INSA Toulouse
GMM 5eme annee, Image

1 Lesson 2 : Optimization of differentiable functions

In this part we consider the following general problem:

$$\min_{x \in E} F(x) \tag{1}$$

where F is a real valued, convex, coercive, differentiable function defined on a Hilbert space $\mathcal{H}$, and where E is a closed, convex set. As F is differentiable, it is lower semi-continuous and proper by definition. These hypotheses ensure that the optimization problem (1) admits at least one solution. Additionally, if F is strictly convex, then the solution is unique.

1.1 Properties of differentiable functions

Convex, differentiable functions have the property of being bounded from below by their affine approximations:

Proposition 1 *Let f be a convex, differentiable function defined from $\mathbb{R}^n$ to $\mathbb{R}$. Then for all $(x, y) \in (\mathbb{R}^n)^2$, we have*

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle \tag{2}$$

Proof:
By convexity of f, we have for all $\theta \in]0, 1[$,

$$f(x + \theta(y - x)) \leqslant (1 - \theta)f(x) + \theta f(y).$$

which can be rewritten

$$\frac{f(x + \theta(y - x)) - f(x)}{\theta} \leqslant f(y) - f(x).$$

By letting θ go to zero we obtain

$$\langle \nabla f(x), y - x \rangle \leqslant f(y) - f(x)$$

By applying this proposition to the points x and y, we deduce that the gradient of f is monotone:

Definition 1 *Let g be a function defined from $\mathbb{R}^n$ to $\mathbb{R}^n$. We say g is monotone if for all $(x, y) \in (\mathbb{R}^n)^2$,*

$$\langle g(x) - g(y), x - y\rangle \geqslant 0.$$

Corrollary 1 *Let f be a convex, differentiable function defined from $\mathbb{R}^n$ to $\mathbb{R}^n$. Then the gradient of f is monotone.*

There exists a different notion of convexity, called *strong convexity*, that allows us to locally control the function in a more precise way:

Definition 2 *Let f be a function defined from E to $\mathbb{R} \cup \{+\infty\}$ and $\alpha > 0$. We say that the function f is $\alpha-$strongly convex (or $\alpha-$convex) if the function g defined by $g(x) = f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.*

We may note (exercice!) that if there exists x_0 such that the function $g_{x_0} = f(x) - \alpha\|x - x_0\|^2$ is convex, then f is $\alpha-$convex, and that a strongly convex function is strictly convex. In particular, such a function admits a unique minimizer.
By definition, if f is convex and y is an element of E, the function $x \mapsto f(x) + \frac{1}{2\gamma}\|x - y\|$ is $\frac{1}{\gamma}$-strongly convex.
The following lemma shows that if we move away from the minimizer of a strongly convex functional, the value of the functional grows at least quadratically. This lemma will be useful in what follows.

Lemma 1 *Let f be an $\alpha-$strongly convex function, and x^* be the minimizer of f. Then for all $x \in E$ we have*

$$f(x^*) + \frac{\alpha}{2}\|x^* - x\|^2 \leqslant f(x) \tag{3}$$

Proof:
We denote by g the function $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$, which is convex by definition. Let $\lambda \in]0, 1[$. By convexity of g we have

$$\begin{aligned}
g(\lambda x^* + (1-\lambda)x) &\leqslant \lambda g(x^*) + (1-\lambda)g(x) \\
f(\lambda x^* + (1-\lambda)x) - \frac{\alpha}{2}\|\lambda x^* + (1-\lambda)x\|^2 &\leqslant \lambda\left(f(x^*) - \frac{\alpha}{2}\|x^*\|^2\right) + (1-\lambda)\left(f(x) - \frac{\alpha}{2}\|x\|^2\right) \\
(1-\lambda)\left(f(x^*) - f(x)\right) &\leqslant \frac{\alpha}{2}\left((\lambda^2 - \lambda)\|x^*\|^2 + 2\lambda(1-\lambda)\langle x^*, x\rangle - \lambda(1-\lambda)\|x\|^2\right) \\
f(x^*) - f(x) &\leqslant \frac{-\lambda\alpha}{2}\|x^* - x\|^2 \\
f(x^*) + \lambda\frac{\alpha}{2}\|x^* - x\|^2 &\leqslant f(x)
\end{aligned}$$

where we used the fact that x^* is a minimizer on the third line. This is true for all $\lambda \in]0,1[$, which concludes the proof of this lemma.

We will also need the definition of the Lipschitz continuity of function:

Definition 3 *A function g defined from $\mathbb{R}^n$ to $\mathbb{R}^n$ is said to be $L-$Lipschitz (or $L-$Lipschitz continuous) if for all $(x,y) \in (\mathbb{R}^n)^2$, we have*

$$\|g(x) - g(y)\| \leqslant L\|x - y\|$$

If g is 1-Lipschitz we say that g is non expansive and if g is $L-$Lipschitz with $L < 1$ we say that g is L-contracting or just a contraction.

We just saw that convex functions are bounded from below by their affine approximations. However, if the gradient of f is L-Lipschitz, we can obtain an upper bound of f regardless of its convexity:

Lemma 2 *Let f be a differentiable function defined from $\mathbb{R}^n$ to $\mathbb{R}^n$, with an $L-$Lipschitz gradient. Then for all $(y,z) \in (\mathbb{R}^n)^2$, we have*

$$f(z) \leqslant f(y) + \langle \nabla f(y), z - y \rangle + \frac{L}{2}\|z - y\|^2. \tag{4}$$

Proof:
Let g be a differentiable function on $\mathbb{R}$ such that g' is $K-$Lipschitz. Then

$$g(1) = g(0) + \int_0^1 g'(t)dt = g(0) + g'(0) + \int_0^1 (g'(t) - g'(0))dt \leqslant g(0) + g'(0) + \frac{K}{2}.$$

Let $(y,z) \in (\mathbb{R}^n)^2$. For all $t \in [0,1]$, we set $v = z - y$, $y_t = y + t(z - y)$ and $g(t) = f(y_t)$. The function g is differentiable and $g'(t) = \langle \nabla f(y_t), v \rangle$. According to the hypotheses on f, g' is K-Lipchitz with $K = L\|v\|^2$.
From this we deduce that for all $(y,z) \in (\mathbb{R}^n)^2$,

$$f(z) \leqslant f(y) + \langle \nabla f(y), z - y \rangle + \frac{L}{2}\|z - y\|^2.$$

We can also establish a more restrictive notion than the non-expansivity, which will be useful in what follows:

Definition 4 *A function f is said to be firmly non-expansive if for all $(x,y) \in (\mathbb{R}^n)^2$,*

$$\|f(x) - f(y)\|^2 + \|x - f(x) - y + f(y)\|^2 \leqslant \|x - y\|^2$$

A classic example of a firmly non-expansive function is the case of a projection onto a closed convex set. We will also see that there are other examples that are just as interesting.

Proposition 2 *Let f be a differentiable, convex function defined on $\mathbb{R}^n$ with an $L-$Lipschitz gradient. If $\gamma < \dfrac{1}{L}$, then the map $Id - \gamma\nabla f$ is firmly non-expansive and thus 1-Lipschitz.*

We will see that this property is crucial when proving the convergence of many optimization methods, as many optimization methods are based on fixed point theorems associated to 1-Lipschitz operators.

Proof:

The proof is based on determining the sign of a scalar product.

Let $(x, y) \in (\mathbb{R}^n)^2$. We set $u = \dfrac{1}{\gamma}(\nabla f(x) - \nabla f(y))$ and $v = x - \dfrac{1}{\gamma}\nabla f(x) - (y - \dfrac{1}{\gamma}\nabla f(y))$. Then

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

In order to show that $Id - \gamma\nabla f$ is firmly non-expansive it is sufficient to show that the scalar product $\langle u, v\rangle$ is non-negative, which follows directly from the fact that $\gamma < \frac{1}{L}$ and from the co-coercivity of ∇f which is defined and proved in the following Lemma.

Lemma 3 *Let f be a differentiable, convex function such that there exists $L > 0$ such that for all $(x, y) \in (\mathbb{R}^n)^2$, $\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|$. We have the following relation:*

$$\forall (x, y) \in (\mathbb{R}^n)^2, \quad \langle \nabla f(x) - \nabla f(y), x - y\rangle \geqslant \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2. \tag{5}$$

This property is called co-coercivity of the function ∇f.

Proof:

We apply the majoration (4), and by letting $z = x$ and $y = x - \dfrac{1}{L}\nabla f(x)$ we obtain

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leqslant f(x) - f(x - \frac{1}{L}\nabla f(x)) \leqslant f(x) - f(x^\star).$$

where $x^\star$ is a minimizer of f. We also have, for all $x \in \mathbb{R}^n$,

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leqslant f(x) - f(x^\star). \tag{6}$$

Let $(x, y) \in (\mathbb{R}^n)^2$. We then introduce the functions $h_1(z) = f(z) - \langle \nabla f(x), z\rangle$ and $h_2(z) = f(z) - \langle \nabla f(y), z\rangle$. These two functions are convex and admit the minimizers x et y respectively. We apply inequality (6) to these two functions and obtain

$$h_1(x) \leqslant h_1(y) - \frac{1}{2L}\|\nabla h_1(y)\|^2 \text{ and } h_2(y) \leqslant h_2(x) - \frac{1}{2L}\|\nabla h_2(y)\|^2.$$

By adding these two inequalities we obtain the result of the lemma.

A classical theoretical framework ensuring $L-$Lipschitz gradients is the case of twice differentiable functions: If we are able to bind the operator norm of the Hessian matrix by L, then we deduce that the gradient of f is $L-$Lipschitz and we may apply the previous results.

1.2 Optimality conditions

The minimizers x^* of a convex, differentiable function f defined on $\mathbb{R}^n$ are simply characterized by the Euler equation:

$$\nabla f(x^*) = 0 \tag{7}$$

A different way to formulate this is that all the critical points of a convex function are global minima.
The difficulties appear when we treat the case of constrained problems, i.e. problems of the form

$$\min_{x \in K} f(x) \tag{8}$$

where K is a closed, convex set.
We characterize the minimizers in the following way:

Theorem 1 *Let f be a differentiable function and K a closed, convex set. x^* is a solution to* (8) *if and only if for all $v \in K$,*

$$\langle \nabla f(x^*), v - x^*\rangle \geqslant 0 \tag{9}$$

1.3 Gradient descent methods

In the case where f is a convex, differentiable function, there exist many gradient descent methods that construct minimizing sequences. In this section we will describe the most classical methods using only the gradient of f to approach solutions of the following problem:

$$\min_{u \in X} f(x). \tag{10}$$

1.3.1 Gradient descent with fixed step

The gradient descent method with fixed step is quite easy to describe. We consider a strictly positive real number γ called *step*, and an element $x_0 \in X$. For all $n \in \mathbb{N}$ we define:

$$x_{n+1} = x_n - \gamma \nabla f(x_n). \tag{11}$$

If we note T the operator defined from X to X by

$$Tx = x - \gamma \nabla f(x) = (I - \gamma \nabla f)(x),$$

the sequence $(x_n)_{n \in \mathbb{N}}$ is simply defined by $x_{n+1} = Tx_n$.

If no hypothesis is made on the gradient of f, a method like this may not converge to a minimizer, as the algorithm may diverge if γ is chosen too large (examples: oscillation and divergence). On the other hand, if the gradient of f is L-Lipschitz, then by choosing γ correctly, this method converges to a minimizer of f:

Theorem 2 *Let f be a convex, differentiable function with an $L-$Lipschitz gradient, admitting a minimizer. If $\gamma < \frac{2}{L}$ then for all $x_0 \in X$, the sequence defined for all $n \in \mathbb{N}$ by $x_{n+1} = Tx_n$ converges to a minimizer of f.*

Proof:
There exist different proofs of this convergence, we will propose one based on the non-expansivity of the operator T. This proof may also be applied to other algorithms. We will thus use the following lemma after having showed that the operator T associated to the explicit gradient descent satisfies the hypotheses of the lemma:

Lemma 4 *Let T be a 1-Lipschitz operator defined from X to X admitting a fixed point. Let $x_0 \in X$ and let $(x_n)_{n\in\mathbb{N}}$ be the sequenced defined by $x_{n+1} = Tx_n$.*
If $\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0$ then the sequence $(x_n)_{n\in\mathbb{N}}$ converges to a fixed point of T.

Proof:
Let y be a fixed point of T. As T is 1-Lipschitz, the sequence $(\|x_n - y\|)_{n\in\mathbb{N}}$ is decreasing and therefore bounded. The sequence $(x_n)_{n\in\mathbb{N}}$ is thus bounded.

As X is of finite dimension, the sequence $(x_n)_{n\in\mathbb{N}}$ admits an adherent point (or closure point) $z \in X$. As $\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0$, this adherent point satisfies $Tz = z$ and is thus a fixed point of T.

The sequence $\|z - x_n\|$ is thus decreasing and admits a subsequence that goes to zero, the sequence thus converges to zero, which concludes the proof.

In order to prove Theorem 2, we will show that the operator T satisfies all the hypotheses of the lemma. The fact that T is 1-Lipschitz is the result of the following lemma. Note that if $\gamma \leqslant \frac{1}{L}$, the operator T satisfies an even more interesting property, the firmly non-expansivity.

Lemma 5 *Let f be a convex, differentiable function with an $L-$Lipschitz gradient. For $\gamma > 0$ we note $T = Id - \gamma\nabla f$. Then*

1. *If $\gamma \leqslant \frac{1}{L}$, the operator T is firmly non expansive, that is:*

$$\forall (x, y) \in X^2 \quad \|Tx - Ty\|^2 + \|x - Tx - y + Ty\|^2 \leqslant \|x - y\|^2. \tag{12}$$

2. *If $\gamma \leqslant \frac{2}{L}$, the operator T is 1-Lipschitz.*

The proof of this lemma (found in Appendix) is based on the cocoercive character of the gradient of a convex, differentiable function. Furthermore, we may note that there is equivalence between being a fixed point of T and being a minimizer of f, as $Tx = x$ is equivalent to $\nabla f(x) = 0$. The hypotheses of Theorem 2 suppose the existence of such a minimizer and thus the existence of a fixed point of T. To finish the proof of the theorem it is sufficient to show that $\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0$. For this we will apply Lemma 2 to the points $y = x_n$ and $z = x_{n+1}$. We then have $z - y = \gamma\nabla f(x_n)$ and thus

$$f(x_{n+1}) \leqslant f(x_n) - \frac{1}{\gamma}\|x_{n+1} - x_n\|^2 + \frac{L}{2}\|x_{n+1} - x_n\|^2 \tag{13}$$

and thus if $\gamma < \frac{2}{L}$,

$$f(x_{n+1}) + \left(\frac{2-\gamma L}{2\gamma}\right) \|x_{n+1} - x_n\|^2 \leqslant f(x_n) \tag{14}$$

As the sequence $(f(x_n))_{n\in\mathbb{N}}$ is bounded from below we deduce that $\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0$, which concludes the proof of the theorem. We may note that the gradient descent method with fixed step is a *descent method*, i.e. the value of $f(x_n)$ decreases. When studying the Forward-Backward algorithm (which is a generalization of this algorithm), we will see that it is possible to control the rate at which $f(x_n) - f(x^*)$ decreases towards zero, where x^* refers to an arbitrary minimizer of f.

1.3.2 Implicit gradient descent with fixed step

The implicit gradient descent with fixed step can be expressed as simply as the explicit method but raises a practical problem of solving an implicit problem at each step.

We consider a strictly positive real number γ called *step* and an element $x_0 \in X$. For all $n \in \mathbb{N}$ we define:

$$x_{n+1} = x_n - \gamma \nabla f(x_{n+1}). \tag{15}$$

Without this hypothesis on f this relation does not guarantee the existance and uniqueness of x_{n+1}, and the resolution of this implicit equation can be practically complicated.

The convexity of f ensures the existance and uniqueness of x_{n+1}, as we may note that x_{n+1} is the unique minimizer to a strongly convex and coercive function:

$$x_{n+1} = \arg\min_{x\in X} f(x) + \frac{1}{2}\|x_n - x\|^2 \tag{16}$$

Thus x_{n+1} is the unique element of X such that $x_{n+1} + \gamma\nabla f(x_{n+1}) = x_n$, which allows us to write $x_{n+1} = (Id + \gamma\nabla f)^{-1}(x_n)$, becoming $x_{n+1} = Tx_n$ by noting $T = (Id + \gamma\nabla f)^{-1}$.

We will later see that this second definition of x_{n+1}, that does not take into account the gradient at all, can be generalized to any convex, proper, lower semi-continuous function without making any hypothesis of differentiability.

We will show that the implicit gradient descent method constructs a minimizing sequence, regardless of the choice of x_0 and $\gamma > 0$. This method has the advantage over the explicit method of not having a condition on γ, but the disadvantage of requiring the resolution of a possibly difficult problem at each iteration.

Theorem 3 *Let f be a convex, differentiable function, admitting a minimizer. Then for all $x_0 \in X$, the sequence defined for all $n \in \mathbb{N}$ by $x_{n+1} = (Id + \gamma\nabla f)^{-1}x_n$ converges to a minimizer of f.*

The proof of this theorem is similar to the one of the theorem ensuring the convergence of the explicit gradient descent. It is based on the non-expansivity of the operator $T = (Id + \gamma\nabla f)^{-1}$. We will here give a proof in the case where f is differentiable but we will later see that this hypothesis is unnecessary.

Let $(x, y) \in E^2$. We define

$$p = \arg\min_{z\in E} f(z) + \frac{1}{2}\|x - z\|^2 \text{ and } q = \arg\min_{z\in E} f(z) + \frac{1}{2}\|y - z\|^2. \tag{17}$$

We will show that

$$\|p - q\|^2 + \|x - p - y + q\|^2 \leqslant \|x - y\|^2 \tag{18}$$

which implies that T is firmly non expansive and thus 1-Lipschitz.
Firstly we note that

$$\|x - y\|^2 - \|p - q\|^2 - \|x - p - y + q\|^2 = 2\langle p - q, x - p - (y - q)\rangle \tag{19}$$

In order to prove the non-expansivity of T, it is thus necessary and sufficient to prove the non-negativity of the scalar product below.

By definition of p and q, both minimizers to convex, differentiable functionals:

$$\nabla f(p) = p - x \text{ and } \nabla f(q) = q - y \tag{20}$$

and by using the fact that f is bounded by its affine approximations, we have for all $z \in E$

$$f(z) \geqslant f(p) + \langle z - p, p - x\rangle \text{ and } f(z) \geqslant f(q) + \langle z - q, q - y\rangle. \tag{21}$$

By applying the first inequality with $z = q$ and the second with $z = p$ and by adding the two inequalities we deduce that $\langle p - q, x - p - (y - q)\rangle \geqslant 0$ which ensures that T is firmly non expansive and thus 1-Lipschitz.

As in the case of the explicit gradient descent, the fixed points of T are characterized by $\nabla f(x) = 0$ and are the minimizers of f. Thus, by hypothesis, there exists at least one such fixed point. It remains to show that $\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0$ which is immediate as by definition of x_{n+1}:

$$f(x_{n+1}) + \frac{1}{2}\|x_{n+1} - x_n\|^2 \leqslant f(x_n) + \frac{1}{2}\|x_n - x_n\|^2 = f(x_n). \tag{22}$$

We thus deduce that $(f(x_n))_{n\in\mathbb{N}}$ is a decreasing sequence, and as it is bounded from below we deduce that $\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0$, which concludes the proof of the theorem.

1.3.3 Gradient descent with optimal step

The gradient descent with optimal step consists as its name implies of adapting the descent step at each iteration in order for the value of functional to decrease the most. We thus define a sequence for $x_0 \in E$ and for all $n \in \mathbb{N}$:

$$x_{n+1} = x_n - \gamma_n \nabla f(x_n) \text{ with } \gamma_n = \arg\min_{\gamma>0} f(x_n - \gamma\nabla f(x_n)) \tag{23}$$

This method converges to the unique minimizer of f if f is $\alpha-$strongly convex.

This method has the advantage of converging faster than the gradient descent method with fixed step in terms of number of iterations, but each iteration requires the resolution of an optimization problem that can be solved in an exact or approximated way.

1.4 Projected gradient, Uzawa Algorithm

If we want to solve the following constrained optimization problem:

$$\min_{x\in K} f(x) \tag{24}$$

where K is a closed, convex set and f is differentiable with an $L-$Lipschitz gradient, it is possible to adapt the gradient descent method with fixed step.

Theorem 4 *Let f be a convex, differentiable function such that ∇f is $L-$Lipschitz. Let K be a closed, convex set and $\gamma < \frac{2}{L}$. Then for allt $x_0 \in E$, the sequence defined by*

$$x_{n+1} = P_K(x_n - \gamma\nabla f(x_n)) \tag{25}$$

converges to a minimizer of f.

We do not prove this result for two reasons. The first is that it is a simple adaptation of the convergence proof for the gradient descent method with fixed step, and the second is that this method of projected gradient is a particular case of a more general algorithm that will be treated later, called Forward-Backward and for which the convergence will be proved.

The main difficulty in the implementation of this algorithm is the projection onto K which is not always explicit. On the other hand, if K is of the form

$$K = \prod_{i=1}^{M} [a_i, b_i] \tag{26}$$

the projection is simply just a truncation. This remark is the basis of the Uzawa algorithm for solving certain augmented Lagrangian problems.
As we have seen previously, if f is a convex function and if the functions $(F_i)_{i \leqslant M}$ are convex, solving the following problem

$$\min_{F_i(x) \leqslant 0, \forall i \in [\![1,M]\!]} f(x) \tag{27}$$

is equivalent to determine a saddle point of the following Lagrangian:

$$\sup_{q \in (\mathbb{R}^+)^M} \inf_{v \in E} \mathcal{L}(v,q) = f(v) + \langle q, F(v) \rangle \tag{28}$$

We set $\mathcal{G}$ to be the function defined by

$$\mathcal{G}(q) = \inf_{v \in E} \mathcal{L}(v,q) = f(v) + \langle q, F(v) \rangle. \tag{29}$$

The pair (u,p) is a saddle point of $\mathcal{L}$ if and only if p is solution to

$$\inf_{q \in (\mathbb{R}^+)^M} -\mathcal{G}(q). \tag{30}$$

This is a minimization problem of a convex function under convex constraints, on which the projection is easy to compute. We can thus use a projected gradient algorithm to solve it if we are able to compute the gradient of $\mathcal{G}$. Under some rather general hypotheses, $\mathcal{G}$ is differentiable and we can express its gradient easily.
We will not go into the theoretical details here, but we can give some formal arguments.

For all $q \in (\mathbb{R}^+)^M$, the functional $u \mapsto f(u) + qF(u)$ is convex and admits a unique minimizer u_q (which is, after all, quite a strong hypothesis), thus $\mathcal{G}(q) = f(u_q) + \langle q, F(u_q) \rangle$. If we derive this relation formally we obtain

$$\mathcal{G}'(q) = F(u_q) + \langle \nabla f(u_q) + \langle q, F'(u_q) \rangle, u'_q \rangle = F(u_q) \tag{31}$$

We thus have a simple expression of the gradient of $\mathcal{G}$ in the point q, simply just the value of F in the point u_q.

We deduce from this the Uzawa algorithm that can be used to solve (27), which is simply just a projected gradient. We consider $p_0 \in (\mathbb{R}^+)^M$, and for all $n \in \mathbb{N}$ we define:

$$\left\{ \begin{array}{ccc} u_n & = & \min_{u \in E} f(u) + \langle p_n, F(u) \rangle \\ p_{n+1} & = & P_{(\mathbb{R}^+)^M}(p_n + \gamma F(u_n)) \end{array} \right. \tag{32}$$

where $\gamma < \frac{2}{L}$ if ∇f is $L-$Lipschitz.

1.5 Conjugate gradient method

The conjugate gradient method is an iterative method used for solving approximately and iteratively the following inverse problem:

$$Ax = b \tag{33}$$

where A is a symmetric definite positive matrix in $\mathbb{R}^{n \times n}$ and b is a vector in $\mathbb{R}^n$.

The idea of the algorithm is to part from a point $x_0 \in \mathbb{R}^n$ and a residual $r_0 = b - Ax_0$, and to iteratively minimize the following function:

$$\min_{x \in K_n} \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle \tag{34}$$

where $K_n = Vect(r_0, Ar_0, \cdots, A^n r_0)$ is the Krylov subspace of dimension $n+1$ associated to r_0.

In order to perform such a minimization on this family of increasing spaces we construct the sequences (x_n) such that the residual $r_{n+1} = Ax_{n+1}$ is orthogonal to the Krylov space K_n. Properly speaking, the algorithm only applies the operator A once per iteration, and can be written in the following way:

Definition 5 *Let A ba a positive definite matrix and $x_0 \in \mathbb{R}^n$. We define the three sequences (x_n, r_n, p_n) in the following way:*

$$p_0 = r_0 = b - Ax_0, \text{ and } \forall k > 0, \begin{cases} x_{k+1} &= x_k + \alpha_k p_k \\ r_{k+1} &= r_k - \alpha_k A p_k \\ p_{k+1} &= r_{k+1} - \beta_k p_k. \end{cases} \tag{35}$$

with

$$\alpha_k = \frac{\|r_k\|^2}{\langle Ap_k, p_k\rangle} \text{ et } \beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2} \tag{36}$$

In practice this method converges rapidly, i.e. the approximated value x_k is quite precise even for small values of k and even more so if x_0 is well chosen. It can be shown that the convergence rate towards the exact value x^* is linear and depends on the conditioning of A.

In the following we will see that the inversion of systems containing symmetric positive definite matrices is an operation that is performed in a large number of optimization algorithms.

1.6 Newton's method

Newton's method was originally intended for approching a regular zero of a function F of class C^2 defined from $\mathbb{R}^n$ to $\mathbb{R}^n$. We say that u is a *regular zero* of F if

$$F(u) = 0 \text{ and } F'(u) \text{ is invertible.} \tag{37}$$

This method is based on a Taylor development in the neighbourhood of a zero of F, and constructs a sequence from an element x_0 in the neighbourhood of u, defined for all $n \in \mathbb{N}$ by

$$x_{n+1} = x_n - (F'(x_n))^{-1} F(x_n). \tag{38}$$

In order for this sequence to be defined we need the matrix $F'(x_n)$ to be invertible which is only guaranteed if the departure point x_0 is close enough to u.

We have the following proposition:

Proposition 3 *Let F be a function of class C^2 from $\mathbb{R}^n$ to $\mathbb{R}^n$ and u a regular zero of F. There exists $\varepsilon > 0$ such that if $\|x_0 - x\| \leqslant \varepsilon$, Newton's method defined by* (38) *converges to u and there exists a constant C such that*

$$\|x_{n+1} - u\| \leqslant C\|x_n - u\|^2. \tag{39}$$

The proof of this theorem will be proposed as an exercise.

2 Lesson 3 : Non-differentiable convex functions

In this part we consider non smooth convex optimization problems, i.e. the function F that we want to minimize is non-differentiable. A particular case that we will consider is the case of the sum of two convex functions where at least one is non-differentiable:

$$\min_{x\in E} F(x) = \min_{x\in E} f(x) + g(x) \tag{40}$$

where g is non-differentiable. The problem of minimizing a differentiable function f under convex constraints can be rewritten on this form, by using an indicator function of the convex set K:

$$\min_{x\in K} f(x) \Longleftrightarrow \min_{x\in E} f(x) + i_K(x). \tag{41}$$

However, this is not the only situation we will consider. For example, in image treatment, it is quite common to minimize functionals of the form of a sum of a sum of two terms; a *data attachment* part and a *regularization* part, the latter chosen to enhance certain features of the image to be reconstructed. It is often non-differentiable, as in the case of an ℓ_1 regularization, that promotes sparsity, or the total variation regularization, that seeks to respect the contours of the images. The LASSO in statistics is an estimator that is also computed by minimizing a functional of the above mentioned form. We will see that certain problems involving deblurring or inpainting may be treated by solving problems of the form:

$$\min_{x\in E} \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1 \tag{42}$$

possibly through a change of variables.

A classical method of denoising is to minimize the total variation of an image:

$$\min_{x\in E} \frac{1}{2}\|y - x\|_2^2 + \lambda\|\nabla x\|_1 \tag{43}$$

an expression that takes into account a non-differentiable term. We will see that this problem can be rewritten as:

$$\min_{p} \frac{1}{2}\|\mathrm{div}\, p - z\|_2^2 + i_{\mathcal{B}_{\ell_\infty}(\lambda)}(z) \tag{44}$$

which we are able to solve using many different algorithms.
Some denoising problems involving non Gaussian noise, speckle or Laplacians take into account non regular data attachment terms:

$$\min_{x \in E} \frac{1}{2}\|y-x\|_1 + \lambda\|D^*x\|_1 \tag{45}$$

where D^* is a well chosen transform.

Sometimes it can be useful to write F as a sum of many functions, even though they are all non-differentiable, as certain algorithms, called *splitting* algorithms, are capable of minimizing a sum of functions by alternating elementary operations using each of the functions separately. We will see that this is the case in the *Douglas Rachford* algorithm and in the *ADMM* (Alternating Direction Method of Multipliers).

2.1 Definitions

Definition 6 *Let f be a function from E to $\mathbb{R} \cup \{+\infty\}$. The subdifferential of f is the multivalued operator that to each $x \in E$ associates the set of slopes*

$$\partial f(x) = \{u \in E \text{ such that } \forall y \in E,\ \langle y-x, u\rangle + f(x) \leqslant f(y)\}$$

Remarks:

- The subdifferential of a function in a point can be either empty or a convex set, possibly reduced to a singleton.
- If f is convex and differentiable in each point $x \in E$, the subdifferential is reduced to the singleton $\{\nabla f(x)\}$, the gradient of f in x.
- The subdifferential of the function $x \mapsto -x^2$ is empty in each point of E, even though f is differentiable in each point. In this example f is not convex.
- It can be shown that in each point x of the relative interior of the domain of a **convex** function f, the subdifferential $\partial f(x)$ is non-empty.
- On the boundary of the domain things are a somewhat more complicated. The function f defined from $\mathbb{R}$ to $\mathbb{R} \cap +\infty$ by

$$f(x) = \begin{cases} -\sqrt{x} & \text{si } x \geqslant 0 \\ +\infty & \text{si } x < 0 \end{cases}$$

is convex, takes a finite value in 0 but does not admit a subdifferential in zero.

The importance of the subdifferential is made clear through Fermat's rule, as defined by

Theorem 5 *Let f be a proper function defined from E to $\mathbb{R} \cup +\infty$. Then*

$$\operatorname{argmin} f = zer\ \partial f = \{x \in E \text{ such that } 0 \in \partial f(x)\}.$$

Proof:

$$x \in \operatorname{argmin} f \Leftrightarrow \forall y \in E,\ \langle y - x, 0\rangle + f(x) \leqslant f(y) \Leftrightarrow 0 \in \partial f(x).$$

In order to minimize a functional we thus seek to find a zero of the subdifferential. If by lower semi-continuity and coercivity, we know that a functional admits at least one minimum, then we know that in this minimum, the subdifferential is non-empty and contains 0.

This rule is an extension of the well known fact that in the global minimum of a differentiable function, the gradient is zero.

We end this paragraph by some rules on summing subdifferentials:

Lemma 6 *Let $(f_i)_{i \leqslant p}$ be a family of proper, lower semi-continuous functions and let $x \in E$. Then*

$$\partial(\sum_{i=1}^{p} f_i)(x) \supset \sum_{i=1}^{p} \partial f_i(x)$$

Additionally, if

$$\bigcap_{1 \leqslant i \leqslant p} \operatorname{inter}(\operatorname{dom}(f_i)) \neq \emptyset$$

then we have equality between the subdifferential of the sum and the sum of subdifferentials.

See Rockafellar [**?**].

The hypothesis assuring equality is not very constraining. In the majority of interesting practical cases, it is satisified.

Remark:

If J is the sum of a convex differentiable function f and a convex function g, then under the hypotheses of Lemma 6,

$$\partial J = \nabla f + \partial g.$$

We may now define the proximity (or proximal) operator of a convex, proper, lower semi-continuous function:

Definition 7 *Let g be a convex, proper, lower semi-continuous function from $\mathbb{R}^n$ to $\mathbb{R} \cup +\infty$. The proximity operator of g, denoted* prox_g *and also called "*prox *of g" is the operator defined from $\mathbb{R}^n$ to $\mathbb{R}^n$ by*

$$\operatorname{prox}_g(x) = \underset{z\in\mathbb{R}^n}{\arg\min}\ g(z) + \frac{1}{2}\|x-z\|_2^2 \tag{46}$$

Remark: The function $z \mapsto g(z) + \frac{1}{2}\|x-z\|_2^2$ that defines the prox is convex, proper, lower semi-continuous and coercive, and thus admits at least one minimizer. The uniqueness of the minimizer comes from the strict convexity of the function $y \mapsto \frac{1}{2}\|x-z\|^2$.

The proximity operator may be seen as a generalization of the concept of projecting onto a convex set in the sense that if f is the indicator of a closed convex set C then $\operatorname{prox}_f(x)$ is the projection of x onto C.

In the litterature there are many functions for which an analytical expression of proximity operator is given. It can also be noted that if the function f is separable, as in $x \mapsto \|x\|_p^p$ for example, the proximity operator is computed component-wise, which simplifies the computations. From now on we will say that **a function f is simple when its proximity opearator may be computed easily.**

We may note that if f is not convex, the proximity operator may not be defined or it may be multivalued. The convexity of f is a sufficient condition for the existence and uniqueness of the proximity operator, but it is not a necessary condition.

The proximity operator has many remarkable properties that make it a tool to be privileged when minimizing convex functions.

Proposition 4 *Let f be a convex function on E. Then for all $(x,p) \in E^2$, we have*

$$p = \operatorname{prox}_f(x) \Leftrightarrow \forall y \in E,\ \langle y-p, x-p\rangle + f(p) \leqslant f(y) \tag{47}$$

Thus $p = \operatorname{prox}_f(x)$ is the unique vector $p \in E$ such that $x - p \in \partial f(p)$.

This last characterization is quite important. Thus the decomposition of $x = p + z$ with $z \in \partial f(p)$ as the sum of one element p from E and one element from the subdifferential of f in the point p is unique. This remarkable property is the source of the notation

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

to which we can give a sense event though the map ∂f is multivalued.
Proof:

- Suppose that $p = \operatorname{prox}_f(x)$. Let $\alpha \in]0,1[$ and $p_\alpha = \alpha y + (1-\alpha)p$. Then, by using the definition of $\operatorname{prox}_f(x)$ and then the convexity of f, we obtain

  $$\begin{aligned} f(p) &\leqslant f(p_\alpha) + \frac{1}{2}\|x - p_\alpha\|^2 - \frac{1}{2}\|x-p\|^2 \\ &\leqslant \alpha f(y) + (1-\alpha)f(p) - \alpha\langle y-p, x-p\rangle + \frac{\alpha^2}{2}\|y-p\|^2 \end{aligned}$$

  and thus
  $$\langle y-p, x-p\rangle + f(p) \leqslant f(y) + \frac{\alpha}{2}\|y-p\|^2.$$

  By letting α go to zero we obtain the direct implication.

- Conversely, suppose that $\langle y-p, x-p\rangle + f(p) \leqslant f(y)$. Then

  $$f(p)+\frac{1}{2}\|x-p\|^2 \leqslant f(y)+\frac{1}{2}\|x-p\|^2+\langle x-p, p-y\rangle+\frac{1}{2}\|p-y\|^2 = f(y)+\frac{1}{2}\|x-y\|^2$$

  from which we deduce that $p = \operatorname{prox}_f(x)$ by using the definition of $\operatorname{prox}_f(x)$.

An important property of the fixed points of the proximity operator is the following:

Proposition 5 *Let f be a proper, convex function defined on E. Then*

$$\operatorname{Fix} \operatorname{prox}_f = \operatorname{argmin} f.$$

Proof:

$$x = \operatorname{prox}_f(x) \Leftrightarrow \forall y \in E,\ \langle y-x, x-x\rangle + f(x) \leqslant f(y) \Leftrightarrow x \in \operatorname{argmin} f$$

This characterization allows for using iterative algorithms with the proximity operator of f in order to minimize a functional f.

An other essential property of the proximity operator is that it is firmly non-expansive:

Proposition 6 *If f is a proper, convex function on E then the maps* prox_f *and* $Id - \operatorname{prox}_f$ *are firmly non-expansive; that is, for all* $(x, y) \in E^2$,

$$\| \operatorname{prox}_f(x) - \operatorname{prox}_f(y)\|^2 + \|(x - \operatorname{prox}_f(x)) - (y - \operatorname{prox}_f(y))\|^2 \leqslant \|x - y\|^2$$

In particular, these two operators are non-expansive (1-Lipschitz).

Proof:
Let $(x, y) \in E^2$. By noting $p = \operatorname{prox}_f(x)$ and $q = \operatorname{prox}_f(y)$, we have, by definition of p and q:

$$\langle q - p, x - p\rangle + f(p) \leqslant f(q) \text{ and } \langle p - q, y - q\rangle + f(q) \leqslant f(p).$$

Adding these two inequalities gives

$$0 \leqslant \langle p - q, (x - p) - (y - q)\rangle$$

and we conclude by noting that

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

In what follows we will need to define the *reflected proximity operator*, as presented below.

Definition 8 *If f is a proper, convex function, we define the reflected prox or* rprox *by*

$$\operatorname{rprox}_f = 2 \operatorname{prox}_f - Id$$

In order to show that rprox is 1-Lipschitz we will need the following lemma:

Lemma 7 *Let T be an operator defined from E to E. The following two statements are equivalent:*

- *T is firmly non-expansive*
- *$R = 2T - Id$ is non-expansive (1-Lipschitz)*

Proof:
We will show that these two properties are equivalent to the non-negativity

of a scalar product. Let $(x, y) \in E^2$. By setting $u = Tx - Ty$ and $v = Tx - x - (Ty - y)$, we have

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

thus T is firmly non-expansive if and only if $\langle u, v\rangle \geqslant 0$. Additionally,

$$\|Rx - Ry\|^2 = \|u + v\|^2 = \|u\|^2 + \|v\|^2 + 2\langle u, v\rangle = \|x - y\|^2 + 4\langle u, v\rangle$$

thus stating that R is non-expansive is equivalent to $\langle u, v\rangle \geqslant 0$ which concludes the proof.

2.2 Proximal point algorithm

The proximal point algorithm, as its name suggests, consists of applying recursively the proximity operator of the function γg in order to minimize g on E. We thus choose $\gamma > 0$ and $x_0 \in E$, and consider the sequence $(x_n)_{n\in\mathbb{N}}$ defined by

$$x_{n+1} = Tx_n = \operatorname{prox}_{\gamma g}(x_n)$$

A sequence on this form converges to a minimizer of g, as can be proven by applying Lemma 4. The fixed points of the proximity operator are minimizers of g, and this operator is 1-Lipschitz. By construction, we have, just like for the implicit gradient descent:

$$g(x_{n+1}) + \frac{1}{2\gamma}\|x_n - x_{n+1}\|^2 \leqslant g(x_n)$$

which implies that the sequence defined by $\|x_n - x_{n+1}\|^2$ goes to zero.

In practice, this algorithm is rarely used, in the sense that computing the proximity operator may be just as difficult as minimizing g directly. We will rather use the proximity operator as an ingredient in more sophisticated algorithms.

2.3 Lesson 4 : Forward-Backward and FISTA

The Forward-Backward algorithm (FB) is an algorithm that may be used to solve the following optimization problem:

$$\min_{x\in E} F(x) = \min_{x\in E} f(x) + g(x)$$

where f is a convex, differentiable function with an $L-$Lipschitz gradient and where g is a function of which we are able to compute the proximity operator. This algorithm consists of alternating an explicit gradient descent step on f an a proximity operator on g. The algortihm called FISTA is an acceleration of the FB algorithm proposed by Beck and Teboulle based on the ideas of Nesterov of using an inertial term. FB was used in the early 2000s to solve problems where g was an ℓ_1-norm, for which case the proximity operator of g is a soft thresholding, and the name of this algorithm was consecutively *ISTA* for Iterative Shrinkage and Thresholding Algorithm. Beck and Teboulle gave the name FISTA (Fast Iterativ Shrinkage and Thresholding Algorithm) to their algorithm, fast compared to ISTA, but technically it is an acceleration of FB that is only used in the case where g is an ℓ_1-norm.

2.3.1 Forward-Backward

The Forward-Backward algorithm, in its most simple version, is defined by an initial point $x_0 \in E$ and a parameter $\gamma > 0$ as follows:

$$\forall n \in \mathbb{N} \quad x_{n+1} = Tx_n = \operatorname{prox}_{\gamma g}(x_n - \gamma \nabla f(x_n)) \tag{48}$$

We will see later that there also exists a version using Krasnoselsky-Mann iterations.

The algorithm results from the following proposition:

Proposition 7 *Let $F = f + g$ be a functional defined from E to $\mathbb{R} \cup +\infty$ having the form of a sum of two convex, proper, lower semi-continuous functions satisfying the hypotheses of Lemma 6 such that f is differentiable and let $\gamma > 0$. Then*

$$\text{zeros}(\partial F) = \text{Fix}(\operatorname{prox}_{\gamma g}(Id - \gamma \nabla f)) \tag{49}$$

Proof:

$$\begin{aligned}0 \in \partial F(x) &\Leftrightarrow 0 \in \partial \gamma F(x)\\ &\Leftrightarrow 0 \in \nabla \gamma f(x) + \partial \gamma g(x)\\ &\Leftrightarrow -\gamma \nabla f(x) \in \partial \gamma g(x)\\ &\Leftrightarrow x - \gamma \nabla f(x) \in (Id + \partial \gamma g)(x)\\ &\Leftrightarrow x = \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x))\end{aligned}$$

which concludes the proof.

Of course, having an operator T for which the fixed points are the minimizers of F does not guarantee that the sequence $(x_n)_{n\in\mathbb{N}}$ converges to one of these minimizers. But under the conditions of the following theorem, this is indeed the case:

Theorem 6 *Let $F = f + g$ be a sum of two convex, coercive, lower semi-continuous functions that are bounded from below. We suppose that f is differentiable with an $L-$Lipschitz gradient. Let $\gamma < \dfrac{2}{L}$ and $x_0 \in E$, and let $(x_n)_{n\in\mathbb{N}}$ be the sequence defined for all $n \in \mathbb{N}$ by*

$$x_{n+1} = \operatorname{prox}_{\gamma g}(Id - \gamma \nabla f)(x_n). \tag{50}$$

Then the sequence $(x_n)_{\in\mathbb{N}}$ converges to a minimizer of F.

Proof:
To proove this convergence, we will procede in the same way as for the gradient descent method. That is; we will show that the operator T is $1-$Lipschitz and that the sequence with the general term $\|x_n - x_{n+1}\|$ goes to zero. We will conclude by using Lemma 4.

The operator T is $1-$Lipschitz as a composition of a proximity operator that is 1-Lipschitz and an operator of the form $Id - \gamma \nabla f$ that is also 1-Lipschitz under the condition that $\gamma \leqslant \frac{2}{L}$, as it has been shown in Lemma 5.

To prove the second point, we will introduce the concept of *surrogate functions*. We can show that at each iteration in the algorithm, at each computation of a new term in the sequence, the value of the functional F decreases. Additionally, this decrease ensures that the sequence with the general term $\|x_{n+1} - x_n\|^2$ is summable and thus goes to zero when n goes to $+\infty$.

Lemma 8 *The sequence* $(x_n)_{n\in\mathbb{N}}$ *defined by* $x_{n+1} = \mathrm{prox}_{\gamma g}(x_n - \gamma\nabla f(x_n))$ *satisfies the following relation:*

$$F(x_{n+1}) + \left(\frac{1}{\gamma} - \frac{L}{2}\right)\|x_{n+1} - x_n\|^2 \leqslant F(x_n). \tag{51}$$

Proof:
By definition of x_{n+1}, we have

$$x_{n+1} = \operatorname*{argmin}_{x\in E} \gamma g(x) + \frac{1}{2}\|x - x_n + \gamma\nabla f(x_n)\|^2$$

We may note that x_{n+1} is also the minimizer of the following functional:

$$x_{n+1} = \operatorname*{argmin}_{x\in E} g(x) + f(x_n) + \langle\nabla f(x_n), x - x_n\rangle + \frac{1}{2\gamma}\|x - x_n\|^2 \tag{52}$$

According to the inequality (4), we have, for all $x \in E$,

$$f(x) \leqslant f(x_n) + \langle\nabla f(x_n), x - x_n\rangle + \frac{L}{2}\|x - x_n\|^2 \tag{53}$$

The right hand side is a function that bounds F from above, and it is equal to F in the point x_n. We call this function a *surrogate function.*

The considered function being $\frac{1}{\gamma}$-strongly convex, we may deduce from (52) that

$$g(x_{n+1})+f(x_n)+\langle\nabla f(x_n), x_{n+1}-x_n\rangle+\frac{1}{2\gamma}\|x_{n+1}-x_n\|^2 \leqslant F(x_n)-\frac{1}{2\gamma}\|x_{n+1}-x_n\|^2$$

By applying (53) to x_{n+1} and adding the previous inequality we obtain

$$F(x_{n+1}) + \frac{1}{\gamma}\|x_{n+1} - x_n\|^2 \leqslant F(x_n) + \frac{L}{2}\|x_{n+1} - x_n\|^2,$$

which concludes the proof of the lemma and the theorem.

In the following section we will see that it is possible to control the speed of the functional's decrease towards its minimum.

2.3.2 Fast Iterative Shrinkage thresholding Algorithm (FISTA)

The idea behind FISTA is to apply the operator $T = \text{prox}_{\gamma g} \circ (Id - \gamma \nabla f)$ by using an inertial term, in order to improve the convergence rates.

FISTA is defined by a sequence $(t_n)_{n\in\mathbb{N}}$ of real numbers greater than 1 and by a point $x_0 \in E$. Let $(t_n)_{n\geqslant 1}$ be a sequence of non-negative real numbers and $x_0 \in E$. For $y_0 = u_0 = x_0$ and $n \geqslant 1$, we define the sequences $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ by

$$x_n = T(y_{n-1}) \tag{54}$$

$$u_n = x_{n-1} + t_n(x_n - x_{n-1}) \tag{55}$$

$$y_n = \left(1 - \frac{1}{t_{n+1}}\right) x_n + \frac{1}{t_{n+1}} u_n. \tag{56}$$

The point y_n may also be defined using the points x_n et x_{n-1} by

$$y_n = x_n + \alpha_n(x_n - x_{n-1}) \text{ with } \alpha_n := \frac{t_n - 1}{t_{n+1}} \tag{57}$$

For an adequate choice of $(t_n)_{n\geqslant 1}$, the sequence $(F(x_n))_{n\in\mathbb{N}}$ converges towards $F(x^*)$, i.e the sequence $(w_n)_{n\in\mathbb{N}}$, defined by

$$w_n := F(x_n) - F(x^*) \geqslant 0 \tag{58}$$

goes to zero when n goes to infinity.

Many proofs use the local variation of the sequence $(x_n)_{n\in\mathbb{N}}$. We willwill denote this by δ_n:

$$\delta_n := \frac{1}{2}\|x_n - x_{n-1}\|_2^2. \tag{59}$$

The sequence $(v_n)_{n\in\mathbb{N}}$ that describes the distance between u_n and a fixed minimizer x^* of F will also be useful:

$$v_n := \frac{1}{2}\|u_n - x^*\|_2^2. \tag{60}$$

In order to complete these notations we will also define the sequence $(\rho_n)_{n\geqslant 2}$ associated to $(t_n)_{n\geqslant 1}$, of which the non-negativity ensures the (quadratic) convergence rate of FISTA:

$$\rho_n := t_{n-1}^2 - t_n^2 + t_n. \tag{61}$$

2.4 Lesson 6 : The Douglas Rachford algorithm.

In this section we wish to solve the problem

$$\min_{x\in E} F(x) = \min_{x\in E} f(x) + g(x)$$

where the functions f and g are convex, proper, lower semi-continuous and where F is bounded from below. We do not make any assumptions on the differentiablity of f. We also assume that we are able to compute the proximity operators of both f and g. A first thing to notice is that under these hypotheses, there exists a minimizer to F.

As for the Forward-Backward algorithm, we will identify an operator that is 1-Lipschitz, for which the fixed points are associated to the minimizers of F. In the case of FB, these fixed point *are* the minimizers of F, while for the Douglas-Rachford algorithm, the minimizers are the images of these fixed points through an operator. Unlike FB, we cannot simply iterate on the operator T itself to approximate a fixed point. However, there exists a way to build a sequence that converges to a fixed point of an operator T that possesses at least one fixed point, and that is 1-Lipschitz by using a mixed procedure. It is called the Krasnoselsky-Mann algorithm, that we will now present. It may also be noted that this algorithm may also be used for FB.

Theorem 7 *The Krasnoselsky-Mann algorithm.*
Let D be a non-empty, closed, convex subset of E and let T, $D \to D$ be a 1-Lipschitz operator such that the set of fixed points of T is non-empty. Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence of real numbers in $[0,1]$ such that $\sum_{n\in\mathbb{N}} \lambda_n(1-\lambda_n) = +\infty$, and let $x_0 \in D$. We define

$$\forall n \in \mathbb{N}, \quad x_{n+1} = x_n + \lambda_n(Tx_n - x_n).$$

Then the sequence $(x_n)_{n\in\mathbb{N}}$ converges to a fixed point of T.

Remark
If E is not finite the convergence towards a fixed point is only weak, while the convergence of $Tx_n - x_n$ towards 0 is strong.

Proof:
As D is convex, the sequence $(x_n)_{n\in\mathbb{N}}$ is well defined. Let y be a fixed point of T and $n \in \mathbb{N}$. Then

$$\begin{aligned}\|x_{n+1} - y\|^2 &= \|(1-\lambda_n)(x_n - y) + \lambda_n(Tx_n - y)\|^2 \\ &= (1-\lambda_n)\|x_n - y\|^2 + \lambda_n\|Tx_n - Ty\|^2 - \lambda_n(1-\lambda_n)\|Tx_n - x_n\|^2 \\ &\leqslant \|x_n - y\|^2 - \lambda_n(1-\lambda_n)\|Tx_n - x_n\|^2.\end{aligned}$$

This implies that

$$\sum_{n\in\mathbb{N}} \lambda_n(1-\lambda_n)\|Tx_n - x_n\|^2 \leqslant \|x_0 - y\|^2. \tag{62}$$

Additionally, we have

$$\begin{aligned}\|Tx_{n+1} - x_{n+1}\| &= \|Tx_{n+1} - Tx_n + (1-\lambda_n)(Tx_n - x_n)\| \\ &\leqslant \|x_{n+1} - x_n\| + (1-\lambda_n)\|Tx_n - x_n\| \\ &\leqslant \|Tx_n - x_n\|\end{aligned}$$

As $\sum_{n\in\mathbb{N}} \lambda_n(1-\lambda_n) = +\infty$ and the sequence $(\|Tx_n - x_n\|)_{n\in\mathbb{N}}$ is decreasing, we deduce from (62) that $(Tx_n - x_n)_{n\in\mathbb{N}}$ goes to zero.

The sequence $\|x_n - y\|$ being decreasing, we deduce that all the elements of the sequence $(x_n)_{n\in N}$ belong to the closed ball centered in y with radius $\|y - x_0\|$, which is a compact of E since E is finite.

We can thus extract a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ that converges to a point $x \in E$.

As $Tx_n - x_n$ goes to zero, the sequence $(Tx_{n_k} - x_{n_k})_{k\in\mathbb{N}}$ goes to zero, which implies that the sequence $(Tx_{n_k})_{k\in\mathbb{N}}$ also goes to x. As this sequence also converges to Tx, we deduce that $x = Tx$ is a fixed point of T.

As x is a fixed point of T, the sequence $(\|x_n - x\|)_{n\in\mathbb{N}}$ is decreasing. Additionally, this sequence admits a subsequence that goes to zero. We thus deduce that the sequence $(x_n)_{n\in\mathbb{N}}$ goes to x a fixed point of T, which concludes the proof of the theorem.

The Douglas-Rachford algorithm is based on the following proposition:

Proposition 8 *Let $F = f + g$ be a functional defined from E to $\mathbb{R} \cup +\infty$, where the two functions f and g are both convex, proper, lower semi-continuous and satisfying the hypotheses of Lemma 6, and let γ be a strictly positive real number. Then*

$$\textit{zeros}\ (\partial F) = \operatorname{prox}_{\gamma g}\left(Fix(\operatorname{rprox}_{\gamma f}\operatorname{rprox}_{\gamma g})\right). \tag{63}$$

Proof:
Under the hypotheses of Lemma 6, we have

$$\begin{aligned}0 \in \partial F(x) &\Leftrightarrow 0 \in \partial\gamma F(x) \\ &\Leftrightarrow 0 \in \partial\gamma f(x) + \partial\gamma g(x) \\ &\Leftrightarrow \exists z \in E \text{ such that } -z \in \partial\gamma f(x) \text{ and } z \in \partial\gamma g(x) \\ &\Leftrightarrow \exists y \in E \text{ such that } x - y \in \partial\gamma f(x) \text{ and } y - x \in \partial\gamma g(x)\end{aligned}$$

We may rewrite $x-y \in \partial\gamma f(x)$ on the form $2x-y \in (Id+\partial\gamma f)(x)$. Similarly, the relation $y-x \in \partial\gamma g(x)$ can be rewritten $y \in (Id+\gamma\partial g)(x)$ and thus $x = \operatorname{prox}_{\gamma g}(y)$, which gives

$$0 \in \partial F(x) \Leftrightarrow \exists y \in E \text{ such that } 2x-y \in (Id+\partial\gamma f)(x) \text{ and } x = \operatorname{prox}_{\gamma g}(y).$$

By using the definition of the operator Rprox we obtain:

$$\begin{aligned} 0 \in \partial F(x) &\Leftrightarrow \exists y \in E \text{ such that } x = \operatorname{prox}_{\gamma f}(\operatorname{rprox}_{\gamma g} y) \text{ and } x = \operatorname{prox}_{\gamma g}(y) \\ &\Leftrightarrow \exists y \in E \text{ such that } y = 2x - \operatorname{rprox}_{\gamma g} y = \operatorname{rprox}_{\gamma f}(\operatorname{rprox}_{\gamma g} y) \text{ and } x = \operatorname{prox}_{\gamma g}(y) \\ &\Leftrightarrow \exists y \in E \text{ such that } y \in Fix\left(\operatorname{rprox}_{\gamma f} \operatorname{rprox}_{\gamma g}\right) \text{ and } x = \operatorname{prox}_{\gamma g}(y) \end{aligned}$$

which concludes the proof of the proposition.

From this proposition we may extract a minimization algorithm for F. If we manage to find a fixed point of $T = \operatorname{rprox}_{\gamma f} \operatorname{rprox}_{\gamma g}$, it is sufficient to apply $\operatorname{prox}_{\gamma f}$ in order to obtain a minimizer of F, as this operator is 1-Lipschitz. We will see that it is possible to use an algorithm called the Kransoselsky-Mann algorithm to find a fixed point.

Theorem 8 *Let f and g be two convex, proper, lower semi-continuous functions, bounded from below. Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of elements in $[0,2]$ such that $\sum_{n\in\mathbb{N}} \mu_n(2-\mu_n) = +\infty$. Let $\gamma > 0$ and $x_0 \in E$. Let $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ and $(z_n)_{n\in\mathbb{N}}$ be the sequences defined by*

$$\forall n \in \mathbb{N} \left\{ \begin{array}{rc} y_n = & \operatorname{prox}_{\gamma g}(x_n), \\ z_n = & \operatorname{prox}_{\gamma f}(2y_n - x_n), \\ x_{n+1} = & x_n + \mu_n(z_n - y_n) \end{array} \right. \tag{64}$$

Then there exists $x \in E$ minimizing $(f+g)$ such that the sequence $(x_n)_{n\in\mathbb{N}}$ converges to x.

Proof:

We set $T = \operatorname{rprox}_{\gamma f} \operatorname{rprox}_{\gamma g}$. We know that this operator is 1-Lipschitz as a composition of two operators that are 1-Lipschitz. We also know that the set of fixed points is equal to the set of minimizers of $f+g$, which is non-empty according to the hypotheses on f and g. We note that $x_{n+1} = x_n + \frac{\mu_n}{2}(Tx_n - x_n)$. We conclude by applying Krasnoselsky-Mann's theorem.

The Douglas-Rachford algorithm can be expressed in many forms in the litterature. One should be vigilant in order to recognize it. For example, it is quite frequent for the parameters μ_n to be fixed to 1; the algorithm is then expressed in the following way:

$$\forall n \in \mathbb{N} \left\{ \begin{array}{rc} y_n = & \mathrm{prox}_{\gamma g}(x_n), \\ z_n = & \mathrm{prox}_{\gamma f}(2y_n - x_n), \\ x_{n+1} = & x_n + z_n - y_n \end{array} \right. \tag{65}$$

By omitting the variable z_n we obtain the following description of the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$:

$$\forall n \geqslant 1 \left\{ \begin{array}{rl} x_n = & x_{n-1} + \mathrm{prox}_{\gamma f}(2y_{n-1} - x_{n-1}) - y_{n-1}, \\ y_n = & \mathrm{prox}_{\gamma g}(x_n) \end{array} \right. \tag{66}$$

We may also introduce the auxiliary variable $u_n = \mathrm{prox}_{\gamma f}(2y_{n-1} - x_{n-1})$, change the order in which we update the variables and rewrite the algorithm:

$$\forall n \geqslant 1 \left\{ \begin{array}{rl} u_n = & \mathrm{prox}_{\gamma f}(2y_{n-1} - x_{n-1}) \\ y_n = & \mathrm{prox}_{\gamma g}(x_{n-1} + u_n - y_{n-1}), \\ x_n = & x_{n-1} + u_n - y_{n-1} \end{array} \right. \tag{67}$$

There are many possibilities; to conclude we may cite a change of variables that sometimes appears in the litterature: $w_n = y_n - x_n$. We then have

$$\forall n \geqslant 1 \left\{ \begin{array}{rl} u_n = & \mathrm{prox}_{\gamma f}(y_{n-1} + w_{n-1}) \\ y_n = & \mathrm{prox}_{\gamma g}(u_n - w_{n-1}), \\ w_n = & w_{n-1} + y_n - u_n \end{array} \right. \tag{68}$$

All of these formulations are equivalent. In the following, we will see how it is possible to adapt this algorithm to the case where the minimization problem is of the form:

$$\min_{x\in E} f(x) + g(Ax)$$

where A is a linear operator.

2.5 Alternating Direction Method of Multipliers (ADMM)

The ADMM is an algorithm to solve optimization problems of the form:

$$\min_{(x_1,x_2)\in E^2,\, A_1x_1+A_2x_2=b} f_1(x_1) + f_2(x_2) \tag{69}$$

where A_1 and A_2 are two linear operators taking values in $\mathbb{R}^m$, b is a vector in $\mathbb{R}^m$ and f_1 and f_2 are two convex, proper, lower semi-continuous functions. It is quite a general formulation, containing the case where $x_2 = x_1$ or even $x_2 = Ax_1$. Here, we present the algorithm without giving a proof of convergence in the general case. We will see that in the particular case where $A_1 = Id$, $A_2 = -Id$ and $b = 0$, i.e. $x_1 = x_2$, this algorithm simply becomes the Douglas-Rachford algorithm.

The ADMM algorithm

Let $(x_1^0, x_2^0) \in E$, $\gamma > 0$ and $z^0 \in \mathbb{R}^m$. We define the sequences $(x_1^n)_{n\in\mathbb{N}}$, $(x_2^n)_{n\in\mathbb{N}}$ and $(z^n)_{n\in\mathbb{N}}$ in the following way, $\forall n \geqslant 1$:

$$\left\{\begin{array}{rl} x_1^n = & \arg\min_x f_1(x) + \langle z^{n-1}, A_1 x\rangle + \frac{1}{2\gamma}\|A_1 x + A_2 x_2^{n-1} - b\|^2 \\ x_2^n = & \arg\min_y f_2(y) + \langle z^{n-1}, A_2 y\rangle + \frac{1}{2\gamma}\|A_1 x_1^n + A_2 y - b\|^2 \\ z^n & = z^{n-1} + \gamma(A_1 x_1^n + A_2 x_2^n - b) \end{array}\right. \tag{70}$$

This algorithm may be used to solve problem (69) in the following sense:

Theorem 9 *Let f_1 and f_2 be two proper, convex, coercive, lower semi-continuous functions. Then*

1. *The sequence $(f_1(x_1^n) + f_2(x_2^n))_{n\in\mathbb{N}}$ converges to the minimum value of $f_1 + f_2$.*
2. *The sequences $(x_1^n)_{n\in\mathbb{N}}$ and $(x_2^n)_{n\in\mathbb{N}}$ converge.*
3. *The sequence $(Ax_1^n + Ax_2^n - b)_{n\in\mathbb{N}}$ goes to zero.*

We do not prove this result here. The simplest proof consists of showing that solving (69) is equivalent to solving a dual problem of (69) using a Douglas-Rachford algorithm.

We may make some remarks on this algorithm:

1. Its name comes from the fact that it can be seen as a variant of an algorithm known as the Augmented Lagrangian Method. If we replace the iterative updates of x_1 and x_2 by a joint update step:

   $$(x_1^n, x_2^n) = \arg\min_{(x,y)\in E} f_1(x)+f_2(y)+\langle z^{n-1}, A_1x+A_2y\rangle+\frac{1}{2\gamma}\|A_1x+A_2y-b\|^2$$

   we obtain the Augmented Lagrangian method that consists of penalizing the constraints with a Lagrange multiplier z and a quadratic term.

One of the problems with this method is that such a conjoint minimization is often difficult to perform. The ADMM separates this problem by optimizing with respect to the first variable x_1 first, and then the second variable x_2. The variable z can be interpreted as a Lagrange multiplier that is updated at each iteration. If $A_1 = Id$, $A_2 = -Id$, $b = 0$ and $\gamma = 1$, the algorithm can be expressed using proximity operators. It becomes:

$$\left\{ \begin{array}{ll} x_1^n = & \operatorname{prox}_{f_1}(x_2^{n-1} - z^{n-1}) \\ x_2^n = & \operatorname{prox}_{f_2}(x_1^n + z^{n-1}) \\ z^n & = z^{n-1} + (x_1^n - x_2^n) \end{array} \right. \tag{71}$$

which is one of the forms of Douglas-Rachford, see (68).

2. In the general case, none of the two first updates are easy to perform. If $A_1 = Id$, the update of x_1 can be expressed by a prox of f, and similarly if $A_2 = Id$ for the second update. If one of the two updates is not the identity operator, we must often use an iterative algorithm to perform the minimization. We then have internal loops. Each iteration has to be considered case-wise, depending on the functions f_1 and f_2 and the operators A_1 and A_2. The performance of the ADMM will then depend on the choices that are made to perform the two minimizations. We may still note that the two minimization problems can be treated by FB or FISTA if we have access to the proximity operators of f_1 and f_2, as the quadratic term is differentiable and its gradient is explicit.

3. It is possible to show that applying an ADMM is equivalent to applying a Douglas-Rachford on a different problem, called the dual problem of (69), and this independently of the choice of the operators A_1 and A_2, but we will not consider this in this course.

2.6 Chambolle-Pock Algorithm

This section is based on a part of the work by Chambolle and Pock published in [?]. Unlike the Krasnoselsky-Mann algorithm, the modification proposed by Chambolle and Pock lies in a non-convex update of x_n. Before presenting the algorithm and its convergence proof, we define the *partial primal-dual* *gap*:

$$\mathcal{G}_{B_1\times B_2}(x,y) = \max_{y'\in B_2} \langle y', Kx\rangle - f^*(y') + g(x) - \min_{x'\in B_1} \langle y, Kx'\rangle + g(x') - f^*(y). \tag{72}$$

If there exists a solution $(\hat{x},\hat{y}) \in B_1 \times B_2$ to equation (**??**), then by using the function h defined in (**??**), we have

$$\begin{aligned}\mathcal{G}_{B_1\times B_2}(x,y) &\geqslant \langle \hat{y}, Kx\rangle - f^*(\hat{y}) + g(x) - (\langle y, K\hat{x}\rangle + g(\hat{x}) + f^*(y))\\ &\geqslant h(x,\hat{y}) - h(\hat{x},y)\\ &\geqslant h(x,\hat{y}) - h(\hat{x},\hat{y}) + h(\hat{x},\hat{y}) - h(\hat{x},y)\\ &\geqslant 0.\end{aligned}$$

and this gap is only zero if (x,y) is solution to (**??**).

Theorem 10 *Let f and g be two convex, proper, lower semi-continuous functions, with f defined from F to $[-\infty,+\infty]$, g defined from E to $[-\infty,+\infty]$ and K a linear operator from F to E. We suppose that the problem (**??**) admits a solution $(\hat{x},\hat{y})$. We note $L = \|K\|$, and choose σ and τ such that $\tau\sigma L^2 < 1$. We choose $(x_0, y_0) \in E \times F$ and set $\bar{x}_0 = x_0$. We define the sequences $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$, and $(\bar{x}_n)_{n\in\mathbb{N}}$ by*

$$\left\{\begin{array}{lcl} y_{n+1} &=& \operatorname{prox}_{\sigma f^*}(y_n + \sigma K\bar{x}_n)\\ x_{n+1} &=& \operatorname{prox}_{\tau g}(x_n - \tau K^* y_{n+1})\\ \bar{x}_{n+1} &=& 2x_{n+1} - x_n \end{array}\right. \tag{73}$$

Then

1. *For all $n \in \mathbb{N}$,*

$$\frac{\|y_n-\hat{y}\|^2}{2\sigma} + \frac{\|x_n-\hat{x}\|^2}{2\tau} \leqslant (1-\tau\sigma L^2)^{-1}\left(\frac{\|y_0-\hat{y}\|^2}{2\sigma} + \frac{\|x_0-\hat{x}\|^2}{2\tau}\right) \tag{74}$$

2. *If we set $x^N = (\sum_{n=1}^N x_n)/N$ and $y^N = (\sum_{n=1}^N y_n)/N$, then for each bounded set $(B_1 \times B_2) \subset E \times F$, the partial primal-dual gap satisfies*

$$\mathcal{G}_{B_1\times B_2}(x^N, y^N) \leqslant \frac{D(B_1,B_2)}{N}$$

where

$$D(B_1,B_2) = \sup_{(x,y)\in B_1\times B_2} \frac{\|x-x_0\|^2}{2\tau} + \frac{\|y-y_0\|^2}{2\sigma}.$$

3. The sequence $(x_n, y_n)_{n\in\mathbb{N}}$ converges to a solution (x^*, y^*) of (**??**).

Proof:
In order to prove this theorem we will rewrite the iterations (73) on the following form:

$$\left\{\begin{array}{lcl} y_{n+1} & = & \operatorname{prox}_{\sigma f^*}(y_n + \sigma K\bar{x}) \\ x_{n+1} & = & \operatorname{prox}_{\tau g}(x_n - \tau K^*\bar{y}) \end{array}\right. \tag{75}$$

The fundamental properties of the proximity operators ensure that

$$\begin{aligned} K\bar{x} + \frac{y_n - y_{n+1}}{\sigma} &\in \partial f^*(y_{n+1}) \\ -K^*\bar{y} + \frac{x_n - x_{n+1}}{\sigma} &\in \partial g(x_{n+1}) \end{aligned}$$

and thus, for all $(x, y) \in E \times F$,

$$\begin{aligned} f^*(y) &\geqslant f^*(y_{n+1}) + \langle \frac{y_n - y_{n+1}}{\sigma}, y - y_{n+1}\rangle + \langle K\bar{x}, y - y_{n+1}\rangle \\ g(x) &\geqslant g(x_{n+1}) + \langle \frac{x_n - y_{n+1}}{\tau}, x - x_{n+1}\rangle + \langle K(x - x_{n+1}), \bar{y}\rangle \end{aligned}$$

By adding these two inequalities we obtain:

$$\begin{aligned} & h(x_{n+1}, y) - h(x, y_{n+1}) \\ & \quad + \frac{\|y - y_{n+1}\|^2}{2\sigma} + \frac{\|x - x_{n+1}\|^2}{2\tau} + \frac{\|y_n - y_{n+1}\|^2}{2\sigma} + \frac{\|x_n - x_{n+1}\|^2}{2\tau} \\ & \qquad + \langle K(x_{n+1} - \bar{x}), y_{n+1} - y\rangle - \langle K(x_{n+1} - x), y_{n+1} - \bar{y}\rangle \\ & \qquad\qquad \leqslant \frac{\|y - y_n\|^2}{2\sigma} + \frac{\|x - x_n\|^2}{2\tau} \end{aligned} \tag{76}$$

Chambolle and Pock propose to choose $\bar{y} = y_{n+1}$ and $\bar{x} = 2x_n - x_{n-1}$. Thus the second last line of the previous inequality can be written

$$\begin{aligned} & \langle K(x_{n+1} - \bar{x}), y_{n+1} - y\rangle - \langle K(x_{n+1} - x), y_{n+1} - \bar{y}\rangle \\ & \quad = \langle K((x_{n+1} - x_n) - (x_n - x_{n-1})), y_{n+1} - y\rangle \\ & \quad = \langle K(x_{n+1} - x_n), y_{n+1} - y\rangle - \langle K(x_n - x_{n-1}), y_n - y\rangle - \langle K(x_n - x_{n-1}), y_{n+1} - y_n\rangle \\ & \quad \geqslant \langle K(x_{n+1} - x_n), y_{n+1} - y\rangle - \langle K(x_n - x_{n-1}), y_n - y\rangle - L\|x_n - x_{n-1}\|\|y_{n+1} - y_n\|. \end{aligned} \tag{77}$$

By using the fact that for all $\alpha > 0$ we have $2ab \leqslant \alpha a^2 + \frac{b^2}{\alpha}$, we obtain

$$L\|x_n - x_{n-1}\|\|y_{n+1} - y_n\| \leqslant \frac{L\alpha\tau}{2\tau}\|x_n - x_{n-1}\|^2 + \frac{L\sigma}{2\alpha\sigma}\|y_{n+1} - y_n\|^2 \tag{78}$$

By reconsidering (76) with the previous majorations with $\alpha = \sqrt{\frac{\sigma}{\tau}}$, we deduce that for all $(x, y) \in E \times F$,

$$\begin{aligned} & h(x_{n+1}, y) - h(x, y_{n+1}) \\ & \qquad + \frac{\|y - y_{n+1}\|^2}{2\sigma} + \frac{\|x - x_{n+1}\|^2}{2\tau} + (1 - \sqrt{\sigma\tau}L)\frac{\|y_n - y_{n+1}\|^2}{2\sigma} \\ & \qquad + \frac{\|x_n - x_{n+1}\|^2}{2\tau} - \sqrt{\sigma\tau}L\frac{\|x_{n-1} - x_n\|^2}{2\tau} \\ & \qquad + \langle K(x_{n+1} - x_n), y_{n+1} - y\rangle - \langle K(x_n - x_{n-1}), y_n - y\rangle \\ & \qquad \leqslant \frac{\|y - y_n\|^2}{2\sigma} + \frac{\|x - x_n\|^2}{2\tau} \end{aligned} \tag{79}$$

By adding the previous inequalities from $n = 0$ to $N-1$, and setting $x_{-1} = x_0$ by convention, we obtain:

$$\begin{aligned} & \sum_{n=1}^{N} h(x_n, y) - h(x, y_n) \\ & \qquad + \frac{\|y - y_N\|^2}{2\sigma} + \frac{\|x - x_N\|^2}{2\tau} + (1 - \sqrt{\sigma\tau}L)\sum_{n=1}^{N} \frac{\|y_n - y_{n-1}\|^2}{2\sigma} \\ & \qquad + (1 - \sqrt{\sigma\tau})\sum_{n=1}^{N-1} \frac{\|x_n - x_{n-1}\|^2}{2\tau} + \frac{\|x_N - x_{N-1}\|^2}{2\tau} \\ & \qquad \leqslant \frac{\|y - y_n\|^2}{2\sigma} + \frac{\|x - x_n\|^2}{2\tau} + \langle K(x_N - x_{N-1}), y_N - y\rangle \end{aligned}$$

and like previously by using the majoration

$$\langle K(x_N - x_{N-1}), y_N - y\rangle \leqslant \|x_N - x_{N-1}\|^2/(2\tau) + (\sigma\tau L^2)\|y - y_N\|^2/(2\sigma)$$

we obtain the following inequality:

$$\begin{aligned} & \sum_{n=1}^{N} h(x_n, y) - h(x, y_n) \\ & + (1 - \sigma\tau L^2)\frac{\|y - y_N\|^2}{2\sigma} + \frac{\|x - x_N\|^2}{2\tau} + (1 - \sqrt{\sigma\tau}L)\sum_{n=1}^{N} \frac{\|y_n - y_{n-1}\|^2}{2\sigma} \\ & \qquad + (1 - \sqrt{\sigma\tau}L)\sum_{n=1}^{N} \frac{\|x_n - x_{n-1}\|^2}{2\tau} \leqslant \frac{\|y - y_0\|^2}{2\sigma} + \frac{\|x - x_0\|^2}{2\tau} \end{aligned} \tag{80}$$

We apply this inequality to a saddle point $(x, y) = (\hat{x}, \hat{y})$ of (**??**). The first line of (80) is the sum of *partial primal-dual gaps* with (x, y) solutions to (**??**). As we have noted previously, this implies that all the terms of the sum from this first line are positive. We deduce the first result of the theorem from the fact that $\tau\sigma L^2 < 1$.
We now consider an arbitrary pair $(x, y) \in B_1 \times B_2$. As

$$\sum_{n=1}^{N} h(x_n, y) - h(x, y_n) \leqslant \frac{\|y - y_0\|^2}{2\sigma} + \frac{\|x - x_0\|^2}{2\tau} \tag{81}$$

we can now use the fact that f^* and g are convex to deduce that

$$h(x^N, y) - h(x, y^N) \leqslant \frac{1}{N}\left(\frac{\|y - y_0\|^2}{2\sigma} + \frac{\|x - x_0\|^2}{2\tau}\right)$$

By taking the supremum on $(x, y) \in B_1 \times B_2$ of the two sides of the inequality, we obtain the second point of the Lemma.
Using the first point of the theorem we know that the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are bounded. As the dimension of E is finite, from each of these sequences we can extract subsequences that converge to some limits x^* et y^*. By taking the limit when N goes to $+\infty$ in the previous inequality we obtain

$$\langle Kx^*, y\rangle - f^*(y) + g(x^*) - (\langle Kx, y^*\rangle - f^*(y^*) + g(x)) \leqslant 0$$

for all $(x, y) \in B_1 \times B_2$. By taking the supremum over the pairs $(x, y) \in B_1 \times B_2$, we deduce that the *partial primal-dual gap* in the point (x^*, y^*) is negative and thus zero, an thus (x^*, y^*) is a solution to (**??**).
As the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are bounded we can extract subsequences $(x_{n_k})_{nk\in\mathbb{N}}$ and $(y_{n_k})_{k\in\mathbb{N}}$ that converge to some points x^∞ et y^∞. As the series having the general term $\|x_n - x_{n-1}\|^2$ and $\|y_n - y_{n-1}\|^2$ are convergent we deduce that the sequences $(\|x_n - x_{n-1}\|^2)_{n\in\mathbb{N}}$ and $(\|y_n - y_{n-1}\|^2)_{n\in\mathbb{N}}$ go to zero when n goes to $+\infty$. We deduce that the sequences $(x_{n_k+1})_{nk\in\mathbb{N}}$ et $(y_{n_k+1})_{k\in\mathbb{N}}$ also converge to x^∞ et y^∞ and thus x^∞ et y^∞ are fixed points of (73).
We make the remark that fixed points of (73) satisfy the relations (**??**), and are thus saddle points, solutions to (**??**).

By adding the inequalities (79) for $n = n_k$ to N we obtain

$$\begin{aligned}
&\frac{\|y^* - y_N\|^2}{2\sigma} + \frac{\|x^* - x_N\|^2}{2\tau} + (1 - \sqrt{\sigma\tau}L) \sum_{n=n_k+1}^{N} \frac{\|y_n - y_{n-1}\|^2}{2\sigma} \\
&\quad - \frac{\|x_{n_k} - x_{n_k-1}\|^2}{2\tau} + (1 - \sqrt{\sigma\tau}L) \sum_{n=n_k}^{N-1} \frac{\|x_n - x_{n-1}\|^2}{2\tau} + \frac{\|x_N - x_{N-1}\|}{2\tau} \\
&\qquad + \langle K(x_N - x_{N-1}), y_N - y^* \rangle - \langle K(x_{n_k} - x_{n_k-1}), y_{n_k} - y^* \rangle \\
&\qquad\qquad \leqslant \frac{\|y^* - y_{n_k}\|^2}{2\sigma} + \frac{\|x^* - x_{n_k}\|^2}{2\tau}.
\end{aligned} \tag{82}$$

By using the fact that $\lim_{n \to +\infty} \|x_n - x_{n-1}\| = \lim_{n \to +\infty} \|y_n - y_{n-1}\| = 0$, we deduce that the sequence $(x_N, y_N)_{N \in \mathbb{N}}$ goes to (x^*, y^*), which concludes the proof of the theorem.

Remark: We can observe that the second point of the theorem gives a convergence rate towards zero of the gap of the Césaro-means of the sequences $(x_n)_{n \in \mathbb{N}}$ and $(y_n)_{n \in \mathbb{N}}$, but not of the sequences themselves.