f} + \mathrm{prox}_{\lambda g} \circ \Big(2\mathrm{prox}_{\lambda f} - \mathrm{Id} - \lambda \nabla h \circ \mathrm{prox}_{\lambda f} \Big),$$

converges to some point $y \in \mathrm{Fix}(\mathcal{T}_{f,g,h})$ and $x := \mathrm{prox}_{\lambda f}(y) \in \mathrm{Argmin}(f+g+h)$.

Proof. Since $\mathrm{Argmin}(f+g+h) \neq \emptyset$, we deduce from Lemma 6.3 that $\mathrm{Fix}(\mathcal{T}_{f,g,h}) \neq \emptyset$. From Lemma 6.5 and Theorem 2.4, we know that the sequence $(y_k)_{k\in\mathbb{N}}$ converges to some point $y \in \mathrm{Fix}(\mathcal{T}_{f,g,h})$. Finally, from Lemma 6.3, we get that $x := \mathrm{prox}_{\lambda f}(y) \in \mathrm{Argmin}(f+g+h)$. $\square$

Remark 6.11. Consider the framework of Proposition 6.4 and let $0 < \lambda < \frac{2}{L}$. A few remarks:

(i) Note that, if $f \equiv 0$, then the Davis-Yin algorithm is nothing else but the forward-backward algoritm.

(ii) Note that, if $g \equiv 0$, then the Davis-Yin algorithm is nothing else but the forward-backward algoritm (by inverting the updates).

(iii) Note also, if $h \equiv 0$, then the Davis-Yin algorithm is nothing else but the Douglas-Rachford algoritm.

(iv) Since the operator $\mathcal{T}_{f,g,h}$ is averaged, one can also propose a relaxed version of the Davis-Yin algorithm by using the KM algorihtm for averaged operators.

Remark 6.12 (Tseng algorithm)**.** In the literature, there are a lot of splitting methods for a large panorama of convex minimization problems. For example, let us discuss the classical *Tseng algorithm*. Let $f \in \Gamma_0(\mathbb{R}^d)$, $g : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function and C be a nonempty closed convex subset of $\mathbb{R}^d$. Assume that $C \cap \mathrm{Argmin}(f+g) \neq \emptyset$ and that $\nabla g : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz continuous for some $L > 0$. Then, for any $0 < \lambda < \frac{1}{L}$, the *Tseng algorithm* given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = \mathcal{T}_{f,g,C}(x_k),$$

where

$$\mathcal{T}_{f,g,C} := \mathrm{proj}_C \circ \Big(\mathrm{Id} - (\mathrm{Id} - \lambda \nabla g) + (\mathrm{Id} - \lambda \nabla g) \circ \mathrm{prox}_{\lambda f} \circ (\mathrm{Id} - \lambda \nabla g) \Big) \Big),$$

converges to some point $x \in C \cap \mathrm{Argmin}(f+g)$. Be careful, in that context, we do not have that $\mathcal{T}_{f,g,C}$ is averaged in general, nor that $C \cap \mathrm{Argmin}(f+g) = \mathrm{Fix}(\mathcal{T}_{f,g,C})$ in general. The proof of the above convergence result is more involved (and beyond the scope of this lecture). Note that the Tseng algorithm involves a forward-backward-forward step. On the other hand, note that $C \cap \mathrm{Argmin}(f+g) \subset \mathrm{Argmin}(f+g+\iota_C)$ but the reverse is not true in general and thus the Davis-Yin algorithm cannot be used to find a point in $C \cap \mathrm{Argmin}(f+g)$. On the contrary, the Tseng algorithm can be used to find a point in $\mathrm{Argmin}(f+g+\iota_C)$. Taking $f \equiv 0$, the Tseng algorithm constitutes an alternative to the projected gradient algorithm.

Chapter 3: Lagrangian duality and primal-dual algorithms

7 Preliminaries

7.1 A proper separation theorem

In the literature on separation of convex sets, several separation notions are very well-known, such as the *large separation*, the *strict separation* or the *strong separation* (see Theorem C.2). In this section, we will discuss the notion of *proper separation* in which at least one of the two separated convex sets is not included in the separating affine hyperplane.

Lemma 7.1 (Proper separation lemma). If R is a nonempty convex subset of $\mathbb{R}^d$ such that $0_{\mathbb{R}^d} \notin R$, then there exists $a \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\}$ such that

$$\forall x \in R, \quad \langle a, x \rangle \geqslant 0 \qquad \text{and} \qquad \exists x' \in R, \quad \langle a, x' \rangle > 0.$$

We say that the linear hyperplane $a^\perp = [\langle a, \cdot \rangle = 0]$ *properly separates* R and $\{0_{\mathbb{R}^d}\}$.

Proof. In this proof, we distinguish several cases:

(i) We first deal with the case $\operatorname{aff}(R) = \mathbb{R}^d$. In that case $\operatorname{int}(R) = \operatorname{rint}(R) \neq \emptyset$. Since $0_{\mathbb{R}^d} \notin R$, we have $0_{\mathbb{R}^d} \notin \operatorname{int}(R)$ and thus we can strictly separate $\{0_{\mathbb{R}^d}\}$ and the nonempty open convex set $\operatorname{int}(R)$. We obtain that there exists $a \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\}$ such that

$$\forall x \in \operatorname{int}(R), \quad \langle a, x \rangle > \langle a, 0_{\mathbb{R}^d} \rangle = 0.$$

Since R is convex with $\operatorname{int}(R) \neq \emptyset$, we have $\operatorname{clos}(R) = \operatorname{clos}(\operatorname{int}(R))$ and thus $\langle a, x \rangle \geqslant 0$ for all $x \in \operatorname{clos}(R)$. Finally, we have proved that

$$\forall x \in R, \quad \langle a, x \rangle \geqslant 0 \qquad \text{and} \qquad \exists x' \in R, \quad \langle a, x' \rangle > 0,$$

by taking some $x' \in \operatorname{int}(R) \neq \emptyset$.

(ii) Now let us deal with the case $\operatorname{aff}(R) \subsetneq \mathbb{R}^d$ and denote by $p := \dim(R) \in \{0, \dots, d-1\}$. We denote by

$$\operatorname{aff}(R) = r_0 + \operatorname{vect}\{r_1, \dots, r_p\},$$

where $r_0 \in \operatorname{aff}(R)$ and $\{r_1, \dots, r_p\}$ is a linearly independent family. Now we distinguish two subcases:

(a) If $0_{\mathbb{R}^d} \notin \operatorname{aff}(R)$, then $r_0 \neq 0_{\mathbb{R}^d}$ and one can easily prove that $\{r_0, r_1, \dots, r_p\}$ is a linearly independent family (exercise) that we can complete to obtain a basis $\{r_0, r_1, \dots, r_{d-1}\}$ of $\mathbb{R}^d$. We define the affine hyperplane

$$H := r_0 + \operatorname{vect}\{r_1, \dots, r_{d-1}\},$$

which contains R (since it contains $\operatorname{aff}(R)$) and does not contain $0_{\mathbb{R}^d}$ (since $\{r_0, r_1, \dots, r_{d-1}\}$ is a basis of $\mathbb{R}^d$). The translation $V := H - r_0$ is a linear hyperplane and we introduce $a \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\}$ such that $a^\perp = V$. Since $0_{\mathbb{R}^d} \notin H$ and $H = r_0 + a^\perp$, we have $r_0 \notin a^\perp$ and thus $\langle a, r_0 \rangle \neq 0$. Up to change the sign of a, we choose $a \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\}$ such that $\langle a, r_0 \rangle > 0$. Thus, for all $x \in R \subset H = a^\perp + r_0$, we have $\langle a, x \rangle = \langle a, r_0 \rangle > 0$ which concludes the proof of this case.

(b) If $0_{\mathbb{R}^d} \in \operatorname{aff}(R)$, then we take $r_0 = 0_{\mathbb{R}^d}$ and $\operatorname{aff}(R) = \operatorname{vect}\{r_1, \dots, r_p\}$ is a linear space. In that case, we can apply the first case (i) by replacing the linear space $\mathbb{R}^d$ by the linear space $\mathcal{S} := \operatorname{aff}(R)$. We can therefore construct a $\mathcal{S}$-linear hyperplane $\tilde{V}$ (of dimension $p-1$) such that R is contained in a $\mathcal{S}$-half-space associated with $\tilde{V}$ but not contained entirely in $\tilde{V}$. Then one has just to introduce the $\mathbb{R}^d$-linear hyperplane $V := \tilde{V} + \mathcal{S}^\perp$ to obtain that R is contained in a $\mathbb{R}^d$-half-space associated with V but not contained entirely in V.

The proof is complete. $\square$

Theorem 7.1 (A proper separation theorem)**.** If K is a nonempty convex subset of $\mathbb{R}^d$ such that $K \cap \mathbb{R}^d_- = \emptyset$, then there exists $b \in \mathbb{R}^d_+ \backslash \{0_{\mathbb{R}^d}\}$ such that

$$\forall x \in K, \quad \langle b, x \rangle \geqslant 0 \qquad \text{and} \qquad \exists \bar{x} \in K, \quad \langle b, \bar{x} \rangle > 0.$$

We say that the linear hyperplane $b^\perp = [\langle b, \cdot \rangle = 0]$ *properly separates* K and $\mathbb{R}^d_-$.

Proof. Consider a maximal subset $\mathcal{I} \subset \{1, \ldots, d\}$ such that

$$\forall i \in \mathcal{I}, \quad \exists \lambda_i > 0, \quad \forall x = (x_1, \ldots, x_d) \in K, \quad \sum_{i \in \mathcal{I}} \lambda_i x_i = 0.$$

Note that $\mathcal{I}$ may be empty or the whole set $\{1, \ldots, d\}$. We also introduce

$$T := \{t = (t_1, \ldots, t_d) \in \mathbb{R}^d_+ \mid \forall i \in \mathcal{I}, \ t_i = 0\} \quad \text{and} \quad R := K + T.$$

Note that R is nonempty (since K and T are nonempty sets), is convex (since K and T are convex) and that $0_{\mathbb{R}^d} \notin R$ (equivalent to $K \cap (-T) = \emptyset$ which is true since $(-T) \subset \mathbb{R}^d_-$ and $K \cap \mathbb{R}^d_- = \emptyset$). We deduce from Lemma 7.1 that there exists $a = (a_1, \ldots, a_d) \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\}$ such that

$$\forall x = (x_1, \ldots, x_d) \in K, \quad \forall t = (t_1, \ldots, t_d) \in T, \quad \sum_{i \in \mathcal{I}} a_i x_i + \sum_{i \notin \mathcal{I}} a_i (x_i + t_i) \geqslant 0,$$

and

$$\exists x' = (x'_1, \ldots, x'_d) \in K, \quad \exists t' = (t'_1, \ldots, t'_d) \in T, \quad \sum_{i \in \mathcal{I}} a_i x'_i + \sum_{i \notin \mathcal{I}} a_i (x'_i + t'_i) > 0.$$

First, note that $a_i \geqslant 0$ for all $i \notin \mathcal{I}$ (otherwise taking $t_i \to +\infty$ would raise a contradiction) and define $b_i := a_i \geqslant 0$ for all $i \notin \mathcal{I}$. Secondly, take $\gamma > 0$ sufficiently large so that $b_i := a_i + \gamma \lambda_i > 0$ for all $i \in \mathcal{I}$. Hence we obtain $b = (b_1, \ldots, b_d) \in \mathbb{R}^d_+$ and, from the above property satisfied by the λ_i, we obtain that

$$\forall x = (x_1, \ldots, x_d) \in K, \quad \forall t = (t_1, \ldots, t_d) \in T, \quad \sum_{i \in \mathcal{I}} b_i x_i + \sum_{i \notin \mathcal{I}} b_i (x_i + t_i) \geqslant 0,$$

and

$$\exists x' = (x'_1, \ldots, x'_d) \in K, \quad \exists t' = (t'_1, \ldots, t'_d) \in T, \quad \sum_{i \in \mathcal{I}} b_i x'_i + \sum_{i \notin \mathcal{I}} b_i (x'_i + t'_i) > 0.$$

Taking $t = 0_{\mathbb{R}^d}$, we obtain that

$$\forall x \in K, \quad \langle b, x \rangle \geqslant 0.$$

Now, by contradiction, assume that

$$\forall x \in K, \quad \langle b, x \rangle = 0.$$

We have two cases:

(i) if $b_{i_0} > 0$ for some $i_0 \notin \mathcal{I}$, then we contradict the maximality of $\mathcal{I}$.

(ii) if $b_i = 0$ for all $i \notin \mathcal{I}$, we obtain that $\langle b, x' \rangle = \sum_{i \in \mathcal{I}} b_i x'_i + \sum_{i \notin \mathcal{I}} b_i (x'_i + t'_i) > 0$ which contradicts the absurd assumption.

Finally, we have proved that

$$\exists \bar{x} \in K, \quad \langle b, \bar{x} \rangle > 0,$$

and in particular that $b \in \mathbb{R}^d_+ \backslash \{0_{\mathbb{R}^d}\}$. $\square$

Remark 7.1. In Theorem 7.1, one can replace $\mathbb{R}^d_-$ by any convex polyhedron. The proof remains identical.

7.2 Lagrangian duality in a general context

In the whole present section, we consider two general nonempty sets $\mathcal{A}$ and $\mathcal{B}$ and a general function $\mathcal{L} : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ which is called the *Lagrangian function*.

Definition 7.1 (Primal problem and dual problem)**.** The *primal problem* associated with the Lagrangian function $\mathcal{L} : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is the minimization problem defined by

$$\underset{a \in \mathcal{A}}{\text{minimize}} \left(\sup_{b \in \mathcal{B}} \mathcal{L}(a,b) \right), \tag{$\mathcal{P}$}$$

whose solutions (if they exist) are called *primal solutions*, and the corresponding *primal value* is defined by

$$\mathrm{val}(\mathcal{P}) := \inf_{a \in \mathcal{A}} \left(\sup_{b \in \mathcal{B}} \mathcal{L}(a,b) \right) \in \mathbb{R} \cup \{\pm\infty\}.$$

The *dual problem* associated with the Lagrangian function $\mathcal{L} : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is the maximization problem defined by

$$\underset{b \in \mathcal{B}}{\text{maximize}} \left(\inf_{a \in \mathcal{A}} \mathcal{L}(a,b) \right), \tag{$\mathcal{D}$}$$

whose solutions (if they exist) are called *dual solutions*, and the corresponding *dual value* is defined by

$$\mathrm{val}(\mathcal{D}) := \sup_{b \in \mathcal{B}} \left(\inf_{a \in \mathcal{A}} \mathcal{L}(a,b) \right) \in \mathbb{R} \cup \{\pm\infty\}.$$

Proposition 7.1 (Weak duality inequality and duality gap)**.** Consider the framework of Definition 7.1. The so-called *weak duality inequality*

$$\mathrm{val}(\mathcal{D}) \leqslant \mathrm{val}(\mathcal{P}),$$

holds true. The *duality gap* is defined (when it has a sense) as the nonnegative difference $\mathrm{val}(\mathcal{P}) - \mathrm{val}(\mathcal{D}) \geqslant 0$.

Proof. It is clear that

$$\forall a' \in \mathcal{A}, \quad \forall b \in \mathcal{B}, \quad \inf_{a \in \mathcal{A}} \mathcal{L}(a,b) \leqslant \mathcal{L}(a',b),$$

and thus, passing to the supremum over $b \in \mathcal{B}$, we get that

$$\forall a' \in \mathcal{A}, \quad \sup_{b \in \mathcal{B}} \left(\inf_{a \in \mathcal{A}} \mathcal{L}(a,b) \right) \leqslant \sup_{b \in \mathcal{B}} \mathcal{L}(a',b),$$

and finally, passing to the infimum over $a' \in \mathcal{A}$, we conclude the proof. □

Definition 7.2 (Saddle-point of the Lagrangian function)**.** A point $(a^*, b^*) \in \mathcal{A} \times \mathcal{B}$ is called a *saddle-point* of the Lagrangian function $\mathcal{L} : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ if

$$\forall (a,b) \in \mathcal{A} \times \mathcal{B}, \quad \mathcal{L}(a^*,b) \leqslant \mathcal{L}(a^*,b^*) \leqslant \mathcal{L}(a,b^*).$$

Remark 7.2. Note that a point $(a^*, b^*) \in \mathcal{A} \times \mathcal{B}$ is a saddle-point of the Lagrangian function $\mathcal{L} : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ if and only if

$$\sup_{b \in \mathcal{B}} \mathcal{L}(a^*,b) \leqslant \mathcal{L}(a^*,b^*) \leqslant \inf_{a \in \mathcal{A}} \mathcal{L}(a,b^*).$$

In that case, note that a^* is a solution to the *partial primal problem* given by

$$\underset{a \in \mathcal{A}}{\text{minimize}}\ \mathcal{L}(a,b^*),$$

and b^* is a solution to the *partial dual problem* given by

$$\underset{b\in\mathcal{B}}{\text{maximize}}\ \mathcal{L}(a^*,b).$$

Proposition 7.2. Consider the framework of Definition 7.1 and let $(a^*,b^*)\in\mathcal{A}\times\mathcal{B}$. The following two properties are equivalent:

(i) (a^*,b^*) is a saddle-point of $\mathcal{L}$.

(ii) a^* is a primal solution, b^* is a dual solution and $\text{val}(\mathcal{P})=\text{val}(\mathcal{D})$.

In that case, $\text{val}(\mathcal{P})=\text{val}(\mathcal{D})=\mathcal{L}(a^*,b^*)\in\mathbb{R}$.

Proof. Let us prove the sufficient condition. Since a^* and b^* are respectively primal solution and dual solution, and since $\text{val}(\mathcal{P})=\text{val}(\mathcal{D})$, we obtain that

$$\inf_{a\in\mathcal{A}}\mathcal{L}(a,b^*)=\text{val}(\mathcal{D})=\text{val}(\mathcal{P})=\sup_{b\in\mathcal{B}}\mathcal{L}(a^*,b).$$

We deduce that

$$\forall (a',b')\in\mathcal{A}\times\mathcal{B},\quad \mathcal{L}(a^*,b')\leqslant\sup_{b\in\mathcal{B}}\mathcal{L}(a^*,b)=\inf_{a\in\mathcal{A}}\mathcal{L}(a,b^*)\leqslant\mathcal{L}(a^*,b^*)\leqslant\sup_{b\in\mathcal{B}}\mathcal{L}(a^*,b)=\inf_{a\in\mathcal{A}}\mathcal{L}(a,b^*)\leqslant\mathcal{L}(a',b^*),$$

and thus (a^*,b^*) is a saddle-point of $\mathcal{L}$. Now, let us prove the necessary condition. Since (a^*,b^*) is a saddle-point of $\mathcal{L}$, we have

$$\sup_{b\in\mathcal{B}}\mathcal{L}(a^*,b)\leqslant\mathcal{L}(a^*,b^*)\leqslant\inf_{a\in\mathcal{A}}\mathcal{L}(a,b^*).$$

From the weak duality inequality, we get that

$$\mathcal{L}(a^*,b^*)\leqslant\inf_{a\in\mathcal{A}}\mathcal{L}(a,b^*)\leqslant\sup_{b\in\mathcal{B}}\left(\inf_{a\in\mathcal{A}}\mathcal{L}(a,b)\right)=\text{val}(\mathcal{D})\leqslant\text{val}(\mathcal{P})=\inf_{a\in\mathcal{A}}\left(\sup_{b\in\mathcal{B}}\mathcal{L}(a,b)\right)\leqslant\sup_{b\in\mathcal{B}}\mathcal{L}(a^*,b)\leqslant\mathcal{L}(a^*,b^*).$$

In particular, we deduce that $\text{val}(\mathcal{P})=\text{val}(\mathcal{D})=\mathcal{L}(a^*,b^*)\in\mathbb{R}$. Furthermore, we have obtained

$$\sup_{b\in\mathcal{B}}\mathcal{L}(a^*,b)=\inf_{a\in\mathcal{A}}\left(\sup_{b\in\mathcal{B}}\mathcal{L}(a,b)\right)\quad\text{and}\quad\inf_{a\in\mathcal{A}}\mathcal{L}(a,b^*)=\sup_{b\in\mathcal{B}}\left(\inf_{a\in\mathcal{A}}\mathcal{L}(a,b)\right).$$

Thus a^* is a primal solution and b^* is a dual solution. □

7.3 Lagrangian duality for general inequality constrained minimization problems

Consider the general *inequality constrained minimization problem* given by

$$\begin{aligned}&\underset{x\in\mathbb{R}^d}{\text{minimize}} && f(x),\\ &\text{subject to} && g_i(x)\leqslant 0,\quad \forall i\in\{1,\dots,p\},\\ & && x\in C,\end{aligned}\tag{$\mathfrak{P}$}$$

where $f:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is a proper (convex or not) function, where $g_i:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is a proper (convex or not) function for all $i\in\{1,\dots,p\}$, for some $p\in\mathbb{N}^*$, and where C is a nonempty (convex or not) subset of $\mathbb{R}^d$ such that $C\subset\text{dom}(f)\cap(\cap_{i=1}^p\text{dom}(g_i))$. The general idea of *Lagrangian dualization* in inequality constrained optimization theory is to rewrite the above problem as a primal problem. To this aim, we introduce the suitable *Lagrangian function* defined by

$$\begin{aligned}\mathfrak{L}:\quad C\times\mathbb{R}^p_+ &\longrightarrow \mathbb{R}\\ (x,\lambda) &\longmapsto \mathfrak{L}(x,y):=f(x)+\sum_{i=1}^p\lambda_i g_i(x).\end{aligned}$$

Note that

$$\forall x \in C, \quad \sup_{\lambda \in \mathbb{R}_+^p} \mathfrak{L}(x,y) = \begin{cases} f(x) & \text{if} \quad g_i(x) \leqslant 0, \quad \forall i \in \{1,\dots,p\}, \\ +\infty & \text{otherwise,} \end{cases}$$

and, therefore, observe that Problem ($\mathfrak{P}$) can be seen as the primal problem

$$\underset{x \in C}{\text{minimize}} \left(\sup_{\lambda \in \mathbb{R}_+^p} \mathfrak{L}(x,\lambda) \right). \tag{$\mathfrak{P}$}$$

The corresponding dual problem is given by

$$\underset{\lambda \in \mathbb{R}_+^p}{\text{maximize}} \left(\inf_{x \in C} \mathfrak{L}(x,\lambda) \right). \tag{$\mathfrak{D}$}$$

From Proposition 7.1, the weak duality inequality $\text{val}(\mathfrak{D}) \leqslant \text{val}(\mathfrak{P})$ holds true. From Proposition 7.2, we get that a point $(x^*,\lambda^*) \in C \times \mathbb{R}_+^p$ is a saddle point of $\mathfrak{L}$ if and only if x^* is a primal solution, λ^* is a dual solution and $\text{val}(\mathfrak{P}) = \text{val}(\mathfrak{D})$. In that case, we have $\text{val}(\mathfrak{P}) = \text{val}(\mathfrak{D}) = \mathfrak{L}(x^*,\lambda^*) \in \mathbb{R}$. Furthermore, from Remark 7.2, we know that x^* is a solution to the partial primal problem

$$\underset{x \in C}{\text{minimize}}\, \mathfrak{L}(x,\lambda^*),$$

where the inequality constraints $g_i(x) \leqslant 0$ have "vanished", and that λ^* is a solution to the partial dual problem

$$\underset{\lambda \in \mathbb{R}_+^p}{\text{maximize}}\, \mathfrak{L}(x^*,\lambda).$$

Remark 7.3. Consider the framework of this section and assume that $(x^*,\lambda^*) \in C \times \mathbb{R}_+^p$ is a saddle point of $\mathfrak{L}$. Then:

(i) One can recover (in particular from the partial dual problem, exercise) the standard *complementary conditions* given by

$$\forall i \in \{1,\dots,p\}, \quad g_i(x^*) \leqslant 0, \quad \lambda_i^* \geqslant 0, \quad \lambda_i^* g_i(x^*) = 0.$$

(ii) In the case where $C = \mathbb{R}^d$ and the functions f, $g_i : \mathbb{R}^d \to \mathbb{R}$ are differentiable, then one can recover (from the partial primal problem, exercise) the standard *Karush-Kuhn-Tucker (KKT) condition* given by

$$\nabla f(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla g(x^*) = 0_{\mathbb{R}^d}.$$

Remark 7.4. Note that *inequality/equality constrained minimization problems* can also be considered. For example, if one aims at considering equality constraints $h_j(x) = 0$ for some functions $h_j : \mathbb{R}^d \to \mathbb{R}$ (without extended-real values), then one has just to consider the two inequality constraints $h_j(x) \leqslant 0$ and $-h_j(x) \leqslant 0$.

8 Strong Lagrangian duality for convex inequality/equality constrained minimization problems

8.1 Alternative lemma

Lemma 8.1. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be an affine function and C be a nonempty convex subset of $\mathbb{R}^d$. Assume that φ is nonnegative over C and that there exists $x_0 \in \operatorname{rint}(C)$ such that $\varphi(x_0) = 0$. Then, φ vanishes all over C.

Proof. By contradiction, assume that there exists $x \in C$ such that $\varphi(x) > 0$. Since $x_0 \in \operatorname{rint}(C)$, there exists $t > 1$ sufficiently close to 1 such that $y := x + t(x_0 - x) \in C$ and thus $0 \leqslant \varphi(y) = (1-t)\varphi(x) < 0$ which is absurd. □

Lemma 8.2 (Alternative lemma)**.** Let $p \in \mathbb{N}^*$ and $q \in \mathbb{N}$. Let $g_i : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function for all $i \in \{1, \dots, p\}$ and $h_j : \mathbb{R}^d \to \mathbb{R}$ be an affine function for all $j \in \{1, \dots, q\}$. Let C be a nonempty convex subset of $\mathbb{R}^d$. Assume that

$$C \subset \bigcap_{i=1}^{p} \operatorname{dom}(g_i) \qquad \text{and} \qquad \exists x_0 \in \operatorname{rint}(C), \quad \forall j \in \{1, \dots, q\}, \quad h_j(x_0) \leqslant 0.$$

Then, exactly one of the two following alternatives holds true:

(i) There exists $\bar{x} \in C$ such that

$$\forall i \in \{1, \dots, p\}, \quad g_i(\bar{x}) < 0 \qquad \text{and} \qquad \forall j \in \{1, \dots, q\}, \quad h_j(\bar{x}) \leqslant 0.$$

(ii) There exists $(\lambda, \mu) \in (\mathbb{R}^p_+ \backslash \{0_{\mathbb{R}^p}\}) \times \mathbb{R}^q_+$ such that

$$\forall x \in C, \quad \sum_{i=1}^{p} \lambda_i g_i(x) + \sum_{j=1}^{q} \mu_j h_j(x) \geqslant 0.$$

Proof. First, since $p \neq 0$, let us note that (i) and (ii) cannot be true together since then

$$0 \leqslant \sum_{i=1}^{p} \lambda_i g_i(\bar{x}) + \sum_{j=1}^{q} \mu_j h_j(\bar{x}) < 0.$$

Therefore, to conclude the proof, assume that (i) is not true and let us prove that (ii) is true. To this aim, let us introduce

$$K := \{(y, z) \in \mathbb{R}^p \times \mathbb{R}^q \mid \exists x \in C, \ \forall i \in \{1, \dots, p\}, \ g_i(x) < y_i \text{ and } \forall j \in \{1, \dots, q\}, \ h_j(x) = z_j\}.$$

Since C is nonempty and satisfies $C \subset \operatorname{dom}(g_i)$ for all $i \in \{1, \dots, p\}$, it is clear that K is nonempty. Furthermore, from the convexity of C and of the g_i and the fact that the h_j are affine, it is clear that K is convex. Since (i) is not true, we derive that $K \cap \mathbb{R}^{p+q}_- = \emptyset$. From the proper separation theorem, we obtain that there exists $(\lambda, \mu) \in \mathbb{R}^{p+q}_+ \backslash \{0_{\mathbb{R}^{p+q}}\}$ such that

$$\forall x \in C, \quad \forall s = (s_i)_{i=1,\dots,p} \in (\mathbb{R}^*_+)^p, \quad \sum_{i=1}^{p} \lambda_i (g_i(x) + s_i) + \sum_{j=1}^{q} \mu_j h_j(x) \geqslant 0,$$

and

$$\exists \bar{x} \in C, \quad \exists \bar{s} = (\bar{s}_i)_{i=1,\dots,p} \in (\mathbb{R}_+^*)^p, \quad \sum_{i=1}^{p} \lambda_i (g_i(\bar{x}) + \bar{s}_i) + \sum_{j=1}^{q} \mu_j h_j(\bar{x}) > 0.$$

Letting $s \to 0_{\mathbb{R}^p}$ in the first above inequality, we obtain that

$$\forall x \in C, \quad \sum_{i=1}^{p} \lambda_i g_i(x) + \sum_{j=1}^{q} \mu_j h_j(x) \geqslant 0.$$

Thus, we only need to prove that $\lambda \neq 0_{\mathbb{R}^p}$. If $q = 0$, it is trivial since $(\lambda, \mu) \in \mathbb{R}_+^{p+q} \backslash \{0_{\mathbb{R}^{p+q}}\}$. Thus, let us deal with the case $q \in \mathbb{N}^*$. By contradiction, assume that $\lambda = 0_{\mathbb{R}^p}$ and let us introduce the affine function $\varphi : \mathbb{R}^d \to \mathbb{R}$ defined by $\varphi(x) := \sum_{j=1}^{q} \mu_j h_j(x)$ for all $x \in \mathbb{R}^d$. Based on the above, we do know that φ is nonnegative over C and from the second assumption of the alternative lemma we have $\varphi(x_0) = 0$ for some $r_0 \in \operatorname{rint}(C)$. We deduce from Lemma 8.1 that φ vanishes all over C which contradicts the fact that $\varphi(\bar{x}) > 0$. The proof is complete. $\square$

Remark 8.1. From the above proof, note that, when $q = 0$, the second part of the assumption of the alternative lemma vanishes and is not required.

8.2 Main theorem for convex inequality constrained minimization problems

In this section, we consider the *convex inequality constrained minimization problem* given by

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \quad & f(x), \\ \text{subject to} \quad & g_i(x) \leqslant 0, \quad \forall i \in \{1, \dots, p\}, \\ & h_j(x) \leqslant 0, \quad \forall j \in \{1, \dots, q\}, \\ & x \in C, \end{aligned} \tag{P}$$

where $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper convex function, where $g_i : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper convex function for all $i \in \{1, \dots, p\}$ for some $p \in \mathbb{N}$, where $h_j : \mathbb{R}^d \to \mathbb{R}$ is an affine function for all $j \in \{1, \dots, q\}$ for some $q \in \mathbb{N}$, and where C is a nonempty convex subset of $\mathbb{R}^d$ such that $C \subset \operatorname{dom}(f) \cap (\cap_{i=1}^{p} \operatorname{dom}(g_i))$. To avoid trivial cases, we assume that $(p, q) \neq (0, 0)$. Consider the associated Lagrangian function given by

$$\begin{array}{rccl} \mathtt{L}: & C \times \mathbb{R}_+^p \times \mathbb{R}_+^q & \longrightarrow & \mathbb{R} \\ & (x, \lambda, \mu) & \longmapsto & \mathtt{L}(x, \lambda, \mu) := f(x) + \displaystyle\sum_{i=1}^{p} \lambda_i g_i(x) + \sum_{j=1}^{q} \mu_j h_j(x), \end{array}$$

and the corresponding dual problem given by

$$\underset{(\lambda,\mu) \in \mathbb{R}_+^p \times \mathbb{R}_+^q}{\text{maximize}} \left(\inf_{x \in C} \mathtt{L}(x, \lambda, \mu) \right). \tag{D}$$

Recall that the weak duality inequality $\operatorname{val}(\mathtt{D}) \leqslant \operatorname{val}(\mathtt{P})$ holds true.

Theorem 8.1. Consider the framework of this section. Assume that $\mathrm{val}(\mathtt{P}) \in \mathbb{R}$ and that the *Slater's condition*

$$\exists x_0 \in \mathrm{rint}(C), \qquad \begin{cases} g_i(x_0) < 0 & \forall i \in \{1,\dots,p\}, \\ h_j(x_0) \leqslant 0 & \forall j \in \{1,\dots,q\}, \end{cases}$$

holds true. Then, there exists a dual solution $(\lambda^*, \mu^*) \in \mathbb{R}^p_+ \times \mathbb{R}^q_+$ and $\mathrm{val}(\mathtt{P}) = \mathrm{val}(\mathtt{D})$ (no duality gap). Furthermore, if x^* is a primal solution, then:

- (x^*, λ^*, μ^*) is a saddle-point of $\mathtt{L}$.
- $\mathrm{val}(\mathtt{P}) = \mathrm{val}(\mathtt{D}) = \mathtt{L}(x^*, \lambda^*, \mu^*) \in \mathbb{R}$.
- x^* is a solution to the partial primal problem

$$\underset{x \in C}{\mathrm{minimize}}\, \mathtt{L}(x, \lambda^*, \mu^*).$$

- (λ^*, μ^*) is a solution to the partial dual problem

$$\underset{(\lambda,\mu) \in \mathbb{R}^p_+ \times \mathbb{R}^q_+}{\mathrm{maximize}}\, \mathtt{L}(x^*, \lambda, \mu).$$

- The *complementary conditions*

$$\forall i \in \{1,\dots,p\}, \quad g_i(x^*) \leqslant 0, \quad \lambda_i^* \geqslant 0, \quad \lambda_i^* g_i(x^*) = 0,$$

$$\forall j \in \{1,\dots,q\}, \quad h_j(x^*) \leqslant 0, \quad \mu_j^* \geqslant 0, \quad \mu_j^* h_j(x^*) = 0,$$

  hold true.

Proof. Consider the alternative lemma by adding the proper convex function $f - \mathrm{val}(\mathtt{P})$ to the proper convex functions g_i. The assumption of the alternative lemma is satisfied (in particular thanks to the second part of the Slater's condition). Furthermore, from the definition of $\mathrm{val}(\mathtt{P})$, the point (i) is not satisfied (since it does not exist a point $\bar{x} \in C$ such that $f(\bar{x}) - \mathrm{val}(\mathtt{P}) < 0$, $g_i(\bar{x}) < 0$ for all $i \in \{1,\dots,p\}$ and $h_j(\bar{x}) \leqslant 0$ for all $j \in \{1,\dots,q\}$). Therefore, from the alternative lemma, there exists $(\lambda_0^*, \lambda^*, \mu^*) \in (\mathbb{R}^{p+1}_+ \backslash \{0_{\mathbb{R}^{p+1}}\}) \times \mathbb{R}^q_+$ such that

$$\forall x \in C, \quad \lambda_0^*(f(x) - \mathrm{val}(\mathtt{P})) + \sum_{i=1}^{p} \lambda_i^* g_i(x) + \sum_{j=1}^{q} \mu_j^* h_j(x) \geqslant 0.$$

If $p = 0$, it is trivial that $\lambda_0^* > 0$. Let us deal with the case $p \in \mathbb{N}^*$ and assume by contradiction that $\lambda_0^* = 0$. Then $(\lambda^*, \mu^*) \in (\mathbb{R}^p_+ \backslash \{0_{\mathbb{R}^p}\}) \times \mathbb{R}^q_+$ and

$$\forall x \in C, \quad \sum_{i=1}^{p} \lambda_i^* g_i(x) + \sum_{j=1}^{q} \mu_j^* h_j(x) \geqslant 0.$$

From the alternative lemma, it means that it does not exist some $\bar{x} \in C$ such that $g_i(\bar{x}) < 0$ for all $i \in \{1,\dots,p\}$ and $h_j(\bar{x}) \leqslant 0$ for all $j \in \{1,\dots,q\}$ which contradicts the Slater's condition. Thus we deduce that $\lambda_0^* > 0$.

Up to dividing by λ_0^* (without loss of generality), we can take $\lambda_0^* = 1$. Finally we obtain that

$$\forall x \in C, \quad \mathrm{val}(\mathtt{P}) \leqslant f(x) + \sum_{i=1}^{p} \lambda_i^* g_i(x) + \sum_{j=1}^{q} \mu_j^* h_j(x) = \mathtt{L}(x, \lambda^*, \mu^*).$$

We get that

$$\mathrm{val}(\mathtt{P}) \leqslant \inf_{x \in C} \mathtt{L}(x, \lambda^*, \mu^*) \leqslant \sup_{(\lambda,\mu) \in \mathbb{R}^p_+ \times \mathbb{R}^q_+} \left(\inf_{x \in C} \mathtt{L}(x, \lambda, \mu) \right) = \mathrm{val}(\mathtt{D}) \leqslant \mathrm{val}(\mathtt{P}).$$

We deduce that val(P) = val(D) and that

$$\inf_{x\in C} \mathtt{L}(x,\lambda^*,\mu^*) = \sup_{(\lambda,\mu)\in\mathbb{R}^p_+\times\mathbb{R}^q_+} \left(\inf_{x\in C} \mathtt{L}(x,\lambda,\mu) \right),$$

which means that (λ^*,μ^*) is a dual solution. The rest of the proof is trivial from results already obtained about general Lagrangian duality. □

Remark 8.2. Consider the framework of Theorem 8.1. Note that the strict inequalities in the Slater's condition concern only the *purely convex* (that is, convex but a priori not affine) constraint functions. On the other hand, let us precise that, when there is no purely convex inequality constraint, that is, when $p=0$, then the Slater's condition reduces to

$$\exists x_0 \in \mathrm{rint}(C), \quad \forall j\in\{1,\dots,q\}, \quad h_j(x_0)\leqslant 0,$$

and is required.

8.3 Corollary for convex inequality/equality constrained minimization problems

In this section, we consider the *convex inequality/equality constrained minimization problem* given by

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}} \quad & f(x), \\ \text{subject to} \quad & g_i(x)\leqslant 0, \quad \forall i\in\{1,\dots,p\}, \\ & h_j(x)\leqslant 0, \quad \forall j\in\{1,\dots,q\}, \\ & \ell_k(x) = 0, \quad \forall k\in\{1,\dots,r\}, \\ & x\in C, \end{aligned} \tag{$\mathtt{P}_0$}$$

where $f:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is a proper convex function, where $g_i:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is a proper convex function for all $i\in\{1,\dots,p\}$ for some $p\in\mathbb{N}$, where $h_j:\mathbb{R}^d\to\mathbb{R}$ is an affine function for all $j\in\{1,\dots,q\}$ for some $q\in\mathbb{N}$, where $\ell_k:\mathbb{R}^d\to\mathbb{R}$ is an affine function for all $k\in\{1,\dots,r\}$ for some $r\in\mathbb{N}$, and where C is a nonempty convex subset of $\mathbb{R}^d$ such that $C\subset\mathrm{dom}(f)\cap(\cap_{i=1}^p\mathrm{dom}(g_i))$. To avoid trivial cases, we assume that $(p,q,r)\neq(0,0,0)$. Note that Problem ($\mathtt{P}_0$) can be rewritten as Problem (P) by replacing the equality $\ell_k(x)=0$ by the two affine inequalities $\ell_k(x)\leqslant 0$ and $-\ell_k(x)\leqslant 0$ for all $k\in\{1,\dots,r\}$. Note that the affine nature of the functions ℓ_k is required (convexity is not enough). Therefore, we consider the associated Lagrangian function given by

$$\begin{aligned} \mathtt{L}_0: \quad C\times\mathbb{R}^p_+\times\mathbb{R}^q_+\times\mathbb{R}^r \quad & \longrightarrow \quad \mathbb{R} \\ (x,\lambda,\mu,\gamma) \quad & \longmapsto \quad \mathtt{L}_0(x,\lambda,\mu,\gamma) := f(x)+\sum_{i=1}^p \lambda_i g_i(x)+\sum_{j=1}^q \mu_j h_j(x)+\sum_{k=1}^r \gamma_k \ell_k(x). \end{aligned}$$

In that context, note that the variable γ lies in $\mathbb{R}^r$ (and not in $\mathbb{R}^r_+$). This is due to the fact that we consider both the affine inequalities $\ell_k(x)\leqslant 0$ and $-\ell_k(x)\leqslant 0$ for all $k\in\{1,\dots,r\}$. On the other hand, note that

$$\forall x\in C, \quad \sup_{(\lambda,\mu,\gamma)\in\mathbb{R}^p_+\times\mathbb{R}^q_+\times\mathbb{R}^r} \mathtt{L}_0(x,\lambda,\mu,\gamma) = \begin{cases} f(x) & \text{if} \quad g_i(x)\leqslant 0, \quad \forall i\in\{1,\dots,p\}, \\ & \qquad h_j(x)\leqslant 0, \quad \forall j\in\{1,\dots,q\}, \\ & \qquad \ell_k(x)=0, \quad \forall k\in\{1,\dots,r\}, \\ +\infty & \text{otherwise}. \end{cases}$$

Hence Problem ($\mathtt{P}_0$) can be rewritten as the primal problem

$$\underset{x\in C}{\text{minimize}} \left(\sup_{(\lambda,\mu,\gamma)\in\mathbb{R}^p_+\times\mathbb{R}^q_+\times\mathbb{R}^r} \mathtt{L}_0(x,\lambda,\mu,\gamma) \right),$$

and the corresponding dual problem is given by

$$\underset{(\lambda,\mu,\gamma)\in\mathbb{R}^p_+\times\mathbb{R}^q_+\times\mathbb{R}^r}{\text{maximize}} \left(\inf_{x\in C} \mathtt{L}_0(x,\lambda,\mu,\gamma) \right). \tag{$\mathtt{D}_0$}$$

Finally, recall that the weak duality inequality $\mathrm{val}(\mathtt{D}_0) \leqslant \mathrm{val}(\mathtt{P}_0)$ holds true.

Corollary 8.1. Consider the framework of this section. Assume that $\mathrm{val}(\mathtt{P}_0) \in \mathbb{R}$ and that the *Slater's condition*

$$\exists x_0 \in \mathrm{rint}(C), \quad \begin{cases} g_i(x_0) < 0 & \forall i \in \{1,\dots,p\}, \\ h_j(x_0) \leqslant 0 & \forall j \in \{1,\dots,q\}, \\ \ell_k(x_0) = 0 & \forall k \in \{1,\dots,r\}, \end{cases}$$

holds true. Then, there exists a dual solution $(\lambda^*, \mu^*, \gamma^*) \in \mathbb{R}^p_+ \times \mathbb{R}^q_+ \times \mathbb{R}^r$ and $\mathrm{val}(\mathtt{P}_0) = \mathrm{val}(\mathtt{D}_0)$ (no duality gap). Furthermore, if x^* is a primal solution, then:

- $(x^*, \lambda^*, \mu^*, \gamma^*)$ is a saddle-point of $\mathtt{L}_0$.
- $\mathrm{val}(\mathtt{P}_0) = \mathrm{val}(\mathtt{D}_0) = \mathtt{L}_0(x^*, \lambda^*, \mu^*, \gamma^*) \in \mathbb{R}$.
- x^* is a solution to the partial primal problem

$$\underset{x \in C}{\text{minimize}}\, \mathtt{L}_0(x, \lambda^*, \mu^*, \gamma^*).$$

- $(\lambda^*, \mu^*, \gamma^*)$ is a solution to the partial dual problem

$$\underset{(\lambda,\mu,\gamma) \in \mathbb{R}^p_+ \times \mathbb{R}^q_+ \times \mathbb{R}^r}{\text{maximize}} \quad \mathtt{L}_0(x^*, \lambda, \mu, \gamma).$$

- The *complementary conditions*

$$\forall i \in \{1,\dots,p\}, \quad g_i(x^*) \leqslant 0, \quad \lambda_i^* \geqslant 0, \quad \lambda_i^* g_i(x^*) = 0,$$

$$\forall j \in \{1,\dots,q\}, \quad h_j(x^*) \leqslant 0, \quad \mu_j^* \geqslant 0, \quad \mu_j^* h_j(x^*) = 0,$$

$$\forall k \in \{1,\dots,r\}, \quad \ell_k(x^*) = 0,$$

  hold true.

Proof. The proof is trivial from Theorem 8.1. □

8.4 Some examples and applications

Remark 8.3. In this remark, our aim is to provide a counterexample showing that, without appropriate assumptions (such as the Slater's condition), then a positive duality gap is possible. Consider the two-dimensional convex equality constrained minimization problem given by

$$\begin{aligned} \underset{x=(x_1,x_2) \in \mathbb{R}^2}{\text{minimize}} \quad & e^{-\sqrt{x_1 x_2}}, \\ \text{subject to} \quad & x_2 = 0, \\ & x = (x_1, x_2) \in \mathbb{R}^2_+. \end{aligned}$$

Note that the objective function is indeed convex (by using Hessian matrix over the open convex $(\mathbb{R}^*_+)^2$, with eigenvalues 0 and $\frac{x_1^2+x_2^2}{4x_1x_2} e^{-\sqrt{x_1x_2}}$, and then by continuity over $\mathbb{R}^2_+$, exercise). It is clear that any point $(x_1, 0)$, with $x_1 \geqslant 0$, is a primal solution and that the primal value is equal to 1. The dual problem is given by

$$\underset{\gamma \in \mathbb{R}}{\text{maximize}} \left(\inf_{x \in \mathbb{R}^2_+} \mathtt{L}_0(x, \gamma) \right)$$

where the Lagrangian function is given by

$$\forall (x, \gamma) \in \mathbb{R}^2_+ \times \mathbb{R}, \quad \mathtt{L}_0(x, \gamma) := e^{-\sqrt{x_1 x_2}} + \gamma x_2,$$

One can prove (not trivial, exercise) that

$$\forall \gamma \in \mathbb{R}, \quad \inf_{x \in \mathbb{R}^2_+} \mathtt{L}_0(x,\gamma) = \begin{cases} 0 & \text{if} \quad \gamma \geqslant 0, \\ -\infty & \text{if} \quad \gamma < 0. \end{cases}$$

We deduce that any $\gamma \geqslant 0$ is a dual solution and that the dual value is equal to zero. Therefore, in that example, the duality gap is positive and note that the Slater's condition is not satisfied.

Remark 8.4. Consider the linear equality constrained minimization problem given by

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \quad & \langle a, x \rangle, \\ \text{subject to} \quad & Mx = b, \\ & x \in \mathbb{R}^d_+, \end{aligned}$$

where $a \in \mathbb{R}^d$, $M \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$. The dual problem is given by

$$\underset{y \in \mathbb{R}^n}{\text{maximize}} \left(\inf_{x \in \mathbb{R}^d_+} \mathtt{L}_0(x,y) \right),$$

where the Lagrangian function is given by

$$\forall (x,y) \in \mathbb{R}^d_+ \times \mathbb{R}^n, \quad \mathtt{L}_0(x,y) := \langle a, x \rangle + \langle y, Mx - b \rangle = \langle M^\top y + a, x \rangle - \langle y, b \rangle.$$

We obtain that

$$\forall y \in \mathbb{R}^n, \quad \inf_{x \in \mathbb{R}^d_+} \mathtt{L}_0(x,y) := \begin{cases} -\langle y, b \rangle & \text{if} & M^\top y + a \in \mathbb{R}^d_+, \\ -\infty & \text{otherwise.} \end{cases}$$

Finally, the dual problem is given by

$$\begin{aligned} \underset{y \in \mathbb{R}^n}{\text{maximize}} \quad & -\langle y, b \rangle, \\ \text{subject to} \quad & M^\top y + a \in \mathbb{R}^d_+. \end{aligned}$$

In the primal problem, note that one can consider the constraint $x \in \mathbb{R}^d_+$ as d affine inequality constraints $-x_i \leqslant 0$ and thus consider $C = \mathbb{R}^d$. In that case, the Lagrangian function is given by

$$\forall (x,y,z) \in \mathbb{R}^d \times \mathbb{R}^n \times \mathbb{R}^d_+, \quad \widetilde{\mathtt{L}_0}(x,y,z) := \langle a, x \rangle + \langle y, Mx - b \rangle - \langle z, x \rangle.$$

With this point of view, we obtain the dual problem given by

$$\begin{aligned} \underset{(y,z) \in \mathbb{R}^n \times \mathbb{R}^d_+}{\text{maximize}} \quad & -\langle y, b \rangle, \\ \text{subject to} \quad & M^\top y + a - z = 0_{\mathbb{R}^d}, \end{aligned}$$

which is the same dual problem.

Proposition 8.1 (Projection onto the intersection of a box and an affine hyperplane)**.** Consider

$$K := \text{Box}[\ell, u] \cap [\langle h, \cdot \rangle = \alpha],$$

for some $(h, \alpha) \in (\mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\}) \times \mathbb{R}$ and $\ell = (\ell_i)_{i=1,\dots,d} \in (\mathbb{R} \cup \{-\infty\})^d$ and $u = (u_i)_{i=1,\dots,d} \in (\mathbb{R} \cup \{+\infty\})^d$ with $\ell_i \leqslant u_i$ for all $i \in \{1, \dots, d\}$. Assume that $\text{rint}(\text{Box}[\ell, u]) \cap [\langle h, \cdot \rangle = \alpha] \neq \emptyset$. Then

$$\forall x \in \mathbb{R}^d, \quad \text{proj}_K(x) = \text{proj}_{\text{Box}[\ell,u]}(x - \gamma^*_x h),$$

where $\gamma^*_x \in \mathbb{R}$ stands for a solution to the scalar equation

$$\langle h, \text{proj}_{\text{Box}[\ell,u]}(x - \gamma h) \rangle = \alpha.$$

Proof. Thanks to the assumption, note that K is a nonempty closed convex subset of $\mathbb{R}^d$. Let $x \in \mathbb{R}^d$ and define $y^* := \mathrm{proj}_K(x)$ which is the unique solution to the convex equality constrained minimization problem given by

$$\begin{aligned} &\underset{y\in\mathbb{R}^d}{\text{minimize}} && \frac{1}{2}\|y-x\|^2, \\ &\text{subject to} && \langle h, y\rangle = \alpha, \\ &&& y \in \mathrm{Box}[\ell, u]. \end{aligned}$$

The corresponding Lagrangian function is given by

$$\forall (y,\gamma) \in \mathrm{Box}[\ell,u]\times\mathbb{R}, \quad \mathtt{L}_0(y,\gamma) := \frac{1}{2}\|y-x\|^2 + \gamma(\langle h,y\rangle - \alpha) = \frac{1}{2}\|y-(x-\gamma h)\|^2 - \frac{\gamma^2}{2}\|h\|^2 + \gamma(\langle h,x\rangle-\alpha).$$

Since the Slater's condition is satisfied (by considering any point $y \in \mathrm{rint}(\mathrm{Box}[\ell,u]) \cap [\langle h,\cdot\rangle = \alpha] \neq \emptyset$), we know that there exists $\gamma_x^* \in \mathbb{R}$ (dual solution) such that y^* is a solution to the partial primal problem

$$\underset{y\in\mathrm{Box}[\ell,u]}{\text{minimize}}\ \mathtt{L}_0(y,\gamma_x^*) = \frac{1}{2}\|y-(x-\gamma_x^* h)\|^2 - \frac{(\gamma_x^*)^2}{2}\|h\|^2 + \gamma_x^*(\langle h,x\rangle - \alpha),$$

which gives exactly

$$y^* = \mathrm{proj}_{\mathrm{Box}[\ell,u]}(x-\gamma_x^* h).$$

Furthermore, since $\langle h, y^*\rangle = \alpha$, the last equality follows. □

Remark 8.5. Consider the framework of Proposition 8.1. Recall that $\mathrm{proj}_{\mathrm{Box}[\ell,u]}$ is very simple to express (see Proposition 3.1), precisely as

$$\forall x \in \mathbb{R}^d, \quad \mathrm{proj}_{\mathrm{Box}[\ell,u]}(x) = \Big(\min\{\max\{x_i,\ell_i\},u_i\}\Big)_{i=1,\dots,d}.$$

Now let us fix $x \in \mathbb{R}^d$ and introduce the function $\varphi_x : \mathbb{R} \to \mathbb{R}$ defined by $\varphi_x(\gamma) := \langle h, \mathrm{proj}_{\mathrm{Box}[\ell,u]}(x-\gamma h)\rangle - \alpha$ for all $\gamma \in \mathbb{R}$. Note that

$$\forall \gamma \in \mathbb{R}, \quad \varphi_x(\gamma) = \sum_{i=1}^d h_i\Big(\min\{\max\{x_i-\gamma h_i,\ell_i\},u_i\}\Big) - \alpha.$$

Separating the two cases $h_i \geqslant 0$ and $h_i < 0$, and considering $\gamma_1 \leqslant \gamma_2$, one can prove (exercise) that each sum term is nonincreasing with respect to γ. Therefore, the function φ_x is a nonincreasing function. Therefore, the scalar equation $\varphi_x(\gamma) = 0$ is easy to solve numerically.

Remark 8.6. Consider $\Delta^{d-1} := \{x = (x_1,\dots,x_d) \in \mathbb{R}^d_+ \mid \sum_{i=1}^d x_i = 1\}$ the unit simplex of $\mathbb{R}^d$ of dimension $d-1$. Then

$$\forall x \in \mathbb{R}^d, \quad \mathrm{proj}_{\Delta^{d-1}}(x) = ((x_i - \gamma_x^*)^+)_{i=1,\dots,d},$$

where $\gamma_x^* \in \mathbb{R}$ is a solution to the scalar equation

$$\sum_{i=1}^d (x_i-\gamma)^+ = 1.$$

To see this, one has just to apply Proposition 8.1 with $\ell_i = 0$, $u_i = +\infty$, $h = (1,\dots,1)$ and $\alpha = 1$.

9 Primal-dual algorithms

When considering a primal problem

$$\inf_{a\in\mathcal{A}}\left(\sup_{b\in\mathcal{B}}\mathcal{L}(a,b)\right),$$

associated with a Lagrangian function $\mathcal{L}:\mathcal{A}\times\mathcal{B}\to\mathbb{R}$, where $\mathcal{A}$ and $\mathcal{B}$ are two nonempty sets, the main idea of *primal-dual algorithms* is, roughly speaking, to solve conjointly the primal problem and the dual problem with an algorithm of the form

$$a_0\in\mathcal{A},\quad b_0\in\mathcal{B},\qquad\text{and}\qquad\forall k\in\mathbb{N},\quad\begin{cases} a_{k+1} & \in & \underset{a\in\mathcal{A}}{\arg\min}\,\mathcal{L}(a,b_k),\\ b_{k+1} & \in & \underset{b\in\mathcal{B}}{\arg\max}\,\mathcal{L}(a_k,b).\end{cases}$$

Unfortunately, it does not work so easily in general. In this section, we are going to introduce and study some very well-known primal-dual algorithms, focusing on a convex linear-equality constrained minimization problem of the form

$$\begin{array}{ll}\underset{x\in\mathbb{R}^d}{\text{minimize}} & f(x),\\ \text{subject to} & Mx=b,\\ & x\in\operatorname{dom}(f),\end{array}\tag{P$_1$}$$

where $f\in\Gamma_0(\mathbb{R}^d)$, $M\in\mathbb{R}^{n\times d}$ and $b\in\mathbb{R}^n$. Extensions to inequality constraints are possible in the literature, but this is beyond the scope of this lecture. The associated Lagrangian function $\mathrm{L}_1:\operatorname{dom}(f)\times\mathbb{R}^n\to\mathbb{R}$ is given by

$$\forall(x,y)\in\operatorname{dom}(f)\times\mathbb{R}^n,\quad \mathrm{L}_1(x,y):=f(x)+\langle y,Mx-b\rangle,$$

and the dual problem is given by

$$\underset{y\in\mathbb{R}^n}{\text{maximize}}\left(\inf_{x\in\operatorname{dom}(f)}\mathrm{L}_1(x,y)\right).\tag{D$_1$}$$

For some $\lambda>0$, a basic algorithm, called *dual ascent method*, is given by

$$y_0\in\mathbb{R}^n\qquad\text{and}\qquad\forall k\in\mathbb{N},\quad\begin{cases} x_{k+1}\in\underset{x\in\operatorname{dom}(f)}{\arg\min}\,\mathrm{L}_1(x,y_k),\\ y_{k+1}=y_k+\lambda(Mx_{k+1}-b).\end{cases}$$

Under several assumptions (no detail here) and if $\lambda>0$ is chosen appropriately, then $(x_k)_{k\in\mathbb{N}}$ converges to a primal solution and $(y_k)_{k\in\mathbb{N}}$ converges to a dual solution. However, these assumptions do not hold in many applications, so dual ascent method often cannot be used. To obtain general convergence results, the main idea of primal-dual algorithms (including the ones exposed in the next subsections) is to rewrite the dual problem (D$_1$) as a convex minimization problem and to invoke proximal algorithms.

9.1 Method of multipliers (or augmented Lagrangian method)

In this section, we consider the convex linear-equality constrained minimization problem given by

$$\begin{array}{ll}\underset{x\in\mathbb{R}^d}{\text{minimize}} & f(x),\\ \text{subject to} & Mx=b,\\ & x\in\operatorname{dom}(f),\end{array}\tag{P$_1$}$$

where $f\in\Gamma_0(\mathbb{R}^d)$, $M\in\mathbb{R}^{n\times d}$ and $b\in\mathbb{R}^n$. The associated Lagrangian function $\mathrm{L}_1:\operatorname{dom}(f)\times\mathbb{R}^n\to\mathbb{R}$ is given by

$$\forall(x,y)\in\operatorname{dom}(f)\times\mathbb{R}^n,\quad \mathrm{L}_1(x,y):=f(x)+\langle y,Mx-b\rangle,$$

and the dual problem is given by

$$\underset{y\in\mathbb{R}^n}{\text{maximize}} \left(\inf_{x\in\mathrm{dom}(f)} \mathrm{L}_1(x,y) \right). \tag{D$_1$}$$

Lemma 9.1. Consider the framework of this section. The dual problem (D$_1$) can be rewritten as the convex minimization problem given by

$$\underset{y\in\mathbb{R}^n}{\text{minimize}}\, f^*(-M^\top y) + \langle b, y\rangle.$$

Now consider the closed convex function $F : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$, defined by $F(y) := f^*(-M^\top y) + \langle b, y\rangle$ for all $y \in \mathbb{R}^n$. Assume that F is proper (that is, $\mathrm{im}(M^\top)\cap\mathrm{dom}(f^*) \neq \emptyset$) and thus $F \in \Gamma_0(\mathbb{R}^n)$. Let $\lambda > 0$. Then, for any $y \in \mathbb{R}^n$ such that

$$\partial\Big(f^* \circ (-M^\top)\Big)(\mathrm{prox}_{\lambda F}(y)) = -M\partial f^*(-M^\top \mathrm{prox}_{\lambda F}(y)),$$

(which is satisfied if $-M^\top \mathrm{prox}_{\lambda F}(y) \in \mathrm{int}(\mathrm{dom}(f^*))$ for example), it holds that

$$\mathrm{prox}_{\lambda F}(y) = y + \lambda(M\hat{x} - b),$$

where

$$\hat{x} \in \underset{x\in\mathrm{dom}(f)}{\arg\min}\, \mathrm{L}^\lambda(x,y),$$

where $\mathrm{L}^\lambda : \mathrm{dom}(f)\times\mathbb{R}^n \to \mathbb{R}$ is the so-called *augmented Lagrangian function* defined by

$$\forall (x,y) \in \mathrm{dom}(f)\times\mathbb{R}^n, \quad \mathrm{L}^\lambda(x,y) := f(x) + \langle y, Mx - b\rangle + \frac{\lambda}{2}\|Mx - b\|^2.$$

Proof. For any $y \in \mathbb{R}^n$, note that

$$\begin{aligned}\inf_{x\in\mathrm{dom}(f)} \mathrm{L}_1(x,y) = -\langle b,y\rangle + \inf_{x\in\mathrm{dom}(f)} \Big(f(x) + \langle M^\top y, x\rangle\Big) &= -\langle b,y\rangle + \inf_{x\in\mathbb{R}^d}\Big(f(x) + \langle M^\top y, x\rangle\Big)\\ &= -\langle b,y\rangle - \sup_{x\in\mathbb{R}^d}\Big(\langle -M^\top y, x\rangle - f(x)\Big) = -\langle b,y\rangle - f^*(-M^\top y),\end{aligned}$$

which concludes the proof of the first assertion. Now assume that F is proper and let $\lambda > 0$. Let $y \in \mathbb{R}^n$ satisfying the given assumption and let us denote by $\hat{z} := \mathrm{prox}_{\lambda F}(y)$ (well-defined since $F \in \Gamma_0(\mathbb{R}^n)$). It holds that

$$\hat{z} = \underset{z\in\mathbb{R}^n}{\arg\min} \left(f^*(-M^\top z) + \langle b, z\rangle + \frac{1}{2\lambda}\|z - y\|^2 \right).$$

Using some subdifferential calculus rules (see Theorem E.5 and Proposition E.3), we have

$$0_{\mathbb{R}^n} \in -M\partial f^*(-M^\top \hat{z}) + b + \frac{1}{\lambda}(\hat{z} - y),$$

and thus there exists $\hat{x} \in \partial f^*(-M^\top \hat{z})$ such that $\hat{z} = y + \lambda(M\hat{x} - b)$. We also derive that $-M^\top \hat{z} \in \partial f(\hat{x})$ and thus $0_{\mathbb{R}^d} \in M^\top \hat{z} + \partial f(\hat{x})$ which gives

$$0_{\mathbb{R}^d} \in M^\top(y + \lambda(M\hat{x} - b)) + \partial f(\hat{x}),$$

which can be rewritten as

$$\begin{aligned}\hat{x} \in \underset{x\in\mathbb{R}^d}{\arg\min}\left(f(x) + \frac{\lambda}{2}\left\| Mx - b + \frac{1}{\lambda}y\right\|^2 \right) &= \underset{x\in\mathrm{dom}(f)}{\arg\min}\left(f(x) + \frac{\lambda}{2}\left\| Mx - b + \frac{1}{\lambda}y\right\|^2 \right)\\ &= \underset{x\in\mathrm{dom}(f)}{\arg\min}\left(\mathrm{L}^\lambda(x,y) + \frac{1}{2\lambda}\|y\|^2\right) = \underset{x\in\mathrm{dom}(f)}{\arg\min}\, \mathrm{L}^\lambda(x,y),\end{aligned}$$

which concludes the proof. $\square$

Remark 9.1. Consider the framework of Lemma 9.1. With $\lambda = 0$, note that the augmented Lagrangian function L^λ is nothing else but the classical Lagrangian function L_1. On the other hand, with $\lambda > 0$, note that the augmented Lagrangian function L^λ corresponds to the (classical) Lagrangian function but associated with the equivalent convex linear-equality constrained minimization problem

$$\begin{array}{ll} \underset{x\in\mathbb{R}^d}{\text{minimize}} & f(x) + \dfrac{\lambda}{2}\|Mx-b\|^2, \\ \text{subject to} & Mx = b, \\ & x \in \mathrm{dom}(f). \end{array}$$

Proposition 9.1 (Method of multipliers)**.** Consider the framework of this section and let $\lambda > 0$. Under appropriate assumptions (not detailed in this lecture, but basically in order to ensure that the assumptions of Lemma 9.1 are satisfied), the so-called *method of multipliers* (or *augmented Lagrangian method*) given by

$$y_0 \in \mathbb{R}^n \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \begin{cases} \hat{x}_{k+1} \in \underset{x\in\mathrm{dom}(f)}{\arg\min}\, \mathrm{L}^\lambda(x, y_k), \\ y_{k+1} = y_k + \lambda(M\hat{x}_{k+1} - b), \end{cases}$$

is well-defined and corresponds to the proximal point algorithm applied to the dual problem (D_1). Therefore, the sequence $(y_k)_{k\in\mathbb{N}}$ converges if and only if a dual solution exists. In that case, the limit of $(y_k)_{k\in\mathbb{N}}$ is a dual solution and, if moreover the sequence $(\hat{x}_k)_{k\in\mathbb{N}}$ converges to some point in $\mathrm{dom}(f)$, then its limit is a primal solution.

Proof. The first assertions are trivial thanks to Lemma 9.1. Let us prove the last assertion. Assume that $(\hat{x}_k)_{k\in\mathbb{N}}$ converges to some $x^* \in \mathrm{dom}(f)$. From the convergence of $(y_k)_{k\in\mathbb{N}}$ to a dual solution y^*, it is clear that $Mx^* = b$ and, from lower semicontinuity, from convergences of $(y_k)_{k\in\mathbb{N}}$ and $(\hat{x}_k)_{k\in\mathbb{N}}$ and from definition of $\hat{x}_{k+1}$, we have

$$\begin{aligned} f(x^*) \leqslant \liminf f(\hat{x}_{k+1}) = \liminf \Big(f(\hat{x}_{k+1}) + \langle y_k, M\hat{x}_{k+1} - b\rangle + \frac{\lambda}{2}\|M\hat{x}_{k+1} - b\|^2 \Big) \\ = \liminf \mathrm{L}^\lambda(\hat{x}_{k+1}, y_k) \leqslant \liminf \mathrm{L}^\lambda(x, y_k), \end{aligned}$$

for all $x \in \mathrm{dom}(f)$ and thus, in particular, we have $f(x^*) \leqslant f(x)$ for all $x \in \mathrm{dom}(f)$ such that $Mx = b$. The proof is complete. ☐

Remark 9.2. In the method of multipliers, the first step $\hat{x}_{k+1} \in \arg\min_{x\in\mathrm{dom}(f)} \mathrm{L}^\lambda(x, y_k)$ is called the *primal step*. The second step $y_{k+1} = y_k + \lambda(M\hat{x}_{k+1} - b)$ is called the *dual step*. Note that the primal step is not trivial to achieve numerically. Usually, solving numerically the primal step is made by invoking an internal algorithm such as the forward-backward algorithm (if $\mathrm{prox}_{\mu f}$ is known for some $\mu > 0$) or even a smooth algorithm if f is smooth.

Remark 9.3. In Proposition 9.1, we did not detail the *appropriate assumptions* allowing to obtain the convergence results of the method of multipliers. Furthermore, we did not prove that the sequence $(\hat{x}_k)_{k\in\mathbb{N}}$ converges. Actually, the literature on convergence results for the method of multipliers is very vast and still in progress. It is not our aim in this lecture to investigate in details this technical question.

9.2 Alternating Direction Method of Multipliers (ADMM)

In this subsection, we will focus on a particular case of Problem (P_1). Precisely, we are going to focus on a convex linear-equality constrained minimization problem given by

$$\begin{array}{ll} \underset{(x,y)\in\mathbb{R}^d\times\mathbb{R}^n}{\text{minimize}} & f(x) + g(y), \\ \text{subject to} & Mx + Ny = c, \\ & (x,y) \in \mathrm{dom}(f)\times\mathrm{dom}(g), \end{array} \tag{P_2}$$

where $f \in \Gamma_0(\mathbb{R}^d)$, $g \in \Gamma_0(\mathbb{R}^n)$, where $M \in \mathbb{R}^{m\times d}$, $N \in \mathbb{R}^{m\times n}$ and $c \in \mathbb{R}^m$. For some $\lambda > 0$, the method of multipliers is given by

$$z_0 \in \mathbb{R}^m \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \begin{cases} (\hat{x}_{k+1}, \hat{y}_{k+1}) \in \underset{(x,y)\in\mathrm{dom}(f)\times\mathrm{dom}(g)}{\arg\min} \mathrm{L}^\lambda(x,y,z_k), \\ z_{k+1} = z_k + \lambda(M\hat{x}_{k+1} + N\hat{y}_{k+1} - c), \end{cases}$$

where the augmented Lagrangian function $\mathrm{L}^\lambda : \mathrm{dom}(f) \times \mathrm{dom}(g) \times \mathbb{R}^m \to \mathbb{R}$ is given by

$$\forall (x,y,z) \in \mathrm{dom}(f) \times \mathrm{dom}(g) \times \mathbb{R}^m, \quad \mathrm{L}^\lambda(x,y,z) := f(x) + g(y) + \langle z, Mx + Ny - c\rangle + \frac{\lambda}{2}\|Mx + Ny - c\|^2.$$

We do know that the primal step is not trivial in general. In addition, here, the primal step requires to minimize the two (a priori unseparable) primal variables (x,y) conjointly. To overcome this difficulty, the *Alternating Direction Method of Multipliers (ADMM)* consists in separating the minimization with respect to one primal variable (x for instance) from the minimization with respect to the other primal variable (y for instance), which gives the name of *alternating direction*. Hence, for some $\lambda > 0$, ADMM consists in the algorithm given by

$$\hat{y}_0 \in \mathbb{R}^n, \quad z_0 \in \mathbb{R}^m \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \begin{cases} \hat{x}_{k+1} \in \underset{x\in\mathrm{dom}(f)}{\arg\min}\, \mathrm{L}^\lambda(x, \hat{y}_k, z_k), \\ \hat{y}_{k+1} \in \underset{y\in\mathrm{dom}(g)}{\arg\min}\, \mathrm{L}^\lambda(\hat{x}_{k+1}, y, z_k), \\ z_{k+1} = z_k + \lambda(M\hat{x}_{k+1} + N\hat{y}_{k+1} - c). \end{cases}$$

Our aim in this section is to prove that ADMM is nothing else but the Douglas-Rachford algorithm applied to the dual problem

$$\underset{z\in\mathbb{R}^m}{\text{maximize}} \left(\inf_{(x,y)\in\mathrm{dom}(f)\times\mathrm{dom}(g)} \mathrm{L}_2(x,y,z) \right), \tag{D$_2$}$$

where the Lagrangian function $\mathrm{L}_2 : \mathrm{dom}(f) \times \mathrm{dom}(g) \times \mathbb{R}^m \to \mathbb{R}$ is given by

$$\forall (x,y,z) \in \mathrm{dom}(f) \times \mathrm{dom}(g) \times \mathbb{R}^m, \quad \mathrm{L}_2(x,y,z) := f(x) + g(y) + \langle z, Mx + Ny - c\rangle.$$

Lemma 9.2. Consider the framework of this section. The dual problem (D$_2$) can be rewritten as the convex minimization problem given by

$$\underset{z\in\mathbb{R}^m}{\text{minimize}}\, f^*(-M^\top z) + g^*(-N^\top z) + \langle c, z\rangle.$$

Proof. This proof is similar to the beginning of the proof of Lemma 9.1. For any $z \in \mathbb{R}^m$, note that

$$\begin{aligned} \inf_{(x,y)\in\mathrm{dom}(f)\times\mathrm{dom}(g)} \mathrm{L}_2(x,y,z) &= -\langle c, z\rangle + \inf_{x\in\mathrm{dom}(f)} \Big(f(x) + \langle M^\top z, x\rangle \Big) + \inf_{y\in\mathrm{dom}(g)} \Big(g(y) + \langle N^\top z, y\rangle \Big) \\ &= -\langle c, z\rangle + \inf_{x\in\mathbb{R}^d} \Big(f(x) + \langle M^\top z, x\rangle \Big) + \inf_{y\in\mathbb{R}^n} \Big(g(y) + \langle N^\top z, y\rangle \Big) \\ &= -\langle c, z\rangle - \sup_{x\in\mathbb{R}^d} \Big(\langle -M^\top z, x\rangle - f(x) \Big) - \sup_{y\in\mathbb{R}^n} \Big(\langle -N^\top z, y\rangle - g(y) \Big) \\ &= -\langle c, z\rangle - f^*(-M^\top z) - g^*(-N^\top z), \end{aligned}$$

which concludes the proof. ∎

Proposition 9.2 (ADMM)**.** Consider the framework of this section and let $\lambda > 0$. Under appropriate assumptions (not detailed in this lecture, see the proof), the Alternating Direction Method of Multipliers (ADMM) given by

$$\hat{y}_0 \in \mathbb{R}^n, \quad z_0 \in \mathbb{R}^m \qquad \text{and} \qquad \forall k \in \mathbb{N}, \qquad \begin{cases} \hat{x}_{k+1} \in \underset{x \in \mathrm{dom}(f)}{\arg\min}\ \mathrm{L}^\lambda(x, \hat{y}_k, z_k), \\ \hat{y}_{k+1} \in \underset{y \in \mathrm{dom}(g)}{\arg\min}\ \mathrm{L}^\lambda(\hat{x}_{k+1}, y, z_k), \\ z_{k+1} = z_k + \lambda(M\hat{x}_{k+1} + N\hat{y}_{k+1} - c), \end{cases}$$

is well-defined and corresponds to the Douglas-Rachford algorithm applied to the dual problem (D_2). Therefore, the sequence $(z_k)_{k\in\mathbb{N}}$ converges if and only if a dual solution exists. In that case, the limit of $(z_k)_{k\in\mathbb{N}}$ is a dual solution and $M\hat{x}_{k+1} + N\hat{y}_{k+1} - c$ tends to $0_{\mathbb{R}^m}$. If furthermore the (nonaugmented) Lagrangian function L_2 admits a saddle-point, then $f(\hat{x}_{k+1}) + g(\hat{y}_{k+1})$ tends to $\mathrm{val}(\mathrm{P}_2)$.

Proof. From Lemma 9.2, the dual problem (D_2) can be rewritten as the convex minimization problem given by

$$\underset{z \in \mathbb{R}^m}{\text{minimize}}\ F(z) + G(z),$$

where $F(z) := f^*(-M^\top z) + \langle c, z\rangle$ and $G(z) := g^*(-N^\top z)$ for all $z \in \mathbb{R}^m$. In the sequel, we assume that F, $G \in \Gamma_0(\mathbb{R}^m)$. Consider the Douglas-Rachford algorithm (see Remark 6.6) given by

$$v_0, w_0 \in \mathbb{R}^m, \qquad \text{and} \qquad \forall k \in \mathbb{N}, \qquad \begin{cases} u_{k+1} = \mathrm{prox}_{\lambda F}(v_k + w_k), \\ v_{k+1} = \mathrm{prox}_{\lambda G}(u_{k+1} - w_k), \\ w_{k+1} = w_k + v_{k+1} - u_{k+1}. \end{cases}$$

Recall that, if $\mathrm{dom}(F) \cap \mathrm{int}(\mathrm{dom}(G)) \neq \emptyset$ and $\mathrm{Argmin}(F + G) \neq \emptyset$, then the sequence $(v_k)_{k\in\mathbb{N}}$ converges to a point of $\mathrm{Argmin}(F + G)$ and thus to a dual solution. From Lemma 9.1, we have $u_{k+1} = \mathrm{prox}_{\lambda F}(v_k + w_k) = v_k + w_k + \lambda(M\hat{x}_{k+1} - c)$ where

$$\begin{aligned} \hat{x}_{k+1} &\in \underset{x \in \mathrm{dom}(f)}{\arg\min} \left(f(x) + \langle v_k + w_k, Mx - c\rangle + \frac{\lambda}{2}\|Mx - c\|^2 \right) \\ &= \underset{x \in \mathrm{dom}(f)}{\arg\min} \left(f(x) + \langle v_k, Mx - c\rangle + \frac{\lambda}{2}\left\|Mx - c + \frac{1}{\lambda}w_k\right\|^2 - \frac{1}{2\lambda}\|w_k\|^2 \right) \\ &= \underset{x \in \mathrm{dom}(f)}{\arg\min} \left(f(x) + \langle v_k, Mx - c\rangle + \frac{\lambda}{2}\left\|Mx - c + \frac{1}{\lambda}w_k\right\|^2 \right), \end{aligned}$$

for all $k \in \mathbb{N}$. In the sequel, we will use the above trick at several occasions (and thus we will not make appear the second line in what follows). Then, still from Lemma 9.1, we obtain that $v_{k+1} = \mathrm{prox}_{\lambda G}(u_{k+1} - w_k) = \mathrm{prox}_{\lambda G}(v_k + \lambda(M\hat{x}_{k+1} - c)) = v_k + \lambda(M\hat{x}_{k+1} - c) + \lambda N\hat{y}_{k+1}$ where

$$\begin{aligned} \hat{y}_{k+1} &\in \underset{y \in \mathrm{dom}(g)}{\arg\min} \left(g(y) + \langle v_k + \lambda(M\hat{x}_{k+1} - c), Ny\rangle + \frac{\lambda}{2}\|Ny\|^2 \right) \\ &= \underset{y \in \mathrm{dom}(g)}{\arg\min} \left(g(y) + \langle v_k, Ny\rangle + \frac{\lambda}{2}\|M\hat{x}_{k+1} + Ny - c\|^2 \right), \end{aligned}$$

for all $k \in \mathbb{N}$. Finally, we obtain that $w_{k+1} = w_k + v_{k+1} - u_{k+1} = \lambda N\hat{y}_{k+1}$ for all $k \in \mathbb{N}$. Since this last equality is true for any $k \in \mathbb{N}$, we get that $w_k = \lambda N\hat{y}_k$ for all $k \in \mathbb{N}^*$ and thus

$$\hat{x}_{k+1} \in \underset{x \in \mathrm{dom}(f)}{\arg\min} \left(f(x) + \langle v_k, Mx - c\rangle + \frac{\lambda}{2}\|Mx + N\hat{y}_k - c\|^2 \right),$$

for all $k \in \mathbb{N}^*$. Separating the variables x and y in the definition of L^λ, one can easily see that we have obtained

$$\hat{x}_{k+1} \in \underset{x \in \mathrm{dom}(f)}{\arg\min}\ \mathrm{L}^\lambda(x, \hat{y}_k, v_k) \quad \text{and} \quad \hat{y}_{k+1} \in \underset{y \in \mathrm{dom}(g)}{\arg\min}\ \mathrm{L}^\lambda(\hat{x}_{k+1}, y, v_k),$$

and $v_{k+1} = v_k + \lambda(M\hat{x}_{k+1} + N\hat{y}_{k+1} - x)$ for all $k \in \mathbb{N}^*$. Using the change of notation $z := v$, we obtain that ADMM (after a first iteration) coincides with the Douglas-Rachford algorithm applied to the dual problem (D$_2$) and that the sequence $(z_k)_{k\in\mathbb{N}}$ converges to a dual solution (when it exists). From this convergence, it is clear that $M\hat{x}_{k+1} + N\hat{y}_{k+1} - c$ tends to $0_{\mathbb{R}^m}$.

The (long) rest of this proof is dedicated to the last assertion. To this aim, let us denote by $r_{k+1} := M\hat{x}_{k+1} + N\hat{y}_{k+1} - c$ (which tends to $0_{\mathbb{R}^m}$) and by $p_{k+1} := f(\hat{x}_{k+1}) + g(\hat{y}_{k+1})$ for all $k \in \mathbb{N}$. Assume that L_2 admits a saddle-point (x^*, y^*, z^*). In particular, (x^*, y^*) is a primal solution (and thus $Mx^* + Ny^* = c$), z^* is a dual solution and $\mathrm{val}(\mathrm{P}_2) = \mathrm{val}(\mathrm{D}_2) = \mathrm{L}_2(x^*, y^*, z^*) \in \mathbb{R}$. Let us denote by $p^* := f(x^*) + g(y^*) = \mathrm{val}(\mathrm{P}_2)$. Our aim is to prove that $(p_{k+1})_{k\in\mathbb{N}}$ tends to p^*. The proof is decomposed in several steps (in which we will omit the "for all $k \in \mathbb{N}$" all along the text):

(i) Since (x^*, y^*) is a solution to the partial primal problem, we have

$$p^* = \mathrm{L}_2(x^*, y^*, z^*) \leqslant \mathrm{L}_2(\hat{x}_{k+1}, \hat{y}_{k+1}, z^*) = p_{k+1} + \langle z^*, r_{k+1} \rangle.$$

(ii) Since $\hat{x}_{k+1}$ minimizes $\mathrm{L}^\lambda(x, \hat{y}_k, z_k)$ over x, we get that

$$0_{\mathbb{R}^d} \in \partial f(\hat{x}_{k+1}) + M^\top z_k + \lambda M^\top (M\hat{x}_{k+1} + N\hat{y}_k - c).$$

Since $z_{k+1} = z_k + \lambda r_{k+1}$, we get that

$$0_{\mathbb{R}^d} \in \partial f(\hat{x}_{k+1}) + M^\top \Big(z_{k+1} - \lambda(N\hat{y}_{k+1} - N\hat{y}_k) \Big),$$

and thus $\hat{x}_{k+1}$ minimizes $f(x) + \langle z_{k+1} - \lambda(N\hat{y}_{k+1} - N\hat{y}_k), Mx \rangle$ over x. We deduce that

$$f(\hat{x}_{k+1}) + \langle z_{k+1} - \lambda(N\hat{y}_{k+1} - N\hat{y}_k), M\hat{x}_{k+1} \rangle \leqslant f(x^*) + \langle z_{k+1} - \lambda(N\hat{y}_{k+1} - N\hat{y}_k), Mx^* \rangle,$$

that is

$$f(\hat{x}_{k+1}) \leqslant f(x^*) + \langle z_{k+1} - \lambda(N\hat{y}_{k+1} - N\hat{y}_k), Mx^* - M\hat{x}_{k+1} \rangle.$$

(iii) Similarly, since $\hat{y}_{k+1}$ minimizes $\mathrm{L}^\lambda(\hat{x}_{k+1}, y, z_k)$ over y, we get that

$$0_{\mathbb{R}^n} \in \partial g(\hat{y}_{k+1}) + N^\top z_k + \lambda N^\top (M\hat{x}_{k+1} + N\hat{y}_{k+1} - c).$$

Since $z_{k+1} = z_k + \lambda r_{k+1}$, we get that

$$0_{\mathbb{R}^n} \in \partial g(\hat{y}_{k+1}) + N^\top z_{k+1},$$

and thus $\hat{y}_{k+1}$ minimizes $g(y) + \langle z_{k+1}, Ny \rangle$ over y. We deduce that

$$g(\hat{y}_{k+1}) + \langle z_{k+1}, N\hat{y}_{k+1} \rangle \leqslant g(y^*) + \langle z_{k+1}, Ny^* \rangle,$$

that is

$$g(\hat{y}_{k+1}) \leqslant g(y^*) + \langle z_{k+1}, Ny^* - N\hat{y}_{k+1} \rangle.$$

(iv) Adding the two inequalities obtained in (ii) and (iii), and rearranging terms (using in particular $Mx^* + Ny^* = c$), we obtain that

$$p_{k+1} \leqslant p^* - \langle z_{k+1}, r_{k+1} \rangle - \lambda \langle N\hat{y}_{k+1} - N\hat{y}_k, N\hat{y}_{k+1} - Ny^* - r_{k+1} \rangle.$$

Together with (i), we have obtained that

$$-\langle z^*, r_{k+1} \rangle \leqslant p_{k+1} - p^* \leqslant -\langle z_{k+1}, r_{k+1} \rangle - \lambda \langle N\hat{y}_{k+1} - N\hat{y}_k, N\hat{y}_{k+1} - Ny^* - r_{k+1} \rangle.$$

To conclude the proof, in the next items, we will prove that $(N\hat{y}_{k+1} - Ny^*)_{k\in\mathbb{N}}$ is bounded (see item (vii)) and that $(N\hat{y}_{k+1} - N\hat{y}_k)_{k\in\mathbb{N}}$ tends to $0_{\mathbb{R}^m}$ (see item (viii)). But, before going to these two items (vii) and (viii), we first need to obtain other crucial inequalities in items (v) and (vi).

(v) Adding the two inequalities obtained in (i) and (iv) (and multiplying by 2), we obtain that

$$2\langle z_{k+1} - z^*, r_{k+1}\rangle - 2\lambda\langle N\hat{y}_{k+1} - N\hat{y}_k, r_{k+1}\rangle + 2\lambda\langle N\hat{y}_{k+1} - N\hat{y}_k, N\hat{y}_{k+1} - Ny^*\rangle \leqslant 0.$$

- Let us rewrite the first term. To do so, we will use the basic fact $2\langle a + b, b\rangle = \|a + b\|^2 - \|a\|^2 + \|b\|^2$. Since $z_{k+1} = z_k + \lambda r_{k+1}$, note that

$$2\langle z_{k+1} - z^*, r_{k+1}\rangle = \frac{1}{\lambda}\Big(2\langle z_k - z^* + \lambda r_{k+1}, \lambda r_{k+1}\rangle\Big) = \frac{1}{\lambda}(\|z_{k+1} - z^*\|^2 - \|z_k - z^*\|^2) + \lambda\|r_{k+1}\|^2.$$

- Let us rewrite the last term of the above inequality. To do so, we will use the basic fact $\|a\|^2 + \|b\|^2 - \|a - b\|^2 = 2\langle a, b\rangle$. We deduce that

$$2\lambda\langle N\hat{y}_{k+1} - N\hat{y}_k, N\hat{y}_{k+1} - Ny^*\rangle = \lambda\|N\hat{y}_{k+1} - N\hat{y}_k\|^2 + \lambda\|N\hat{y}_{k+1} - Ny^*\|^2 - \lambda\|N\hat{y}_k - Ny^*\|^2.$$

- Now, let us rewrite the last two terms of the above inequality, plus $\lambda\|r_{k+1}\|^2$ from the first item. We get that

$$\begin{aligned}
&\lambda\|r_{k+1}\|^2 - 2\lambda\langle N\hat{y}_{k+1} - N\hat{y}_k, r_{k+1}\rangle + 2\lambda\langle N\hat{y}_{k+1} - N\hat{y}_k, N\hat{y}_{k+1} - Ny^*\rangle \\
&\quad = \lambda\|r_{k+1} - (N\hat{y}_{k+1} - N\hat{y}_k)\|^2 - \lambda\|N\hat{y}_{k+1} - N\hat{y}_k\|^2 + 2\lambda\langle N\hat{y}_{k+1} - N\hat{y}_k, N\hat{y}_{k+1} - Ny^*\rangle \\
&\quad = \lambda\|r_{k+1} - (N\hat{y}_{k+1} - N\hat{y}_k)\|^2 + \lambda(\|N\hat{y}_{k+1} - Ny^*\|^2 - \|N\hat{y}_k - Ny^*\|^2),
\end{aligned}$$

  where the last equality comes from the previous second item.

- Finally, we have obtained

$$\begin{aligned}
&\frac{1}{\lambda}(\|z_{k+1} - z^*\|^2 - \|z_k - z^*\|^2) \\
&\qquad + \lambda\|r_{k+1} - N(\hat{y}_{k+1} - \hat{y}_k)\|^2 + \lambda(\|N\hat{y}_{k+1} - Ny^*\|^2 - \|N\hat{y}_k - Ny^*\|^2) \leqslant 0.
\end{aligned}$$

(vi) Now, let us define

$$V_k := \frac{1}{\lambda}\|z_k - z^*\|^2 + \lambda\|N\hat{y}_k - Ny^*\|^2 \in \mathbb{R}_+.$$

Our aim at this step (vi) is to prove that

$$\lambda\|r_{k+1}\|^2 + \lambda\|N\hat{y}_{k+1} - N\hat{y}_k\|^2 \leqslant V_k - V_{k+1}.$$

From the previous step (v), we have

$$\lambda\|r_{k+1} - (N\hat{y}_{k+1} - N\hat{y}_k)\|^2 \leqslant V_k - V_{k+1}.$$

To obtain what is wanted, we expand the above left term and we only need to prove that the middle term is nonnegative, that is

$$-2\lambda\langle N\hat{y}_{k+1} - N\hat{y}_k, r_{k+1}\rangle \geqslant 0.$$

To see this, recall that $\hat{y}_{k+1}$ minimizes $g(y) + \langle z_{k+1}, Ny\rangle$ (see step (iii)) and similarly $\hat{y}_k$ minimizes $g(y) + \langle z_k, Ny\rangle$, both over y. Thus we have

$$g(\hat{y}_{k+1}) + \langle z_{k+1}, N\hat{y}_{k+1}\rangle \leqslant g(\hat{y}_k) + \langle z_{k+1}, N\hat{y}_k\rangle \quad \text{and} \quad g(\hat{y}_k) + \langle z_k, N\hat{y}_k\rangle \leqslant g(\hat{y}_{k+1}) + \langle z_k, N\hat{y}_{k+1}\rangle,$$

and adding these two inequalities leads to $\langle z_{k+1} - z_k, N\hat{y}_{k+1} - N\hat{y}_k\rangle \leqslant 0$. Using the equality $z_{k+1} - z_k = \lambda r_{k+1}$, we have finished this step.

(vii) From the inequality obtained in (vi), we deduce that $(V_k)_{k\in\mathbb{N}}$ is a decreasing nonnegative sequence and thus converges. We deduce that the sequence $(N\hat{y}_k - Ny^*)_{k\in\mathbb{N}}$ is bounded.

(viii) From the inequality obtained in (vi), we deduce that

$$\lambda \sum_{k \in \mathbb{N}} \left(\|r_{k+1}\|^2 + \|N\hat{y}_{k+1} - N\hat{y}_k\|^2 \right) \leqslant V_0,$$

and thus the sequence $(N\hat{y}_{k+1} - N\hat{y}_k)_{k\in\mathbb{N}}$ tends to $0_{\mathbb{R}^m}$.

Finally, recall from item (iv) that

$$-\langle z^*, r_{k+1} \rangle \leqslant p_{k+1} - p^* \leqslant -\langle z_{k+1}, r_{k+1} \rangle - \lambda \langle N\hat{y}_{k+1} - N\hat{y}_k, N\hat{y}_{k+1} - Ny^* - r_{k+1} \rangle.$$

Since $(r_{k+1})_{k\in\mathbb{N}}$ tends to $0_{\mathbb{R}^m}$, since $(z_k)_{k\in\mathbb{N}}$ converges (to a dual solution that is not necessarily z^*), and from the conclusions of (vii) and (viii), we obtain that $(p_{k+1})_{k\in\mathbb{N}}$ converges to p^*, which concludes the proof. $\square$

Remark 9.4. In Proposition 9.2, we did not detail the *appropriate assumptions* allowing to obtain the convergence results for ADMM (see the proof for some clues). Also note that we did not obtain the convergences of $(\hat{x}_k)_{k\in\mathbb{N}}$ and of $(\hat{y}_k)_{k\in\mathbb{N}}$. Furthermore, we did not prove that, when they converge (to some limits denoted respectively by x^* and y^*), then (x^*, y^*) is a primal solution. Actually, the literature on convergence results for ADMM is very vast and still in progress. It is not our aim in this lecture to investigate in details these technical questions.

Remark 9.5. Consider the framework of Proposition 9.2. When $m = d$ and $M = \mathrm{Id}_d$, then the first step in ADMM

$$\hat{x}_{k+1} \in \operatorname*{arg\,min}_{x \in \mathrm{dom}(f)} \mathrm{L}^\lambda(x, \hat{y}_k, z_k)$$

can simply be rewritten as

$$\begin{aligned}
\hat{x}_{k+1} \in \operatorname*{arg\,min}_{x\in\mathrm{dom}(f)} & \left(f(x) + \langle z_k, x\rangle + \frac{\lambda}{2}\|x + N\hat{y}_k - c\|^2 \right) \\
&= \operatorname*{arg\,min}_{x\in\mathrm{dom}(f)} \left(f(x) + \frac{\lambda}{2}\left\| x + N\hat{y}_k - c + \frac{1}{\lambda} z_k \right\|^2 - \frac{1}{2\lambda}\|z_k\|^2 - \langle z_k, N\hat{y}_k - c\rangle \right) \\
&= \operatorname*{arg\,min}_{x\in\mathrm{dom}(f)} \left(f(x) + \frac{\lambda}{2}\left\| x + N\hat{y}_k - c + \frac{1}{\lambda} z_k \right\|^2 \right) \\
&= \operatorname*{arg\,min}_{x\in\mathrm{dom}(f)} \left(\frac{1}{\lambda} f(x) + \frac{1}{2}\left\| x + N\hat{y}_k - c + \frac{1}{\lambda} z_k \right\|^2 \right).
\end{aligned}$$

Thus, in that case, the first step can simply be rewritten as the proximal step

$$\hat{x}_{k+1} = \mathrm{prox}_{f/\lambda}\left(c - N\hat{y}_k - \frac{1}{\lambda} z_k \right).$$

Remark 9.6. Consider the framework of Proposition 9.2. Similarly to Remark 9.5:

(i) When $m = d$ and $M = \mathrm{Id}_d$, then the first step in ADMM can simply be rewritten as the proximal step

$$\hat{x}_{k+1} = \mathrm{prox}_{f/\lambda}\left(c - N\hat{y}_k - \frac{1}{\lambda} z_k \right).$$

(ii) When $m = d$ and $M = -\mathrm{Id}_d$, then the first step in ADMM can simply be rewritten as the proximal step

$$\hat{x}_{k+1} = \mathrm{prox}_{f/\lambda}\left(N\hat{y}_k - c + \frac{1}{\lambda} z_k \right).$$

(iii) When $m = n$ and $N = \mathrm{Id}_n$, then the second step in ADMM can simply be rewritten as the proximal step

$$\hat{y}_{k+1} = \mathrm{prox}_{g/\lambda}\left(c - M\hat{x}_{k+1} - \frac{1}{\lambda} z_k \right).$$

(iv) When $m = n$ and $N = -\mathrm{Id}_n$, then the second step in ADMM can simply be rewritten as the proximal step

$$\hat{y}_{k+1} = \mathrm{prox}_{g/\lambda}\left(M\hat{x}_{k+1} - c + \frac{1}{\lambda} z_k \right).$$

Remark 9.7. Consider the framework of this section. From Proposition 9.2, the ADMM corresponds to the Douglas-Rachford algorithm applied to the dual problem (D_2). But, actually, one can recover the classical (primal) Douglas-Rachford algorithm (with $\lambda = 1$) by applying the ADMM. Indeed, consider the convex minimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f(x) + g(x),$$

with f, $g \in \Gamma_0(\mathbb{R}^d)$. This problem can be rewritten as

$$\begin{array}{ll} \underset{(x,y)\in\mathbb{R}^d\times\mathbb{R}^d}{\text{minimize}} & f(x) + g(y), \\ \text{subject to} & x - y = 0_{\mathbb{R}^d}, \\ & (x,y) \in \mathrm{dom}(f) \times \mathrm{dom}(g). \end{array}$$

Applying ADMM (with $\lambda = 1$, $m = n = d$, $M = \mathrm{Id}_d$, $N = -\mathrm{Id}_d$ and $c = 0_{\mathbb{R}^d}$), we obtain from Remark 9.6 the algorithm given by

$$\hat{y}_0, z_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \left\{ \begin{array}{l} \hat{x}_{k+1} = \mathrm{prox}_f(\hat{y}_k - z_k), \\ \hat{y}_{k+1} = \mathrm{prox}_g(\hat{x}_{k+1} + z_k), \\ z_{k+1} = z_k + \hat{x}_{k+1} - \hat{y}_{k+1}, \end{array} \right.$$

which exactly corresponds to one writting of the Douglas-Rachford algorithm with $\lambda = 1$ (see Remark 6.6).

Remark 9.8. The ADMM is also well-known in order to numerically solve convex minimization problem of the form

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f(x) + g(Mx),$$

where $f \in \Gamma_0(\mathbb{R}^d)$, $g \in \Gamma_0(\mathbb{R}^m)$ and $M \in \mathbb{R}^{m\times d}$. Indeed, this problem can be rewritten as

$$\begin{array}{ll} \underset{(x,y)\in\mathbb{R}^d\times\mathbb{R}^m}{\text{minimize}} & f(x) + g(y), \\ \text{subject to} & Mx - y = 0_{\mathbb{R}^m}, \\ & (x,y) \in \mathrm{dom}(f) \times \mathrm{dom}(g). \end{array}$$

From Remark 9.6 and for some $\lambda > 0$, the ADMM gives the algorithm

$$\hat{y}_0, z_0 \in \mathbb{R}^m \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \left\{ \begin{array}{l} \hat{x}_{k+1} \in \underset{x\in\mathrm{dom}(f)}{\arg\min}\, \mathrm{L}^\lambda(x, \hat{y}_k, z_k), \\ \hat{y}_{k+1} = \mathrm{prox}_{g/\lambda}\left(M\hat{x}_{k+1} + \dfrac{1}{\lambda} z_k \right), \\ z_{k+1} = z_k + \lambda(M\hat{x}_{k+1} - \hat{y}_{k+1}). \end{array} \right.$$

The first step can numerically be solved via an internal algorithm based on the forward-backward algorithm (if $\mathrm{prox}_{\mu f}$ is known for some $\mu > 0$) or even based on a smooth algorithm if f is smooth.

Appendix: Basics and recalls from convex analysis

A Mathematical framework and notations

• In the whole lecture, we will work in the finite-dimensional $\mathbb{R}$-vector space $\mathbb{R}^d$ with $d \in \mathbb{N}^*$. We will denote by $\langle \cdot, \cdot \rangle$ the usual *scalar product* defined on $\mathbb{R}^d$ by

$$\forall x = (x_1, \dots, x_d) \in \mathbb{R}^d, \quad \forall y = (y_1, \dots, y_d) \in \mathbb{R}^d, \quad \langle x, y \rangle := \sum_{i=1}^{d} x_i y_i,$$

and we denote by $\|\cdot\|$ the corresponding *Euclidean norm*[2] associated and defined by

$$\forall x \in \mathbb{R}^d, \quad \|x\| := \sqrt{\langle x, x \rangle}.$$

Lemma A.1 (Parallelogram identities and classical inequalities)**.** The following properties hold true.

1. *Parallelogram identities:* it holds that

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle,$$
$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2,$$
$$\|x+y\|^2 - \|x-y\|^2 = 4\langle x, y \rangle,$$

for all x, $y \in \mathbb{R}^d$.

2. *Cauchy-Schwarz inequality:* it holds that

$$|\langle x, y \rangle| \leqslant \|x\|\|y\|,$$

for all x, $y \in \mathbb{R}^d$, with equality if and only if x, y are linearly dependent.

3. *Young inequality*: it holds that

$$2\langle x, y \rangle \leqslant \delta\|x\|^2 + \frac{1}{\delta}\|y\|^2,$$

for all x, $y \in \mathbb{R}^d$ and all $\delta > 0$.

Proof. We only prove the third item, based on the inequality $ab \leqslant \frac{1}{2}(a^2+b^2)$. Let x, $y \in \mathbb{R}^d$ and $\delta > 0$. From Cauchy-Schwarz inequality, we get $\langle x, y \rangle \leqslant \|x\|\|y\| = (\sqrt{\delta}\|x\|)(\frac{1}{\sqrt{\delta}}\|y\|) \leqslant \frac{1}{2}(\delta\|x\|^2 + \frac{1}{\delta}\|y\|^2)$. □

Lemma A.2 (Basic properties)**.** The following properties hold true.

1. It holds that

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

for all x, $y \in \mathbb{R}^d$ and all $\lambda \in \mathbb{R}$.

2. It holds that

$$\langle x, y \rangle \geqslant 0 \qquad \Longleftrightarrow \qquad \forall \delta \geqslant 0, \quad \|x\| \leqslant \|x + \delta y\| \qquad \Longleftrightarrow \qquad \forall \delta \in [0,1], \quad \|x\| \leqslant \|x + \delta y\|,$$

for all x, $y \in \mathbb{R}^d$.

Proof. The first item follows from basic computations. Let us prove the second item. Let x, $y \in \mathbb{R}^d$. The implication (i) $\Rightarrow$ (ii) is easy, by proving that $\|x\|^2 \leqslant \|x+\delta y\|^2$ for all $\delta \geqslant 0$, with simple computations. The

[2] At some occasions we will denote the Euclidean norm by $\|\cdot\|_2$, to differentiate it from other norms such as the one-norm $\|\cdot\|_1$ or the infinite-norm $\|\cdot\|_\infty$ (or other general norm denoted by $\mathcal{N}(\cdot)$). Let us take this occasion to recall that all norms on $\mathbb{R}^d$ are equivalent.

implication (ii) $\Rightarrow$ (iii) is trivial. Now, let us prove the implication (iii) $\Rightarrow$ (i). We have $\|x\|^2 \leqslant \|x+\delta y\|^2$, which gives $\delta^2\|y\|^2 + 2\delta\langle x, y\rangle \geqslant 0$, for all $\delta \in [0,1]$. The result is obtained by dividing by $\delta > 0$ and by making $\delta \to 0^+$. $\square$

• When E and F are two subsets of $\mathbb{R}^d$, we denote by $E+F$ and $E-F$ the *Minkowski's sum* and *difference* respectively defined by

$$E+F := \{x+y \mid x \in E,\ y \in F\} \quad \text{and} \quad E-F := \{x-y \mid x \in E,\ y \in F\}.$$

Moreover, when Λ is a subset of $\mathbb{R}$, we will denote by ΛE the *product* defined by

$$\Lambda E := \{\lambda x \mid \lambda \in \Lambda,\ x \in E\}.$$

The above notations will be simplified when the sets are reduced to singletons. For example, if $E = \{x\}$ is a singleton, we will simply denote the above Minkowski's sum $E+F$ by $x+F$, instead of $\{x\}+F$. Similarly, when $\Lambda = \{\lambda\}$ is reduced to a singleton, we will denote the product ΛF by λF, instead of $\{\lambda\}F$.

• In this lecture, we will consider acquired all basic notions from linear algebra such as *linear combinations* of elements, *linear subspaces* and *spanned linear subspaces*, *generating families* and *linearly independent families*, *bases* and *dimensions* of linear subspaces, *supplementary linear subspaces*, *linear hyperplanes*, *linear maps*, *matrices*, etc. Some remarks and recalls:

1. Let E be a subset of $\mathbb{R}^d$. Recall that E is a linear subspace of $\mathbb{R}^d$ if and only if $E \neq \emptyset$ and $\mathbb{R}E + \mathbb{R}E \subset E$. In that case, $0_{\mathbb{R}^d} \in E$ and E is a closed subset of $\mathbb{R}^d$.

2. Recall that linear subspaces of $\mathbb{R}^d$ are stable under intersection. The linear subspace of $\mathbb{R}^d$ spanned by a subset E of $\mathbb{R}^d$, denoted by $\mathrm{span}(E)$, is defined as the intersection of all linear subspaces of $\mathbb{R}^d$ containing E. It is the smallest linear subspace of $\mathbb{R}^d$ containing E.

3. Two elements x, $y \in \mathbb{R}^d$ are said to be *orthogonal* if $\langle x, y\rangle = 0$. In that case, we denote by $x \perp y$. Recall that the *orthogonal set* of a subset E of $\mathbb{R}^d$ is defined by

$$E^\perp := \{y \in \mathbb{R}^d \mid \forall x \in E,\ \langle x, y\rangle = 0\}.$$

   It is a linear subspace of $\mathbb{R}^d$. Furthermore, $\mathrm{span}(E)^\perp = E^\perp$. Also recall that $\mathbb{R}^d = \mathrm{span}(E) \oplus E^\perp$ et $(E^\perp)^\perp = \mathrm{span}(E)$.

4. By definition, a *linear hyperplane* of $\mathbb{R}^d$ is a linear subspace of $\mathbb{R}^d$ of dimension $d-1$. Let E be a subset of $\mathbb{R}^d$. Recall that E is a linear hyperplane of $\mathbb{R}^d$ if and only if there exists $h \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\}$ such that $E = h^\perp$.

• In this lecture, we will occasionally work with other spaces $\mathbb{R}^m$ and $\mathbb{R}^n$ where $m \in \mathbb{N}^*$ and $n \in \mathbb{N}^*$. In such a case, we will denote the scalar products by $\langle \cdot, \cdot \rangle_{\mathbb{R}^d}$, $\langle \cdot, \cdot \rangle_{\mathbb{R}^m}$ and $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ in order to avoid confusions. Similarly, the Euclidean norms will be denoted by $\|\cdot\|_{\mathbb{R}^d}$, $\|\cdot\|_{\mathbb{R}^m}$ and $\|\cdot\|_{\mathbb{R}^n}$.

B Standard notions and basic results on convex sets

B.1 Affine subspaces

Definition B.1 (Line)**.** Let x, $y \in \mathbb{R}^d$. The *line* (x, y) is the subset of $\mathbb{R}^d$ defined by

$$(x, y) := \{(1-\lambda)x + \lambda y \mid \lambda \in \mathbb{R}\}.$$

Definition B.2 (Affine subspace)**.** A subset E of $\mathbb{R}^d$ is said to be an *affine subspace* of $\mathbb{R}^d$ if

$$\forall x, y \in E, \quad (x, y) \subset E,$$

in other words, if

$$\forall x, y \in E, \quad \forall \lambda \in \mathbb{R}, \quad (1-\lambda)x + \lambda y \in E.$$

Remark B.1. If $(A_i)_{i\in\mathcal{I}} \subset \mathbb{R}^d$ is a family of affine subspaces of $\mathbb{R}^d$, then $E := \cap_{i\in\mathcal{I}} A_i$ is an affine subspace of $\mathbb{R}^d$. We say that the affine subspaces of $\mathbb{R}^d$ are stable under intersection.

Definition B.3 (Spanned affine subspace)**.** The affine subspace of $\mathbb{R}^d$ *spanned* by a subset E of $\mathbb{R}^d$, denoted by $\operatorname{aff}(E)$, is defined as the intersection of all affine subspaces of $\mathbb{R}^d$ containing E. It is the smallest affine subspace of $\mathbb{R}^d$ containing E.

Lemma B.1. If A is a nonempty affine subspace of $\mathbb{R}^d$, then, for all $x \in A$, the translated set $A - x$ is a linear subspace of $\mathbb{R}^d$ (which is independent of the element $x \in A$ chosen).

Definition B.4 (Direction and dimension of a nonempty affine subspace)**.** Let A be a nonempty affine subspace of $\mathbb{R}^d$. The linear space A_{dir}, defined by $A_{\text{dir}} := A - x$ for any $x \in A$ (see Lemma B.1), is called *direction* of A. The *dimension* of A is defined as the dimension of A_{dir}.

Definition B.5 (Affine hyperplane)**.** An *affine hyperplane* of $\mathbb{R}^d$ is an affine subspace of $\mathbb{R}^d$ of dimension $d-1$ (that is an affine subspace of $\mathbb{R}^d$ whose direction is a linear hyperplane of $\mathbb{R}^d$).

Proposition B.1. A subset H of $\mathbb{R}^d$ is an affine hyperplane of $\mathbb{R}^d$ if and only if there exists $(h, \alpha) \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\} \times \mathbb{R}$ such that $H = \{x \in \mathbb{R}^d \mid \langle h, x\rangle = \alpha\}$. In that case, we denote by $H = [\langle h, \cdot\rangle = \alpha]$ and its direction H_{dir} satisfies $H_{\text{dir}} = [\langle h, \cdot\rangle = 0] = h^\perp$.

Definition B.6 (Affine half-spaces)**.** Let $H = [\langle h, \cdot\rangle = \alpha]$ be an affine hyperplane of $\mathbb{R}^d$ with $(h, \alpha) \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\} \times \mathbb{R}$. The *closed affine half-spaces* (resp. *open affine half-spaces*) of $\mathbb{R}^d$ associated with H are the two sets $\{x \in \mathbb{R}^d \mid \langle h, x\rangle \leqslant \alpha\}$ and $\{x \in \mathbb{R}^d \mid \langle h, x\rangle \geqslant \alpha\}$ (resp. the two sets $\{x \in \mathbb{R}^d \mid \langle h, x\rangle < \alpha\}$ and $\{x \in \mathbb{R}^d \mid \langle h, x\rangle > \alpha\}$). We symply denote them by $[\langle h, \cdot\rangle \leqslant \alpha]$ and $[\langle h, \cdot\rangle \geqslant \alpha]$ (resp. $[\langle h, \cdot\rangle < \alpha]$ and $[\langle h, \cdot\rangle > \alpha]$).

B.2 Convex sets

Definition B.7 (Segment)**.** Let x, $y \in \mathbb{R}^d$. The *segment* $[x, y]$ is the subset of $\mathbb{R}^d$ defined by

$$[x, y] := \{(1 - \lambda)x + \lambda y \mid \lambda \in [0, 1]\}.$$

The *open segment* $]x, y[$ is the subset of $\mathbb{R}^d$ defined by

$$]x, y[:= \{(1 - \lambda)x + \lambda y \mid \lambda \in]0, 1[\}.$$

We define accordingly the *half-open segments* $]x, y]$ and $[x, y[$.

Definition B.8 (Convex set)**.** A subset E of $\mathbb{R}^d$ is said to be *convex* if

$$\forall x, y \in E, \quad [x, y] \subset E,$$

in other words, if

$$\forall x, y \in E, \quad \forall \lambda \in [0, 1], \quad (1 - \lambda)x + \lambda y \in E.$$

Remark B.2. If $(C_i)_{i \in \mathcal{I}} \subset \mathbb{R}^d$ is a family of convex subsets of $\mathbb{R}^d$, then $E := \cap_{i \in \mathcal{I}} C_i$ is convex. We say that the convex sets are stable under intersection.

Definition B.9 (Convex hull)**.** The *convex hull* of a subset E of $\mathbb{R}^d$, denoted by $\operatorname{conv}(E)$, is defined as the intersection of all convex subsets of $\mathbb{R}^d$ containing E. It is the smallest convex set of $\mathbb{R}^d$ containing E.

B.3 Conic sets and convex conic sets

Definition B.10 (Half-lines)**.** Let x, $y \in \mathbb{R}^d$. The *half-line* $[x, y)$ is the subset of $\mathbb{R}^d$ defined by

$$[x, y) := \{(1 - \lambda)x + \lambda y \mid \lambda \geqslant 0\}.$$

The *open half-line* $]x, y)$ is the subset of $\mathbb{R}^d$ defined by

$$[x, y) := \{(1 - \lambda)x + \lambda y \mid \lambda > 0\}.$$

Definition B.11 (Cone)**.** A subset E of $\mathbb{R}^d$ is said to be a *cone* (or to be *conic*) if

$$\forall x \in E, \quad [0_{\mathbb{R}^d}, x) \subset E,$$

in other words, if

$$\forall x \in E, \quad \forall \lambda \geqslant 0, \quad \lambda x \in E.$$

Remark B.3. Let $(S_i)_{i \in \mathcal{I}} \subset \mathbb{R}^d$ be a family of subsets of $\mathbb{R}^d$ and $S := \cap_{i \in \mathcal{I}} S_i$. If all S_i are conic, then S is conic. Hence, if all S_i are convex conic, then S is convex conic.

Definition B.12 (Convex conic hull)**.** The *convex conic hull* of a subset E of $\mathbb{R}^d$, denoted by $\operatorname{cone}(E)$, is defined as the intersection of all convex conic subsets of $\mathbb{R}^d$ containing E. It is the smallest convex conic subset of $\mathbb{R}^d$ containing E.

B.4 Topology basics

In the sequel, $\overline{\mathrm{B}}(x,\varepsilon)$ (resp. $\mathrm{B}(x,\varepsilon)$) will stand for the *closed ball* (resp. *open ball*) of center $x \in \mathbb{R}^d$ with radius $\varepsilon > 0$, associated to the Euclidean norm $\|\cdot\|$.

Definition B.13 (Bounded sequence)**.** A sequence $(x_k)_{k\in\mathbb{N}}$ of elements of $\mathbb{R}^d$ is said to be *bounded* if

$$\exists M \geqslant 0, \quad \forall k \in \mathbb{N}, \quad \|x_k\| \leqslant M.$$

Definition B.14 (Limit of a sequence)**.** A sequence $(x_k)_{k\in\mathbb{N}}$ of elements of $\mathbb{R}^d$ is said to be *convergent* if

$$\exists x \in \mathbb{R}^d, \quad \forall \varepsilon > 0, \quad \exists K \in \mathbb{N}, \quad \forall k \geqslant K, \quad \|x_k - x\| < \varepsilon.$$

In that case, x is unique and is called the *limit* of the sequence $(x_k)_{k\in\mathbb{N}}$. We denote by $\lim_{k\to\infty} x_k = x$. Otherwise, the sequence $(x_k)_{k\in\mathbb{N}}$ is said to be *divergent*.

Definition B.15 (Cauchy sequence)**.** A sequence $(x_k)_{k\in\mathbb{N}}$ of elements $\mathbb{R}^d$ is said to be a *Cauchy sequence* if

$$\forall \varepsilon > 0, \quad \exists K \in \mathbb{N}, \quad \forall p \geqslant q \geqslant K, \quad \|x_p - x_q\| < \varepsilon.$$

Proposition B.2. Any convergent sequence is a Cauchy sequence, and any Cauchy sequence is bounded.

Theorem B.1. The normed linear space $(\mathbb{R}^d, \|\cdot\|)$ is *complete*, that is, any Cauchy sequence is convergent.

Definition B.16 (Bounded set)**.** A subset E of $\mathbb{R}^d$ is said to be *bounded* if

$$\exists M \geqslant 0, \quad \forall x \in E, \quad \|x\| \leqslant M.$$

Definition B.17 (Open and closed sets)**.** Let E be a subset of $\mathbb{R}^d$. We say that E is *open* if

$$\forall x \in E, \quad \exists \varepsilon > 0, \quad \mathrm{B}(x,\varepsilon) \subset E.$$

We say that E is *closed* if its complement $\mathbb{R}^d \backslash E$ is open.

Proposition B.3 (Sequential characterization of closed sets)**.** Let E be a subset of $\mathbb{R}^d$. Then E is closed if and only if any convergent sequence of elements of E has its limit in E.

Remark B.4. If $(E_i)_{i\in\mathcal{I}} \subset \mathbb{R}^d$ is a family of closed subsets of $\mathbb{R}^d$, then $E := \cap_{i\in\mathcal{I}} E_i$ is closed. We say that the closed sets are stable under intersection.

Definition B.18 (Interior point and adherent point)**.** Let E be a subset of $\mathbb{R}^d$ and $x \in \mathbb{R}^d$. We say that x is an *interior point* of E if

$$\exists \varepsilon > 0, \quad \mathrm{B}(x, \varepsilon) \subset E.$$

We say that x is an *adherent point* of E if

$$\forall \varepsilon > 0, \quad \mathrm{B}(x, \varepsilon) \cap E \neq \emptyset.$$

We denote by $\mathrm{int}(E)$ the set of all interior points of E (called *interior* of E) and by $\mathrm{clos}(E)$ the set of all adherent points of E (called *closure* of E).

Proposition B.4 (Sequential characterization of adherent points)**.** Let E be a subset of $\mathbb{R}^d$ and $x \in \mathbb{R}^d$. Then $x \in \mathrm{clos}(E)$ if and only if there exists a sequence $(x_k)_{k \in \mathbb{N}}$ of elements of E converging to x.

Remark B.5. Let E be a subset of $\mathbb{R}^d$. One can prove that $\mathrm{clos}(E)$ corresponds to the intersection of all closed subsets of $\mathbb{R}^d$ containing E. It is the smallest closed subset of $\mathbb{R}^d$ containing E.

Definition B.19 (Compact set)**.** Let E be a subset of $\mathbb{R}^d$. We say that E is *compact* if, for all sequences of elements of E, we can extract a convergent subsequence whose limit belongs to E.

Proposition B.5. Let E be a subset of $\mathbb{R}^d$. Then, E is compact if and only if E is closed and bounded.

B.5 Topology and convex sets

Proposition B.6. Let C be a convex subset of $\mathbb{R}^d$. Then:

1. $\mathrm{int}(C)$ and $\mathrm{clos}(C)$ are convex.
2. If $\mathrm{int}(C) \neq \emptyset$, then $\mathrm{clos}(\mathrm{int}(C)) = \mathrm{clos}(C)$ and $\mathrm{int}(\mathrm{clos}(C)) = \mathrm{int}(C)$.

Definition B.20 (Closed convex hull)**.** The *closed convex hull* of a subset E of $\mathbb{R}^d$, denoted by $\overline{\mathrm{conv}}(E)$, is defined as the intersection of all closed convex subsets of $\mathbb{R}^d$ containing E. It is the smallest closed convex subset of $\mathbb{R}^d$ containing E.

Definition B.21 (Closed convex conic hull)**.** The *closed convex conic hull* of a subset E of $\mathbb{R}^d$, denoted by $\overline{\mathrm{cone}}(E)$, is defined as the intersection of all closed convex cones of $\mathbb{R}^d$ containing E.

Proposition B.7. Let E be a subset of $\mathbb{R}^d$. It holds that $\overline{\mathrm{conv}}(E) = \mathrm{clos}(\mathrm{conv}(E))$ and $\overline{\mathrm{cone}}(E) = \mathrm{clos}(\mathrm{cone}(E))$.

Definition B.22 (Relative interior)**.** The *relative interior* of a subset E of $\mathbb{R}^d$, denoted by $\mathrm{rint}(E)$, is defined as the interior of E relatively to the induced topology over $\mathrm{aff}(E)$. In other words

$$\mathrm{rint}(E) := \{x \in E \mid \exists \varepsilon > 0,\ \mathrm{B}(x, \varepsilon) \cap \mathrm{aff}(E) \subset E\}.$$

Remark B.6. Let E be a subset of $\mathbb{R}^d$. If $\operatorname{aff}(E) = \mathbb{R}^d$, then $\operatorname{rint}(E) = \operatorname{int}(E)$. Otherwise, $\operatorname{int}(E) = \emptyset$, but it may possible that $\operatorname{rint}(E) \neq \emptyset$.

Proposition B.8. Let E be a subset of $\mathbb{R}^d$. If $x \in \operatorname{rint}(E)$, then, for all $y \in E$, there exists $t > 1$ such that $tx + (1-t)y \in E$.

Theorem B.2. If C is a nonempty convex subset of $\mathbb{R}^d$, then $\operatorname{rint}(C) \neq \emptyset$.

B.6 Normal cone to a convex set

Definition B.23 (Normal cone to a convex set)**.** Let C be a convex subset of $\mathbb{R}^d$ and $x \in C$. The *normal cone* to C at x is defined by

$$\mathrm{N}_C[x] := \{y \in \mathbb{R}^d \mid \forall x' \in C,\ \langle x' - x, y\rangle \leqslant 0\} = \bigcap_{x' \in C} [\langle x' - x, \cdot\rangle \leqslant 0].$$

Note that $\mathrm{N}_C[x]$ is a nonempty closed convex cone of $\mathbb{R}^d$.

C Projection theorem and separation theorems

Definition C.1 (Distance function)**.** The *distance function* $\mathrm{d}_E : \mathbb{R}^d \to \mathbb{R}$ to a nonempty subset E of $\mathbb{R}^d$ is defined by

$$\mathrm{d}_E(x) := \inf_{x' \in E} \|x' - x\|,$$

for all $x \in \mathbb{R}^d$.

Remark C.1. Let E be a nonempty subset of $\mathbb{R}^d$. If E is closed, then the above infimum is a minimum and, for any $x \in \mathbb{R}^d$, we have $x \in E$ if and only if $\mathrm{d}_E(x) = 0$.

Theorem C.1 (Projection theorem)**.** Let C be a nonempty closed convex subset of $\mathbb{R}^d$. Then, for all $x \in \mathbb{R}^d$, there exists a unique $y \in C$ such that

$$\mathrm{d}_C(x) = \|y - x\|.$$

Furthermore, y is characterized by the property

$$y \in C \qquad \text{and} \qquad \forall x' \in C, \quad \langle x - y, x' - y \rangle \leqslant 0.$$

In that case, the element y is called *projection* of x onto C and is denoted by $y = \mathrm{proj}_C(x)$. In particular, we have $\mathrm{d}_C(x) = \|x - \mathrm{proj}_C(x)\|$ for all $x \in \mathbb{R}^d$.

Proposition C.1. Let A be a nonempty affine subspace of $\mathbb{R}^d$. Then, for all $x \in \mathbb{R}^d$, the element $y := \mathrm{proj}_A(x)$ is characterized by the property

$$y \in A \quad \text{and} \quad x - y \in A_{\mathrm{dir}}^{\perp}.$$

Furthermore, the projection operator $\mathrm{proj}_A : \mathbb{R}^d \to A$ is affine. When $A = V$ is a linear subspace of $\mathbb{R}^d$, then the projection operator $\mathrm{proj}_V : \mathbb{R}^d \to V$ is linear and it holds that

$$\forall x \in \mathbb{R}^d, \quad x = \mathrm{proj}_V(x) + \mathrm{proj}_{V^\perp}(x).$$

Definition C.2 (Separating affine hyperplane)**.** Let E and F be two nonempty subsets of $\mathbb{R}^d$.

1. We say that E and F are *separable in the large sense* if there exists $(h, \alpha) \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\} \times \mathbb{R}$ such that
$$E \subset [\langle h, \cdot \rangle \leqslant \alpha] \quad \text{and} \quad F \subset [\langle h, \cdot \rangle \geqslant \alpha].$$
2. We say that E and F are *separable in the strict sense* if there exists $(h, \alpha) \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\} \times \mathbb{R}$ such that
$$\Big[E \subset [\langle h, \cdot \rangle < \alpha] \quad \text{and} \quad F \subset [\langle h, \cdot \rangle \geqslant \alpha]\Big] \qquad \text{or} \qquad \Big[E \subset [\langle h, \cdot \rangle \leqslant \alpha] \quad \text{and} \quad F \subset [\langle h, \cdot \rangle > \alpha]\Big].$$
3. We say that E and F are *separable in the strong sense* if there exist $(h, \alpha) \in \mathbb{R}^d \backslash \{0_{\mathbb{R}^d}\} \times \mathbb{R}$ and $\varepsilon > 0$ such that
$$E \subset [\langle h, \cdot \rangle \leqslant \alpha - \varepsilon] \quad \text{and} \quad F \subset [\langle h, \cdot \rangle \geqslant \alpha + \varepsilon].$$

In these cases, the affine hyperplane $H = [\langle h, \cdot \rangle = \alpha]$ is called *separating affine hyperplane* of E and F (in the large sense, strict or strong, accordingly).

Theorem C.2 (Separation theorems (geometric Hahn-Banach theorems)). Let C_1 and C_2 be two disjoint nonempty convex sets of $\mathbb{R}^d$. Then:

1. They are separable in the large sense.
2. If one is open, then they are separable in the strict sense.
3. If one is compact and the other is closed, then they are separable in the strong sense.

D Standard notions and basic results on convex functions

D.1 Extended-real-valued functions and basics

• In this lecture, we will work with *extended-real-valued functions*, that is, with functions of the form $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$. For example, for any subset E of $\mathbb{R}^d$, we will consider the *indicator function* of E, denoted by $\iota_E : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$, defined by

$$\forall x \in \mathbb{R}^d, \quad \iota_E(x) := \begin{cases} 0 & \text{if} \quad x \in E, \\ +\infty & \text{if} \quad x \notin E. \end{cases}$$

The possibility of working with extended-real-valued functions is very useful in optimization theory, in particular in order to reduce a constrained optimization problem into an unconstrained optimization problem. For example, the generic constrained optimization problem given by

$$\min_{x \in E} f(x),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a real-valued function and $E \subset \mathbb{R}^d$ stands for the constraints set, can be reduced to the unconstrained optimization problem given by

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

where $f + \iota_E : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is an extended-real-valued function. Of course, we loose a lot of regularity properties with this process. Nevertheless, when f and E are convex, numerous results can be easily derived from this simple rewritting.

• In order to manipulate extented-real-valued functions, we need to introduce some conventions:

$$(+\infty) + (+\infty) = +\infty, \qquad \forall \lambda \in \mathbb{R},\ \lambda + (+\infty) = +\infty, \qquad \forall \lambda > 0,\ \lambda(+\infty) = +\infty, \qquad \forall \lambda < 0,\ \lambda(+\infty) = -\infty.$$

$$(-\infty) + (-\infty) = -\infty, \qquad \forall \lambda \in \mathbb{R},\ \lambda + (-\infty) = -\infty, \qquad \forall \lambda > 0,\ \lambda(-\infty) = -\infty, \qquad \forall \lambda < 0,\ \lambda(-\infty) = +\infty.$$

We will also use the conventions $0(+\infty) = 0(-\infty) = 0$. However, note that the expression $(+\infty) + (-\infty)$ has no sense. Finally, we will also need the following conventions:

$$\inf \emptyset = \inf\{+\infty\} = +\infty \qquad \text{and} \qquad \sup \emptyset = \sup\{-\infty\} = -\infty.$$

$$\forall E \subset \mathbb{R} \cup \{\pm\infty\} \text{ nonempty and not bounded below by a real number}, \quad \inf E = -\infty.$$

$$\forall E \subset \mathbb{R} \cup \{\pm\infty\} \text{ nonempty and not bounded above by a real number}, \quad \sup E = +\infty.$$

Finally, when a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is constant, we denote by $f \equiv c$ with $c \in \mathbb{R} \cup \{\pm\infty\}$.

Definition D.1 (Minorant and majorant functions)**.** Let $f, g : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be two functions such that $g(x) \leqslant f(x)$ for all $x \in \mathbb{R}^d$. In that case, we denote by $g \leqslant f$ and we say that g is a *minorant* of f and f is a *majorant* of g.

We introduce similarly the notation $g < f$ when $g(x) < f(x)$ for all $x \in \mathbb{R}^d$.

Definition D.2 (Supremum of a family of functions)**.** The *infimum* and the *supremum* of a family $(f_i)_{i \in \mathcal{I}}$ of functions $f_i : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ for all $i \in \mathcal{I}$ are the functions $\inf_{i \in \mathcal{I}} f_i : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ and $\sup_{i \in \mathcal{I}} f_i : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ respectively defined by

$$\forall x \in \mathbb{R}^d, \quad \left(\inf_{i \in \mathcal{I}} f_i\right)(x) := \inf_{i \in \mathcal{I}} f_i(x) \quad \text{and} \quad \left(\sup_{i \in \mathcal{I}} f_i\right)(x) := \sup_{i \in \mathcal{I}} f_i(x).$$

In particular, we have $\inf_{i \in \mathcal{I}} f_i \leqslant f_i \leqslant \sup_{i \in \mathcal{I}} f_i$ for all $i \in \mathcal{I}$.

Definition D.3 (Conic combination of functions)**.** Let $p \in \mathbb{N}^*$, let $(f_i)_{i=1,\dots,p}$ be a finite family of functions $f_i : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ for all $i \in \{1, \dots, p\}$ and let $(\lambda_i)_{i=1,\dots,p}$ such that $\lambda_i \geqslant 0$ for all $i \in \{1, \dots, p\}$. The *conic combination* $\sum_{i=1}^p \lambda_i f_i : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is the function defined by

$$\forall x \in \mathbb{R}^p, \quad \left(\sum_{i=1}^p \lambda_i f_i\right)(x) := \sum_{i=1}^p \lambda_i f_i(x).$$

Note that the functions $f_i : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ do not take the value $-\infty$ in order to define correctly $\sum_{i=1}^p \lambda_i f_i$.

Definition D.4 (Domain, epigraph and sublevels of a function)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be a function.

1. The *domain* of f is defined by
$$\operatorname{dom}(f) := \{x \in \mathbb{R}^d \mid f(x) < +\infty\}.$$
2. The *epigraph* of f is defined by
$$\operatorname{epi}(f) := \{(x, r) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leqslant r\}.$$
3. The *sublevels* of f are defined by
$$\forall r \in \mathbb{R}, \quad \operatorname{lev}_r(f) := \{x \in \mathbb{R}^d \mid f(x) \leqslant r\} = f^{-1}([-\infty, r]).$$

D.2 Proper, convex and closed functions

Definition D.5 (Proper function)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is said to be *proper* if f does not take the value $-\infty$ and $\operatorname{dom}(f) \neq \emptyset$. In other words, a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is said to be *proper* if $-\infty < f \not\equiv +\infty$.

Definition D.6 (Convex function)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is said to be *convex* if

$$\forall x, y \in \operatorname{dom}(f), \quad \forall \lambda \in [0, 1], \quad f((1-\lambda)x + \lambda y) \leqslant (1-\lambda)f(x) + \lambda f(y).$$

Remark D.1. The distance function $\mathrm{d}_C : \mathbb{R}^d \to \mathbb{R}$ to a nonempty convex subset C of $\mathbb{R}^d$ is convex and 1-Lipschitz continuous.

Remark D.2. Recall that a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ is continuous if and only if any inverse image of a closed subset of $\mathbb{R}$ under f is a closed subset of $\mathbb{R}^d$. Let us introduce a weaker continuity notion, which is moreover valid for extended-real-valued functions.

Definition D.7 (Closed function)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is said to be *closed* (or *lower semicontinuous*) if its sublevels are closed.

Proposition D.1. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be a function. We have that f is convex (resp. closed) if and only if $\operatorname{epi}(f)$ is convex (resp. closed).

Furthermore, when f is closed, for any convergent sequence $(x_k)_{k\in\mathbb{N}}$ of $\mathbb{R}^d$, whose limit is denoted by x, it holds that $\liminf_{k\to\infty} f(x_k) \geqslant f(x)$.

Remark D.3. In the literature, we denote by $\Gamma_0(\mathbb{R}^d)$ the set of all functions $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ that are proper closed and convex. For example, affine functions $f : \mathbb{R}^d \to \mathbb{R}$, that is, of the form $f(x) := \langle a, x\rangle + b$ for all $x \in \mathbb{R}^d$, with $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, belong to $\Gamma_0(\mathbb{R}^d)$. Also, the indicator function ι_E belongs to $\Gamma_0(\mathbb{R}^d)$ if and only if E is a nonempty closed convex subset of $\mathbb{R}^d$.

Proposition D.2. Let $(f_i)_{i\in\mathcal{I}}$ be a family of functions $f_i : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ for all $i \in \mathcal{I}$. Then

$$\operatorname{epi}\left(\sup_{i\in\mathcal{I}} f_i\right) = \bigcap_{i\in\mathcal{I}} \operatorname{epi}(f_i).$$

Thus, if all functions f_i are convex (resp. closed), then $\sup_{i\in I} f_i$ is convex (resp. closed).

Proposition D.3. Let $p \in \mathbb{N}^*$, let $(f_i)_{i=1,\dots,p}$ be a finite family of functions $f_i : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ for all $i \in \{1, \dots, p\}$ and let $(\lambda_i)_{i=1,\dots,p}$ such that $\lambda_i \geqslant 0$ for all $i \in \{1, \dots, p\}$. If all f_i are convex (resp. closed), then $\sum_{i=1}^{p} \lambda_i f_i$ is convex (resp. closed).

D.3 Hulls of a function

Definition D.8 (Convex hull of a function)**.** The *convex hull* of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is the function $\operatorname{conv}(f) : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$\operatorname{conv}(f) := \sup\Big\{g : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\} \mid g \leqslant f \text{ and } g \text{ convex}\Big\}.$$

It is the largest convex minorant of f.

Definition D.9 (Closed hull of a function)**.** The *closed hull* of a function $f : \mathbb{R}^d \to \mathbb{R}\cup\{\pm\infty\}$ is the function $\operatorname{clos}(f) : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$\operatorname{clos}(f) := \sup\Big\{g : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\} \mid g \leqslant f \text{ and } g \text{ closed}\Big\}.$$

It is the largest closed minorant of f.

Definition D.10 (Closed convex hull of a function)**.** The *closed convex hull* of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is the function $\overline{\operatorname{conv}}(f) : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$\overline{\operatorname{conv}}(f) := \sup\Big\{g : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\} \mid g \leqslant f \text{ and } g \text{ closed convex}\Big\}.$$

It is the largest closed convex minorant of f.

Remark D.4. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be a function. It holds that $\operatorname{epi}(\operatorname{clos}(f)) = \operatorname{clos}(\operatorname{epi}(f))$ and $\operatorname{epi}(\overline{\operatorname{conv}}(f)) = \overline{\operatorname{conv}}(\operatorname{epi}(f))$. In particular, $\overline{\operatorname{conv}}(f) = \operatorname{clos}(\operatorname{conv}(f))$. However, it holds that $\operatorname{conv}(\operatorname{epi}(f)) \subset \operatorname{epi}(\operatorname{conv}(f))$ but the inclusion can be strict.

Definition D.11 (Affine hull of a function)**.** The *affine hull* of a function $f : \mathbb{R}^d \to \mathbb{R}\cup\{\pm\infty\}$ is the function $\operatorname{aff}(f) : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$\operatorname{aff}(f) := \sup\Big\{g : \mathbb{R}^d \to \mathbb{R} \mid g \leqslant f \text{ and } g \text{ affine}\Big\}.$$

Remark D.5. Be careful with the notion of affine hull of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$. Indeed, $\mathrm{aff}(f)$ is a closed convex minorant of f, but $\mathrm{aff}(f)$ is not affine in general.

Remark D.6. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be a function. Note that f admits an affine minorant if and only if $\mathrm{aff}(f) \not\equiv -\infty$. In that case, we have $f \geqslant \mathrm{aff}(f) > -\infty$.

Proposition D.4. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be a function. Then

$$\mathrm{aff}(f) \leqslant \overline{\mathrm{conv}}(f) \leqslant \min(\mathrm{clos}(f), \mathrm{conv}(f)) \leqslant \max(\mathrm{clos}(f), \mathrm{conv}(f)) \leqslant f.$$

Theorem D.1. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. Then

$$-\infty < \mathrm{aff}(f) = \overline{\mathrm{conv}}(f) = \mathrm{clos}(f) \leqslant \mathrm{conv}(f) = f \not\equiv +\infty.$$

In particular, f admits an affine minorant. If furthermore f is closed (and thus $f \in \Gamma_0(\mathbb{R}^d)$), then

$$-\infty < \mathrm{aff}(f) = \overline{\mathrm{conv}}(f) = \mathrm{clos}(f) = \mathrm{conv}(f) = f \not\equiv +\infty.$$

E Regularity of proper convex functions

E.1 Proper convex functions and classical regularity notions

Theorem E.1 (Continuity of convex functions)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. Then, $\mathrm{int}(\mathrm{dom}(f)) \neq \emptyset$ if and only if $\mathrm{int}(\mathrm{epi}(f)) \neq \emptyset$. In that case, f is locally Lipschitz continuous on $\mathrm{int}(\mathrm{dom}(f))$.

Definition E.1 (Gâteaux-differentiability)**.** A proper function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is said to be *Gâteaux-differentiable* at a point $x \in \mathrm{dom}(f)$ if there exists $\ell \in \mathbb{R}^d$ such that

$$\forall v \in \mathbb{R}^d, \qquad \lim_{\delta \to 0} \frac{f(x+\delta v) - f(x)}{\delta} = \langle \ell, v \rangle.$$

In that case, we denote by $\ell = \nabla f(x)$ called the *gradient* of f at x.

Theorem E.2. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function that is Gâteaux-differentiable over $\mathrm{dom}(f)$. Then, the following properties are equivalent:

1. f is convex.
2. $\mathrm{dom}(f)$ is convex and
$$\forall x \in \mathrm{dom}(f), \quad \forall y \in \mathbb{R}^d, \quad f(y) - f(x) \geqslant \langle \nabla f(x), y - x \rangle.$$
3. $\mathrm{dom}(f)$ is convex and $\nabla f : \mathrm{dom}(f) \to \mathbb{R}^d$ is *monotone* in the sense that
$$\forall x, y \in \mathrm{dom}(f), \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geqslant 0.$$

Definition E.2 (Fréchet-differentiability)**.** We say that a proper function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is *Fréchet-differentiable* at a point $x \in \mathrm{dom}(f)$ if there exists $\ell \in \mathbb{R}^d$ such that

$$\lim_{\|v\| \to 0} \frac{f(x+v) - f(x) - \langle \ell, v \rangle}{\|v\|} = 0.$$

In that case, f is continuous and Gâteaux-differentiable at x and $\ell = \nabla f(x)$.

Theorem E.3. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. If f is Gâteaux-differentiable at a point $x \in \mathrm{int}(\mathrm{dom}(f))$, then f is Fréchet-differentiable at x.

E.2 Subdifferentiability of proper convex functions

Definition E.3 (Subdifferentiability)**.** A proper function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is said to be *subdifferentiable* at some point $x \in \operatorname{dom}(f)$ if there exists $\ell \in \mathbb{R}^d$ such that

$$\forall y \in \mathbb{R}^d, \quad f(y) - f(x) \geqslant \langle \ell, y - x \rangle.$$

In that case, the element ℓ is called *subgradient* of f at x. We denote by $\partial f(x)$ the set of all subgradients of f at x (called *subdifferential* of f at x).

Remark E.1. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function and $x \in \operatorname{dom}(f)$. Three remarks:

1. Note that $\partial f(x)$ is a closed convex subset of $\mathbb{R}^d$ (possibly empty if f is not subdifferentiable at x).
2. Note that f admits a minimum at x if and only if $0_{\mathbb{R}^d} \in \partial f(x)$.
3. Note that $\ell \in \partial f(x)$ if and only if $\langle \ell, \cdot \rangle - (\langle \ell, x \rangle - f(x))$ is an affine minorant of f.

Theorem E.4. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function with $\operatorname{int}(\operatorname{dom}(f)) \neq \emptyset$. Then, f is subdifferentiable at any $x \in \operatorname{int}(\operatorname{dom}(f))$ and $\partial f(x)$ is a nonempty compact convex subset of $\mathbb{R}^d$.

Proposition E.1. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function.

1. If f is Gâteaux-differentiable at some point $x \in \operatorname{dom}(f)$, then f is subdifferentiable at x with $\partial f(x) = \{\nabla f(x)\}$.
2. If f is subdifferentiable at some $x \in \operatorname{int}(\operatorname{dom}(f))$ with $\partial f(x) = \{\ell\}$ reduced to a singleton, then f is Fréchet-differentiable at x with $\nabla f(x) = \ell$.

E.3 Some subdifferential calculus rules

Proposition E.2. Two basic results:

1. It holds that $\partial \| \cdot \| (0_{\mathbb{R}^d}) = \overline{\mathrm{B}}(0_{\mathbb{R}^d}, 1)$.
2. Let C be a nonempty convex subset of $\mathbb{R}^d$. Then, $\iota_C : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is subdifferentiable at any $x \in C = \operatorname{dom}(\iota_C)$ with $\partial \iota_C(x) = \mathrm{N}_C[x]$.

Theorem E.5. Let f, $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be two proper functions. Then $\operatorname{dom}(f+g) = \operatorname{dom}(f) \cap \operatorname{dom}(g)$ and the two following properties are satisfied:

1. If f and g are both subdifferentiable at some $x \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$, then $f + g$ is subdifferentiable at $x \in \operatorname{dom}(f+g)$ with
$$\partial f(x) + \partial g(x) \subset \partial (f+g)(x).$$
2. Assume furthermore that f and g are convex and satisfy $\operatorname{int}(\operatorname{dom}(f)) \cap \operatorname{dom}(g) \neq \emptyset$. If $f+g$ is subdifferentiable at some point $x \in \operatorname{dom}(f+g)$, then f and g are both subdifferentiable at $x \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$ and
$$\partial (f+g)(x) = \partial f(x) + \partial g(x).$$

Proposition E.3. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function and $\Phi : \mathbb{R}^n \to \mathbb{R}^d$ be an affine map given by $\Phi(x) := \mathcal{M}x + b$ for all $x \in \mathbb{R}^n$, where $\mathcal{M} \in \mathbb{R}^{d\times n}$ is a matrix and $b \in \mathbb{R}^d$ is a vector. The following properties are satisfied:

1. If f is subdifferentiable at some $\mathcal{M}x + b \in \mathrm{dom}(f)$, with $x \in \mathbb{R}^n$, then $f \circ \Phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper function subdifferentiable at $x \in \mathrm{dom}(f \circ \Phi)$ with

$$\mathcal{M}^\top \Big(\partial f(\mathcal{M}x + b)\Big) \subset \partial(f \circ \Phi)(x).$$

2. If moreover f is convex and $\mathcal{M}x + b \in \mathrm{int}(\mathrm{dom}(f))$, then $f \circ \Phi$ is convex and

$$\partial(f \circ \Phi)(x) = \mathcal{M}^\top \Big(\partial f(\mathcal{M}x + b)\Big).$$

Proposition E.4 (Case $d = 1$)**.** Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function with $\mathrm{int}(\mathrm{dom}(f)) \neq \emptyset$. Then

$$\forall x \in \mathrm{int}(\mathrm{dom}(f)), \quad \partial f(x) = [f'_-(x), f'_+(x)],$$

where $f'_-(x)$ and $f'_+(x)$ stand respectively for the left and right derivatives of f at x.

F Legendre-Fenchel conjugate function

F.1 Definition and main theorems

Definition F.1 (Conjugate function)**.** The *(Legendre-Fenchel) conjugate* of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is the function $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$\forall \ell \in \mathbb{R}^d, \quad f^*(\ell) := \sup_{x \in \mathbb{R}^d} \langle \ell, x \rangle - f(x).$$

Note that f^* is a closed convex function (even if f is not).

Proposition F.1 (Properness of the conjugate function)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be a function and $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ being its conjugate. Then, f^* is proper if and only if f is proper and has (at least) one affine minorant. In that case, $f^* \in \Gamma_0(\mathbb{R}^d)$.

Note that, if f is convex, then f^* is proper if and only if f is proper.

Theorem F.1 (Biconjugate theorem)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be a function. Then

$$f^{**} = \operatorname{aff}(f).$$

Remark F.1. If $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ is a function, then $f^{**} \leqslant f$. When $f \in \Gamma_0(\mathbb{R}^d)$, then $f^{**} = f$. Actually, the Legendre-Fenchel conjugation induces a self-bijection of $\Gamma_0(\mathbb{R}^d)$. In particular, when $f \in \Gamma_0(\mathbb{R}^d)$, then $f^* \in \Gamma_0(\mathbb{R}^d)$.

Proposition F.2. Let $f \in \Gamma_0(\mathbb{R}^d)$. Then

$$\forall x \in \mathbb{R}^d, \quad \forall \ell \in \mathbb{R}^d, \quad x \in \operatorname{dom}(f) \text{ and } \ell \in \partial f(x) \Longleftrightarrow \ell \in \operatorname{dom}(f^*) \text{ and } x \in \partial f^*(\ell).$$

Theorem F.2 (Fenchel-Rockafellar)**.** Let $f, g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be two proper convex functions such that $\operatorname{int}(\operatorname{dom}(f)) \cap \operatorname{dom}(g) \neq \emptyset$. Then

$$\inf_{x \in \mathbb{R}^d} f(x) + g(x) = -\min_{\ell \in \mathbb{R}^d} f^*(\ell) + g^*(-\ell).$$

F.2 Some conjugate calculus rules

Proposition F.3. Let $p > 1$. Then

$$\left(\frac{1}{p}\|\cdot\|^p\right)^* = \frac{1}{q}\|\cdot\|^q,$$

where $q := \frac{p}{p-1} > 1$ is the so-called *adjoint* of p.

Proposition F.4. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be a function. Then

1. For $\lambda > 0$, it holds that $(\lambda f)^* = \lambda f^*(\frac{\cdot}{\lambda})$.
2. For $\lambda \neq 0$, it holds that $(f(\lambda \cdot))^* = f^*(\frac{\cdot}{\lambda})$.
3. For $c \in \mathbb{R}$, it holds that $(f + c)^* = f^* - c$.
4. For $a \in \mathbb{R}^d$, it holds that $(f(\cdot + a))^* = f^* - \langle \cdot, a \rangle$.
5. For $a \in \mathbb{R}^d$, it holds that $(f - \langle a, \cdot \rangle)^* = f^*(\cdot + a)$.
6. If $f = \langle a, \cdot \rangle$ is linear, for some $a \in \mathbb{R}^d$, then $f^* = \iota_{\{a\}}$.

F.3 Dual norm and support function

Definition F.2 (Dual norm)**.** Let $\mathcal{N} : \mathbb{R}^d \to \mathbb{R}$ stand for a general norm on $\mathbb{R}^d$. The corresponding *dual norm* is the map $\mathcal{N}^D : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\forall \ell \in \mathbb{R}^d, \quad \mathcal{N}^D(\ell) := \sup_{x \in \mathbb{R}^d \setminus \{0_{\mathbb{R}^d}\}} \frac{|\langle \ell, x \rangle|}{\mathcal{N}(x)} = \max_{\mathcal{N}(x) \leqslant 1} |\langle \ell, x \rangle| = \max_{\mathcal{N}(x) = 1} |\langle \ell, x \rangle|.$$

Note that $\mathcal{N}^D$ is a norm on $\mathbb{R}^d$ and that the *generalized Cauchy-Schwarz inequality*

$$\forall x, \ell \in \mathbb{R}^d, \quad |\langle \ell, x \rangle| \leqslant \mathcal{N}^D(\ell)\mathcal{N}(x),$$

holds true.

Theorem F.3 (Bidual norm theorem)**.** Let $\mathcal{N} : \mathbb{R}^d \to \mathbb{R}$ stand for a general norm on $\mathbb{R}^d$. Then

$$\mathcal{N}^{DD} = \mathcal{N}.$$

Remark F.2. We have $\|\cdot\|_2^D = \|\cdot\|_2$, $\|\cdot\|_1^D = \|\cdot\|_\infty$ and $\|\cdot\|_\infty^D = \|\cdot\|_1$.

Definition F.3 (Support function)**.** The *support function* associated with a subset E of $\mathbb{R}^d$ is the function $\sigma_E : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$\forall \ell \in \mathbb{R}^d, \quad \sigma_E(\ell) := (\iota_E)^*(\ell) = \sup_{x \in E} \langle \ell, x \rangle.$$

Proposition F.5. Let $\mathcal{N} : \mathbb{R}^d \to \mathbb{R}$ stand for a general norm on $\mathbb{R}^d$ and $\overline{\mathrm{B}}_{\mathcal{N}}(0_{\mathbb{R}^d}, 1)$ stand for the corresponding closed unit ball. Denote by $\mathcal{N}^D$ the dual norm associated and by $\overline{\mathrm{B}}_{\mathcal{N}^D}(0_{\mathbb{R}^d}, 1)$ the corresponding closed unit ball. Then:

$$\sigma_{\overline{\mathrm{B}}_{\mathcal{N}}(0_{\mathbb{R}^d}, 1)} = \mathcal{N}^D \qquad \text{and} \qquad \mathcal{N}^* = \iota_{\overline{\mathrm{B}}_{\mathcal{N}^D}(0_{\mathbb{R}^d}, 1)}.$$

Furthermore

$$\forall x \in \mathbb{R}^d, \quad \partial \mathcal{N}(x) = \operatorname*{arg\,max}_{\mathcal{N}^D(\ell) \leqslant 1} \langle \ell, x \rangle.$$

Remark F.3. Note that the previous result extends the equality $\partial \|\cdot\|(0_{\mathbb{R}^d}) = \overline{\mathrm{B}}(0_{\mathbb{R}^d}, 1)$ mentioned in Proposition E.2.

Some exercises on Chapter 1

G On set-valued maps and monotonicity

Exercise G.1. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map.

1. If E_1 and E_2 are two subsets of $\mathbb{R}^d$ such that $E_1 \subset E_2$, prove that $A(E_1) \subset A(E_2)$.
2. If $(E_i)_{i\in\mathcal{I}}$ is a family of subsets of $\mathbb{R}^d$, prove that
$$A\left(\bigcup_{i\in\mathcal{I}} E_i\right) = \bigcup_{i\in\mathcal{I}} A(E_i) \quad \text{and} \quad A\left(\bigcap_{i\in\mathcal{I}} E_i\right) \subset \bigcap_{i\in\mathcal{I}} A(E_i).$$
Provide a counterexample showing that the reverse inclusion in the second above property is not true in general.
3. If E is a subset of $\mathbb{R}^d$, prove that
$$A(\mathbb{R}^d)\backslash A(E) \subset A(\mathbb{R}^d\backslash E).$$
Provide a counterexample showing that the reverse inclusion is not true in general.
4. Justify why all the above properties are still true when replacing A by A^{-1}.

Correction G.1. (1) Trivial. (2) Simple. For a counterexample, take $d = 1$, $A(x) := \{x^2\}$ for all $x \in \mathbb{R}$, $E_1 = \mathbb{R}_-$ and $E_2 = \mathbb{R}_+$. (3) Simple. For a counterexample, take $d = 1$, $A(x) := \{x^2\}$ for all $x \in \mathbb{R}$ and $E = [0, 1]$. (4) Because $A^{-1} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a set-valued map (just as A).

Exercise G.2. Let A, $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be two set-valued maps.

1. If E is a subset of $\mathbb{R}^d$, prove that
$$(A \cup B)(E) = A(E) \cup B(E), \quad (A \cap B)(E) \subset A(E) \cap B(E) \quad \text{and} \quad (A \circ B)(E) = A(B(E)).$$
Provide a counterexample showing that the reverse inclusion in the second above property is not true in general.
2. Prove that $(A \cup B)^{-1} = A^{-1} \cup B^{-1}$, $(A \cap B)^{-1} = A^{-1} \cap B^{-1}$ and $(A \circ B)^{-1} = B^{-1} \circ A^{-1}$.

Correction G.2. (1) Simple. For a counterexample, take $d = 1$, $E = \mathbb{R}$, $A = \mathrm{Id}$ and $B = -\mathrm{Id}$. (2) Simple.

Exercise G.3. Let $f : \mathbb{R} \to \mathbb{R}$ be an increasing function.

1. Prove that f admits left and right limits at each point $x \in \mathbb{R}$, denoted respectively by $f(x^-)$ and $f(x^+)$, satisfying $f(x^-) \leqslant f(x) \leqslant f(x^+)$.
2. Prove that, if $x_1 < x_2$, then $f(x_1^+) \leqslant f(x_2^-)$.
3. Define the set-valued map $A : \mathbb{R} \rightrightarrows \mathbb{R}$ by
$$\forall x \in \mathbb{R}, \quad A(x) := [f(x^-), f(x^+)].$$
Prove that A is monotone.

Correction G.3. (1) Let $x \in \mathbb{R}$. Let us prove that f admits a left limit at x. Take a strictly increasing sequence $(x_k)_{k\in\mathbb{N}}$ converging to x. Since f is increasing, we know that $(f(x_k))_{k\in\mathbb{N}}$ is increasing and bounded above by $f(x)$ and thus converges to some $\ell \leqslant f(x)$. Taking another strictly increasing sequence $(y_k)_{k\in\mathbb{N}}$ converging to x, we obtain that $(f(y_k))_{k\in\mathbb{N}}$ converges to some $\ell' \leqslant f(x)$. Let us prove that $\ell = \ell'$. By contradiction assume that $\ell < \ell'$. There exists $K \in \mathbb{N}$ such that $\ell < f(y_K) \leqslant \ell'$. For $k \in \mathbb{N}$ sufficiently large we have $f(x_k) \leqslant \ell < f(y_K)$

and $x_k > y_K$ and thus $f(x_k) \geqslant f(y_K)$ which is absurd. We conclude that f admits a left limit at x satisfying $f(x^-) \leqslant f(x)$. The reasonning is similar for the right limit $f(x^+)$ at x. (2) Take a sequence $(y_k)_{k\in\mathbb{N}}$ strictly decreasing converging to x_1 and $(z_k)_{k\in\mathbb{N}}$ strictly increasing converging to x_2. For $k \in \mathbb{N}$ sufficiently large, we have $y_k < z_k$ and thus $f(y_k) \leqslant f(z_k)$. Passing to the limit we get $f(x_1^+) \leqslant f(x_2^-)$. (2) Let (x_1, y_1), $(x_2, y_2) \in \mathrm{Gr}(A)$ and let us prove that $(y_2 - y_1)(x_2 - x_1) \geqslant 0$. If $x_1 = x_2$, the result is trivial. Now assume that $x_1 < x_2$ and thus $y_1 \leqslant f(x_1^+) \leqslant f(x_2^-) \leqslant y_2$ and thus $(y_2 - y_1)(x_2 - x_1) \geqslant 0$. The reasonning is similar for $x_2 < x_1$.

Exercise G.4. Some questions about monotonicity of set-valued maps:

1. Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a maximal monotone set-valued map. We know that λA is maximal monotone for all $\lambda > 0$. What happens for $\lambda = 0$?

2. Provide an example of maximal monotone set-valued map $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ such that $\mathrm{Dom}(A) \subsetneq \mathbb{R}^d$.

3. Let A, $B : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be two maximal monotone set-valued maps. We know that $A + B$ is monotone. Is $A + B$ maximal monotone?

Correction G.4. (1) For this exercise, let us denote by $N : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ the null operator defined by $N(x) := \{0_{\mathbb{R}^d}\}$ for all $x \in \mathbb{R}^d$. Note that N is trivially monotone, and even maximal monotone by Minty's theorem. Finally, when $\lambda = 0$, the set-valued map λA is reduced to $A(x) := \{0_{\mathbb{R}^d}\}$ when $x \in \mathrm{Dom}(A)$. Hence, if $\mathrm{Dom}(A) \subsetneq \mathbb{R}^d$, then λA is not maximal monotone, since it is strictly included in N. On the contrary, when $\mathrm{Dom}(A) = \mathbb{R}^d$, then $\lambda A = N$ is maximal monotone. (2) Take $d = 1$ and $A : \mathbb{R} \rightrightarrows \mathbb{R}$ defined by $A(x) := \{0\}$ for $x < 0$, $A(0) := \mathbb{R}_+$ and $A(x) := \emptyset$ for $x > 0$. This set-valued map is trivially monotone, and maximal monotone by Minty's theorem. Moreover it satisfies $\mathrm{Dom}(A) = \mathbb{R}_- \subsetneq \mathbb{R}$. (3) For a counterexample, one has just to consider $A : \mathbb{R} \rightrightarrows \mathbb{R}$ as in the previous question and $B : \mathbb{R} \rightrightarrows \mathbb{R}$ defined by $B(x) := \emptyset$ for $x < 1$, $B(1) := \mathbb{R}_-$ and $B(x) := \{0\}$ for $x > 1$. In that context $A + B$ is reduced to $(A + B)(x) := \emptyset$ for all $x \in \mathbb{R}$ which is clearly monotone but not maximal monotone.

H On general nonexpansive operators

Exercise H.1. Let $T : D \to \mathbb{R}^d$ be an operator.

1. Prove that, if D is convex and T is quasinonexpansive, then $\mathrm{Fix}(T)$ is convex.
2. Prove that, if D is closed and T is continuous, then $\mathrm{Fix}(T)$ is closed.
3. Prove that, if D is closed convex and T is nonexpansive, then $\mathrm{Fix}(T)$ is closed and convex.

Correction H.1. (1) Let $x, y \in \mathrm{Fix}(T)$ and $0 < \lambda < 1$. Let us prove that $z := \lambda x + (1-\lambda)y \in \mathrm{Fix}(T)$. From the first item of Lemma A.2, we have

$$\begin{aligned}\|T(z) - z\|^2 = \|\lambda(T(z) - x) + (1-\lambda)(T(z) - y)\|^2 &= \lambda\|T(z) - x\|^2 + (1-\lambda)\|T(z) - y\|^2 - \lambda(1-\lambda)\|x - y\|^2 \\ &\leqslant \lambda\|z - x\|^2 + (1-\lambda)\|z - y\|^2 - \lambda(1-\lambda)\|x - y\|^2 = \|\lambda(z - x) + (1-\lambda)(z - y)\|^2 = 0,\end{aligned}$$

and thus $z \in \mathrm{Fix}(T)$. (2) Trivial. (3) Trivial from the two previous items.

Exercise H.2. Let $T : D \to \mathbb{R}^d$ be a firmly nonexpansive operator such that D is closed convex. Prove that

$$\mathrm{Fix}(T) = \bigcap_{x \in D} \{y \in D \mid \langle y - T(x), x - T(x)\rangle \leqslant 0\}.$$

In particular, we recover that $\mathrm{Fix}(T)$ is closed convex.

Correction H.2. Denote by E the right set and let us prove that $\mathrm{Fix}(T) = E$. Let $y \in \mathrm{Fix}(T)$ and $x \in D$. From Proposition 2.1, we get that $0 \leqslant \langle T(y) - T(x), (\mathrm{Id} - T)(y) - (\mathrm{Id} - T)(x)\rangle = \langle y - T(x), T(x) - x\rangle$ and thus $y \in E$. Conversely, let $y \in E$. Taking $x = y \in D$ in the intersection, we get that $\|y - T(y)\|^2 \leqslant 0$ and thus $y \in \mathrm{Fix}(T)$.

Exercise H.3 (Browder–Göhde–Kirk theorem)**.** Let $T : D \to D$ be a nonexpansive operator such that D is compact convex. Prove that $\mathrm{Fix}(T) \neq \emptyset$.

To this aim, one can consider $T_k : D \to D$ defined by $T_k(x) := \lambda_k z + (1-\lambda_k)T(x)$ for all $x \in D$ and all $k \in \mathbb{N}$, for some $z \in D$ and some sequence $(\lambda_k)_{k\in\mathbb{N}} \subset (0,1)$ tending to zero.

Correction H.3. From the Picard fixed point theorem, for all $k \in \mathbb{N}$, we denote by $x_k \in D$ the unique fixed point of T_k. From compactness of D, up to a subsequence, the sequence $(x_k)_{k\in\mathbb{N}}$ converges to some $x^* \in D$. Moreover, it holds that $\|x_k - T(x_k)\| = \|T_k(x_k) - T(x_k)\| = \lambda_k\|z - T(x_k)\| \leqslant \lambda_k \mathrm{diam}(D)$ and thus $x_k - T(x_k)$ tends to $0_{\mathbb{R}^d}$. From continuity of T, we get that $T(x^*) = x^*$ and thus $\mathrm{Fix}(T) \neq \emptyset$.

Exercise H.4. Let $(x_k)_{k\in\mathbb{N}}$ be a Fejér monotone sequence with respect to a set E with a nonempty interior. Prove that $(x_k)_{k\in\mathbb{N}}$ converges.

To this aim, one can introduce the sequence $(y_k)_{k\in\mathbb{N}} \subset E$ defined by

$$\forall k \in \mathbb{N}, \quad y_k := \begin{cases} z & \text{if } x_{k+1} = x_k, \\ z - \varepsilon \dfrac{x_{k+1} - x_k}{\|x_{k+1} - x_k\|} & \text{otherwise.} \end{cases}$$

and prove that $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence by establishing that $\sum_{k\in\mathbb{N}} \|x_{k+1} - x_k\| \leqslant \frac{1}{2\varepsilon}\|x_0 - z\|^2$, for some $z \in$ $\mathrm{int}(E)$ *and $\varepsilon > 0$ such that $\overline{\mathrm{B}}_{\mathbb{R}^d}(z, \varepsilon) \subset E$.*

Correction H.4. From Fejér monotonicity, we know that $\|x_{k+1} - y_k\|^2 \leqslant \|x_k - y_k\|^2$ and thus, in both cases and after computations, we get that

$$\|x_{k+1} - z\|^2 \leqslant \|x_k - z\|^2 - 2\varepsilon\|x_{k+1} - x_k\|,$$

for all $k \in \mathbb{N}$. We deduce that

$$\|x_{k+1} - x_k\| \leqslant \frac{1}{2\varepsilon}(\|x_k - z\|^2 - \|x_{k+1} - z\|^2),$$

for all $k \in \mathbb{N}$, and thus $\sum_{k\in\mathbb{N}} \|x_{k+1} - x_k\| \leqslant \frac{1}{2\varepsilon}\|x_0 - z\|^2$. We conclude that $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence and thus converges in $\mathbb{R}^d$.

Exercise H.5. In this exercise, a linear map $M : \mathbb{R}^d \to \mathbb{R}^n$ will be directly considered as a matrix $M \in \mathbb{R}^{n\times d}$ (and conversely). In that context, recall that the *matrix norm* of M is defined by

$$\|M\| := \sup_{x\in\mathbb{R}^d\setminus\{0_{\mathbb{R}^d}\}} \frac{\|Mx\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^d}} = \sup_{\|x\|_{\mathbb{R}^d}\leqslant 1} \|Mx\|_{\mathbb{R}^n} = \sup_{\|x\|_{\mathbb{R}^d}=1} \|Mx\|_{\mathbb{R}^n}.$$

Note that the three above suprema are maxima.

1. Let $M \in \mathbb{R}^{n\times d}$. Prove that $\|M\| = \|M^\top\|$.
2. Let $M \in \mathbb{R}^{d\times d}$. Prove that M is nonexpansive if and only if $\|M\| \leqslant 1$.
3. Let $M \in \mathbb{R}^{d\times d}$. Prove that the following properties are equivalent:
   (a) M is firmly nonexpansive.
   (b) $\|2M - \mathrm{Id}\| \leqslant 1$.
   (c) $\|Mx\|^2 \leqslant \langle Mx, x\rangle$ for all $x \in \mathbb{R}^d$.
   (d) $M^\top$ is firmly nonexpansive.
   (e) $M + M^\top - 2M^\top M$ is positive semidefinite.
   (f) $M + M^\top - 2MM^\top$ is positive semidefinite.
4. Let $M \in \mathbb{R}^{d\times n}$ be a nonzero matrix and let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a β-cocoercive operator, with $\beta > 0$. Prove that $M^\top \circ T \circ M : \mathbb{R}^n \to \mathbb{R}^n$ is $\frac{\beta}{\|M\|^2}$-cocoercive.
5. Let $M \in \mathbb{R}^{d\times n}$ be such that $\|M\| \leqslant 1$ and let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a firmly nonexpansive operator. Prove that $M^\top \circ T \circ M : \mathbb{R}^n \to \mathbb{R}^n$ is firmly nonexpansive.

Correction H.5. (1) We have

$$\begin{aligned}\|M\| = \max_{\|x\|=1} \|Mx\| = \max_{\|x\|=1} \sqrt{\langle Mx, Mx\rangle} = \max_{\|x\|=1} \sqrt{\langle M^\top Mx, x\rangle} \\ = \sqrt{\max_{\|x\|=1} \langle M^\top Mx, x\rangle} = \sqrt{\rho(M^\top M)} = \sqrt{\rho(MM^\top)} = \|M^\top\|,\end{aligned}$$

where $\rho(S)$ denotes the maximal eigenvalue of a symmetric matrix $S \in \mathbb{R}^{d\times d}$. To prove this, we used two facts:

1. $\max_{\|x\|=1}\langle Sx, x\rangle = \rho(S)$ for a symmetric matrix $S \in \mathbb{R}^{d\times d}$. The inequality $\geqslant$ is simple and the inequality $\leqslant$ can be proved by diagonalizing S in an orthonormed basis (which is possible since S is symmetric).
2. $M^\top M$ and $MM^\top$ have the same eigenvalues. Indeed, if $\lambda \neq 0$ is a nonzero eigenvalue of $M^\top M$, then there exists $v \neq 0$ such that $M^\top Mv = \lambda v$ and we obtain that $Mv \neq 0$ is a eigenvector of $MM^\top$ associated with the eigenvalue λ. Otherwise, if $\lambda = 0$ is a eigenvalue of $M^\top M$, then $0 = \det(M^\top M) = \det(M^\top)\det(M) = \det(M)\det(M^\top) = \det(MM^\top)$ and thus 0 is a eigenvalue of $MM^\top$.

(2) Trivial from linearity.

(3) The equivalences $(a) \Leftrightarrow (b) \Leftrightarrow (c)$ are trivial from linearity and Proposition 2.1. The equivalence $(a) \Leftrightarrow (d)$ is trivial since $\|2M^\top - \mathrm{Id}\| = \|(2M - \mathrm{Id})^\top\| = \|2M - \mathrm{Id}\|$ and from the equivalence $(a) \Leftrightarrow (b)$. Now let us prove

the equivalence $(c) \Leftrightarrow (e)$. Assuming (c), we get $\langle (M - M^\top M)x, x\rangle \geqslant 0$ and thus $\langle (M^\top - M^\top M)x, x\rangle \geqslant 0$ for all $x \in \mathbb{R}^d$. Adding them, we get $\langle (M + M^\top - 2M^\top M)x, x\rangle \geqslant 0$ for all $x \in \mathbb{R}^d$. Conversely, assuming (e), one can easily recover (c). Finally the last equivalence $(a) \Leftrightarrow (f)$ comes from the equivalences $(a) \Leftrightarrow (d)$ and $(a) \Leftrightarrow (e)$.

(4) Since $\|M^\top\| = \|M\|$, we have

$$\begin{aligned}\langle M^\top \circ T \circ M(x_2) - M^\top \circ T \circ M(x_1), x_2 - x_1\rangle &= \langle T(M(x_2)) - T(M(x_1)), M(x_2) - M(x_1)\rangle \\ &\geqslant \beta \|T(M(x_2)) - T(M(x_1))\|^2 \geqslant \frac{\beta}{\|M\|^2}\|M^\top \circ T \circ M(x_2) - M^\top \circ T \circ M(x_1)\|^2,\end{aligned}$$

for all x_1, $x_2 \in \mathbb{R}^d$.

(5) If M is the zero matrix, then the result is trivial (since $2 * 0 - \mathrm{Id} = -\mathrm{Id}$ is nonexpansive). If M is not the zero matrix, then we use the previous item with $\beta = 1$ (since T is firmly nonexpansive). We then obtain that $M^\top \circ T \circ M : \mathbb{R}^n \to \mathbb{R}^n$ is $\frac{1}{\|M\|^2}$-cocoercive and thus, since $\|M\| \leqslant 1$, is 1-cocoercive, that is, is firmly nonexpansive.

I On projections on nonempty closed convex sets

Exercise I.1. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a firmly nonexpansive operator and C be a nonempty closed convex subset of $\mathbb{R}^d$.

1. Justify that $\operatorname{proj}_C \circ T \circ \operatorname{proj}_C$ is $\frac{3}{4}$-averaged.
2. If $C = V$ is a linear subspace of $\mathbb{R}^d$, justify that $\operatorname{proj}_V \circ T \circ \operatorname{proj}_V$ is firmly nonexpansive.

Correction I.1. (1) We know that $\operatorname{proj}_C : \mathbb{R}^d \to \mathbb{R}^d$ is firmly nonexpansive, as well as T, and thus they are both $\frac{1}{2}$-averaged. From Proposition 2.6, we know that the composition $\operatorname{proj}_C \circ T \circ \operatorname{proj}_C$ is κ-averaged with $\kappa = \frac{3}{3+\frac{1}{\alpha}-1} = \frac{3}{4}$ where $\alpha = \max\{\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\} = \frac{1}{2}$. (2) When $C = V$ is a linear subspace, we know that $M := \operatorname{proj}_V$ is a nonexpansive linear map, satisfying thus $\|M\| \leqslant 1$. From Exercise H.5, we deduce that $M^\top \circ T \circ M$ is firmly nonexpansive. Thus we only need to guarantee that $M^\top = M$. To see this, note that

$$\begin{aligned}\langle Mx, y\rangle = \langle \operatorname{proj}_V(x), \operatorname{proj}_V(y) + \operatorname{proj}_{V^\perp}(y)\rangle &= \langle \operatorname{proj}_V(x), \operatorname{proj}_V(y)\rangle \\ &= \langle \operatorname{proj}_V(x) + \operatorname{proj}_{V^\perp}(x), \operatorname{proj}_V(y)\rangle = \langle x, \operatorname{proj}_V(y)\rangle = \langle x, My\rangle,\end{aligned}$$

for all x, $y \in \mathbb{R}^d$.

Exercise I.2. Consider some $0 < \mu < 1$ and the three operators defined by

$$\begin{array}{rccl} T_1 : & \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ & x & \longmapsto & T_1(x) := \begin{cases} \left(1 - \dfrac{1}{\|x\|}\right)x & \text{if} \quad \|x\| > 1, \\ 0_{\mathbb{R}^d} & \text{if} \quad \|x\| \leqslant 1 \end{cases} \end{array}$$

$$\begin{array}{rccl} T_2 : & \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ & x & \longmapsto & T_2(x) := \begin{cases} \mu x & \text{if} \quad \|x\| > 1, \\ 0_{\mathbb{R}^d} & \text{if} \quad \|x\| \leqslant 1 \end{cases} \end{array}$$

$$\begin{array}{rccl} T_3 : & \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ & x & \longmapsto & T_3(x) := \begin{cases} \left(1 - \dfrac{2}{\|x\|}\right)x & \text{if} \quad \|x\| > 1, \\ -x & \text{if} \quad \|x\| \leqslant 1 \end{cases} \end{array}$$

Prove that:

1. T_1 is firmly nonexpansive.
2. T_2 is not nonexpansive but is quasinonexpansive. Is this true when $\mu = 1$?
3. T_3 is nonexpansive but not firmly nonexpansive.

Correction I.2. (1) Let us denote by $\overline{\mathrm{B}}$ the unit ball in $\mathbb{R}^d$ (with respect to the Euclidean norm). Since $\overline{\mathrm{B}}$ is a nonempty closed convex subset of $\mathbb{R}^d$, we know that $\operatorname{proj}_{\overline{\mathrm{B}}}$ is well defined and firmly nonexpansive, and thus $T_1 = \mathrm{Id} - \operatorname{proj}_{\overline{\mathrm{B}}}$ is firmly nonexpansive. (2) T_2 is not continuous, and thus is not nonexpansive. Since $\operatorname{Fix}(T_2) = \{0_{\mathbb{R}^d}\}$, one can easily see that T_2 is quasinonexpansive. But, when $\mu = 1$, it holds that $\operatorname{Fix}(T_2) = (\mathbb{R}^d \backslash \overline{\mathrm{B}}) \cup \{0_{\mathbb{R}^d}\}$ and T_2 is not quasinonexpansive. To see this, one has just to consider a sequence $(x_k)_{k\in\mathbb{N}}$ outside of $\overline{\mathrm{B}}$ and a sequence $(y_k)_{k\in\mathbb{N}}$ inside $\overline{\mathrm{B}}$ such that $y_k - x_k \to 0_{\mathbb{R}^d}$. Then we have $\|T_2(y_k) - x_k\| > \|y_k - x_k\|$ for $k \in \mathbb{N}$ sufficiently large, while $x_k \in \operatorname{Fix}(T_2)$. (3) Since T_1 is firmly nonexpansive and $T_3 = 2T_1 - \mathrm{Id}$, we deduce that T_3 is nonexpansive. But it is not firmly nonexpansive since the inequality

$$\|T_3(y) - T_3(x)\|^2 + \|(\mathrm{Id} - T_3)(y) - (\mathrm{Id} - T_3)(x)\|^2 \leqslant \|y - x\|^2,$$

fails with x, $y \in \mathbb{R}^d$ such that $\|x\| = 1$ and $y = -x$.

Exercise I.3. Answer the following questions:

1. Let $T : D \to \mathbb{R}^d$ be an operator. Prove that $\mathrm{Fix}(T) = \mathrm{Fix}(2T - \mathrm{Id})$.
2. Let $T_1, T_2 : \mathbb{R}^d \to \mathbb{R}^d$ be two firmly nonexpansive operators. Define $T := T_1 \circ (2T_2 - \mathrm{Id}) + \mathrm{Id} - T_2$. Noting that $2T - \mathrm{Id} = (2T_1 - \mathrm{Id}) \circ (2T_2 - \mathrm{Id})$, prove that T is firmly nonexpansive and that $\mathrm{Fix}(T) = \mathrm{Fix}((2T_1 - \mathrm{Id}) \circ (2T_2 - \mathrm{Id}))$.
3. Consider the framework of the second item. Prove that, if $T_1 = \mathrm{proj}_A$, where A is a nonempty affine subspace of $\mathbb{R}^d$, then $\mathrm{Fix}(T) = \{x \in \mathbb{R}^d \mid T_1(x) = T_2(x)\}$.
4. Consider the framework of the second item. Prove that, if $T_1 = \mathrm{proj}_V$, where V is a linear subspace of $\mathbb{R}^d$, then $\mathrm{Fix}(R) = \{x \in \mathbb{R}^d \mid T_1(x) = T_2(x)\}$, where $R := T_1 \circ T_2 + (\mathrm{Id} - T_1) \circ (\mathrm{Id} - T_2)$.

Correction I.3. (1) Trivial. (2) Since T_1, T_2 are firmly nonexpansive, then $2T_1 - \mathrm{Id}$, $2T_2 - \mathrm{Id}$ are nonexpansive. Thus $2T - \mathrm{Id}$ is nonexpansive and thus T is firmly nonexpansive. The rest is trivial. (3) Recall that proj_A is an affine map. Let $x \in \mathrm{Fix}(T)$. Then $x = T(x) = \mathrm{proj}_A(2T_2(x) - x) + x - T_2(x) = 2\mathrm{proj}_A(T_2(x)) - \mathrm{proj}_A(x) + x - T_2(x)$, then $T_2(x) = 2\mathrm{proj}_A(T_2(x)) - \mathrm{proj}_A(x) \in A$ and thus $T_2(x) = \mathrm{proj}_A(x)$ which concludes the first inclusion. Conversely, let $x \in \mathbb{R}^d$ such that $T_2(x) = \mathrm{proj}_A(x)$. Then $T_2(x) \in A$ and $\mathrm{proj}_A(T_2(x)) = T_2(x)$ and thus $T_2(x) = 2\mathrm{proj}_A(T_2(x)) - \mathrm{proj}_A(x)$. We deduce that $x = 2\mathrm{proj}_A(T_2(x)) - \mathrm{proj}_A(x) + x - T_2(x) = \mathrm{proj}_A(2T_2(x) - x) + x - T_2(x) = T(x)$ and thus $x \in \mathrm{Fix}(T)$. (4) Since $T_1 = \mathrm{proj}_V$ is a linear map, one can easily prove that $T = R$ and the result follows from the previous item.

Exercise I.4. Let $p \in \mathbb{N}^*$ and $(C_i)_{i=1,\ldots,p}$ be a finite family of nonempty closed convex sets. Assume that there exists $j \in \{1, \ldots, p\}$ such that C_j is bounded. Prove that $\mathrm{Fix}(\mathrm{proj}_{C_1} \circ \ldots \circ \mathrm{proj}_{C_p}) \neq \emptyset$ (and, therefore, that the POCS algorithm can be used).

To this aim, one can define $T := \mathrm{proj}_{C_j} \circ \ldots \circ \mathrm{proj}_{C_p} \circ \mathrm{proj}_{C_1} \circ \ldots \circ \mathrm{proj}_{C_{j-1}}$ *and apply the Browder-Göhde-Kirk theorem (see Exercise H.3) on the restriction of* T *to* C_j.

Correction I.4. Define $T := \mathrm{proj}_{C_j} \circ \ldots \circ \mathrm{proj}_{C_p} \circ \mathrm{proj}_{C_1} \circ \ldots \circ \mathrm{proj}_{C_{j-1}}$ which is nonexpansive by composition. Consider the restriction $\tilde{T} : C_j \to C_j$. Since C_j is a nonempty compact convex set and $\tilde{T}$ is nonexpansive, we deduce from the Browder-Göhde-Kirk theorem that $\mathrm{Fix}(\tilde{T}) \neq \emptyset$. Denote by $z \in C_j$ a fixed point of $\tilde{T}$. One can easily prove that $\mathrm{proj}_{C_1} \circ \ldots \circ \mathrm{proj}_{C_{j-1}}(z) \in \mathrm{Fix}(\mathrm{proj}_{C_1} \circ \ldots \circ \mathrm{proj}_{C_p})$.

Exercise I.5. Answer the following questions.

1. Let E be a nonempty subset of $\mathbb{R}^d$. Prove that, if $(x_k)_{k\in\mathbb{N}}$ is a Fejér monotone sequence with respect to E, then the sequence $(\mathrm{d}_E(x_k))_{k\in\mathbb{N}}$ is decreasing and converges.
2. Let C be a nonempty closed convex subset of $\mathbb{R}^d$. Take a sequence $(x_k)_{k\in\mathbb{N}}$ that is Fejér monotone with respect to C.
   (a) Prove that $(\mathrm{proj}_C(x_k))_{k\in\mathbb{N}}$ converges to some element of C.
   To this aim, one can prove that $\|\mathrm{proj}_C(x_q) - \mathrm{proj}_C(x_p)\|^2 \leqslant \mathrm{d}_C(x_q)^2 - \mathrm{d}_C(x_p)^2$ *for all* $q \geqslant p$.
   (b) Prove that $(x_k)_{k\in\mathbb{N}}$ converges to some element of C if and only if $(\mathrm{d}_C(x_k))_{k\in\mathbb{N}}$ converges to zero.
   (c) Prove that, if there exists $0 \leqslant \kappa < 1$ such that $\mathrm{d}_C(x_{k+1}) \leqslant \kappa \mathrm{d}_C(x_k)$ for all $k \in \mathbb{N}$, then the sequence $(x_k)_{k\in\mathbb{N}}$ converges *linearly* to some element $x \in C$, in the sense that
   $$\|x_k - x\| \leqslant 2\kappa^k \mathrm{d}_C(x_0),$$
   for all $k \in \mathbb{N}$.

3. Let A be a nonempty affine subspace of $\mathbb{R}^d$. Take a sequence $(x_k)_{k\in\mathbb{N}}$ that is Fejér monotone with respect to A. Prove that $\mathrm{proj}_A(x_k) = \mathrm{proj}_A(x_0)$ for all $k \in \mathbb{N}$.

   To this aim, one can prove that $\lambda^2\|\mathrm{proj}_A(x_k)-\mathrm{proj}_A(x_0)\|^2 = \|\mathrm{proj}_A(x_k)-y_\lambda\|^2 \leqslant \mathrm{d}_A(x_0)^2+(1-\lambda)^2\|\mathrm{proj}_A(x_k)-\mathrm{proj}_A(x_0)\|^2$ *for all* $k \in \mathbb{N}$ *and all* $\lambda \in \mathbb{R}$, *where* $y_\lambda := \lambda\mathrm{proj}_A(x_0) + (1-\lambda)\mathrm{proj}_A(x_k) \in A$.

Correction I.5. (1) We know that

$$\forall k \in \mathbb{N}, \quad \forall x \in E, \quad \|x_{k+1} - x\| \leqslant \|x_k - x\|.$$

Passing to the infimum, we deduce that $0 \leqslant \mathrm{d}_E(x_{k+1}) \leqslant \mathrm{d}_E(x_k)$ for all $k \in \mathbb{N}$. Therefore the nonnegative real sequence $(\mathrm{d}_E(x_k))_{k\in\mathbb{N}}$ is decreasing and converges.

(2)(a) Since C is closed, we only need to prove that the sequence $(\mathrm{proj}_C(x_k))_{k\in\mathbb{N}}$ is a Cauchy sequence. Since $(x_k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to C, we get that

$$\begin{aligned}
&\|\mathrm{proj}_C(x_q) - \mathrm{proj}_C(x_p)\|^2 = \|\mathrm{proj}_C(x_q) - x_q\|^2 + \|x_q - \mathrm{proj}_C(x_p)\|^2 + 2\langle \mathrm{proj}_C(x_q) - x_q, x_q - \mathrm{proj}_C(x_p)\rangle \\
&\leqslant \mathrm{d}_C(x_q)^2 + \|x_p - \mathrm{proj}_C(x_p)\|^2 + 2\langle \mathrm{proj}_C(x_q) - x_q, x_q - \mathrm{proj}_C(x_q)\rangle + 2\langle \mathrm{proj}_C(x_q) - x_q, \mathrm{proj}_C(x_q) - \mathrm{proj}_C(x_p)\rangle \\
&\leqslant \mathrm{d}_C(x_p)^2 - \mathrm{d}_C(x_q)^2,
\end{aligned}$$

for all $q \geqslant p$. Since the sequence $(\mathrm{d}_C(x_k))_{k\in\mathbb{N}}$ decreases and converges, it is a Cauchy sequence, as well as the sequence $(\mathrm{proj}_C(x_k))_{k\in\mathbb{N}}$. (2)(b) Let us prove the necessary condition. If $(x_k)_{k\in\mathbb{N}}$ converges to some $x \in C$, then $\mathrm{d}_C(x_k) = \|x_k - \mathrm{proj}_C(x_k)\|$ converges to zero when $k \to +\infty$. Conversely, from the previous item, we know that $\mathrm{proj}_C(x_k)$ converges to some $y \in C$. We deduce, from $\|x_k - \mathrm{proj}_C(x_k)\| = \mathrm{d}_C(x_k)$ converging to zero, that x_k converges to y when $k \to +\infty$. (2)(c) We know that $\mathrm{d}_C(x_k) \leqslant \kappa^k \mathrm{d}_C(x_0)$ for all $k \in \mathbb{N}$ which tends to zero. We deduce from the previous item that $(x_k)_{k\in\mathbb{N}}$ converges to some element $x \in C$. From Fejér monotonicity, we get that

$$\|x_q - x_k\| \leqslant \|x_q - \mathrm{proj}_C(x_k)\| + \|\mathrm{proj}_C(x_k) - x_k\| \leqslant \|x_k - \mathrm{proj}_C(x_k)\| + \mathrm{d}_C(x_k) = 2\mathrm{d}_C(x_k) \leqslant 2\kappa^k \mathrm{d}_C(x_0),$$

for all $k \leqslant q$. Passing to the limit $q \to +\infty$ yields the result.

(3) Let $k \in \mathbb{N}$ and $\lambda \in \mathbb{R}$. Since $A = a + V$ is an affine subspace of $\mathbb{R}^d$, we know that $y_\lambda := \lambda\mathrm{proj}_A(x_0) + (1-\lambda)\mathrm{proj}_A(x_k) \in A$. Since $x_k - \mathrm{proj}_A(x_k) \in V^\perp$ (and $\mathrm{proj}_A(x_k) - y_\lambda \in V$) and from Fejér monotonicity, we get that

$$\begin{aligned}
\lambda^2\|\mathrm{proj}_A(x_k) - \mathrm{proj}_A(x_0)\|^2 &= \|\mathrm{proj}_A(x_k) - y_\lambda\|^2 \\
&\leqslant \|x_k - \mathrm{proj}_A(x_k)\|^2 + \|\mathrm{proj}_A(x_k) - y_\lambda\|^2 = \|x_k - y_\lambda\|^2 \\
&\leqslant \|x_0 - y_\lambda\|^2 = \|x_0 - \mathrm{proj}_A(x_0)\|^2 + \|\mathrm{proj}_A(x_0) - y_\lambda\|^2 \\
&= \mathrm{d}_A(x_0)^2 + (1-\lambda)^2\|\mathrm{proj}_A(x_k) - \mathrm{proj}_A(x_0)\|^2.
\end{aligned}$$

We obtain that

$$(2\lambda - 1)\|\mathrm{proj}_A(x_k) - \mathrm{proj}_A(x_0)\|^2 \leqslant \mathrm{d}_A(x_0)^2.$$

Making tend $\lambda \to +\infty$, we obtain that $\mathrm{proj}_A(x_k) = \mathrm{proj}_A(x_0)$.

Exercise I.6 (von Neumann-Halperin algorithm)**.** Let $p \in \mathbb{N}^*$ and $(A_i)_{i=1,\ldots,p}$ be a finite family of affine subspaces of $\mathbb{R}^d$ such that $A := \cap_{i=1}^p A_i \neq \emptyset$. Prove that the POCS algorithm given by

$$x_0 \in \mathbb{R}^d \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad x_{k+1} = \mathrm{proj}_{A_1} \circ \ldots \circ \mathrm{proj}_{A_p}(x_k),$$

converges to $\mathrm{proj}_A(x_0)$.

To this aim, one can use Exercise I.5 (last item).

Correction I.6. From the POCS algorithm, we know that $(x_k)_{k\in\mathbb{N}}$ converges to some point $x^* \in A$. Let us prove that $x^* = \mathrm{proj}_A(x_0)$. From Exercise I.5, we have $\mathrm{proj}_A(x_k) = \mathrm{proj}_A(x_0)$ for all $k \in \mathbb{N}$ which allows to conclude by taking $k \to +\infty$.

Exercise I.7. Let $\phi : \mathbb{R}^d \to \mathbb{R}^d$ be a nonexpansive linear map and let V be the linear subspace $V := \mathrm{Fix}(\phi)$. Take $x_0 \in \mathbb{R}^d$ and consider $x_{k+1} := \phi(x_k)$ for all $k \in \mathbb{N}$. Prove that $(x_k)_{k\in\mathbb{N}}$ converges to $\mathrm{proj}_V(x_0)$ if and only if $x_k - x_{k+1} \to 0_{\mathbb{R}^d}$.

To this aim, one can prove that $(x_k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to V and use Exercise I.5 (last item) and Lemma 2.6.

Correction I.7. The necessary condition is trivial. Let us prove the sufficient condition. Firstly, since ϕ is nonexpansive, it is clear that $(x_k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(\phi) = V$. From Exercise I.5 (since V is a linear subspace), we have $\mathrm{proj}_V(x_k) = \mathrm{proj}_V(x_0)$ for all $k \in \mathbb{N}$. Furthermore, note that ϕ is continuous and quasinonexpansive and that $\mathrm{Fix}(\phi) = V \neq \emptyset$ (since it contains $0_{\mathbb{R}^d}$). Then, since $x_k - \phi(x_k) = x_k - x_{k+1} \to 0_{\mathbb{R}^d}$, we know from Lemma 2.6 that $(x_k)_{k\in\mathbb{N}}$ converges to some element $v \in \mathrm{Fix}(\phi) = V$. Passing to the limit in the equality $\mathrm{proj}_V(x_k) = \mathrm{proj}_V(x_0)$, we get that $v = \mathrm{proj}_V(x_0)$.

Exercise I.8. Let $T : D \to D$ be a nonexpansive operator with D closed convex.

1. Let $0 < \lambda < 1$ and $x \in D$. Justify that

$$\begin{array}{rccl} \mathfrak{T}_{(\lambda,x)} : & D & \longrightarrow & D \\ & z & \longmapsto & \mathfrak{T}_{(\lambda,x)}(z) := \lambda x + (1-\lambda)T(z), \end{array}$$

   is well-defined and admits a unique fixed point denoted by $x_\lambda \in D$.

2. For all $0 < \lambda < 1$, we introduce the map

$$\begin{array}{rccl} T_\lambda : & D & \longrightarrow & D \\ & x & \longmapsto & T_\lambda(x) := x_\lambda. \end{array}$$

   Answer the following questions:

   (a) Prove that

$$x - x_\lambda = (1-\lambda)(x - T(x_\lambda)) = \frac{1-\lambda}{\lambda}(x_\lambda - T(x_\lambda)),$$

   for all $x \in D$ and all $0 < \lambda < 1$. Deduce that T_λ is firmly nonexpansive for all $0 < \lambda < 1$.

   (b) Prove that $\mathrm{Fix}(T_\lambda) = \mathrm{Fix}(T)$ for all $0 < \lambda < 1$.

   (c) By contradiction, prove that, if $\mathrm{Fix}(T) = \emptyset$, then $\lim_{\lambda\to 0^+} \|x_\lambda\| = +\infty$ for all $x \in D$.

   (d) Prove that

$$\|x - x_\lambda\|^2 + \|x_\lambda - y\|^2 \leqslant \|x - y\|^2,$$

   for all $x \in D$, $y \in \mathrm{Fix}(T)$ and all $0 < \lambda < 1$.

   (e) Prove that, if $\mathrm{Fix}(T) \neq \emptyset$, then $\lim_{\lambda\to 0^+} x_\lambda = \mathrm{proj}_{\mathrm{Fix}(T)}(x)$ for all $x \in D$.

   To this aim, one can use Lemma 2.1 and the characterization of projection theorem.

Correction I.8. (1) Trivial since D is convex and since $\mathfrak{T}_{(\lambda,x)}$ is a contractive map on D complete. (2)(a) Let $0 < \lambda < 1$ and $x \in D$. The two equalities are trivial from simple computations. Let $0 < \lambda < 1$. To prove that T_λ is firmly nonexpansive, let us prove that

$$0 \leqslant \langle T_\lambda(y) - T_\lambda(x), (\mathrm{Id} - T_\lambda)(y) - (\mathrm{Id} - T)(x) \rangle,$$

for all x, $y \in D$. From the two above equalities and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\langle T_\lambda(y) - T_\lambda(x), (\mathrm{Id} - T_\lambda)(y) - (\mathrm{Id} - T)(x)\rangle &= \langle y_\lambda - x_\lambda, (y - y_\lambda) - (x - x_\lambda)\rangle \\ &= \frac{1-\lambda}{\lambda}\langle y_\lambda - x_\lambda, y_\lambda - T(y_\lambda) - (x_\lambda - T(x_\lambda))\rangle \\ &= \frac{1-\lambda}{\lambda}(\|y_\lambda - x_\lambda\|^2 - \langle y_\lambda - x_\lambda, T(y_\lambda) - T(x_\lambda)\rangle) \\ &\geqslant \frac{1-\lambda}{\lambda}(\|y_\lambda - x_\lambda\|)(\|y_\lambda - x_\lambda\| - \|T(y_\lambda) - T(x_\lambda)\|) \geqslant 0,\end{aligned}$$

since T is nonexpansive, for all x, $y \in D$.

(2)(b) Let $0 < \lambda < 1$. Let us prove that $\mathrm{Fix}(T_\lambda) = \mathrm{Fix}(T)$. Let $y \in \mathrm{Fix}(T_\lambda)$. Then $y_\lambda = T_\lambda(y) = y$ and thus, from the first equality in the previous item, we get that $0_{\mathbb{R}^d} = y - T(y_\lambda) = y - T(y)$ and thus $y \in \mathrm{Fix}(T)$. Conversely, let $y \in \mathrm{Fix}(T)$. Then $\mathfrak{T}_{(\lambda,y)}(y) = \lambda y + (1-\lambda)T(y) = \lambda y + (1-\lambda)y = y$ and thus $y \in \mathrm{Fix}(\mathfrak{T}_{(\lambda,y)})$, which means that $y = y_\lambda = T_\lambda(y)$, that is, $y \in \mathrm{Fix}(T_\lambda)$.

(2)(c) By contradiction, assume that it is not true. Then there exists a decreasing sequence $(\lambda_k)_{k\in\mathbb{N}} \subset (0,1)$ tending to zero such that $(x_{\lambda_k})_{k\in\mathbb{N}}$ is bounded and thus admits a cluster point $x^* \in D$ (since D is closed). From the equality $x_\lambda = \mathfrak{T}_{(\lambda,x_\lambda)}(x_\lambda) = \lambda x_\lambda + (1-\lambda)T(x_\lambda)$ (which is true for all $0 < \lambda < 1$ and all $x \in D$), we get by passing to the limit $k \to +\infty$ (and using the continuity of T) that $T(x^*) = x^*$ and thus $x^* \in \mathrm{Fix}(T) \neq \emptyset$ which is absurd.

(2)(d) Let $x \in D$, $y \in \mathrm{Fix}(T)$ and $0 < \lambda < 1$. Since T_λ is firmly nonexpansive and $y \in \mathrm{Fix}(T) = \mathrm{Fix}(T_\lambda)$, we obtain that

$$\|T_\lambda(y) - T_\lambda(x)\|^2 + \|(\mathrm{Id} - T_\lambda)(y) - (\mathrm{Id} - T_\lambda)(x)\|^2 \leqslant \|y - x\|^2,$$

which exactly gives the result desired.

(2)(e) Since T is nonexpansive and D is closed convex, we know that $\mathrm{Fix}(T)$ is closed convex (see Exercise H.1). Since moreover $\mathrm{Fix}(T) \neq \emptyset$, we know that $\mathrm{proj}_{\mathrm{Fix}(T)} : \mathbb{R}^d \to \mathbb{R}^d$ is well-defined. Let $x \in D$ and $(\lambda_k)_{k\in\mathbb{N}} \subset (0,1)$ be a decreasing sequence tending to zero and let us define $z_k := x_{\lambda_k}$ for all $k \in \mathbb{N}$. Our aim is to prove that $(z_k)_{k\in\mathbb{N}}$ converges to $\mathrm{proj}_{\mathrm{Fix}(T)}(x)$. To this aim (from Lemma 2.1), we only have to prove that $(z_k)_{k\in\mathbb{N}}$ is bounded and admits at most one cluster point (which has to be $\mathrm{proj}_{\mathrm{Fix}(T)}(x)$). Firstly, from the previous item, it is clear that $(z_k)_{k\in\mathbb{N}}$ is bounded since

$$\|z_k\| \leqslant \|x - z_k\| + \|x\| \leqslant \|x - y\| + \|x\|,$$

for any fixed $y \in \mathrm{Fix}(T) \neq \emptyset$, and for all $k \in \mathbb{N}$. Now consider some $z^* \in D$ a cluster point of $(z_k)_{k\in\mathbb{N}}$ (recall that D is closed). If we prove that $z^* = \mathrm{proj}_{\mathrm{Fix}(T)}(x)$, then we have finished. To this aim, we will use the characterization of projection theorem. Passing to the limit in the first equality in the second item (on the subsequence of $(z_k)_{k\in\mathbb{N}}$ converging to z^*) and the continuity of T, we obtain that $z^* \in \mathrm{Fix}(T)$. From the previous item, we obtain that

$$\|x - z_k\|^2 + \|z_k - y\|^2 \leqslant \|x - y\|^2,$$

for all $y \in \mathrm{Fix}(T)$ and all $k \in \mathbb{N}$. Passing to the limit (on the subsequence of $(z_k)_{k\in\mathbb{N}}$ converging to z^*), we get that

$$\|x - z^*\|^2 + \|z^* - y\|^2 \leqslant \|x - y\|^2,$$

which gives

$$\langle x - z^*, y - z^*\rangle \leqslant 0,$$

for all $y \in \mathrm{Fix}(T)$. From characterization of projection theorem, we obtain that $z^* = \mathrm{proj}_{\mathrm{Fix}(T)}(x)$.

References

The present lecture of "Splitting methods for convex optimization" has been elaborated thanks to many references, among which (in alphabetical order):

- Qamrul Hasan Ansari, Elisabeth Köbis and Jen-Chih Yao. *Vector variational inequalities and vector optimization.* Book.
- Jean-Pierre Aubin and Hélène Frankowska. *Set-valued analysis.* Book.
- Heinz Bauschke and Patrick Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Book.
- Amir Beck. *First-order methods in optimization.* Book.
- Loïc Bourdin. *Analyse convexe.* Master's lecture (in french).
- Stephen Boyd and Neal Parikh. *Proximal algorithms.* Article.
- Stephen Boyd, Neal Parikh, Eric Chu Borja Peleato and Jonathan Eckstein. *Distributed optimization and statistical learning via the Alternating Direction Method of Multipliers.* Article.
- Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Book.
- Haïm Brezis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert.* Book (in french).
- Luis Briceno-Arias. *Splitting algorithms for solving composite convex optimisation problems: a survey and new results.* Slides for oral presentation.
- Giovanni Chierchia, Emilie Chouzenoux, Patrick Combettes and Jean-Christophe Pesquet. *The proximity operator repository.* User's guide of proximity-operator.net.
- Patrick Combettes. *Monotone operator theory in convex optimization.* Article.
- Patrick Combettes and Jean-Christophe Pesquet. *Proximal splitting methods in signal processing.* Article.
- Charles Dossal. *Optimisation convexe.* Master's lecture (in french).
- Damek Davis and Wotao Yin. *A three-operator splitting scheme and its optimization applications.* Article.
- Jonathan Eckstein. *Augmented Lagrangian and Alternating Direction Methods for convex optimization: a tutorial and some illustrative computational results.* Research report.
- Jean-Charles Gilbert. *Fragments d'optimisation différentiable.* Book (in french).
- Osman Güller. *Foundations of optimization.* Book.
- Niao He. *Splitting algorithms.* Lecture.
- Ralph Tyrrell Rockafellar and Roger Wets. *Variational analysis.* Book.