Continuous optimization, an introduction

Antonin Chambolle (and many others, ...)

December 13, 2017

Contents

1 Introduction

These lectures notes have been prepared by A. Chambolle for the M2 course "Continuous optimization" given between Oct. and Dec. 2016 in Paris 6, in the Master "modélisation mathématique" of Ecole Polytechnique, Université Pierre-et-Marie Curie (Paris 6) and Ecole National des Ponts et Chaussées. Much of the material is taken from [9], or/and inspired by famous textbooks [29, 25, 32, 13, 2].

They will be updated for the 2017 course. They gather various matherial mostly on first order optimisation and iterative algorithms for generally convex problems, including operator splitting, acceleration, etc.

2 (First order) Descent methods, rates

Most of what we describe in this section is in finite dimension, although extension to Hilbert spaces is in general easy. We will discuss rates of convergence, in particular, which we try to make independent on the dimension. The complexity of the iterations, on the other hand, are usually very dimension-dependent, and this is the reason for which high order descent methods are not practical for modern high dimensional problems (imaging, data analysis...). If time permits, the last lectures of this course will address coordinate descent or stochastic techniques which allow to reduce the complexity of one iteration.

2.1 Gradient descent

The main source for this section is the excellent book of Polyak [29]. Consider the problem of minimising

$$\min_{x\in X} f(x)$$

with X a finite dimensional vector step (or Hilbert) and f a real valued, C^1 function (or at least differentiable).

We introduce and analyse the gradient descent algorithm with fixed step:

$$x^{k+1} = x^k - \tau\nabla f(x^k) =: T_\tau(x^k).$$

Remark: $-\nabla f(x^k)$ is a descent direction. Near x^k, indeed,

$$f(x) = f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + o(x - x^k)$$

so that

$$f(x^{k+1}) = f(x^k) - \tau|\nabla f(x^k)|^2 + o(\tau).$$

One can use various strategies to choose τ:

- optimal step: $\min_\tau f(x^k - \tau\nabla f(x^k))$ (with a "line search", such as for instance for the "conjugate gradient method");
- Armijo-type rule: find $i \geq 0$ such that $f(x^k - \tau\rho^i\nabla f(x^k)) \leq f(x^k) - c\tau\rho^i|\nabla f(x^k)|^2$, $\rho < 1, c < 1$ fixed;
- "Frank-Wolfe"-type method[1]: $\min_{x\in X^k} f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle$ where X^k is appropriately defined. For instance $X^k := \{x : \|x - x^k\| \leq \varepsilon\}$, in which case $\tau = \varepsilon/|\nabla f(x^k)|$;
- Gradient with fixed step: $\min_x f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\tau}\|x - x^k\|^2$

Convergence analysis: if τ is too large with respect to the Lipschitz gradient of ∇f, or ∇f is not Lipschitz, easy to build infinitely oscillating examples (ex: $f(x) = \|x\|$).

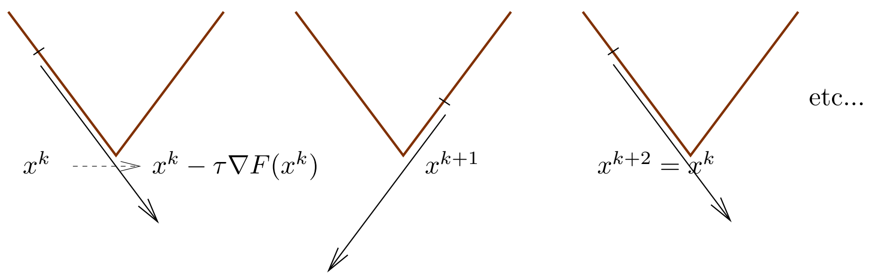


Figure 1: The gradient descent may never converge if the step is too large or the function not smooth enough

If f is C^1, ∇f is L-Lipschitz, $0 < \tau < 2/L$, $\inf f > -\infty$ then the method converges (in $\mathbb{R}^N$) in the following sense: $\nabla f(x^k) \to 0$.

Proof:

$$\begin{aligned} f(x^{k+1}) = f(x^k) - \int_0^\tau &\left\langle \nabla f(x^k + s\nabla f(x^k)), \nabla f(x^k) \right\rangle \\ &= f(x^k) - \tau\|\nabla f(x^k)\|^2 + \int_0^\tau \left\langle \nabla f(x^k) - \nabla f(x^k + s\nabla f(x^k)), \nabla f(x^k) \right\rangle \\ &\leq f(x^k) - \tau(1 - \tfrac{L\tau}{2})\|\nabla f(x^k)\|^2. \quad (1) \end{aligned}$$

[1] or "conditional gradient".

Observe that we just need that $D^2 f$ is bounded from above (if f C^2).

Then letting $\theta = \tau(1 - \tau L/2) > 0$, one finds that

$$f(x^n) + \theta \sum_{k=0}^{n-1} \|\nabla f(x^k)\|^2 \le f(x^0).$$

This shows the claim. IF in addition f is "infinite at infinity" (coercive) then x^k has subsequences which converge, therefore to a stationary point.

Remark 2.1. If $\tau = 0$, the iteration does nothing (and hence converges to the initial point...). If $\tau = 2/L$, the iteration might oscillate forever, as shows the example of the function $f(x) := L|x|^2/2$.

Remark 2.2. Taking x^* a minimizer, $\tau = 1/L$, we deduce that

$$\frac{1}{2L}\|\nabla f(x^k)\|^2 \le f(x^k) - f(x^{k+1}) \le f(x^k) - f(x^*).$$

The convex case If f is convex we have the following additional property:

Theorem 2.3 (Baillon-Haddad[2]). *If f is convex and ∇f is L-Lipschitz, then for all x, y,*

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

(∇f is said to be "$(1/L)$-co-coercive".)

We will see later a general proof of this result based on convex analysis. In finite dimension, if f is C^2, then the proof is easy: one has $0 \le D^2 f \le LI$ (because f is convex, and because ∇f is L-Lipschitz). Then

$$\begin{aligned}\langle \nabla f(x) - \nabla f(y), x - y\rangle = \int_0^1 \left\langle D^2 f(y + s(x-y))(x-y), x - y\right\rangle ds \\ =: \langle A(x-y), x-y\rangle .\end{aligned}$$

with $A = \int_0^1 D^2 f(y + s(x-y))ds$ symmetric, $0 \le A \le LI$. Hence:

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\|^2 = \|A(x-y)\|^2 &= \left\langle AA^{1/2}(x-y), A^{1/2}(x-y)\right\rangle \\ &\le L\left\langle A^{1/2}(x-y), A^{1/2}(x-y)\right\rangle \le L\langle A(x-y), x-y\rangle ,\end{aligned}$$

hence the result. If f is not C^2, one could smooth f by convolution with a smooth, compactly supported kernel, derive the result and then pass to the limit.

Lemma 2.4. *If f is convex with L-Lipschitz gradient, then the mapping $T_\tau = I - \tau\nabla f$ is a weak contraction when $0 \le \tau \le 2/L$ (that is, T_τ is 1-Lipschitz).*

Proof:

$$\begin{aligned}\|T_\tau x - T_\tau y\|^2 &= \|x-y\|^2 - 2\tau\langle x - y, \nabla f(x) - \nabla f(y)\rangle + \tau^2\|\nabla f(x) - \nabla f(y)\|^2 \\ &\le \|x-y\|^2 - 2\tau\Big(1 - \frac{\tau L}{2}\Big)\|\nabla f(x) - \nabla f(y)\|^2.\end{aligned}$$

Remark: It is "averaged" for $0 < \tau < 2/L \to$ convergence (proved later on).

[2]This is part of a more general study of operators which satisfy a similar condition, see [1].

Convergence rate in the convex case. In general, one needs further information on the second order behaviour of f to derive convergence rates for algorithms. In general f is assumed to be globally convex. Then, using that, for x^* a minimizer,

$$f(x^*) \geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle$$

(we will see this is a general property of convex functions), we find

$$\frac{f(x^k) - f(x^*)}{\|x^* - x^k\|} \leq \|\nabla f(x^k)\| \tag{2}$$

And using Lemma 2.4 which implies that $\|x^k - x^*\| \leq \|x^0 - x^*\|$, it follows $(f(x^k) - f(x^*))/\|x^0 - x^k\| \leq \|\nabla f(x^k)\|$. Hence from (1) we derive, letting $\Delta_k = f(x^k) - f(x^*)$, $\theta = \tau(1 - \tau L/2)$, that

$$\Delta_{k+1} \leq \Delta_k - \frac{\theta}{\|x^0 - x^*\|^2} \Delta_k^2$$

We can show the following:

Lemma 2.5. *Let $(a_k)_k$ be a sequence of nonnegative numbers satisfying for $k \geq 0$:*

$$a_{k+1} \leq a_k - c^{-1} a_k^2$$

Then, for all $k \geq 0$,

$$a_k \leq \frac{c}{k+1}$$

Proof: First observe that if we replace a_k with a_k/c, the property becomes $a_{k+1} \leq a_k - a_k^2$: hence it is enough to prove it for $c = 1$. Then, as $a_k(1 - a_k) \geq a_{k+1} \geq 0$, one has $0 \leq a_k \leq 1$ for all $k \geq 0$. We show the inequality by induction: for $k = 0$, $a_0 \leq 1$. If $k \geq 1$ and the property is true for $k - 1$, we write that

$$(k+1)a_k \leq (k+1)a_{k-1} - (k+1)a_{k-1}^2 \leq k a_{k-1} + a_{k-1}(1 - (k+1)a_{k-1}) \leq k a_{k-1} + a_{k-1}(1 - (k+1)a_k)$$

since $a_k \leq a_{k-1}$. We see therefore that $a_k \leq 1/(k+1)$, since if not, $1 - (k+1)a_k < 0$ implies that $(k+1)a_k < k a_{k-1} \leq 1$ by the induction hypothesis, and we find a contradiction. We deduce:

Theorem 2.6. *The gradient descent with fixed step satisfies*

$$\Delta_k \leq \frac{\|x^0 - x^*\|^2}{\theta(k+1)}$$

Observe that this rate is not very good and also a bit pessimistic (it should improve if $x^k \to x^*$ because (2) improves). On the other hand, it does not prove, a priori, anything on the sequence (x^k) itself. Observe also, to conclude that $\theta = \tau(1 - \tau L/2) = (2/L)[(\tau L/2)(1 - \tau L/2)]$ is maximal for $\tau L/2 = 1/2$, that is, $\tau = 1/L$. In that case, $\theta = 1/(2L)$ and the rate is bounded by

$$\Delta_k \leq 2L \frac{\|x^0 - x^*\|^2}{k+1}.$$

Strongly convex case. A function f is γ-strongly convex if and only if $f(x) - \gamma\|x\|^2/2$ is convex: if f is C^2, it is equivalent to $D^2 f \geq \gamma I$. We will discuss more precisely this definition in Section 4.1.

In this, case if x^* is the minimizer (which in this case always exists)

$$x^{k+1} - x^* = x^k - x^* - \tau(\nabla f(x^k) - \nabla f(x^*)) = \int_0^1 (I - \tau D^2 f(x^* + s(x^k - x^*))(x^k - x^*)$$

hence

$$\|x^{k+1} - x^*\| \leq \max\{1 - \tau\gamma, \tau L - 1\}\|x^k - x^*\|.$$

If f is not C^2 one can still show this by smoothing. The best constant is for $\tau = 2/(L+\gamma)$ and gives, for $q = (L-\gamma)/(L+\gamma) \in [0,1]$

$$\|x^k - x^*\| \leq q^k \|x^0 - x^*\|.$$

One can easily deduce the following:

Theorem 2.7. *Let f be C^2, x^* be a strict local minimum of f where $D^2 f$ is definite negative. Then if x^0 is close enough to x^*, the gradient descent method with optimal step (obtained with a line search) will converge linearly. (Or with fixed step small enough.)*

2.2 What can we achieve?

This paragraph contains a very elementary introduction to lower bounds and complexity. We follow the description in [9], were we essentially give elementary variants of deeper results found in [22, 25].)

Idea: consider a "hard problem", for instance, for $x \in \mathbb{R}^n$, $L > 0$, $\gamma \geq 0$, $1 \leq p \leq n$, functions of the form:

$$f(x) = \frac{L-\gamma}{8}\left((x_1 - 1)^2 + \sum_{i=2}^{p}(x_i - x_{i-1})^2\right) + \frac{\gamma}{2}\|x\|^2, \tag{3}$$

which is tackled by a "first order method", which is such that the iterates x^k are restricted to the subspace spanned by the gradients of already computed iterates, i.e. for $k \geq 0$

$$x^k \in x^0 + \left\{\nabla f(x^0), \nabla f(x^1), \ldots, \nabla f(x^{k-1})\right\}, \tag{4}$$

where x^0 is an arbitrary starting point.

Starting from an initial point $x^0 = 0$, any first order method of the considered class can transmit the information of the data term only at the speed of one index per iteration. This makes such problems very hard to solve by any first order methods in the considered class of algorithms. Indeed if one starts from $x^0 = 0$ in the above problem (whose solution is given by $x^*_l = 1$, $k = 1, \ldots, p$, and 0 for $l > p$), then at the first iteration, only the first component x^1_1 will be updated (since $\partial_i f(x^0) = 0$ for $i \geq 2$), and by induction one can check that at iteration k, $x^k_l = 0$ for $l \geq k+1$.

The solution satisfies $\nabla f = 0$, therefore is characterized by

$$x_i = \frac{L-\gamma}{L+\gamma}\frac{x_{i+1} + x_{i-1}}{2}$$

with $x_0 = 1$. The best possible point at iteration k satisfies this equation for $i \leq k$, and $x_{k+1} = 0$. In case $\gamma = 0$ we find that this point x is affine: $x_i = (1 - i/(k+1))^+$, and $x_i - x_{i-1} = -1/(k+1)$ for $i \leq k+1$. Hence

$$f(x) = \frac{L}{8}\sum_{i=1}^{k+1}\frac{1}{(k+1)^2} = \frac{L}{8}\frac{1}{k+1}$$

is the best possible value which can be reached at step k.

If one looks for a bound independent on the dimension with (here for homogeneity reasons) $f(x^k) \sim L\|x^0 - x^*\|a_k$ (for a sequence (a_k)), using here that $x_i^* = 1$ for $i \le p$ and 0 for $i > p$, $x^0 = 0$, and $f(x^*) = 0$, one obtains

$$f(x^k) - f(x^*) \ge \frac{L}{8p(k+1)}\|x^0 - x^*\|^2$$

($k < p$) (while if $k = p$, $x^k = x^*$). For $k = p - 1$ one finds

$$f(x^k) - f(x^*) \ge \frac{L}{8}\frac{\|x^0 - x^k\|^2}{(k+1)^2}$$

hence no first order method can reach a bound of the considered form which is better than this. (*It does not contradict* a bound of the form $f(x^k) - f(x^*) = o(1/k^2)$, for instance!)

It follows a variant of the results in [25] (where a slightly different function is used), see Theorems 2.1.7 and 2.1.13.

Theorem 2.8. *For any $n \ge 2$, any $x^0 \in \mathbb{R}^n$, $L > 0$, and $k < n$, there exists a convex, one times continuously differentiable function f with L-Lipschitz continuous gradient, such that for any first-order algorithm satisfying* (4)*, it holds that*

$$f(x^k) - f(x^*) \ge \frac{L\|x^0 - x^*\|^2}{8(k+1)^2}, \tag{5}$$

where x^* denotes a minimiser of f.

Observe that the above lower bound is valid only if number of iterates k is less than the problem size. We can not improve this with a quadratic function, as the conjugate gradient method (which is a first-order method) is then known to find the global minimiser here after at most p steps.

But practical problems are often so large that it is not possible to perform as many iterations as the dimension of the problem, and will always fulfill similar assumptions.

If choosing $\gamma > 0$ so that the function (3) becomes γ-strongly convex, a lower bound for first order methods is given Theorem 2.1.13 in [25]. It is hard to derive precisely for p finite, however in $\mathbb{R}^{\mathbb{N}} \simeq \ell^2(\mathbb{N})$, for $p = +\infty$, one finds that the solution is given by $x = q^i$, $q = (\sqrt{Q} - 1)/(\sqrt{Q} + 1)$ where $Q = L/\gamma$ is the condition number of the problem (q satisfies $2 = (L-\gamma)/(L+\gamma)(q + 1/q)$). If $x^0 = 0$,

$$\|x^0 - x^*\|^2 = \sum_{i=1}^{\infty} q^{2i} = \frac{q^2}{1-q^2}, \text{ while}$$

$$\|x^k - x^*\|^2 \ge \sum_{i=k+1}^{\infty} q^{2i} = q^{2k}\|x^0 - x^*\|^2.$$

The strong convexity of f shows that

$$f(x^k) \ge f(x^*) + \frac{\gamma}{2}q^{2k}\|x^0 - x^*\|^2$$

and it follows:

Theorem 2.9. *For any $x^0 \in \mathbb{R}^\infty \simeq \ell_2(\mathbb{N})$ and $\gamma, L > 0$ there exists a γ-strongly convex, one times continuously differentiable function f with L-Lipschitz continuous gradient, such that for any algorithm in the class of first order algorithms defined through* (4) *it holds that for all k,*

$$f(x^k) - f(x^*) \ge \frac{\gamma}{2}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k}\|x^0 - x^*\|^2 \tag{6}$$

where $Q = L/\gamma \ge 1$ is the condition number, and x^* a minimiser of f.

In finite dimension, a similar result will hold for k small enough (with respect to n).

2.3 Second order methods: Newton's method

The idea of Newton's method relies on using second order information to improve the precision of the approximation of the function at step k. (In practice, one solves the equation $\nabla f(x) = 0$ using Newton's method.) We have

$$f(x) = f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \tfrac{1}{2} \left\langle D^2 f(x^k)(x - x^k), x - x^k \right\rangle + o(\|x - x^k\|^2).$$

If we are near a minimizer, we can assume $D^2 f(x^k) > 0$ (hopefully), and hence find x^{k+1} by solving

$$\min_x f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \tfrac{1}{2} \left\langle D^2 f(x^k)(x - x^k), x - x^k \right\rangle$$

Compare with the Gradient descent with step τ in a metric defined by a symmetric positive definite matrix $A > 0$, which would be:

$$\min_x f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \tfrac{1}{2\tau} \left\langle A(x - x^k), x - x^k \right\rangle$$

hence we can see Newton's method as a gradient descent in the metric which best approximates the function. We find that x^{k+1} is given by

$$\nabla f(x^k) + D^2 f(x^k)(x^{k+1} - x^k) = 0 \;\Leftrightarrow\; x^{k+1} = x^k - D^2 f(x^k)^{-1} \nabla f(x^k).$$

We have the following "quadratic" convergence rate.

Theorem 2.10. *Assume f is C^2, $D^2 f$ is M-Lipschitz, and $D^2 f \geq \gamma$ (strong convexity). Let $q = M/(2\gamma^2)\|\nabla f(x^0)\|$ and assume x^0 is close enough to the minimizer x^*, so that $q < 1$. Then $\|x^k - x^*\| \leq (2\gamma/M) q^{2^k}$.*

This is extremely fast (the precision is doubled at each iteration, this is called a quadratic rate), but there are strong conditions, and the algorithm can be hard to implement.

Proof: first see that

$$\nabla f(x+h) = \nabla f(x) + \int_0^1 D^2 f(x + sh) h ds = \nabla f(x) + D^2 f(x) h + \int_0^1 (D^2 f(x+sh) - D^2 f(x)) h ds$$

so that

$$\|\nabla f(x+h) - \nabla f(x) - D^2 f(x) h\| \leq \frac{M}{2} \|h\|^2.$$

Hence

$$\begin{aligned} \|\nabla f(x^{k+1}) - \overbrace{\nabla f(x^k) - D^2 f(x^k)(x^{k+1} - x^k)}^{0} \| &\leq \frac{M}{2} \|x^{k+1} - x^k\|^2 \\ \Rightarrow \|\nabla f(x^{k+1})\| \leq \frac{M}{2} \|D^2 f(x^k)^{-1}\|^2 \|\nabla f(x^k)\|^2 &\leq \frac{M}{2\gamma^2} \|\nabla f(x^k)\|^2 \end{aligned}$$

Hence letting $g_k = \|\nabla f(x^k)\|$, for all k one has

$$\log g_{k+1} \leq 2 \log g_k + \log \frac{M}{2\gamma^2} \;\Rightarrow\; \log g_k \leq 2^k \log g_0 + (2^k - 1) \log \frac{M}{2\gamma^2}$$

so that

$$\|\nabla f(x^k)\| \leq \frac{2\gamma^2}{M} q^{2^k}.$$

As f is strongly convex, $\langle \nabla f(x^k), x^k - x^* \rangle \geq \gamma \|x^k - x^*\|^2$, and we can conclude.

The main issue with this is that it is very important to have $q < 1$, otherwise the method could not work.

There are quite a few very important variants of Newton's method, which are designed so that one does not have to explicitly evaluate $D^2 f(x^k)^{-1}$, usually called "Quasi-Newton" type methods: one replaces $D^2 f(x^k)$ with a metric H_k which is improved at each iteration, hoping that $H_k \to D^2 f(x^*)$ in the limit. Very efficient variants: "BFGS" (Broyden-Fletcher-Goldfarb-Shanno) (4 papers of 1970) and improvements (limited memory "L-BFGS") [8, 21]. This topic is covered extensively in [26, Chap. 6].

2.4 Multistep first order methods

2.4.1 Heavy ball method

This description follows Polyak's book [28] where the method is introduced. The idea is to iterate:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}),$$

$\alpha, \beta \geq 0$. This mimicks the equation $\ddot{x} = -\nabla f(x) - \dot{x}$ of a heavy ball in a potential $f(x)$ with friction. Requires that f is C^2, γ-convex, L-Lipschitz (at least near a solution x^*), that is:

$$\gamma I \leq D^2 f \leq LI.$$

Then (see [29])

Theorem 2.11. *Let x^* be a (local) minimizer of f such that $\gamma I \leq D^2 f(x^*) \leq LI$, and choose α, β with $0 \leq \beta < 1$, $0 < \alpha < 2(1+\beta)/L$. There exists $q < 1$ such that if $q < q' < 1$ and if x^0, x^1 are close enough to x^*, one has*

$$\|x^k - x^*\| \leq c(q') q'^k.$$

Moreover, this is almost optimal in the sense of Theorem 6: if

$$\alpha = \frac{4}{(\sqrt{L} + \sqrt{\gamma})^2}, \beta = \left(\frac{\sqrt{L} - \sqrt{\gamma}}{\sqrt{L} + \sqrt{\gamma}}\right)^2 \quad \textit{then } q = \frac{\sqrt{L} - \sqrt{\gamma}}{\sqrt{L} + \sqrt{\gamma}}.$$

Proof: this is an example of a proof where one analyses the iteration of a linearized system near the optimum. Close enough to x^*, one has

$$x^{k+1} = x^k - \alpha D^2 f(x^*)(x^k - x^*) + o(\|x^k - x^*\|) + \beta(x^k - x^{k-1}),$$

and one can write that $z^k = (x^k - x^*, x^{k-1} - x^*)^T$ satisfies, for $B = D^2 f(x^*)$,

$$z^{k+1} = \begin{pmatrix} (1+\beta)I - \alpha B & -\beta I \\ I & 0 \end{pmatrix} z^k + o(z^k).$$

We study the eigenvalues of the matrix A which appears in this iteration: We have

$$A \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} (1+\beta)I - \alpha B & -\beta I \\ I & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \rho \begin{pmatrix} x \\ y \end{pmatrix}$$

if and only if

$$(1+\beta)x - \alpha Bx - \beta y = \rho x, \quad x = \rho y$$

(and $x, y \neq 0$) hence if $(1+\beta)x - \alpha Bx - \beta/\rho x = \rho x$. We find that

$$Bx = \frac{1}{\alpha}\left(1+\beta-\rho-\frac{\beta}{\rho}\right)x$$

hence $\frac{1}{\alpha}\left(1+\beta-\rho-\frac{\beta}{\rho}\right) = \mu \in [\gamma, L]$ is an eigenvalue of B. We derive the equation

$$\rho^2 - (1+\beta-\alpha\mu)\rho + \beta = 0$$

which gives two eigenvalues with product β and sum $1+\beta-\alpha\mu$. If $\beta \in [0,1)$ and $-(1+\beta) < 1+\beta-\alpha\mu < (1+\beta)$ (extreme cases where $\pm(1,\beta)$ are solutions) then $|\rho| < 1$, that is, if $0 < \alpha < (2+\beta)/\mu$. Since $\mu < L$ one deduces that if $0 \le \beta < 1$, $0 < \alpha < (2+\beta)/L$, the eigenvalues of A are all in $(-1,1)$ (incidentally, it has $2n$ eigenvalues).

We well use here the following fundamental classical lemma [18]:

Lemma 2.12. *Let A be a $N \times N$ matrix and assume that all its eigenvalues (complex or real) have modulus $\le \rho$. Then for any $\rho' > \rho$, there exists a norm $\|\cdot\|_*$ in $\mathbb{C}^N$ such that $\|A\|_* := \sup_{\|\xi\|_* \le 1} \|A\xi\|_* < \rho'$.*

This is an important result of linear algebra. The proof is as follows: up to a change of a basis, A is triangular: there exists P such that

$$P^{-1}AP = T$$

with $T = (t_{i,j})_{i,j}$, $t_{i,i} = \lambda_i$, an eigenvalue, and $t_{i,j} = 0$ if $i > j$. Then, if $D_s = \mathrm{diag}(s, s^2, s^3, \ldots, s^N) = (s^i \delta_{i,j})_{i,j}$, $D_s P^{-1} A P D_s^{-1} = (x^s_{i,j})$ with

$$x^s_{i,j} = \sum_{k,l} s^i \delta_{i,k} t_{k,l} s^{-l} \delta_{l,j} = s^{i-j} t_{i,j}$$

and (since $t_{i,j} = 0$ for $i > j$), $x^s_{i,j} \to \lambda_i \delta_{i,j}$ as $s \to +\infty$. Hence, if s is large enough, denoting $\|\xi\|_1 = \sum_i |\xi_i|$ the 1-norm,

$$\max_{\|\xi\|_1 \le 1} \|D_s P^{-1} A P D_s^{-1} \xi\|_1 \le \max_i (|\lambda_i| + (\rho' - \rho)) \le \rho'$$

if s is large. Hence, if $\|\xi\|_* := \|D_s P^{-1}\xi\|_1$, one has

$$\|A\|_* = \sup_{\|\xi\|_* \le 1} \|A\xi\|_* \le \rho'.$$

It follows, in particular, that if $\rho' < 1$, $\|A^k\|_* \le \|A\|_*^k \le \rho'^k \to 0$ as $k \to \infty$. Applying this to our problem, we see that (choosing $\rho' < 1$)

$$\|z^{k+1}\|_* = \|Az^k + o(z^k)\|_* \le (\rho' + \varepsilon)\|z^k\|_*$$

if $\|z^k\|_*$ is small enough. Starting from z^0 such that this holds for ε with $\rho' + \varepsilon < 1$, we find that it holds for all $k \ge 0$ and that $\|z^{k+1}\|_* \le (\rho' + \varepsilon)^k \|z^0\|_*$, showing the linear convergence.

2.4.2 The conjugate gradient method

(We refer still to Polyak [29].)

The conjugate gradient is "the best" two-steps method, in the sense that it can be defined as follows: given x^k, x^{k-1}, we let $x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k(x^k - x^{k-1})$ where α_k, β_k are minimizing

$$\min_{\alpha,\beta} f(x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})).$$

In particular, we deduce that

$$\left\langle \nabla f(x^{k+1}), \nabla f(x^k)\right\rangle = 0 \quad \text{and} \quad \left\langle \nabla f(x^{k+1}), x^k - x^{k-1}\right\rangle = 0 \tag{7}$$

and it also follows

$$\left\langle \nabla f(x^{k+1}), x^{k+1} - x^k\right\rangle = 0. \tag{8}$$

Notice moreover that

$$\nabla f(x^{k+1}) = \nabla f(x^k) - \alpha_k D^2 f(x^k + s(x^{k+1} - x^k))\nabla f(x^k) \\ + \beta_k D^2 f(x^k + s(x^{k+1} - x^k))(x^k - x^{k-1}) \tag{9}$$

for some $s \in [0,1]$.

However this method is in general "conceptual". Except when f is quadratic: $f(x) = (1/2)\langle Ax, x\rangle - \langle b, x\rangle + c$ (A symmetric). Denoting then the gradients $p^k = Ax^k - b$ and the residuals $r^k = x^k - x^{k-1}$, we find that (*cf* (9))

$$p^{k+1} = p^k - \alpha_k A p^k + \beta_k A r^k$$

and using the orthogonality formulas (7),

$$0 = \|p^k\|^2 - \alpha_k \left\langle Ap^k, p^k\right\rangle + \beta_k \left\langle Ar^k, p^k\right\rangle, \quad 0 = \left\langle p^k, r^k\right\rangle - \alpha_k \left\langle Ap^k, r^k\right\rangle + \beta_k \left\langle Ar^k, r^k\right\rangle$$

we can compute explicitly the values of α_k, β_k (exercise).

Lemma 2.13. *The gradients* (p^i) *are all orthogonal.*

Proof: we start from $x^{k+1} = x^k - \alpha_k p^k + \beta_k(x^k - x^{k-1})$ and deduce (since ∇f is affine)

$$p^{k+1} = p^k - \alpha_k A p^k + \beta_k(p^k - p^{k-1}).$$

Assume that $(p^0, \dots, p^i)$ are orthogonal, and that α_l, $l = 0, \dots, i-1$, do not vanish (or we have found the solution, why?). Then

$$\left\langle Ap^k, p^l\right\rangle = \frac{1}{\alpha_k}\left\langle p^{k+1} - p^k - \beta_k(p^k - p^{k-1}), p^l\right\rangle = 0$$

if $l \le k-2$, $k \le i-1$ or if $i \ge l \ge k+2$. In particular, $\left\langle Ap^k, p^i\right\rangle = 0$ if $k \le i-2$. Hence:

$$\left\langle p^{i+1}, p^k\right\rangle = \left\langle p^i, p^k\right\rangle - \alpha_k \left\langle Ap^i, p^k\right\rangle + \beta_k \left\langle p^i - p^{i-1}, p^k\right\rangle = 0$$

if $k \le i-2$. It remains therefore to check that $\left\langle p^{i-1}, p^{i+1}\right\rangle = 0$ and $\left\langle p^i, p^{i+1}, =\right\rangle 0$. The latter is already known (7), hence we are left with the case $k = i-1$. If $k = i-1$: we use again $x^{k+1} = x^k - \alpha_k p^k + \beta_k(x^k - x^{k-1})$ to derive (with $r^0 = 0$)

$$r^{k+1} = -\alpha_k p^k + \beta_k r^k$$

so that $\forall k$, $r^k \in \operatorname{vect}\{p^0, \dots, p^{k-1}\}$. Knowing (7) that $\left\langle p^{i+1}, r^i\right\rangle = 0$, one sees that

$$0 = -\alpha_{i-1}\left\langle p^{i+1}, p^{i-1}\right\rangle + \beta_{i-1}\left\langle p^{i-1}, r^{i-1}\right\rangle = -\alpha_{i-1}\left\langle p^{i+1}, p^{i-1}\right\rangle$$

which shows that $\left\langle p^{i+1}, p^{i-1}\right\rangle = 0$. Hence $(p^0, \dots, p^{i+1})$ are orthogonal. This holds as long as x^{i+1} is not a solution (then $p^{i+1} = 0$).

Corollary 2.14. *The solution is found in $k = \operatorname{rk} A$ iterations.*

Indeed, if $p^{k+1} \neq 0$ then $p^i = Ax^i - b$, $i = 0, \dots, k+1$ are $k+2$ orthogonal vectors in $\operatorname{Im} A - b$ which is an affine space of dimension k and contains at most $k+1$ independent points.

One important point is that also the directions r_i satisfy an orthogonality conditions: they are A-orthogonal: $\langle Ar_i, r_j \rangle = 0$ for all $i \neq j$, hence the name "conjugate directions".

Variants One can show that the following rules defines the same points (for quadratic functions)

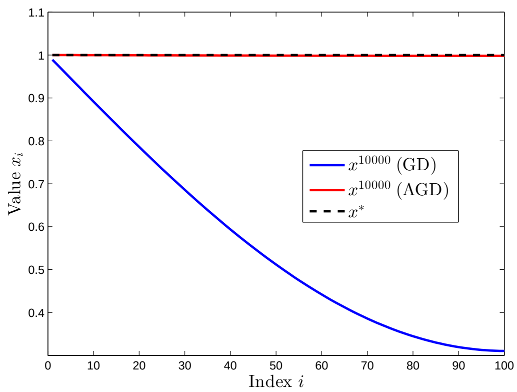
$$\begin{cases} p^k = \nabla f(x^k) \\ \beta_k = \frac{\|p^k\|^2}{\|p^{k-1}\|^2} & (\beta_0 = 0) \\ r^k = -p^k + \beta_k r^{k-1} \\ \alpha_k = \arg\min_{\alpha \geq 0} f(x^k + \alpha p^k), x^{k+1} = x^k + \alpha_k p^k \end{cases}$$

A variant replaces the 2nd line with $\beta_k = \langle p^k, p^k - p^{k-1} \rangle / \|p^{k-1}\|^2$. If f not quadratic, these variants can be implemented.

Optimality The conjugate gradient computes x^k as the minimum of f in the space generated by the orthogonal gradients $(p^0, \dots, p^k)$. It is then possible to prove if $\gamma I \leq A \leq LI$ that

$$\|x^k - x^*\| \leq 2\sqrt{Q} q^k \|x^0 - x^*\|$$

with $q = (\sqrt{Q} - 1)/(\sqrt{Q} + 1)$, $Q = L/\gamma$ the condition number. This is the same rate as the Heavy-Ball.



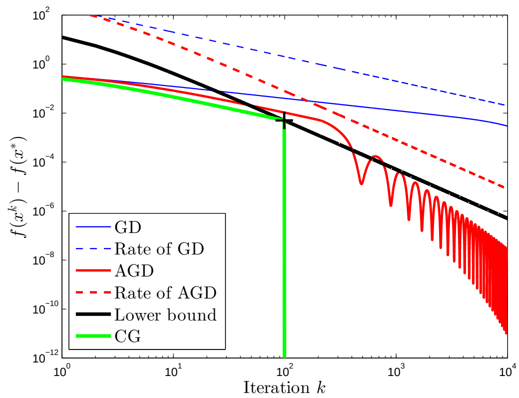


Figure 2: Comparison between accelerated vs non-accelerated gradient schemes. Top: Comparisons of the solutions x of GD and AGD after 10000(!) iterations. Bottom: Rate of convergence for GD, AGD together with their theoretical worst case rates, and the lower bound for smooth optimization. For comparison we also provide the rate of convergence for CG. Note that CG exactly touches the lower bound at $k = 99$ (problem (3) with $\gamma = 0$, $p = n = 100$)

2.4.3 Accelerated algorithm: Nesterov 83

We rapidly mention the "Accelerated Gradient Descent" (AGD) Algorithm by Yu. Nesterov [24]. The idea is of a Gradient Descent with memory.

Algorithm: $x^0 = x^{-1}$ given, x^{k+1} defined by:

$$\begin{cases} y^k = x^k + \frac{t_k - 1}{t_{k+1}}(x^k - x^{k-1}) \\ x^{k+1} = y^k - \tau \nabla f(y^k) \end{cases}$$

where $\tau = 1/L$ and for instance $t_k = 1 + k/2$. Then,

$$f(x^k) - f(x^*) \leq \frac{2L}{(k+1)^2} \|x^0 - x^*\|^2$$

Proof: comes later on in these notes... (easy but requires a bit of convexity.)

2.5 Nonsmooth problems?

2.5.1 Subgradient descent

The first basic approach to tackle nonsmooth problems (or more generally problems where the (local) Lipschitz constant of the gradient is unknown and possibly rapidly varying) is called a "subgradient descent". The idea, given f convex, is to iterate:

$$x^{k+1} = x^k - h_k \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}.$$

In practice, the gradient here can be replaced with any selection of the subgradient if f is not differentiable at x^k, see Section 4.1 for the technical details.

Then if x^* is a solution,

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - 2\frac{h_k}{\|\nabla f(x^k)\|} \left\langle \nabla f(x^k), x^k - x^* \right\rangle + h_k^2 \\ &\leq \|x^k - x^*\|^2 - 2\frac{h_k}{\|\nabla f(x^k)\|}(f(x^k) - f(x^*)) + h_k^2 \end{aligned}$$

Hence, assuming in addition f is M-Lipschitz (near x^* at least)

$$\min_{0 \leq i \leq k} f(x^i) - f(x^*) \leq M \frac{\|x^0 - x^*\|^2 + \sum_{i=0}^k h_i^2}{2\sum_{i=0}^k h_i}$$

and choosing $h_i = C/\sqrt{k+1}$ for k iterations, we obtain

$$\min_{0 \leq i \leq k} f(x^i) - f(x^*) \leq M \frac{C^2 + \|x^0 - x^*\|^2}{2C\sqrt{k+1}}$$

(the best choice is $C \sim \|x^0 - x^*\|$ but this is of course unknown).

In general, one should choose steps such that $\sum_i h_i^2 < +\infty$, $\sum_i h_i = +\infty$, such as $h_i = 1/i$. This is very slow!

2.5.2 Implicit descent

Consider a gradient descent where instead of using the gradient at x^k, one is able to evaluate the gradient in x^{k+1}:

$$x^{k+1} = x^k - \tau \nabla f(x^{k+1}).$$

This is of course often "conceptual"... It says that x^{k+1} is a critical point of (and one can ask that it minimises)

$$f(x) + \frac{1}{2\tau}\|x - x^k\|^2.$$

Observe that if one lets

$$f_\tau(x) := \min_y f(y) + \frac{1}{2\tau}\|y - x\|^2 \tag{10}$$

(this defines an "inf-convolution") which is well-defined if f is bounded from below (or $\geq -\alpha\|x\|^2$ and $\tau < 1/\alpha$) and lower-semicontinuous (if not, the min has to be replaced with an inf), then one can show that f_τ is semi-concave and when differentiable, $\nabla f_\tau(x) = (x - y_x)/\tau$ where y_x solves (10) (and is thus, in this case, unique).

Proof: first we observe that

$$\begin{aligned} &f_\tau(x-h) - 2f_\tau(x) + f_\tau(x+h) \\ &\qquad \leq f(y_x) + \frac{1}{2\tau}\|x-h-y_x\|^2 - 2f(y_x) - \frac{1}{\tau}\|x-y_x\|^2 + f(y_x) + \frac{1}{2\tau}\|x+h-y_x\|^2 \leq \frac{1}{\tau}\|h\|^2 \end{aligned}$$

showing that $f_\tau(x) - \|x\|^2/(2\tau)$ is concave This shows that f_τ is "semi-concave". Hence f_τ is differentiable a.e. (even twice, Aleksandrov's theorem [14]), and if $\nabla f_\tau(x)$ exists, one has

$$f_\tau(x+h) \leq f(y_x) + \frac{1}{2\tau}\|x+h-y_x\|^2, \text{ hence}$$

$$f_\tau(x+h) - f_\tau(x) \leq \frac{1}{\tau}\langle x - y_x, h\rangle + \frac{\|h\|^2}{2\tau},$$

so that for all h,

$$\nabla f_\tau(x)\cdot h \leq \frac{1}{\tau}\langle x - y_x, h\rangle$$

showing the claim. Then, $y_x = x - \tau\nabla f_\tau(x)$.

Conversely, if y_x is unique, then $\nabla f_\tau(x)$ exists and is $(x - y_x)/\tau$. This follows from the observation that if $x_n \to x$ and y_{x_n} is a minimizer for x, as

$$f(y_{x_n}) + \frac{1}{2\tau}\|x_n - y_{x_n}\|^2 \leq f(y_x) + \frac{1}{2\tau}\|x_n - y_x\|^2$$

showing that (f being bounded from below) (y_{x_n}) is a bounded sequence. If $(y_{x_{n_k}})$ is a subsequence which converges to some $\bar{y}$ passing to the limit in

$$f(y_{x_{n_k}}) + \frac{1}{2\tau}\|x_{n_k} - y_{x_{n_k}}\|^2 \leq f(y) + \frac{1}{2\tau}\|x_{n_k} - y\|^2$$

and using the semi-continuity of f, we find that $\bar{y}$ is a minimizer for x, hence $\bar{y} = y_x$ and $y_{x_n} \to y_x$: the multivalued mapping $x \mapsto y_x$ is thus continuous at points where the argument is unique. Now, we can write that

$$f_\tau(x+h) \leq f_\tau(x) + \frac{1}{\tau}\langle x - y_x, h\rangle + \frac{\|h\|^2}{2\tau}$$

and in the same way (exchanging x and $x+h$)

$$\begin{aligned} f_\tau(x) &\leq f_\tau(x+h) - \frac{1}{\tau}\langle x + h - y_{x+h}, h\rangle + \frac{\|h\|^2}{2\tau} \\ &= f_\tau(x+h) - \frac{1}{\tau}\langle x - y_{x+h}, h\rangle - \frac{\|h\|^2}{2\tau} \end{aligned}$$

hence for $t > 0$, small:

$$\frac{1}{\tau}\langle x - y_{x+th}, h\rangle \le \frac{f_\tau(x+th) - f_\tau(x)}{t} \le \frac{1}{\tau}\langle x - y_x, h\rangle + O(t)$$

and in the limit $t \to 0$ we recover the claim.

This proof is finite-dimensional, we will however see later on for convex functions in Hilbert spaces that the same result is true.

We find that

$$x^{k+1} = x^k - \tau\nabla f(x^{k+1}) \Leftrightarrow x^{k+1} = x^k - \tau\nabla f_\tau(x^k)$$

hence the implicit descent is an explicit descent on f_τ! Which has the same minimisers. It converges to critical points of f_τ (as $D^2 f_\tau \le I/\tau$), as before (and under the same assumptions). These are local minimizers of $f(\cdot) + \|\cdot - x\|^2/(2\tau)$.

Example 2.15 (Lasso problem)**.** Consider:

$$\min_x \|x\|_1 + \frac{1}{2}\|Ax - b\|^2$$

If $\|x\|_M^2 = \langle Mx, x\rangle$ and $M = I/\tau - A^*A$, $\tau < 1/\|A\|^2$, then

$$\min_x \frac{1}{2}\|x - x^k\|_M^2 + \|x\|_1 + \frac{1}{2}\|Ax - b\|^2$$

is solved by

$$x^{k+1} = S_\tau(x^k - \tau A^*(Ax^k - b))$$

where $S_\tau\xi$ is the unique minimizer of

$$\min_x \|x\|_1 + \frac{1}{2\tau}\|x - \xi\|^2,$$

called the "shrinkage" operator. This converges with rate $O(1/k)$ to a solution.

3 Krasnoselskii-Mann's convergence theorem

3.1 A "general" convergence theorem

We show here a general form of a convergence theorem of Krasnoselskii and Mann for the iterates of weak contractions (or nonexpansive operators) (it is found in all convex optimisation books, cf for instance [4, 2]). We state first a simple form. Consider (in a Hilbert space $\mathcal{X}$) an operator $T : \mathcal{X} \to \mathcal{X}$ which is 1-Lipschitz:

$$\|Tx - Ty\| \le \|x - y\| \quad \forall\, x, y \in \mathcal{X}.$$

If in addition it is ρ-Lipschitz with $\rho < 1$, then Picard's classical fixed point theorem shows that the iterates $x^k = T^k x^0$, $k \ge 1$, form a Cauchy sequence and therefore converge to a fixed point, necessarily unique. This relies on the fact that the space is complete.

However, for $\rho = 1$, this does not always work ($\rho = 1$ does not provide much relevant information, as when $T = I$). For instance, if $Tx = -x$, there is only one fixed point but the iterates never converge, unless $x^0 = 0$. The simplest statement of Krasnoselskii-Mann's theorem shows that if T is *averaged* and has fixed points, then the iterates weakly converge to a fixed point. For $\theta \in]0,1[$, we define the (θ-)averaged operator T_θ by letting

$$T_\theta x = (1-\theta)x + \theta Tx.$$

We also let $T_0 = I$, $T_1 = T$, and $F = \{x \in \mathcal{X} : Tx = x\}$. Observe that for any $\theta \in]0,1]$, F is the set of fixed point of T_θ.

Theorem 3.1. *Let $x \in \mathcal{X}$, $0 < \theta < 1$, and assume $F \neq \emptyset$. Then $(T_\theta^k x)_{k\geq 1}$ weakly converges to some point $x^* \in F$.*

Proof in four simple steps. We denote $x^0 = x$, $x^k = T^k x$, $k \geq 1$.

Step 1 first, we see that since T_θ is a weak contraction, then for any $x^* \in F$, $\|T_\theta x^k - x^*\| \leq \|x^k - x^*\|$ and the sequence $(\|x^k - x^*\|)_k$ is nonincreasing. $(x_k)_k$ is said to be "Fejér-monotone" with respect to F, see [2, Chap. 5] for details and interesting properties.

It follows that one can define $m(x^*) = \inf_k \|x^k - x^*\| = \lim_k \|x^k - x^*\|$. If there exists $x^* \in F$ such that $m(x^*) = 0$ then the theorem is proved (with strong convergence), otherwise we proceed to the next step. We will see later on what happens if the sequence is "quasi-Fejér-monotone", which happens for instance if T is computed with errors.

Step 2 We assume that $m(x^*) > 0$ for all $x^* \in F$. We will show that in this case, we still can show that $x^{k+1} - x^k \to 0$ strongly. The operator is said to be "asymptotically regular". We need the following result.

Lemma 3.2. $\forall \varepsilon > 0, \theta \in (0,1)$, $\exists \delta > 0$ *such that for all $x, y \in \mathcal{X}$ with $\|x\| \leq 1$, $\|y\| \leq 1$ and $\|x-y\| \geq \varepsilon$,*

$$\|\theta x + (1-\theta)y\| \leq (1-\delta)\max\{\|x\|, \|y\|\}$$

Proof: just observe that (parallelogram identity/strong convexity)

$$\begin{aligned}\|\theta x + (1-\theta)y\|^2 &= \theta^2\|x\|^2 + (1-\theta)^2\|y\|^2 + 2\theta(1-\theta)\langle x, y\rangle \\ &= \theta\|x\|^2 + (1-\theta)\|y\|^2 - \theta(1-\theta)\|x-y\|^2\end{aligned}$$

and the result follows. We see that a similar result also would hold in reflexive Banach spaces, where the unit ball is uniformly convex: this allows to extend this theorem to some of these spaces.

Proof of $x^{k+1} - x^k \to 0$: assume that along a subsequence, one has $\|x^{k_l+1} - x^{k_l}\| \geq \varepsilon > 0$. Observe that

$$x^{k_l+1} - x^* = (1-\theta)(x^{k_l} - x^*) + \theta(T_1 x^{k_l} - x^*) \tag{11}$$

and that

$$(x^{k_l} - x^*) - (T_1 x^{k_l} - x^*) = x^{k_l} - T_1 x^{k_l} = -\frac{1}{\theta}(x^{k_l+1} - x^{k_l})$$

so that $\|(x^{k_l} - x^*) - (T_1 x^{k_l} - x^*)\| \geq \varepsilon/\theta > 0$. Hence we can invoke the lemma (remember that $(x^k - x^*)_k$ is globally bounded since its norm is nonincreasing), and we obtain that for some $\delta > 0$,

$$m(x^*) \leq \|x^{k_l+1} - x^*\| \leq (1-\delta)\max\{\|x^{k_l} - x^*\|, \|T_1 x^{k_l} - x^*\|\}$$

but since $\|T_1 x^{k_l} - x^*\| \leq \|x^{k_l} - x^*\|$, it follows

$$m(x^*) \leq (1-\delta)\|x^{k_l} - x^*\|.$$

As $k_l \to \infty$, we get a contradiction if $m(x^*) > 0$.

Remark 3.3. This way to write the proof emphasizes the fact that the uniform convexity of the unit ball is enough to get a similar property in a more general Banach space; however in the Hilbertian setting it is a bit stupid. Indeed, one can apply directly the parallelogram identity to (11) to find that for all k,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= (1-\theta)\|x^k - x^*\|^2 + \theta\|T_1 x^k - x^*\|^2 - \theta(1-\theta)\|T_1 x^k - x^k\|^2 \\ &\leq \|x^k - x^*\|^2 - \tfrac{1-\theta}{\theta}\|x^{k+1} - x^k\|^2\end{aligned}$$

from which one deduces that $\sum_k \|x^{k+1} - x^k\|^2 < \infty$, hence the result. In addition, one observes that the sequence $(1-\theta)/\theta\|x^{k+1} - x^k\|^2$ (which is nonincreasing) is controlled in the following way:

$$\tfrac{1-\theta}{\theta}(k+1)\|x^{k+1} - x^k\|^2 \le \tfrac{1-\theta}{\theta}\sum_{i=0}^{k}\|x^{i+1} - x^i\|^2 \le \|x^0 - x^*\|^2 - \|x^{k+1} - x^*\|^2.$$

As $x^{k+1} - x^k = \theta(T_1 x^k - x^k)$ we obtain a rate for the error $T_1 x^k - x^k$, in the Hilbertian setting, given by:

$$\|T_1 x^k - x^k\| \le \frac{\|x^0 - x^*\|}{\sqrt{\theta(1-\theta)}\sqrt{k+1}}. \tag{12}$$

Step 3. Assume now that $\bar{x}$ is the weak limit of some subsequence $(x^{k_l})_l$. Then, we claim it is a fixed point.

We use Opial's lemma:

Lemma 3.4 ([27, Lem. 1])**.** *If in a Hilbert space $\mathcal{X}$ the sequence $(x_n)_n$ is weakly convergent to x_0 then for any $x \neq x_0$,*

$$\liminf_n \|x_n - x\| > \liminf_n \|x_n - x_0\|$$

Proof of Opial's lemma (obvious): one has

$$\|x_n - x\|^2 = \|x_n - x_0\|^2 + 2\langle x_n - x_0, x_0 - x\rangle + \|x_0 - x\|^2.$$

Since $\langle x_n - x_0, x_0 - x\rangle \to 0$ by weak convergence, we deduce

$$\liminf_n \|x_n - x\|^2 = \liminf_n(\|x_n - x_0\|^2 + \|x_0 - x\|^2) = \|x_0 - x\|^2 + \liminf_n \|x_n - x_0\|^2$$

and the claim follows.

Proof that $\bar{x}$ is a fixed point: since T_θ is a contraction, we observe that for each k,

$$\begin{aligned}\|x^k - \bar{x}\| &\ge \|T_\theta x^k - T_\theta \bar{x}\| \\ &= \|x^{k+1} - x^k + x^k - T_\theta \bar{x}\| \ge \|x^{k+1} - x^k\| - \|x^k - T_\theta \bar{x}\|\end{aligned}$$

and we deduce (thanks to the previous Step 2):

$$\liminf_l \|x^{k_l} - \bar{x}\| \ge \liminf_l \|x^{k_l} - T_\theta \bar{x}\|.$$

Opial's lemma implies that $T_\theta \bar{x} = \bar{x}$.

One advantage of this approach is that it can be extended to Banach spaces [27] where "Opial's property" (the statement of the Lemma) holds (in the norm for which T is a contraction).

Remark 3.5. Another classical approach to prove this claim is to use "Minty's trick" to study the limit of "monotone" operators: Since T_θ is a contraction, for each $y \in \mathcal{X}$ we have (thanks to Cauchy-Schwarz's inequality)

$$\langle (I - T_\theta)x_{n_k} - (I - T_\theta)y, x_{n_k} - y\rangle \ge 0$$

and as we have just proved that $(I - T_\theta)x_{n_k} \to 0$ (strongly), then

$$\langle -(I - T_\theta)y, \bar{x} - y\rangle \ge 0.$$

Choose $y = \bar{x} + \varepsilon z$ for $z \in \mathcal{X}$ and $\varepsilon > 0$: it follows after dividing by ε that

$$\langle (I - T_\theta)(\bar{x} + \varepsilon z), z\rangle \ge 0.$$

and since T_θ is Lipschitz, sending $\varepsilon \to 0$ we recover $\langle (I - T_\theta)\bar{x}, z\rangle \ge 0$ for any z, which shows that $\bar{x} \in F$.

Step 4. To conclude, assume that a subsequence $(x^{m_l})_l$ of $(x^k)_k$ converges weakly to another fixed point $\bar{y}$. Then it must be that $\bar{y} = \bar{x}$, otherwise Opial's lemma 3.4 again would imply both that $m(\bar{x}) < m(\bar{y})$ and $m(\bar{y}) < m(\bar{x})$:

$$m(\bar{y}) = \liminf_l \|x^{m_l} - \bar{y}\| < \liminf_l \|x^{m_l} - \bar{x}\| = m(\bar{x}).$$

It follows that the whole sequence (x^k) must weakly converge to $\bar{x}$.

3.2 Varying steps

One can consider more generally iterations of the form

$$x^{k+1} = x^k + \tau_k(T_1 x^k - x^k)$$

with varying steps τ_k. Then, if $0 < \underline{\tau} \leq \tau_k \leq \overline{\tau} < 1$, the convergence still holds, with almost the same proof. (This is obvious in the Hilbertian setting, *cf* Remark 3.3.)

Remark: a sufficient condition is that $\sum_k \tau_k(1-\tau_k) = \infty$, see [30]. In addition, a slight improvement to Remark 3.3 shows that

$$\sum_{i=0}^{k} (1-\tau_i)\tau_i \|T_1 x^i - x^i\|^2 \leq \|x^0 - x^*\|^2 - \|x^{k+1} - x^*\|^2$$

so that $\min_{0 \leq i \leq k} \|T_1 x^i - x^i\| \leq \|x^0 - x^*\| / \sqrt{\sum_{i=0}^{k}(1-\tau_i)\tau_i}$. In fact, in a general normed space one has the estimate

$$\|T_1 x^k - x^k\| \leq \frac{1}{\sqrt{\pi}} \frac{\|x^0 - x^*\|}{\sqrt{\sum_{i=0}^{k} \tau_i(1-\tau_i)}}$$

which improves (12), see [11].

3.3 A variant with errors

Assume now the sequence (x_k) is an inexact iteration of T_θ:

$$\|x^{k+1} - T_\theta x^k\| \leq \varepsilon_k.$$

Then one has the following result:

Theorem 3.6 (Variant of Thm 3.1). *If $\sum_k \varepsilon_k < \infty$, then $x^k \to \bar{x}$ a fixed point of T.*

Proof: now, x^k is "quasi-Fejér monotone": denoting $e_k = x^{k+1} - T_\theta x^k$ so that $\|e_k\| \leq \varepsilon_k$,

$$\|x^{k+1} - x^*\| = \|T_\theta x^k - T_\theta x^* + e_k\| \leq \|x^k - x^*\| + \varepsilon_k$$

for all k, and any $x^* \in F$. Hence, $\|x^{k+1} - x^*\| \leq \|x^0 - x^*\| + \sum_{i=0}^{k} \varepsilon_i$ is bounded. Letting $a_k = \sum_{i=k}^{\infty} \varepsilon_i$ which is finite and goes to 0 as $k \to \infty$, this can be rewritten

$$\|x^{k+1} - x^*\| + a_{k+1} \leq \|x^k - x^*\| + a_k$$

so that once more one can define

$$m(x^*) := \lim_{k \to \infty} \|x^k - x^*\| = \inf_{k \geq 0} \|x^k - x^*\| + a_k$$

Again, if $m(x^*) = 0$ the theorem is proved, otherwise, one can continue the proof as before: now,

$$x^{k_l+1} - x^* = (1-\theta)(x^{k_l} - x^* + e_{k_l}) + \theta(T_1 x^{k_l} - x^* + e_{k_l})$$

while

$$(x^{k_l} - x^* + e_{k_l}) - (T_1 x^{k_l} - x^* + e_{k_l}) = x^{k_l} - T_1 x^{k_l} = -\frac{1}{\theta}(x^{k_l+1} - x^{k_l} - e_{k_l})$$

so that $\|(x^{k_l} - x^*) - (T_1 x^{k_l} - x^*)\| \geq (\varepsilon - \varepsilon_{k_l})/\theta > \varepsilon/(2\theta) > 0$ if l is large enough, and one can invoke again Lemma 3.2 to find that

$$\begin{aligned} m(x^*) \leq \|x^{k_l+1} - x^*\| \leq (1-\delta)\max\{\|x^{k_l} - x^* + e_{k_l}\|, \|T_1 x^{k_l} - x^* + e_{k_l}\|\} \\ \leq (1-\delta)\left(\|x^{k_l} - x^*\| + \varepsilon_{k_l}\right) \end{aligned}$$

and again sending $l \to \infty$ we obtain that $m(x^*) \leq (1-\delta)m(x^*)$, a contradiction if $m(x^*) > 0$. The rest of the proof (steps 3, 4) is almost identical.

Remark 3.7. What do you think about the condition $\sum_k \varepsilon_k < \infty$ in practice?

3.4 Examples

3.4.1 Gradient descent

It follows the convergence for the explicit and implicit gradient descent for convex functions. Consider indeed the iteration $x^{k+1} = T_\tau(x^k) := x^k - \tau\nabla f(x^k)$, for f convex with L-Lipschitz gradient. Then, Lemma 2.4 claims that

$$T_{2/L}(x) = x - \frac{2}{L}\nabla f(x)$$

is a weak contraction (1-Lipschitz or "nonexpansive" operator).

We observe that if $0 < \tau < 2/L$, one has

$$T_\tau(x) = x - \frac{\tau L}{2}\frac{2}{L}\nabla f(x) = \frac{\tau L}{2}T_{2/L}(x) + \left(1 - \frac{\tau L}{2}\right)x$$

is an averaged operator (with here $\theta = L\tau/2 \in]0,1[$). Theorem 3.1 yields the convergence of the iterates. Moreover, one still has convergence if one uses varying steps τ_k with $0 < \inf_k \tau_k \leq \sup_k \tau_k < 2/L$. One can also consider (summable) errors. Eventually, thanks to Remark 3.3, one has the rate

$$\|\tfrac{2}{L}\nabla f(x^k)\| \leq \frac{\|x^0 - x^*\|}{\sqrt{(1 - L\tau/2)L\tau/2}\sqrt{k+1}}.$$

(Compare this with (2), Theorem 2.6, Remark 2.2.)

For the implicit descent, we can use the fact that it is an explicit descent on the function f_τ, which has $1/\tau$-Lipschitz gradient, to get a similar result: Let $x^{k+1} = x^k - \lambda\nabla f_\tau(x^k) = x^k + (\lambda/\tau)(y_{x_k} - x_k)$ (where y_x solves (10)) for $0 < \lambda < 2\tau$, then x^k converges (weakly) to a minimizer of f_τ (which is also a minimizer of f)...

3.4.2 Composition of averaged operators

An important remark is the following: Let T_θ, S_λ be averaged operators: $T_\theta = (1-\theta)I + \theta T_1$, $S_\lambda = (1-\lambda)I + \lambda S_1$. Then $T_\theta \circ S_\lambda$ is also averaged: letting $\mu = \theta + \lambda(1-\theta) \in]0,1[$, one has

$$T_\theta \circ S_\lambda = (1-\mu)I + \mu\frac{(1-\theta)\lambda S_1 + \theta T_1 \circ ((1-\lambda)I + \lambda S_1)}{\theta + (1-\theta)\lambda}.$$

An important application is the following: consider the problem

$$\min_x f(x) + g(x) \tag{13}$$

where f, g convex, lsc, f has L-Lipschitz gradient and g is such that one knows how to compute, for all y and all $\tau > 0$:

$$g_\tau(x) := \min_y g(y) + \frac{1}{2\tau}\|x - y\|^2. \tag{14}$$

Then one can compose the averaged operators

$$T_\tau x := x - \tau \nabla f(x),$$

$0 < \tau < 2/L$, and

$$S_\tau x := y_x$$

which solves (14) (and is (1/2)-averaged, as it is $x - \tau\nabla g_\tau(x)$ where ∇g_τ is $1/\tau$-Lipschitz). Hence, if one defines the iterates $x^{k+1} := S_\tau \circ T_\tau x^k$, $k \geq 0$, then $x^k \rightharpoonup x^*$ (weakly) where x^* is a fixed point if $S_\tau \circ T_\tau$. As $S_\tau x$ satisfies

$$\nabla g(S_\tau x) + \frac{1}{\tau}(S_\tau x - x) = 0,$$

one has

$$0 = \nabla g(S_\tau(T_\tau x^*)) + \frac{1}{\tau}(S_\tau(T_\tau x^*) - T_\tau x^*)$$
$$= \nabla g(x^*) + \frac{1}{\tau}(x^* - (x^* - \tau\nabla f(x^*))) = \nabla g(x^*) + \nabla f(x^*)$$

so that x^* is a minimizer of (13). We deduce the following:

Theorem 3.8. *The iterates of the "forward-backward" algorithm $x^{k+1} := S_\tau \circ T_\tau x^k$ weakly converge to a minimizer of* (13).

We will see later on that one can say much more about this approach. Compare this with Example 2.15.

Remark 3.9. What about the "explicit-explicit" ("forward-forward") iteration

$$x^{k+1} = x^k - \tau\nabla f(x^k) - \tau\nabla g(x^k - \tau\nabla f(x^k)),$$

with $\tau < \min\{2/L_f, 2/L_g\}$ where L_f, L_g are the Lipschitz constants of the gradients of f, g, respectively?

We will see later on other useful examples of composition of averaged operators.

4 An introduction to convex analysis and monotone operators

Most of this section is in Hilbert spaces, though many results are also valid in more general vector spaces, but often with more involved proofs.

4.1 Convexity

See for instance: [32, 13] for a general introduction. We discuss here the following notions: Convex function; Subgradients; Inf-convolution; sum of subgradients; Convex Conjugate (Legendre-Fenchel); Fenchel-Rockafellar duality; Moreau-Yosida's regularization (inf-convolution); Moreau's identity.

4.1.1 Convex functions

An extended-valued function $f : \mathcal{X} \to [-\infty, +\infty]$ is said to be *convex* if and only if its *epigraph*

$$\operatorname{epi} f := \{(x, \lambda) \in \mathcal{X} \times \mathbb{R} : \lambda \geq f(x)\}$$

is a convex set, that is, if when $\lambda \geq f(x)$, $\mu \geq f(y)$, and $t \in [0,1]$, one has $t\lambda + (1-t)\mu \geq f(tx + (1-t)y)$.[3] It is *proper* if it is not identically $+\infty$ and nowhere $-\infty$: in this case, it is convex if and only if for all $x, y \in \mathcal{X}$ and $t \in [0,1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

It is *strictly convex* if the above inequality is strict whenever $x \neq y$ and $0 < t < 1$. It is *strongly convex* (or μ-convex) if in addition, there exists $\mu > 0$ such that for all $x, y \in \mathcal{X}$ and $t \in [0,1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \mu\frac{t(1-t)}{2}\|x - y\|^2.$$

Thanks to the parallelogram identity, in the Hilbertian setting, one easily checks that this is equivalent to require that $x \mapsto f(x) - \mu/2\|x\|^2$ is still convex. The function is also said to be, in this case, "μ-convex". The archetypical example of a μ-convex function is a quadratic plus affine function $\mu\|x\|^2/2 + \langle b, x\rangle + c$.

The *domain* of a proper convex function f is the set $\operatorname{dom} f = \{x \in \mathcal{X} : f(x) < +\infty\}$. It is obviously a convex set.

We say that f is *lower semi-continuous* (l.s.c.) if for all $x \in \mathcal{X}$, if $x_n \to x$, then

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

It is easy to see that f is l.s.c. if and only if $\operatorname{epi} f$ is closed.

A trivial but important example is the *characteristic function* or *indicator function* of a set C:

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{else,} \end{cases}$$

which is convex, l.s.c., and proper as soon as C is convex, closed and nonempty. The minimisation of such functions will allow to easily model convex constraints in our problems.

One can show the following result:

Lemma 4.1. *If there exists $B \subset \operatorname{dom} f$ an open ball where the proper convex function f is bounded from above, then f is locally Lipschitz in the interior of* $\operatorname{dom} f$*. In finite dimension, a proper convex function f is locally Lipschitz in the relative interior of* $\operatorname{dom} f$*,* $\operatorname{ri}\operatorname{dom} f$*.*

In finite dimension, the relative interior is defined as the interior of $\operatorname{dom} f$ in the space $x + \operatorname{vect}(\operatorname{dom} f - x)$ for any $x \in \operatorname{dom} f$; unless this space has dimension zero it is never empty.

Remark 4.2. Note that in infinite dimension (using the axiom of choice) one can possibly find noncontinuous linear forms hence noncontinuous convex functions. However, one can show that a convex proper *lower semi-continuous* function is always locally bounded in the interior of its domain, and therefore locally Lipschitz (as if 0 is an interior point and one considers the convex closed set $C = \{x : f(x) \leq 1 + f(0)\}$, one can check that $\cup_{n\geq 1} nC = \mathcal{X}$, as if $x \in \mathcal{X}$, $t \mapsto f(tx)$ is locally Lipschitz near $t = 0$. Hence $\mathring{C} \neq \emptyset$ by Baire's property: it follows that there is an open ball where f is bounded, as requested), *cf* [13, Cor. 2.5].

[3] This definition avoids the embarrassing expression $(+\infty) + (-\infty)$, see for instance [32, Sec. 4].

Proof of the lemma: we assume that $B = B(0,\delta)$, $\delta > 0$, and let $M = \sup_B f < \infty$. Observe also that for $x \in B$, by convexity $f(x) \geq 2f(0) - f(-x) \geq 2f(0) - M$ so that $|f| \leq M + 2|f(0)|$. We prove that f is Lipschitz in $B(0,\delta/2)$: indeed, if $x, y \in B(0,\delta/2)$, there is $z \in B(0,\delta)$ such that $y = (1-t)x + tz$ for some $t \in [0,1]$, and $\|z - x\| \geq \delta/2$. In particular by convexity, $f(y) - f(x) \leq t(f(z) - f(x)) \leq t2(M - f(0))$. Now, $t(z-x) = y-x$ so that $t \leq \|y-x\|/\|z-x\| \leq 2\|y-x\|/\delta$ hence: $f(y)-f(x) \leq \big(4(M-f(0))/\delta\big)\|y-x\|$ which shows the claim (one could show in fact in the same way that f is Lipschitz in any ball contained in $B(0,\delta)$).

Now, let $\bar{x}$ in the interior of $\operatorname{dom} f$. Observe that for some $\lambda > 1$, $\lambda\bar{x} \in \operatorname{dom} f$ and as a consequence $B' = 1/\lambda(\lambda\bar{x}) + (1-1/\lambda)B(0,\delta) = B(\bar{x}, \delta(1-1/\lambda)) \subset \operatorname{dom} f$; moreover, if $x \in B'$, $x = 1/\lambda(\lambda\bar{x}) + (1-1/\lambda)z$ for some z with $f(z) \leq M$ hence $f(x) \leq 1/\lambda f(\lambda\bar{x}) + (1 - 1/\lambda)M$, so that $\sup_{B'} f < \infty$. Hence as before f is Lipschitz in a smaller ball.

In finite dimension, assume $0 \in \operatorname{dom} f$ and let d be the dimension of $\operatorname{vect}\operatorname{dom} f$. It means there exist $x_1, \dots, x_d$ independent points in $\operatorname{dom} f$. Now, the d-dimensional set $\{\sum_i t_i x_i : t_i > 0, \sum_i t_i \leq 1\}$ (the interior of the convex envelope of $\{0, x_1, \dots, x_d\}$) is an open set in $\operatorname{vect}\operatorname{dom} f$, moreover if $x = \sum_i t_i x_i$, $f(x) \leq \sum_i t_i f(x_i) + (1 - \sum_i t_i) f(0) \leq M := \max\{f(0), f(x_1), \dots, f(x_d)\}$. Hence we can apply the first part of the theorem, and f is locally Lipschitz in the relative interior of the domain.

4.1.2 Separation of convex sets

Let us mention the two following theorems:

Theorem 4.3. *Let $\mathcal{X}$ be a (real) Hilbert space, $C \subset \mathcal{X}$ a closed, convex set and $x \notin C$. Then there exists a closed hyperplane which "separates" strictly x and C: precisely, in the Hilbertian setting, one can find $v \in X, \alpha \in \mathbb{R}$ such that*

$$\langle v, x \rangle > \alpha \geq \langle v, y \rangle \ \ \forall\ y \in C$$

Proof: introduce the projection $z = \Pi_C(x)$ defined by $\|x - z\| = \min_{y \in C} \|x - y\|$ (existence is classically shown by proving that any minimizing sequence is a Cauchy sequence, thanks to the parallelogram identity [or strong convexity of $\|x - \cdot\|^2$]). The first order optimality condition for z is found by writing that for any $y \in C$, $\|x - z\|^2 \leq \|x - (z + t(y - z))\|^2$ for $t \in (0,1]$ and then sending $t \to 0$. We find

$$\langle x - z, y - z \rangle \leq 0 \ \forall\ y \in C.$$

It follows that if $v = x - z \neq 0$, $y \in C$,

$$\langle v, x \rangle = \langle x - z, x \rangle = \|x - z\|^2 + \langle x - z, z \rangle \geq \|x - z\|^2 + \langle x - z, y \rangle = \|v\|^2 + \langle v, y \rangle .$$

The result follows (letting for instance $\alpha = \|v\|^2/2 + \sup_{y \in C} \langle v, y \rangle$). The proof can easily be extended to the situation where $\{x\}$ is replaced with a compact convex set not intersecting C.

Corollary 4.4. *In a real Hilbert space $\mathcal{X}$, a closed convex set C is weakly closed.*

Indeed, if $x \notin C$, one finds v, α with $\langle v, x \rangle > \alpha \geq \langle v, y \rangle$ for all $y \in C$ and this defines a neighborhood $\{\langle v, \cdot \rangle > \alpha\}$ of x for the weak topology which does not intersect C.

Theorem 4.5. *Let $\mathcal{X}$ be a (real) Hilbert space, $C \subset \mathcal{X}$ an open convex set and $C' \subset \mathcal{X}$ a convex set with $C' \cap C = \emptyset$. Then Then there exists a closed hyperplane which "separates" C and C': precisely, in the Hilbertian setting, one can find $v \in X, \alpha \in \mathbb{R}$, $v \neq 0$, such that*

$$\langle v, x \rangle \geq \alpha \geq \langle v, y \rangle \ \ \forall\ x \in C, y \in C'$$

Proof: first assume that $C' = \{\bar{x}\}$ is a singleton. Then as $x \notin C$, one can find $x_n \to \bar{x}$ such that $x_n \notin \overline{C}$. By Theorem 4.3 there exists v_n such that for all $x \in \overline{C}$,

$$\langle v_n, x_n \rangle \leq \langle v_n, x \rangle$$

and we can assume $\|v_n\| = 1$. Up to a subsequence, we may then assume that $v_n \rightharpoonup v$ weakly in $\mathcal{X}$. In the limit, (using that $x_n \to \bar{x}$ strongly) we obtain $\langle v, \bar{x} \rangle \leq \langle v, x \rangle \ \forall\, x \in C$, which is our claim if $v \neq 0$.

Observe now that since there exists a ball $B(y, \delta)$ in C, one has for any $\|z\| \leq 1$

$$\langle v_n, x_n \rangle \leq \langle v_n, y - \delta z \rangle$$

so that $\langle v_n, y - x_n \rangle \geq \delta \langle v_n, z \rangle$: and taking the supremum over all possible z we find $\langle v_n, y - x_n \rangle \geq \delta$. In the limit we deduce $\langle v, y - \bar{x} \rangle \geq \delta$ which shows that $v \neq 0$.

Now, to show the general case, one lets $A = C' - C = \{y - x : y \in C', x \in C\}$: this is an open convex set and by assumption, $0 \notin A$. Hence by the previous part, there exists $v \neq 0$ such that $\langle v, y - x \rangle \leq \langle v, 0 \rangle = 0$ for all $y \in C'$, $x \in C$, which is the thesis of the Theorem.

These are simple examples of separation theorems, which are geometric versions of the Hahn-Banach theorem and are valid in fact in a much more general setting, see [7, 13].

4.1.3 Subgradient

Given a convex, extended valued, $f : \mathcal{X} \to [-\infty, +\infty]$, its subgradient at a point x is defined as the set

$$\partial f(x) := \{p \in \mathcal{X} \,:\, f(y) \geq f(x) + \langle p, y - x \rangle \ \forall y \in \mathcal{X}\}.$$

This is a closed, convex set.

If f is (Fréchet-)differentiable at x, then it is easy to see that $\partial f(x) = \{\nabla f(x)\}$: one has

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + o(|y - x|)$$

so that if $p \in \partial f(x)$,

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

Taking $y = x + th$, for $h \in \mathcal{X}$ and $t > 0$ small, we find after dividing by t and sending $t \to 0$ that $\langle \nabla f(x) - p, h \rangle \geq 0$. Hence $p = \nabla f(x)$. We leave to the reader the proof that $\nabla f(x) \in \partial f(x)$ (hence $\nabla f(x) \neq \emptyset$), which follows from convexity.

Observe that if $x \in \operatorname{dom} f$, $v \in \mathcal{X}$, $t > s > 0$ one has

$$f(x + sv) = f((s/t)(x + tv) + (1 - s/t)x) \leq \tfrac{s}{t} f(x + tv) + (1 - \tfrac{s}{t}) f(x)$$

so that

$$\frac{f(x + sv) - f(x)}{s} \leq \frac{f(x + tv) - f(x)}{t}.$$

It follows that

$$f'(x; v) := \lim_{t \downarrow 0^+} \frac{f(x + tv) - f(x)}{t} = \inf_{t > 0} \frac{f(x + tv) - f(x)}{t} \tag{15}$$

is well defined (in $[-\infty, \infty]$), and $< +\infty$ as soon as $\{x + tv : t > 0\} \cap \operatorname{dom} f \neq \emptyset$. If $x \in \overset{\circ}{\overbrace{\operatorname{dom} f}}$, then $f'(x; v) < \infty$ for all v, moreover as $f'(x; 0) = 0 \leq f'(x; v) + f'(x; -v)$ it is not $-\infty$ either. In fact, $f'(x; \cdot)$ is a limit of convex functions, and hence convex, moreover, it is clearly positively 1-homogeneous: $f'(x; \lambda v) = \lambda f'(x; v)$ for all $\lambda \geq 0$ and all v.

If this quantity is finite, then the function has a Gâteaux derivative in the direction v (however, usual definitions of Gâteaux differentiability require that this derivative be a continuous linear form of v).

By definition, one easily sees that $f'(x;v) \geq \langle p, v\rangle$ if and only if $p \in \partial f(x)$. ($f'(x;v) \geq \langle p, v\rangle \Rightarrow f(x+tv) - f(x) \geq t\,\langle p, v\rangle$ for all $t > 0$, $v \in \mathcal{X}$.) This means that

$$\partial f'(x;\cdot)(0) = \partial f(x). \tag{16}$$

If in addition, f is locally bounded near $x \in \overbrace{\mathrm{dom}\, f}^{\circ}$ (for this, as we have seen in Lemma 4.1, it is enough that f be locally bounded near one point of the domain, or that f be lsc, *cf* Remark 4.2), then one can easily deduce that also $f'(x;\cdot)$ is, and in particular it is Lipschitz (globally, as it is 1-homogeneous).

(In addition, in finite dimension, the convergence in (15) is uniform for $\|v\| \leq 1$ because of Ascoli-Arzelà's theorem: In fact, if $t \leq t_0$ small enough and $\|v\| \leq 2$,

$$h_x^t(v) := \frac{f(x+tv)-f(x)}{t} \leq \frac{f(x+t_0v)-f(x)}{t_0} \leq M$$

for some M and the proof of Lemma 4.1 shows that the h_x^t are uniformly Lipschitz in B_1.)

We will see later on (Sections 4.2, 4.2.2) that since in these cases, $f'(x;\cdot)$ is continuous, $f'(x;v) = \sup_{p\in\partial f(x;\cdot)(0)} \langle p, x\rangle$, so that $f'(x;\cdot)$ is the *support function* of $\partial f(x)$ which in particular cannot be empty.

Moreover, we deduce that if $\partial f(x) = \{p\}$ is a singleton, then $f'(x;v) = \langle p, x\rangle$. In finite dimension (as the convergence $h_x^t \to f'(x;\cdot)$ is uniform) we deduce that f is differentiable at x. In infinite dimension, we deduce that f is Gâteaux-differentiable. It is not necessarily Fréchet-differentiable: for instance in $\ell^2(\mathbb{N})$, the convex function

$$f(x) = \sup_{i\geq 0}\left(\sqrt{\tfrac{1}{i+1} + x_i^2} - \sqrt{\tfrac{1}{i+1}}\right)$$

which is bounded near 0 ($\|x\| \leq 1 \Leftrightarrow \sum_i |x_i|^2 \leq 1 \Rightarrow |x_i| \leq 1\ \forall\, i \geq 0$) satisfies $\partial f(0) = \{0\}$, however if $v = e_i = (\delta_{i,j})_{j\geq 0}$, then

$$\frac{f(0+tv)-f(0)}{t} = \frac{1}{t}\left(\sqrt{\tfrac{1}{i+1} + t^2} - \sqrt{\tfrac{1}{i+1}}\right) = \sqrt{2} - 1$$

if $t = 1/\sqrt{i+1}$, showing that the differentiability is only Gâteaux. (It is, as for any v and $t > 0$,

$$\frac{f(tv)-f(0)}{t} = \sup_{i\geq 0}\frac{1}{t}\left(\sqrt{\tfrac{1}{i+1} + t^2v_i^2} - \sqrt{\tfrac{1}{i+1}}\right)$$

and for each i, the quantity in the sup is less than $|v_i|$. Given ε, one can find i_0 such that $|v_i| \leq \varepsilon$ for $i > i_0$, while for $i = 0, \ldots, i_0$, if t is small enough one can make the quantity below the sup less than ε. Hence the Gâteaux derivative exists and is zero.)

Using Lemma 4.1, we can deduce the following two results:

Lemma 4.6. *Let f be proper, convex. Assume it is lsc, or continuous in one point. Then, in the interior of the domain, $\partial f(x) \neq \emptyset$. In finite dimension, f has a nonempty subdifferential everywhere in* $\mathrm{ri}\,\mathrm{dom}\, f$.

Lemma 4.7. *Let f be proper, convex. Then if f is Gâteaux-differentiable at x, $\partial f(x) = \{\nabla f(x)\}$. Conversely if x is in the interior of* $\mathrm{dom}\, f$ *and f is continuous at some point, then if $\partial f(x)$ is a singleton, f is Gâteaux-differentiable at x.*

In finite dimension, ∂f is a singleton if and only if f is differentiable at x.

An obvious remark which stems from the definition of a subgradient is that this notion allows to generalise the Euler-Lagrange stationary conditions ($\nabla f(x) = 0$ if x is a minimiser of f) to nonsmooth convex functions: we have indeed

$$x \in \mathcal{X} \text{ is a global minimiser of } f \text{ if and only if } 0 \in \partial f(x). \tag{17}$$

In the same way, one has:

Lemma 4.8. *if $x \in \operatorname{dom} f$ is a local minimiser of $f + g$, f convex, g C^1 near x, then for all $y \in \mathcal{X}$,*

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

and $-\nabla g(x) \in \partial f(x)$.

Indeed, one just writes that for $t > 0$ small enough,

$$f(x) + g(x) \leq f(x + t(y - x)) + g(x + t(y - x)) \leq f(x) + t(f(y) - f(x)) + g(x + t(y - x))$$

so that

$$\frac{g(x) - g(x + t(y - x))}{t} \leq f(y) - f(x)$$

and we recover the claim in the limit $t \to 0$.

If the function f is strongly convex or "μ-convex" and $p \in \partial f(x)$, then x is by definition a minimiser of $y \mapsto f(y) - \langle p, y - x \rangle$ which is also μ-convex. In particular, letting $h(y) = f(y) - \langle p, y - x \rangle - \mu \|y - x\|^2/2$, one has that x is a minimizer of $h(y) + \mu \|y - x\|^2/2$ and h is convex. Hence, by Lemma 4.8,

$$0 = -\nabla \left(\frac{\mu}{2} \| \cdot - x\|^2 \right)(x) \in \partial h(x).$$

Hence, $h(y) \geq h(x)$ for all $y \in \mathcal{X}$, that is

$$f(y) - \langle p, y - x \rangle - \mu \|y - x\|^2/2 \geq f(x).$$

We deduce that for any $x, y \in \mathcal{X}$ and $p \in \partial f(x)$:

$$f(y) \geq f(x) + \langle p, y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \tag{18}$$

An equivalent (but important) remark is that if f is strongly convex and x is a minimiser, then one has (since $0 \in \partial f(x)$)

$$f(y) \geq f(x) + \frac{\mu}{2} \|y - x\|^2 \tag{19}$$

for all $y \in \mathcal{X}$.

The domain of ∂f is the set $\operatorname{dom} \partial f = \{x \in \mathcal{X} : \partial f(x) \neq \emptyset\}$. Clearly, $\operatorname{dom} \partial f \subset \operatorname{dom} f$, in fact if f is convex, l.s.c. and proper, we will see later on (see Prop 4.17 or [13]) that $\operatorname{dom} \partial f$ is dense in $\operatorname{dom} f$ (even when $\operatorname{dom} f$ has empty interior, as for instance when $f(u) = \int_\Omega |\nabla u|^2 dx$ for $u \in L^2(\Omega)$). The fact it is not empty will also follow.

In finite dimension, one has seen that for a proper convex function, $\operatorname{dom} \partial f$ contains at least the relative interior of $\operatorname{dom} f$ (that is, the interior in the vector subspace which is generated by $\operatorname{dom} f$).

4.1.4 Subdifferential calculus

Theorem 4.9. *Assume f, g are convex, proper. Then for all x, $\partial f(x)+\partial g(x) \subset \partial(f+g)(x)$. Moreover if there exists $\bar{x} \in \operatorname{dom} f$ where g is continuous, then $\partial f(x)+\partial g(x) = \partial(f+g)(x)$. In finite dimension, if $x \in \operatorname{ri}\operatorname{dom} g \cap \operatorname{ri}\operatorname{dom} f$, this is also true.*

Proof: the inclusion is obvious from the definition. For the reverse inclusion, we assume $p \in \partial(f+g)(x)$ and want to show that it can be decomposed as $q+r$ with $q \in \partial f(x)$ and $r \in \partial g(x)$. For this, we use that $f(y)+g(y) \geq f(x)+g(x)+\langle p, y-x\rangle$.

Thanks to the assumption that g is continuous at $\bar{x}$, $\operatorname{epi}(g(\cdot)-\langle p,\cdot\rangle)$ contains a ball centered at $(\bar{x}, g(\bar{x})-\langle p,\bar{x}\rangle+1)$ and has non empty interior. Denote E this interior, and F the following translation/flip of $\operatorname{epi} f$:

$$F = \{(y,t) : -t \geq f(y) - [f(x)+g(x)-\langle p,x\rangle]\},$$

which is convex. For $(y,t) \in F$, one has $-t \geq f(y)-[f(x)+g(x)-\langle p,x\rangle] \geq -[g(y)-\langle p,y\rangle]$, that is $t \leq [g(y)-\langle p,y\rangle]$ so that $(y,t) \notin E$. Hence by Theorem 4.5 there exists $(q,\lambda) \neq (0,0)$, such that for all $(y,t) \in E$, $(y',t') \in F$,

$$\langle q,y\rangle + \lambda t \geq \langle q,y'\rangle + \lambda t'.$$

As t' can be sent to $-\infty$ (or t to $+\infty$), $\lambda \geq 0$. Moreover since $\bar{x}$ is in $\operatorname{dom} f$, if $\lambda = 0$ one finds that $\langle q, y-\bar{x}\rangle \leq 0$ for all $y \in \operatorname{dom} g$ which contains a ball centered in $\bar{x}$, so that $q = 0$, which is a contradiction. Hence $\lambda > 0$ so that without loss of generality we can assume $\lambda = 1$.

In particular choosing $t' = f(x)+g(x)-\langle p,x\rangle - f(y')$,

$$\langle q,y\rangle + t \geq \langle q,y'\rangle + f(x)+g(x)-\langle p,x\rangle - f(y').$$

for all $(y,t) \in E$. The closure of E contains $\operatorname{epi}(g(\cdot)-\langle p,\cdot\rangle)$, hence it follows that for all y, y',

$$\begin{aligned}
\langle q,y\rangle + g(y) - \langle p,y\rangle \geq \langle q,y'\rangle + f(x)+g(x)-\langle p,x\rangle - f(y') \\
\Leftrightarrow f(y')+g(y) \geq f(x)+g(x)+\langle p,y-x\rangle + \langle q,y'-y\rangle \\
= f(x)+g(x)+\langle p-q,y-x\rangle + \langle q,y'-x\rangle
\end{aligned}$$

showing that $q \in \partial f(x)$ and $r = p-q \in \partial g(x)$, as requested.

In finite dimension, we only sketch the proof: up to a translation, we may assume $0 \in \operatorname{ri}\operatorname{dom} g \cap \operatorname{ri}\operatorname{dom} f$ so that $\operatorname{ri}\operatorname{dom} g$ is the interior of $\operatorname{dom} g$ in $V = \operatorname{vect}\operatorname{dom} g$ and $\operatorname{ri}\operatorname{dom} f$ is the interior of $\operatorname{dom} f$ in $W = \operatorname{vect}\operatorname{dom} f$.

Let $p \in \partial(f+g)(x)$. Let $\tilde{f}, \tilde{g}$ be the restrictions of f, g to V: one has for all $y \in V$

$$\tilde{f}(y)+\tilde{g}(y) = f(y)+g(y) \geq f(x)+g(x)+\langle p,y-x\rangle = \tilde{f}(x)+\tilde{g}(x)+\langle \tilde{p},y-x\rangle$$

where $\tilde{p}$ is the orthogonal projection of p on $V \ni y-x$.

As $x \in \operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g$, $\tilde{g}$ is continuous at x and the previous result shows that there exists $\tilde{q} \in \partial\tilde{f}(x)$, $\tilde{r} \in \partial\tilde{g}$ with $\tilde{p} = \tilde{r}+\tilde{q}$.

As $x \in \operatorname{ri}\operatorname{dom} f$, $\partial f(x)$ is not empty. (using for instance again the separation theorem between $\{(y, f(x)+\langle\tilde{q},y-x\rangle) : y \in V \cap W\}$ and the interior $\operatorname{epi} f$ in W, the affine subspace generated by $\operatorname{dom} f$) that there is $q' \perp V$ such that $q = \tilde{q}+q' \in \partial f(x)$, then $r' = p-\tilde{r}-\tilde{q}-q' \perp V$ and $r+r' \in \partial g(x)$...

Consider $C' = \{(y, f(x)+\langle\tilde{q},y-x\rangle) : y \in V \cap W\}$, and C the interior of $\operatorname{epi} f$ in W. As $\tilde{q} \in \partial\tilde{f}$, $C \cap C' = \emptyset$. Then by Theorem 4.5, there exists $v \in W$, $\lambda \in \mathbb{R}$ with $(v,\lambda) \neq (0,0)$, such that

$$\langle v,y\rangle + \lambda(f(x)+\langle\tilde{q},y-x\rangle) \leq \langle v,z\rangle + \lambda t$$

for all $y \in V \cap W$ and (z,t) in the relative interior of $W \cap \operatorname{epi} f$ (and passing to the limit, any $(z,t) \in \operatorname{epi} f$). As usual, $\lambda \geq 0$ (as one can send $t \to +\infty$). Then, $\lambda \neq 0$, since as $\operatorname{dom} f$ is relatively open in W, and by assumption, contains a small ball $B(0,\delta)$, it would yield taking $z = -\delta v/|v|$ that

$$\langle v, y \rangle \leq -\delta |v|$$

for any $y \in V \cap W \ni 0$, hence that $v = 0$, a contradiction. Hence we may assume $\lambda = 1$ and we find for all z and all $y \in V \cap W$:

$$f(z) + \langle v, z \rangle \geq \langle v, y \rangle + f(x) + \langle \tilde{q}, y - x \rangle$$

This is possible only if $v + \tilde{q} \in W \cap V^\perp$ (otherwise the supremum of the right-hand side over all $v \in V \cap W$ would be $+\infty$). We can thus write $v = -(\tilde{q} + q')$ for some $q' \in W$, $q' \perp V$. Then, we are left with

$$f(z) \geq f(x) + \langle \tilde{q} + q', z \rangle - \langle \tilde{q}, x \rangle = f(x) + \langle \tilde{q} + q', z - x \rangle$$

using that $\langle q', x \rangle = 0$ since $x \in V$. Hence $\tilde{q} + q' \in \partial f(x)$. Letting then then $r' = p - \tilde{r} - \tilde{q} - q' \perp V$ one has that $\tilde{r} + r' \in \partial g(x)$ (as $\tilde{r} \in \partial g(x)$ and $r' \perp \operatorname{dom} g$). This concludes the proof.

Theorem 4.10. *Let $A : \mathcal{X} \to \mathcal{Y}$ be a continuous operator between two Hilbert spaces and f a proper, convex function on $\mathcal{Y}$. Let $g = f(Ax)$, then if there is $\bar{x}$ such that f is continuous at $A\bar{x}$, $\partial g(x) = A^* \partial f(Ax)$. In finite dimension, one can just require that $A\bar{x} \in \operatorname{ri} \operatorname{dom} f$.*

Proof: $A^* \partial f(Ax) \subset \partial g(x)$ is easy. If $p \in \partial g(x)$, one has for all z,

$$f(Az) \geq f(Ax) + \langle p, z - x \rangle . \tag{20}$$

Hence $\overbrace{\operatorname{epi} f}^{\circ}$ (which is non empty because f is continuous at some point) and

$$E = \{(Az, f(Ax) + \langle p, z - x \rangle) : z \in \mathcal{X}\} \subset \mathcal{Y} \times \mathbb{R}$$

have no common point: if $(y,t) \in E$, then by (20) $t \leq f(y)$. Then there exists by Theorem 4.5 (q,λ) such that

$$-\langle q, y \rangle + \lambda t \geq -\langle q, y' \rangle + \lambda t'$$

for all $(y,t) \in \operatorname{epi} f$ and $(y',t') \in E$. Again, $\lambda \geq 0$, and if $\lambda = 0$ one can find a contradiction as in the previous proof. Then, assuming $\lambda = 1$, one obtains for all $z \in \mathcal{X}$, $y \in \mathcal{Y}$,

$$-\langle q, y \rangle + f(y) \geq -\langle q, Az \rangle + f(Ax) + \langle p, z - x \rangle = f(Ax) + \langle p - A^* q, z \rangle - \langle p, x \rangle .$$

This is possible only if $p = A^* q$, otherwise one can send the right-hand side to $+\infty$. Hence, $p = A^* q$, $\langle p, x \rangle = \langle q, Ax \rangle$ and

$$f(y) \geq f(Ax) + \langle q, y - Ax \rangle$$

for all y, so that $q \in \partial f(Ax)$.

In finite dimension, we leave the proof to the reader (see also [32, Thm 23.9]).

4.1.5 Remark: KKT's theorem

Theorem 4.11 (Karush-Kuhn-Tucker)**.** *Let f, g_i, $i = 1, \ldots, m$ be C^1, convex and assume*

$$\exists\, \bar{x},\ (g_i(\bar{x}) < 0\, \forall\, i = 1, \ldots, m) \qquad \textit{(Slater's condition)}$$

Then x^* is a solution of

$$\min_{g_i(x) \leq 0, i=1,\ldots,m} f(x)$$

if and only if there exists $(\lambda_i)_{i=1}^m$, $\lambda_i \geq 0$ *such that*

$$\nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla g_i(x^*) = 0 \tag{21}$$

and for all $i = 1, \dots, m$*:*

$$\lambda_i g_i(x^*) = 0 \qquad \textit{(complementary slackness condition)}$$

Proof: first, if (21) holds together with the complementary slackness condition, then it is easy to show that x^*, which is a minimizer of the convex function $f + \sum_i \lambda_i g_i$, is a solution of the constrained problem: if x satisfies the constraints, then

$$f(x) \geq f(x) + \sum_i \lambda_i g_i(x) \geq f(x^*) + \sum_i \lambda_i g_i(x^*) = f(x^*).$$

Conversely, consider for all i the function

$$\delta_i(x) = \begin{cases} 0 & \text{if } g_i(x) \leq 0, \\ +\infty & \text{else.}, \end{cases}$$

then the problem is equivalent to $\min_x f(x) + \sum_i \delta_i(x)$. By Slater's condition, we know that there exists $\bar{x}$ where all functions f, δ_i are continuous. Hence by Thm. 4.9,

$$0 \in \partial(f + \sum_i \delta_i)(x^*) = \nabla f(x^*) + \sum_{i=1}^m \partial\delta_i(x^*).$$

It remains to characterize $\partial\delta_i(x^*)$: if $g_i(x^*) < 0$ then it is negative in a neighborhood of x^* and $\partial\delta_i(x^*) = \{0\}$. If $g_i(x^*) = 0$, then we need to characterize the vectors p such that for all y with $g_i(y) \leq 0$,

$$0 \geq \langle p, y - x^* \rangle .$$

Let $v \perp \nabla g_i(x^*)$, and consider $y = x^* - t(\nabla g_i(x^*) + v)$: then

$$g_i(y) = -t \langle \nabla g_i(x^*), \nabla g_i(x^*) + v \rangle + o(t) = -t \|\nabla g_i(x^*)\|^2 + o(t) < 0$$

if t is small enough, hence

$$0 \leq \langle p, \nabla g_i(x^*) + v \rangle .$$

We easily deduce that we must have $p = \lambda_i \nabla g_i(x^*)$, for some $\lambda_i \geq 0$ (in other words, $\partial\delta_i(x^*) = \mathbb{R}_+ \nabla g_i(x^*)$). The theorem follows.

4.2 Convex duality

4.2.1 Legendre-Fenchel conjugate

Given a function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, we introduce the *Legendre-Fenchel conjugate*

$$f^*(y) := \sup_{x \in \mathcal{X}} \langle y, x \rangle - f(x)$$

which is defined for all $p \in \mathcal{X}$, as a supremum of continuous linear forms: in particular, it is obviously a convex, lsc function. Observe that here we rely on the Riesz theorem to define the conjugate, in a more general vector space E, the proper definition should be as a function defined in a dual space E', see for instance [13].

Obviously for all x, y,

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

and in particular $f(x) \geq \langle y, x \rangle - f^*(y)$. Thus, the *biconjugate* f^{**}, defined as f^* by $f^{**}(y) = \sup_{x \in \mathcal{X}} \langle y, x \rangle - f^*(y)$, clearly satisfies

$$f^{**} \leq f.$$

The following is the most important result about the Legendre-Fenchel conjugate (it is also elementary in our Hilbertian setting):

Theorem 4.12. *If f has an affine minorant, then f^{**} is the largest convex lsc function below f, called the convex lsc envelope of f (sometimes also the Γ-regularization, or the convex relaxation). In this case, either $f \equiv +\infty$, or f^*, f^{**} are proper. If f has no affine minorant, $f^{**} \equiv -\infty$.*

This is a consequence of the separation theorem. Observe that the convex lsc envelope of a function f is always well defined as the sup of all the convex lsc functions below f, or $-\infty$ if there is none. Observe also that it is the function whose epigraph is the closed convex envelope of epi f.

Proof: if $f \equiv +\infty$ then $f^* \equiv -\infty$ and $f^{**} \equiv +\infty$: the theorem is trivial. So we assume there exists x with $f(x) < +\infty$. If f has no affine minorant, then $f^* \equiv +\infty$ as for any p, $\langle p, x \rangle - f(x)$ cannot be bounded. Hence $f^{**} \equiv -\infty$.

If f has an affine minorant and is finite at some point, then for some p, $f(x) - \langle p, x \rangle$ is bounded from below: we deduce that $-\infty < f^*(p) = \sup_x \langle p, x \rangle - f(x) < +\infty$. It follows that $f^{**}(x) \geq \langle p, x \rangle - f^*(p)$ hence $f^{**} > -\infty$ everywhere.

Assuming that we are in this case, let g be convex, lsc with $g \leq f$. To show that f^{**} is maximal among such functions, we must show that $g \leq f^{**}$. Since $g \leq f$, then $f^* \leq g^*$, so that $g^{**} \leq f^{**}$. Hence it is enough to show that $g^{**} = g$. Moreover, since $f^{**}(x) \geq \langle p, x \rangle - f^*(p)$, we can consider only functions g with $g(x) \geq \langle p, x \rangle - f^*(p)$ $\forall x$ (otherwise replace g with $\max\{g(x), \langle p, x \rangle - f^*(p)\}$). So it is enough to show $g^{**} = g$ for convex lsc functions with an affine minorant. Hence, without loss of generality, we can assume that f is convex and lower-semicontinuous.

The next simplification consists in replacing f with $f'(x) = f(x) - \langle p, x \rangle + f^*(p) \geq 0$ where p is such that $-\infty < f^*(p) < +\infty$ ($\langle p, x \rangle - f^*(p)$ is an affine minorant). Indeed,

$$(f')^*(y) = \sup_x \langle y, x \rangle - f(x) + \langle p, x \rangle - f^*(p)$$
$$= -f^*(p) + \sup_y \langle y + p, x \rangle - f(x) = f^*(y + p) - f^*(p),$$

so that

$$(f')^{**}(x) = \sup_y \langle y, x \rangle - f^*(y + p) + f^*(p)$$
$$= f^*(p) - \langle p, x \rangle + \sup_y \langle y + p, x \rangle - f^*(y + p) = f^{**}(x) - \langle p, x \rangle + f^*(p).$$

Hence $f = f^{**} \Leftrightarrow f' = (f')^{**}$ and it is enough to show the result for nonnegative functions.

Assume therefore that f is convex, lsc, with $0 \leq f \not\equiv +\infty$. If $f^{**} \neq f$, then there exists x with $f^{**}(x) < f(x)$. That is, $(x, f^{**}(x)) \notin$ epi f and from Theorem 4.3, there exists p, λ, α with

$$\langle p, x \rangle - \lambda f^{**}(x) > \alpha \geq \langle p, y \rangle - \lambda s$$

for all $y \in \operatorname{dom} f$ and $s \geq f(y)$. In particular, as $\operatorname{dom} f \neq \emptyset$, letting $s \to +\infty$ we see that $\lambda \geq 0$.

Case 1: $\lambda > 0$: then we can divide the inequality and assume that $\lambda = 1$. It follows that $f^{**}(x) < -\alpha + \langle p, x\rangle$, while $\alpha \geq \langle p, y\rangle - f(y)$ for all y, hence taking the sup over y, $\alpha \geq f^*(p)$. Hence, $f^{**}(x) < \langle p, x\rangle - f^*(p)$, a contradiction. *Case 2:* $\lambda = 0$: then $\langle p, x\rangle > \alpha \geq \langle p, y\rangle$ for all $y \in \operatorname{dom} f$. Observe then that for $t > 0$, using that $f \geq 0$ in $\operatorname{dom} f$ and $f = +\infty$ outside,

$$f^*(tp) = \sup_y t\,\langle p, y\rangle - f(y) \leq t \sup_{y \in \operatorname{dom} f} \langle p, y\rangle \leq t\alpha,$$

so that

$$f^{**}(x) = \sup_q \langle q, x\rangle - f^*(q) \geq \sup_{t>0} \langle tp, x\rangle - f^*(tp) \geq \sup_{t>0} t(\langle p, x\rangle - \alpha) = +\infty$$

which is again a contradiction.

Remark 4.13. If x realizes the sup in $f^*(y) = \sup_x \langle y, x\rangle - f(x)$ then for all z,

$$\langle y, x\rangle - f(x) \geq \langle y, z\rangle - f(z) \;\Leftrightarrow\; f(z) \geq f(x) + \langle y, z - x\rangle$$

which means that $y \in \partial f(x)$. Conversely if $y \in \partial f(x)$, by definition one easily deduces that $f^*(y) \leq \langle y, x\rangle - f(x)$, and moreover that $f^{**}(x) = f(x)$, $y \in \partial f^{**}(x)$, and f is lsc at x. In particular we see that $\partial f^{**}(x) \supseteq \partial f(x)$ for all x.

One derives the celebrated *Legendre-Fenchel identity*:

$$y \in \partial f(x) \Leftrightarrow \langle x, y\rangle = f(x) + f^*(y) \Rightarrow x \in \partial f^*(y), \tag{22}$$

the latter being also an equivalence if f is lsc, convex (if $f = f^{**}$).

One also can check that conversely, if the convex function f is lsc at x, then $f^{**}(x) = f(x)$. This is true because f^{**} is the lsc envelope of f (since it is convex), which can be defined by $z \mapsto \inf_{z_n \to z} \liminf_n f(z_n)$.

4.2.2 Examples

1. $f(x) = \|x\|^2/(2\alpha)$, $\alpha > 0$: $f^*(y) = \alpha\|y\|^2/2$;
2. $f(x) = |x|^p/p$: $f^*(y) = |y|^{p'}/p'$, $1/p + 1/p' = 1$;
3. $F(f) = \|f\|^p_{L^p}/p$: $F^*(g) = \|g\|^{p'}_{L^{p'}}/p'$ (the duality is in L^2, however this is also true in the $(L^p, L^{p'})$ duality, see [13]);
4. $f(x) = \delta_{B(0,1)}(x) = 0$ if $x \in B(0,1)$, $+\infty$ else: $f^*(p) = |p|$.

The last example is a particular case of the following situation: if f is convex, 1-homogeneous, then

$$f^*(y) = \sup_x \langle y, x\rangle - f(x) = \sup_{t>0} \sup_x \langle y, tx\rangle - f(tx) = \sup_{t>0} t f^*(y) \in \{0, +\infty\}$$

and precisely

$$f^*(y) = \begin{cases} 0 & \text{if } \langle y, x\rangle \leq f(x)\ \forall\, x \in \mathcal{X}, \\ +\infty & \text{if } \exists\, x \in \mathcal{X},\ \langle y, x\rangle > f(x). \end{cases}$$

Letting $C = \{y : \langle y, x\rangle \leq f(x)\ \forall\, x \in \mathcal{X}\} = \partial f(0)$, one has $f^* = \delta_C$ (C is clearly closed and convex, and f^* convex lsc). Eventually, observe that if f is lsc, then $f^{**} = f$ which shows that in this case

$$f(x) = \sup_{y \in \partial f(0)} \langle y, x\rangle .$$

Observe in particular that $\partial f(x) = \{y \in \partial f(0) : \langle y, x\rangle = f(x)\}$.

This example, in turn, is a particular case of the following: if f is β-homogeneous, $\beta > 1$, then

$$f^*(ty) = \sup_x \langle ty, x\rangle - f(x) = t^\alpha \sup_x \left\langle y, t^{1-\alpha}x\right\rangle - f(t^{-\alpha/\beta}x) = t^\alpha f^*(y)$$

if $1 - \alpha = -\alpha/\beta$, hence if $1/\alpha + 1/\beta = 1$.

4.2.3 Relationship between the growth of f and f^*

Lemma 4.14. *If f is finite everywhere, then $f^*(tp)/t \to +\infty$ as $t \to +\infty$ for all $p \in \mathcal{X}$ (f^* is superlinear). The converse is true in finite dimension if f is convex, lsc.*

Proof: if f^* is not superlinear, there exists p, $c < \infty$, such that $f^*(tp) \le ct$ for all $t > 0$: hence $f^{**}(x) \ge \sup_{t\ge0} t\langle p, x\rangle - f^*(tp) \ge \sup_{t\ge0} t(\langle p, x\rangle - c) = +\infty$ as soon as x is such that $\langle p, x\rangle > c$. Of course then, $f(x) \ge f^{**}(x) = +\infty$.

Conversely, in finite dimension, let f be convex, lsc and assume that there is x with $f(x) = +\infty$. We can assume without loss of generality that $f \ge 0$ (*cf* proof of Thm 4.12).

Then, since $\overline{\mathrm{dom}\, f} \neq \mathcal{X}$ (in finite dimension only, in infinite dimension $\mathrm{dom}\, f$ could be dense, for instance think of $f(u) = \int |\nabla u|^2 dx$ for $u \in L^2$) one can consider $x \notin \overline{\mathrm{dom}\, f}$. Then, there exists by Theorem 4.3 p, α with $\langle p, x\rangle > \alpha \ge \langle p, y\rangle$ $\forall y \in \mathrm{dom}\, f$. We have

$$f^*(tp) = \sup_y \langle tp, y\rangle - f^*(y) \le \sup_{y\in\mathrm{dom}\, f} t\langle p, y\rangle \le t\alpha$$

for $t > 0$, so that $f^*(tp)/t \le \alpha$ and f^* is not superlinear.

Remark 4.15. *In infinite dimension, one needs to strengthen a bit the assumption, for instance if $f \ge g(|p|)$ with g superlinear then f^* is finite everywhere.*

Proposition 4.16. *Let f a convex, lsc function: then f is μ-convex if and only if f^* has $(1/\mu)$-Lipschitz gradient.*

Proof: observe that if f is μ-convex one has in particular, given $x \in \mathrm{dom}\, \partial f$, for $p \in \partial f(x)$, that (18) holds:

$$f(y) \ge f(x) + \langle p, y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \tag{18}$$

for all y, hence taking the conjugate (*cf* Example 1 in the previous Section), we find for all q:

$$\begin{aligned}
f^*(q) &\le \sup_y \langle q, y\rangle - f(x) - \langle p, y - x\rangle - \frac{\mu}{2}\|y - x\|^2 \\
&= \langle q, x\rangle - f(x) + \sup_y \langle q - p, y - x\rangle - \frac{\mu}{2}\|y - x\|^2 = \langle q, x\rangle - f(x) + \frac{1}{2\mu}\|q - p\|^2 \\
&= \langle p, x\rangle - f(x) + \langle q - p, x\rangle + \frac{1}{2\mu}\|q - p\|^2 = f^*(p) + \langle q - p, x\rangle + \frac{1}{2\mu}\|q - p\|^2. \qquad (23)
\end{aligned}$$

We have used that $\langle p, x\rangle - f(x) = f^*(p)$ which follows from (22). In particular we see that f^* has at most a quadratic growth, and we deduce that it is locally Lipschitz (Lemma 4.1), and its subgradient is not empty everywhere. Moreover, we deduce from (23) that when $x \in \partial f^*(p) \Leftrightarrow p \in \partial f(x)$ (*cf* (22)),

$$f^*(p) + \langle q - p, x\rangle \le f^*(q) \le f^*(p) + \langle q - p, x\rangle + \frac{1}{2\mu}\|q - p\|^2,$$

in other words, $f^*(q) = f^*(p) + \langle q - p, x\rangle + o(q - p)$ which shows that f^* is (Fréchet)-differentiable and $x = \nabla f^*(p)$.

Eventually, given $p, q \in \mathcal{X}$ and $x = \nabla f^*(p)$, $y = \nabla f^*(q)$, one has by (22) that $p \in \partial f(x)$, $q \in \partial f(y)$ and by strong convexity, using (18) and the same with x, y switched and p replaced with q, and summing, we find

$$\langle q - p, y - x \rangle \geq \mu \|y - x\|^2$$

so that in particular, $\|\nabla f^*(q) - \nabla f^*(p)\| \leq (1/\mu)\|q - p\|$: ∇f^* is $(1/\mu)$-Lipschitz. In fact, we see that

$$\langle q - p, \nabla f^*(q) - \nabla f^*(p) \rangle \geq \mu \|\nabla f^*(q) - \nabla f^*(p)\|^2,$$

which expresses that ∇f^* is "μ-co-coercive", a property which is stronger than being $(1/\mu)$-Lipschitz.

Conversely, if f^* has $(1/\mu)$-Lipschitz gradient, let us show that f is μ-convex. Observe that

$$\begin{aligned} f^*(q) = f^*(p) + \int_0^1 \langle \nabla f^*(p + s(q-p)), q - p \rangle \, ds \\ = f^*(p) + \langle \nabla f^*(p), q - p \rangle + \int_0^1 \langle \nabla f^*(p + s(q-p)) - \nabla f^*(p), q - p \rangle \, ds \\ \leq f^*(p) + \langle \nabla f^*(p), q - p \rangle + \tfrac{1}{\mu}\|q - p\|^2 \int_0^1 s ds. \end{aligned}$$

If $p \in \partial f(x)$, so that $x = \nabla f^*(p)$, we deduce

$$f^*(q) \leq f^*(p) + \langle q - p, x \rangle + \frac{1}{2\mu}\|q - p\|^2.$$

Hence taking the conjugate:

$$\begin{aligned} f(y) = f^{**}(y) \geq \sup_q \langle q, y \rangle - \left(f^*(p) + \langle q - p, x \rangle + \frac{1}{2\mu}\|q - p\|^2 \right) \\ = \langle p, x \rangle - f^*(p) + \sup_q \langle p - q, x - y \rangle - \frac{1}{2\mu}\|q - p\|^2 = \langle p, y \rangle - f^*(p) + \frac{\mu}{2}\|x - y\|^2. \end{aligned}$$

By (22), $\langle p, x \rangle - f^*(p) = f(x)$ (as f is convex lsc), and we find

$$f(y) \geq f(x) + \langle p, y - x \rangle + \frac{\mu}{2}\|x - y\|^2,$$

showing that f is strongly convex. Notice in particular that we have found another proof of Theorem 2.3, valid also in Hilbert spaces for convex lsc functions.

4.3 Proximity operator

(Or *Proximal map.*) Given f convex lsc, proper, observe that for any $\tau > 0$, $x \in \mathcal{X}$, $y \mapsto f(y) + \|y - x\|^2/(2\tau)$ is strongly convex and hence has a unique minimizer. We define

$$f_\tau(x) := \min_{y \in \mathcal{X}} f(y) + \frac{1}{2\tau}\|y - x\|^2, \tag{24}$$

which is a convex function. As we have seen before (Lemma 4.8), one has at the minimizer y_x

$$\partial f(y_x) + \frac{1}{\tau}(y_x - x) \ni 0. \tag{25}$$

This characterizes the unique minimizer of (24) and in particular

$$y_x = (I + \tau\partial f)^{-1}(x) =: \mathrm{prox}_{\tau f}(x)$$

is uniquely defined.

As already shown, $(x - y_x)/\tau = \nabla f_\tau(x)$. Actually, in the convex case, there is a direct proof: one has, letting $\eta = \mathrm{prox}_{\tau f}(y)$ and $\xi = \mathrm{prox}_{\tau f}(x)$,

$$\begin{aligned} f_\tau(y) &= f(\eta) + \tfrac{\|\eta-y\|^2}{2\tau} = f(\eta) + \tfrac{\|(\eta-x)+(x-y)\|^2}{2\tau} \\ &= f(\eta) + \tfrac{\|\eta-x\|^2}{2\tau} + \left\langle \tfrac{x-\eta}{\tau}, y-x\right\rangle + \tfrac{\|x-y\|^2}{2\tau} \\ &\geq f(\xi) + \tfrac{\|\xi-x\|^2}{2\tau} + \tfrac{\|\eta-\xi\|^2}{2\tau} + \left\langle \tfrac{x-\xi}{\tau}, y-x\right\rangle + \left\langle \tfrac{\xi-\eta}{\tau}, y-x\right\rangle + \tfrac{\|x-y\|^2}{2\tau} \\ &= f_\tau(x) + \left\langle \tfrac{x-\xi}{\tau}, y-x\right\rangle + \tfrac{\tau}{2}\|\tfrac{y-\eta}{\tau} - \tfrac{x-\xi}{\tau}\|^2. \end{aligned}$$

In the third line, we have used the fact that ξ is the minimiser of a $(1/\tau)$-strongly convex problem, so that $f(\eta) + \|\eta - x\|^2/(2\tau) \geq f(\xi) + \|\eta - x\|^2/(2\tau) + \|\eta - \xi\|^2/(2\tau)$ for all η. We deduce from the inequality

$$f_\tau(y) \geq f_\tau(x) + \left\langle \tfrac{x-\xi}{\tau}, y-x\right\rangle + \tfrac{\tau}{2}\|\tfrac{y-\eta}{\tau} - \tfrac{x-\xi}{\tau}\|^2$$

both that $(x-\xi)/\tau$ is a subgradient of f_τ at x, and that the map $x \mapsto (x - \mathrm{prox}_{\tau f}(x))/\tau$ is τ-co-coercive, hence $(1/\tau)$-Lipschitz: indeed, writing the same inequality after having swapped x and y, and summing the two inequalities, we obtain

$$\left\langle \tfrac{y-\eta}{\tau} - \tfrac{x-\xi}{\tau}, y-x\right\rangle \geq \tau\|\tfrac{y-\eta}{\tau} - \tfrac{x-\xi}{\tau}\|^2.$$

In particular, f_τ is C^1. Also, we find that

$$\mathrm{prox}_{\tau f}(x) = x - \tau\nabla f_\tau(x)$$

is a $(1/2)$-averaged operator (it is $(1/2)I + (1/2)(x - 2\tau\nabla f_\tau(x))$, see Lemma 2.4).

Thanks to (22), (25) yields

$$y_x \in \partial f^*(\tfrac{x-y_x}{\tau}) \Leftrightarrow \tfrac{x-y_x}{\tau} + \tfrac{1}{\tau}\partial f^*(\tfrac{x-y_x}{\tau}) \ni \tfrac{x}{\tau} \Leftrightarrow \tfrac{x-y_x}{\tau} = (I + \tfrac{1}{\tau}\partial f^*)^{-1}(\tfrac{x}{\tau}).$$

We deduce the *Moreau Identity*

$$x = \mathrm{prox}_{\tau f}(x) + \tau\mathrm{prox}_{\frac{1}{\tau}f^*}(\tfrac{x}{\tau}) \tag{26}$$

One also can show the following:

Proposition 4.17. *Let f be proper, convex, lsc: then* $\mathrm{dom}\,\partial f$ *is dense in* $\mathrm{dom}\,F$.

Indeed, let $x \in \mathrm{dom}\, f$: then $f_\tau(x) \leq f(x)$. In particular, denoting $x_\tau = \mathrm{prox}_{\tau f}(x)$,

$$f_\tau(x) = f(x_\tau) + \frac{1}{2\tau}\|x - x_\tau\|^2 \leq f(x).$$

We use again that f, being proper, is larger than some affine function: hence there is p, c such that $\langle p, x_\tau\rangle + c + \frac{1}{2\tau}\|x - x_\tau\|^2 \leq f(x)$ from which it follows that $\|x_\tau - x\| \leq c'\sqrt{\tau}$ for some constant $c' > 0$. Hence $x_\tau \to x$. Now, $\partial f(x_\tau) \ni (x - x_\tau)/\tau \neq \emptyset$ hence $x_\tau \in \mathrm{dom}\, f$, which shows the proposition. As a by-product of the proof, one sees that:

Proposition 4.18. *Let f be proper, lsc, convex and f_τ defined by* (24)*. Then for all x, $f_\tau(x) \to f(x)$ as $\tau \to 0$.*

(We leave to the reader the proof that if $f(x) = +\infty$, $f_\tau(x) \to +\infty$, which is easy using that f is lsc.)

Examples: $f(x) = \|x\|_1 = \sum_i |x_i|$, $x \in \mathbb{R}^d$:

$$\operatorname{prox}_{\tau f}(x) = ((|x_i| - \tau)^+ \operatorname{sign}(x_i))_{i=1}^d.$$

If $f(x) = \delta_{|x_i| \le 1}$, $\operatorname{prox}_{\tau f}(x) = (\max\{-1, \min\{1, x_i\}\})_{i=1}^d$.
If $f(x) = \|x\|^2/2$, $\operatorname{prox}_{\tau f}(x) = x/(1+\tau)$.

4.3.1 Inf-convolution

Formula (24) is a particular case of an *inf-convolution*: letting f, g be convex, lsc functions we define

$$f \Box g(x) = \inf_y f(x-y) + g(y), \tag{27}$$

which is convex but not necessarily lsc. One can show the following:

Lemma 4.19. *If there is $p \in \mathcal{X}$ where f^* is continuous and g^* is finite, then the inf is reached in* (27) *and $f \Box g$ is convex, lsc. In finite dimension, it is enough to have $p \in \operatorname{ri} \operatorname{dom} f^* \cap \operatorname{ri} \operatorname{dom} g^*$.*

Proof: consider indeed $x_n \to x$ and y_n such that

$$f \Box g(x_n) \ge f(x_n - y_n) + g(y_n) - \frac{1}{n}.$$

Consider a subsequence with

$$\lim_k f(x_{n_k} - y_{n_k}) + g(y_{n_k}) = \liminf_n f(x_n - y_n) + g(y_n) \le \liminf_n f \Box g(x_n)$$

Observe that if f^* is continuous at p, then it means that there is a constant c such that

$$f^*(q) \le c + \delta_{B(0,\varepsilon)}(q-p)$$

(where δ_C is the characteristic function of C which is zero in C and $+\infty$ elsewhere) while $g^*(p) < +\infty$: so that for all z

$$f(z) = f^{**}(z) \ge \langle p, z \rangle - c + \varepsilon \|z\|, \quad g(z) \ge \langle p, z \rangle - g^*(p).$$

Hence,

$$\begin{aligned} f(x_{n_k} - y_{n_k}) + g(y_{n_k}) &\ge \langle p, x_{n_k} - y_{n_k} \rangle - c + \varepsilon \|x_{n_k} - y_{n_k}\| + \langle p, y_{n_k} \rangle - g^*(p) \\ &= \langle p, x_{n_k} \rangle + \varepsilon \|x_{n_k} - y_{n_k}\| - (c + g^*(p)) \end{aligned}$$

so that $(x_{n_k} - y_{n_k})_k$ is a bounded sequence, hence there exists y and a subsequence of (y_{n_k}) (not relabelled) with $y_{n_k} \rightharpoonup y$. In the limit (as, f, g are weakly lsc),

$$f \Box g(x) \le \liminf_k f(x_{n_k} - y_{n_k}) + g(y_{n_k}) \le \liminf_n f \Box g(x_n).$$

Eventually, we observe that if the sequence $x_n \equiv x$, then this proves that there is a minimizer y in (27). We can derive a second, more precise variant of Theorem 4.9:

Corollary 4.20. *Let f, g be convex, lsc: if there exists $x \in \operatorname{dom} f \cap \operatorname{dom} g$ such that f is continuous at x (in finite dimension, $x \in \operatorname{ri} \operatorname{dom} f \cap \operatorname{ri} \operatorname{dom} g$), then*

- $(f+g)^* = f^* \Box g^*$,
- $\partial(f+g) = \partial f + \partial g$.

The first point is clear: as by our assumption, $f^*\Box g^*$ is lsc, and:

$$
\begin{aligned}
(f^*\Box g^*)^*(x) &= \sup_{p,q} \langle x, p\rangle - f^*(q) - g^*(p-q) \\
&= \sup_{p,q} \langle x, q\rangle - f^*(q) + \langle x, p-q\rangle - g^*(p-q) = f(x) + g(x).
\end{aligned}
$$

The second point is because if $p \in \partial(f+g)(x)$, using that $x \in \partial(f^*\Box g^*)(p)$ and

$$f^*\Box g^*(p) = f^*(q) + g^*(p-q)$$

for some q, one obtains letting $p - q = r$:

$$
\begin{aligned}
f^*(s) + g^*(t) \geq f^*\Box g^*(s+t) &\geq f^*\Box g^*(p) + \langle x, s+t-p\rangle \\
&\geq f^*(q) + \langle x, s-q\rangle + g^*(r) + \langle x, t-r\rangle
\end{aligned}
$$

for all s, t. Hence $x \in \partial f^*(q) \cap \partial g^*(r)$, which shows that $p = q + r \in \partial f(x) + \partial g(x)$.

4.3.2 A useful variant

Consider now the modified inf-convolution problem

$$h(x) = \inf_{y \in \mathcal{Y}} f(x - Ky) + g(y)$$

where $K : \mathcal{Y} \to \mathcal{X}$ is a continuous operator and f, g are convex, lsc, proper. Then one can show similarly that if there exists p such that $f^*(p) < +\infty$ and g^* is continuous at K^*p, h is lsc and since

$$
\begin{aligned}
h^*(q) &= \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \langle q, x\rangle - f(x - Ky) - g(y) \\
&= \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \langle q, x - Ky\rangle + \langle K^*q, y\rangle - f(x - Ky) - g(y) = f^*(q) + g^*(K^*q)
\end{aligned}
$$

it follows that $h = [f^*(\cdot) + g^*(K^*\cdot)]^*$.

The proof is exactly as the proof of Lemma 4.19, but now one uses that $g^* \leq a + \delta_{B(K^*p,\varepsilon)}$ for some $a \in \mathbb{R}$ and $\varepsilon > 0$, so that $g(y) \geq -a + \langle p, Ky\rangle + \varepsilon\|y\|$ and $f^*(p) \in \mathbb{R}$ so that $f(x) \geq \langle p, x\rangle - f^*(p)$.

Then, if $x_n \to x$ and y_n is such that $f(x_n - Ky_n) + g(y_n) \leq h(x_n) + 1/n$, and if $\liminf_n h(x_n) < +\infty$, one find that along a subsequence $\|y_n\|$ is bounded, hence we may assume it converges weakly to some y (and as a consequence Ky_n converges weakly to Ky). Hence

$$h(x) \leq f(x - Ky) + g(y) \leq \liminf_n f(x_n - Ky_n) + g(y_n) \leq \liminf_n h(x_n)$$

and the semicontinuity follows. In addition, we deduce that the “inf” is in fact a “min”.

A useful application is the following: let g be convex, lsc and proper and K a continuous operator, and define

$$g^K(x) := \inf_{y: Ky = x} g(y).$$

Then, if there exists p where g^* is continuous at K^*p, g^K is lsc and $g^K = [g^*(K^*\cdot)]^*$. It is enough to apply the previous result with $f = \delta_{\{0\}}$, so that $f^* \equiv 0$ and $p \in \operatorname{dom} f^*$.

4.3.3 Fenchel-Rockafellar duality

Consider now a minimization problem of the form

$$\min_{x\in\mathcal{X}} f(Kx) + g(x) \tag{28}$$

where $K:\mathcal{X}\to\mathcal{Y}$ is a continuous linear map and f,g are convex, lsc. Then, clearly

$$\begin{aligned}(\mathcal{P}) = \min_x f(Kx)+g(x) &= \min_x \sup_y \langle y, Kx\rangle - f^*(y) + g(x)\\ &\geq \sup_y \inf_x \langle K^*y, x\rangle + g(x) - f^*(y) = \sup_y -\left(g^*(-K^*y)+f^*(y)\right) = (\mathcal{D})\end{aligned}$$

A natural question is when there is equality: this is true under various criteria: we will give a simple example below.

The problem "$(\mathcal{P})$" is usually called the *primal problem* and "$(\mathcal{D})$" the *dual problem* (observe though that there is a symmetry between these problems...) Notice that the *primal-dual gap*

$$\mathcal{G}(x,y) = f(Kx) + g(x) + g^*(-K^*y) + f^*(y)$$

is a measure of optimality. If it vanishes at (x^*,y^*), then $(\mathcal{P}) = (\mathcal{D})$, and (x^*,y^*) is a *saddle point* of the *Lagrangian*

$$\mathcal{L}(x,y) = \langle y, Kx\rangle - f^*(y) + g(x), \tag{29}$$

as one has

$$\mathcal{L}(x^*,y) \leq \mathcal{L}(x^*,y^*) \leq \mathcal{L}(x,y^*) \tag{30}$$

for all $x\in\mathcal{X}$, $y\in\mathcal{Y}$. [Indeed, for all y,x, $\mathcal{L}(x^*,y)\leq f(Kx^*)+g(x^*) = -f^*(y^*)-g^*(-K^*y^*)\leq\mathcal{L}(x,y^*)$.]

Theorem 4.21. *If there exists $\bar{x}\in\operatorname{dom} g$ with f continuous at $K\bar{x}$, then $(\mathcal{P}) = (\mathcal{D})$. Moreover under these assumptions, $(\mathcal{D})$ has a solution.*

We show the result following a classical approach, see [13, (4.21)] for more general variants. In finite dimension, it is shown in [32, Cor 31.2.1] that equality holds if there exists $x\in\operatorname{ri}\operatorname{dom} g$ with $Kx\in\operatorname{ri}\operatorname{dom} f$, or even more generally that $0\in\operatorname{ri}(\operatorname{dom} f - K\operatorname{dom} g)$ (the proof works as below).

Proof: the method is called the "perturbation method": We introduce, for $z\in\mathcal{Y}$,

$$\Phi(z) := \inf_{x\in\mathcal{X}} f(Kx+z)+g(x).$$

Assume $\Phi(0) > -\infty$ (otherwise there is nothing to prove), then by assumption, one can find M and ε such that for $|z|<\varepsilon$, $\Phi(z)\leq f(K\bar{x}+z)+g(\bar{x})\leq M<+\infty$. Being Φ convex, we deduce that it is locally Lipschitz and in particular, $\Phi(0)=\Phi^{**}(0)=\sup_y -\Phi^*(y)$. We compute:

$$\begin{aligned}\Phi^*(y) &= \sup_{z\in\mathcal{Y}} \langle y,z\rangle - \inf_{x\in\mathcal{X}}(f(Kx+z)+g(x))\\ &= \sup_{x,z} \langle y, z+Kx\rangle - \langle K^*y, x\rangle - f(Kx+z) - g(x) = f^*(y)+g^*(-K^*y).\end{aligned}$$

The claim follows. Moreover, since Φ is Lipschitz near 0 it is also subdifferentiable: there exists $y\in\partial\Phi(0)$. This subdifferential provides a solution to the "dual" problem $\max_y -\Phi^*(y)$.
Exercise: show the result in finite dimension if $0\in\operatorname{ri}(\operatorname{dom} f - K\operatorname{dom} g)$ (one needs to show again that Φ is lsc at 0).

Observe that one has by optimality in (30) that $Kx^* - \partial f^*(y^*) \ni 0$, $K^*y^* + \partial g(x^*) \ni 0$, which may be written

$$0 \in \begin{pmatrix} \partial g(x) \\ \partial f^*(y) \end{pmatrix} + \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{31}$$

meaning the solution is found by finding the "zero" of the sum of two *monotone operators* (see Section 4.4).

Example Consider the problem

$$\min_x \lambda \|Dx\|_1 + \frac{1}{2}\|x - x^0\|^2$$

where $D : \mathbb{R}^n \to \mathbb{R}^m$ is a continuous operator, $x^0 \in \mathbb{R}^n$, $\|\cdot\|_1$ is the ℓ^1-norm. One has

$$f = \lambda\|\cdot\|_1, \quad K = D, \quad g = \frac{1}{2}\|\cdot - x^0\|^2.$$

Then the Lagrangian is

$$\mathcal{L}(x, y) = \langle y, Dx\rangle - f^*(y) + g(x)$$

where $f^*(y) = 0$ if $|y_i| \leq \lambda$ for $i = 1, \ldots, n$, and $+\infty$ else. To find the dual problem, we compute $g^*(z) = \left\langle z, x^0 \right\rangle + \|z\|^2/2$, and we obtain

$$\max \left\{ \left\langle D^*y, x^0 \right\rangle - \frac{1}{2}\|D^*y\|^2 : |y_i| \leq \lambda, i = 1, \ldots, n \right\}.$$

This can be rewritten as a projection problem:

$$\min_{|y_i| \leq \lambda} \|D^*y - x^0\|^2.$$

4.4 Elements of monotone operators theory

For more results, see [6]. We mostly mention the main properties, which extend the properties shown so far for subgradients.

Observe that if f is convex, one has for all x, y, $p \in \partial f(x)$, $q \in \partial f(y)$

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

so that, summing,

$$\langle p - q, x - y\rangle \geq 0.$$

This leads to introduce the class of operators which satisfy such an inequality, which share many properties with subgradients. Consider in the Hilbert space $\mathcal{X}$ a multi-valued operator $A : \mathcal{X} \to \mathcal{P}(\mathcal{X})$. By a slight abuse of notation, we will also denote A the graph $\{(x, y) : x \in \mathcal{X}, y \in Ax\}$.

We introduce the following definitions:

Definition 1. *The operator A is said monotone if for all $x, y \in \mathcal{X}$, $p \in Ax$, $q \in Ay$,*

$$\langle p - q, x - y\rangle \geq 0.$$

It is (μ-)strongly monotone if

$$\langle p - q, x - y\rangle \geq \mu\|x - y\|^2.$$

It is (μ-)co-coercive if

$$\langle p - q, x - y\rangle \geq \mu\|p - q\|^2.$$

It is maximal if the graph $\{(x, p) : p \in Ax\} \subset \mathcal{X} \times \mathcal{X}$ is maximal with respect to inclusion, among all the graphs of monotone operators.

In dimension 1, monotone graphs are graphs of nondecreasing functions.

One sees that the subgradient of a convex function f is monotone, strongly monotone if f is strongly convex, co-coercive if ∇f is Lipschitz (*cf* Theorem 2.3).

A subgradient is maximal if and only it is the subgradient of a lower-semicontinuous function. A simple proof is due to Rockafellar: if f is lsc, to show that ∂f is maximal we must show that if $x \in \mathcal{X}$ and $p \notin \partial f(x)$ then one can find $y, q \in \partial f(y)$ with $\langle p - q, x - y\rangle < 0$. Replacing f with $f(x) - \langle p, x\rangle$ we can assume that $p = 0$. Saying that $0 \notin \partial f(x)$ is precisely saying that x is not a minimizer, that is, there exists $y \in \mathcal{X}$ with $f(y) < f(x)$.

Consider now $y = \operatorname{prox}_f(x)$, the minimizer of $f(y) + \|y - x\|^2/2$. As we have seen, $q = x - y \in \partial f(y)$. One has

$$\langle p - q, x - y\rangle = \langle -q, x - y\rangle = -\|x - y\|^2 < 0,$$

unless $y = x$. But $y = x$ would imply that $q = 0 \in \partial f(x)$, a contradiction. Hence ∂f is maximal. (The proof can be extended to non-Hilbert spaces, see [33].)

Conversely if ∂f is maximal, since $\partial f^{**} \supset \partial f$, then this operator is also the subgradient of the convex, lsc function f^{**}. We are *not* proving here that $f = f^{**}$, only that ∂f is *also* the subgradient of the convex, lsc function f^{**}.

A monotone operator is not necessarily a subgradient: for instance, in $\mathbb{R}^2$, the linear operator

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

is monotone but not the subgradient of a convex function. In order for a monotone operator to be (included in) the subgradient of a convex function, it needs to be *cyclically monotone* [31]: for any $x_0, x_1 \dots, x_n = x_0$ and $p_i \in Ax_i$, $p_0 = p_n$,

$$\sum_{i=0}^{n-1} \langle p_i, x_{i+1} - x_i\rangle \leq 0.$$

An important case of monotone operator is obtained from nonexpansive (1-Lipschitz mappings) T, as in Section 3. Indeed, it is obvious to check that $I - T$ is maximal monotone:

$$\langle (x - Tx) - (x - Ty), x - y\rangle = \|x - y\|^2 - \langle Tx - Ty, x - y\rangle \geq 0$$

thanks to Cauchy-Schwartz inequality and the fact T is 1-Lipschitz.

Given A a monotone operator, its inverse is simply $A^{-1} : p \mapsto \{x : Ax \ni p\}$, with graph $\{(p, x) : p \in Ax\}$. It is therefore maximal if A is maximal, co-coercive if A is strongly monotone (*cf* Prop. 4.16). Clearly, $(\partial f)^{-1} = \partial f^*$ (see (22)).

Theorem 4.22 (Minty [20])**.** *The resolvent of a maximal-monotone operator A, defined by*

$$x \mapsto y = (I + A)^{-1}x =: J_A x \Leftrightarrow y + Ay \ni x$$

is a well (everywhere) defined single-valued nonexpansive mapping. (Conversely, for a monotone operator A if $(I + A)$ is surjective then A is maximal.)

One will see that the resolvent is also a (1/2)-averaged operator (and any (1/2)-averaged operator has this form).

Proof: Let us introduce the graph $G = \{(y + x, y - x) : x \in \mathcal{X}, y \in Ax\}$. If (a, b), $(a', b') \in G$, with $a = y + x, b = y - x$ and $a' = y' + x', b = y' - x'$, then

$$\begin{aligned}\|b - b'\|^2 = \|y - y'\|^2 - 2\langle y - y', x - x'\rangle + \|y + y'\|^2 \\ = \|a - a'\|^2 - 4\langle y - y', x - x'\rangle \leq \|a - a'\|^2\end{aligned}$$

showing that G is the graph of a 1-Lipschitz function. Moreover, if $G' \supseteq G$ is also the graph of a 1-Lipschitz function, then defining $A' = \{((a-b)/2, (a+b)/2) : (a,b) \in G'\}$ the same computation shows that $A' \supseteq A$ is the graph of a monotone operator, hence $A' = A$ if A is maximal. (Conversely, if G is defined for all a then clearly G and therefore A are maximal, as being 1-Lipschitz G is necessarily single-valued.)

So the theorem is equivalent to the question whether a 1-Lipschitz function which is not defined in the whole of $\mathcal{X}$ can be extended. This result (which is true only in Hilbert spaces) is known as Kirszbraun-Valentine's theorem (1935). [The proof we give is derived from [15, 2.10.43].]

The basic brick is the following extension from n to $n+1$ points:

Lemma 4.23. *If $(x_i)_{i=1}^n$, $(y_i)_{i=1}^n$ are points in Hilbert spaces respectively $\mathcal{X}, \mathcal{Y}$ such that $\forall i,j$, $\|y_i - y_j\| \leq \|x_i - x_j\|$, then for any $x \in \mathcal{X}$ there exists $y \in \mathcal{Y}$ with $\|y_i - y\| \leq \|x_i - x\|$ for all $i = 1, \dots, n$.*

It is enough to prove this for $x = 0$: we need to find a common point to $\bar{B}(y_i, \|x_i\|)$. There is nothing to prove if $x = x_i$ for some i, so we assume $x_i \neq 0$, $i = 1, \dots, n$. We define

$$\bar{c} = \min \left\{ c \geq 0 : \bigcap_{i=1}^n \bar{B}(y_i, c\|x_i\|) \neq \emptyset \right\} > 0$$

(if the y_i are distinct, which we may also assume). This is a min because the closed balls are weakly compact, and we can consider y such that $\|y - y_i\| \leq \bar{c}\|x_i\|$, $i = 1, \dots, n$. Then we observe that y must be a convex combination of the points $(y_i)_{i \in I}$ such that $\|y - y_i\| = \bar{c}\|x_i\|$. Indeed, if not, let y' be the projection of y onto $\overline{\mathrm{co}}\,\{y_i : i \in I\}$. As for any $i \in I$, $\langle y_i - y', y - y' \rangle \leq 0$ one has, letting $y_t = (1-t)y + ty'$, that for any $i \in I$:

$$\begin{aligned} \|y_i - y_t\|^2 = \|y_i - y + t(y - y')\|^2 &= \|y_i - y\|^2 + 2t \langle y_i - y, y - y' \rangle + t^2 \|y - y'\|^2 \\ &= \|y_i - y\|^2 + 2t \langle y_i - y', y - y' \rangle - 2t\|y - y'\|^2 + t^2 \|y - y'\|^2 \\ &\leq \|y_i - y\|^2 - t(2-t)\|y - y'\|^2 < \|y_i - y\|^2 \end{aligned}$$

if $t \in (0,2)$. Hence if $t > 0$ is small enough, one sees that $\|y_i - y_t\| < \|y_i - y\| = \bar{c}\|x_i\|$ for $i \in I$, while since for $i \notin I$, $\|y_i - y\| < \bar{c}\|x_i\|$, one can still guarantee the same strict inequality for y_t if t is small enough. But this contradicts the definition of $\bar{c}$, since then there would exists $c < \bar{c}$ such that $y_t \in \bigcap_{i=1}^n \bar{B}(y_i, c\|x_i\|)$.

We therefore can write $y = \sum_{i \in I} \theta_i y_i$ as a convex combination ($\theta_i \in [0,1], \sum_{i \in I} \theta_i = 1$). Then since $2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$,

$$\begin{aligned} 0 = \|\sum_{i \in I} \theta_i y_i - y\|^2 = \sum_{i,j \in I} \theta_i \theta_j \langle y_i - y, y_j - y \rangle \\ &= \frac{1}{2} \sum_{i,j \in I} \theta_i \theta_j \left(\|y_i - y\|^2 + \|y_j - y\|^2 - \|y_i - y_j\|^2 \right) \\ &\geq \frac{1}{2} \sum_{i,j \in I} \theta_i \theta_j \left(\bar{c}^2 \|x_i\|^2 + \bar{c}^2 \|x_j\|^2 - \|x_i - x_j\|^2 \right) \\ &= \bar{c}^2 \sum_{i,j \in I} \theta_i \theta_j \langle x_i, x_j \rangle - \frac{1 - \bar{c}^2}{2} \|x_i - x_j\|^2 \end{aligned}$$

which shows that

$$(1 - \bar{c}^2) \sum_{i,j \in I} \theta_i \theta_j \|x_i - x_j\|^2 \geq 2\bar{c}^2 \|\sum_{i \in I} \theta_i x_i\|^2$$

so that $\bar{c} \leq 1$. Hence, y satisfies $\|y - y_i\| \leq \|x_i\|$, as requested, which shows Lemma 4.23.

We can conclude the proof of Theorem 4.22: if there exists $x \in \mathcal{X}$ such that $\{x\} \times \mathcal{X} \cap G = \emptyset$, consider the set

$$K = \bigcap_{(a,b)\in G} \bar{B}(b, \|x-a\|)$$

which is an intersection of weakly compact sets.

We show that because the compact sets defining K have the "*finite intersection property*", K can not be empty: Choosing $(a_0, b_0) \in G$, if $\bar{B}_0 = \bar{B}(b_0, \|x - b_0\|)$, we see that

$$K = \bar{B}_0 \cap \left(\bigcap_{(a,b)\in G} \bar{B}(b, \|x-a\|) \right)$$

hence $\bar{B}_0 \setminus K = \bar{B}_0 \cap \bigcup_{(a,b)\in G} \bar{B}(b, \|x-a\|)^c$. If this is $\bar{B}_0$, by compactness one can extract a finite covering $\bigcup_{i=1}^n \bar{B}(b_i, \|x - a_i\|)^c$ for $(a_i, b_i) \in G$, $i = 1, \dots, n$. We find that

$$\bar{B}_0 \cap \bigcup_{i=1}^n \bar{B}(b_i, \|x-a_i\|)^c = \bar{B}_0$$

or equivalently that

$$\bar{B}_0 \cap \bigcap_{i=1}^n \bar{B}(b_i, \|x-a_i\|) = \emptyset$$

which contradicts Lemma 4.23. Hence, $\bar{B}_0 \setminus K \neq \bar{B}_0$ which means that $K \neq \emptyset$. Choosing $y \in K$, we find that $G \cup \{(x,y)\}$ is the graph of a 1-Lipschitz function and is strictly larger than G, which contradicts the maximality of A.

The non-expansiveness of $(I + A)^{-1}$ follows from, if $y + Ay \ni x$, $y' + Ay' \ni x'$, $p = x - y \in Ax$, $p' = x' - y' \in Ay'$:

$$\|x - x'\|^2 = \|y - y'\|^2 + 2\langle p - p', y - y'\rangle + \|p - p'\|^2 \geq \|y - y'\|^2 + \|p - p'\|^2,$$

that is, for $T = (I + A)^{-1}$:

$$\|Tx - Tx'\|^2 + \|(I - T)x - (I - T)x'\|^2 \leq \|x - x'\|^2. \tag{32}$$

An operator which satisfies (32) is *firmly non-expansive.*

Let us now consider the "*reflexion operator*"

$$R_A = 2J_A - I = 2(I + A)^{-1} - I \tag{33}$$

Lemma 4.24. *R_A is nonexpansive, and in particular, $J_A = I/2 + R_A/2$ is $(1/2)$-averaged.*

More generally we prove the following: *An operator T is firmly non-expansive if and only if it is $1/2$-averaged, that is, $R = 2T - I$ is non-expansive* (so that indeed $T = I/2 + R/2$ is 1/2-averaged).

It follows in an obvious way from the parallelogram idendity, since for any x, x',

$$\begin{aligned}
\|Rx - Rx'\|^2 &= \|(Tx - x) - (Tx' - x') + Tx - Tx'\|^2 \\
&= 2\|(I - T)x - (I - T)x'\|^2 + 2\|Tx - Tx'\|^2 - \|x - x'\|^2 \leq \|x - x'\|^2 \\
&\qquad \Leftrightarrow \quad \|(I - T)(x) - (I - T)(x')\|^2 + \|Tx - Tx'\|^2 \leq \|x - x'\|^2.
\end{aligned}$$

We have shown that if A is maximal monotone, then $J_A = (I+A)^{-1}$ is defined everywhere and single-valued, then that it is firmly non-expansive, and eventually that an operator is firmly non-expansive if and only if it is (1/2)-averaged. We conclude by showing that if an operator $T = I/2 + R/2$ is (1/2)-averaged (R is non-expansive), then there exists a maximal monotone operator A such that $T = J_A$.

The proof follows by the same (or reverse) construction as in the beginning of the proof of Minty's theorem: we consider the graph

$$G = \{((x+y)/2, (x-y)/2) : x \in \mathcal{X}, y = Rx\} = \{(Tx, (I-T)x) : x \in \mathcal{X}\}$$

and denote by A the corresponding operator ($y \in Ax \Leftrightarrow (x, y) \in G$). Then A is monotone: if $(\xi, \eta), (\xi', \eta') \in G$, then for some $x, x' \in \mathcal{X}$, $\xi = (x + Rx)/2$, $\eta = (x - Rx)/2$, etc., and we find:

$$\begin{aligned}\langle \xi - \xi', \eta - \eta' \rangle = \frac{1}{4} \langle x + Rx - x' - Rx', x - Rx - x' + Rx' \rangle \\ = \frac{1}{4} \left(\|x - x'\|^2 - \|Rx - Rx'\|^2 \right) \geq 0.\end{aligned}$$

Moreover, A is maximal, if not, one could build as before from $A' \supset A$ a non-expansive graph $\{(\xi + \eta, \xi - \eta) : \eta \in A'\xi\}$ strictly larger than the graph $\{(x, Rx) : x \in \mathcal{X}\}$, which is of course impossible. By construction, $ATx \ni (I - T)x$ for all x, hence $(I + A)Tx \ni x \Leftrightarrow Tx = (I + A)^{-1}x$.

To sum up, we have shown the following result:

Theorem 4.25. *Let T be an operator, then the following are equivalent:*

- *$T = (I + A)^{-1}$ for some maximal operator A;*
- *T is firmly non-expansive;*
- *T is (1/2)-averaged ($2T - I$ is nonexpansive).*

A consequence is that if $x \in \mathcal{X}$ and $x^{k+1} = (I + A)^{-1}x^k$, $k \geq 0$, then $x^k \rightharpoonup x$ where x is a fixed point of $(I + A)^{-1}$, that is, $Ax = 0$, if such a point exists (Theorem 3.1). We will return soon to these iterations.

Another way to interpret Theorem 4.22 is to observe that it says that a strongly monotone maximal operator has a well-defined single-valued inverse everywhere. Indeed, if A is maximal μ-monotone, then $A' = A/\mu - I$ is maximal monotone hence $I + A'$ is surjective with single-valued inverse, and so is A. From

$$\langle p - q, x - y \rangle \geq \mu \|x - y\|^2, \ p \in Ax, q \in Ay$$

we deduce if $B = A^{-1}$ that

$$\langle p - q, Bp - Bq \rangle \geq \mu \|Bp - Bq\|^2,$$

showing that B is co-coercive and $(1/\mu)$-Lipschitz.

The maximal monotone operator $A_\tau = [x - (I + \tau A)^{-1}x]/\tau$ is called the *Yosida* approximation of A_τ: it is a $(1/\tau)$-Lipschitz-continuous mapping, with full domain (in case $A = \partial f$, $A_\tau = \partial f_\tau$). τA_τ is firmly non-expansive, since $I - \tau A_\tau$ is. It has very important properties, see in particular Brézis' book [6]. We mention in particular Theorems 2.2, Prop. 2.5, and Cor. 2.7 in that book: the first two say that for a maximal monotone operator A, $C = \overline{\operatorname{dom} A}$ is convex and $\lim_{\tau \to 0} J_{\tau A} x$ is the orthogonal projection of x onto C, in addition if $x \in \operatorname{dom} A$, $A_\tau x \to A^0 x$, the element of Ax with minimal norm, while if not, $|A_\tau x| \to \infty$. The last shows that if for A, B two maximal monotone operators $\overbrace{\operatorname{dom} A}^{\circ} \cap \operatorname{dom} B \neq \emptyset$, then also $A + B$ is maximal monotone. The Yosida approximation is used in [6] to show the existence of solutions to $\dot{x} + Ax \ni 0$ for A maximal-monotone, by showing it is obtained as the limit of the solutions of $\dot{x} + A_\tau x \ni 0$ (which trivially exist because of Cauchy-Lipschitz's theorem). This allows to define properly the "gradient flow" of a convex lsc function, which is the time-continuous equivalent of the gradient descent algorithms. An exhaustive study of maximal monotone operators in Hilbert spaces is found in [2].

We will use the generalization of Moreau's identity (26):

$$x = (I + \tau A)^{-1}(x) + \tau(I + \tfrac{1}{\tau}A^{-1})^{-1}(\tfrac{x}{\tau}). \tag{34}$$

5 Algorithms. Operator splitting

We introduce here the "Forward-Backward splitting" technique. We discuss convergence rates and introduce acceleration, in particular the famous "FISTA / Nesterov acceleration".

We also introduce other splitting: Douglas-Rachford (DR), Alternating directions method of multipliers (ADMM), Primal-Dual.

5.1 Abstract algorithms for monotone operators

In this section, we describe rapidly general algorithms for solving the equations

$$0 \in Ax \quad \text{or} \quad 0 \in Ax + Bx$$

where A, B are maximal monotone operators (sometimes subgradients, sometimes not). The idea is to generalise algorithms already seen, and then to have at hand general results which will be useful for studying more concrete algorithms.

5.1.1 Explicit algorithm

Let us first consider the equivalent of the "gradient descent":

$$x^{k+1} = x^k - \tau p^k \,, p^k \in Ax^k.$$

Even if A is single-valued and continuous, then this might not converge. For instance, if $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ then

$$x^k = \begin{pmatrix} 1 & -\tau \\ \tau & 1 \end{pmatrix}^k x^0.$$

But the eigenvalues of this matrix are $1 + \pm\tau i$ and have modulus $\sqrt{1+\tau^2}$, so that the iteration always diverges.

So one needs to require a further condition on A. We recall (Baillon-Haddad) that the gradient descent works for convex functions with Lipschitz gradient, whose gradient is a co-coercive monotone operator. We can show here the same:

Theorem 5.1. *Let A maximal monotone be μ-co-coercive (in particular, single-valued):*

$$\langle Ax - Ay, x - y\rangle \geq \mu\|Ax - Ay\|^2.$$

Assume there exists a solution to $Ax = 0$. Then the iteration $x^{k+1} = x^k - \tau Ax^k$ converges to x^* with $Ax^* = 0$ if $0 < \tau < 2\mu$.

For the proof we just show that $I - \tau A$ is an averaged operator. Let us compute

$$\begin{aligned}\|(I-\tau A)x-(I-\tau A)y\|^2 + \|\tau Ax - \tau Ay\|^2 \\ &= \|x-y\|^2 - 2\tau\langle x-y, Ax-Ay\rangle + 2\tau^2\|Ax-Ay\|^2 \\ &= \|x-y\|^2 - 2\tau(\mu-\tau)\|Ax-Ay\|^2.\end{aligned}$$

This shows that if $0 \le \tau \le \mu$, τA and $(I - \tau A)$ are firmly non-expansive hence (1/2)-averaged. It follows that for $0 \le \tau < 2\mu$, $(I - \tau A)$ is averaged. Hence by Theorem 3.1 the iterates weakly converge, as $k \to \infty$, to a fixed point of $(I - \tau A)$ (if it exists). If $\tau = 0$ this is not interesting, if $0 < \tau < 2\mu$, then it is a zero of A, which exists by assumption.

5.1.2 Proximal point algorithm

Then we consider the "implicit descent" $x^{k+1} \in x^k - \tau A x^{k+1}$. This is precisely which is solved by $x^{k+1} = (I + \tau A)^{-1}x^k$, which is well-posed if A is maximal monotone (Th. 4.22). The corresponding iteration

$$x^{k+1} = (I+\tau A)^{-1}x^k$$

is known as the *proximal point algorithm.* It obviously converges to a fixed point as the operator is (1/2)-averaged (if the fixed point, that is a point with $Ax = 0$, exists). Moreover, as we have seen, one can consider more generally, if $R_{\tau A} = 2(I + \tau A)^{-1} - I$,

$$x^{k+1} = (1-\theta_k)x^k + \theta_k R_{\tau A}x^k = x^k + 2\theta_k\left((I+\tau A)^{-1}x^k - x^k\right) = x^k - 2\theta_k \tau A_\tau x^k,$$

for $0 < \underline{\theta} \le \theta_k \le \overline{\theta} < 1$ and still get convergence. More generally, we prove:

Theorem 5.2 (PPA Algorithm)**.** *Let $x^0 \in \mathcal{X}$, $\tau_k > 0$, $0 \le \underline{\lambda} \le \lambda_k \le \overline{\lambda} \le 2$, and let*

$$x^{k+1} = x^k + \lambda_k((I+\tau_k A)^{-1}x^k - x^k). \tag{35}$$

If there exists x with $Ax \ni 0$, then x^k weakly converges to a zero of A.

Proof: the proof follows the lines of the proof of Thm 3.1. We have:

$$\begin{aligned}\|x^{k+1}-x\|^2 = \|x^k - x\|^2 + \lambda_k^2\|J_{\tau_k A}x^k - x^k\|^2 + 2\lambda_k\left\langle x^k - x, J_{\tau_k A}x^k - x^k\right\rangle \\ \le \|x^k - x\|^2 - \lambda_k(2-\lambda_k)\|J_{\tau_k A}x^k - x^k\|^2 \\ + 2\lambda_k\left(\|J_{\tau_k A}x^k - x^k\|^2 + \left\langle x^k - x, J_{\tau_k A}x^k - x^k\right\rangle\right).\end{aligned}$$

If $Ax = 0$ we have $J_{\tau_k A}x = (I + \tau_k A)^{-1}x = x$, and using that $J_{\tau_k A}$ is firmly non-expansive we find

$$\|(I - J_{\tau_k A})x^k - (I - J_{\tau_k A})x\|^2 + \|J_{\tau_k A}x^k - J_{\tau_k A}x\|^2 \le \|x^k - x\|^2$$

which is

$$\|J_{\tau_k A}x^k - x^k\|^2 + \|J_{\tau_k A}x^k - x\|^2 \le \|x^k - x\|^2.$$

Hence

$$\begin{aligned}\|J_{\tau_k A}x^k - x^k\|^2 + \left\langle x^k - x, J_{\tau_k A}x^k - x^k\right\rangle \\ \le \tfrac{1}{2}\|J_{\tau_k A}x^k - x^k\|^2 + \left\langle x^k - x, J_{\tau_k A}x^k - x^k\right\rangle + \tfrac{1}{2}\left(\|x^k - x\|^2 - \|J_{\tau_k A}x^k - x\|^2\right) = 0\end{aligned}$$

and we deduce that

$$\|x^{k+1} - x\|^2 \le \|x^k - x\|^2 - \lambda_k(2-\lambda_k)\|J_{\tau_k A}x^k - x^k\|^2.$$

Letting $c = \underline{\lambda}(2-\overline{\lambda}) > 0$, we deduce that $(x^k)_k$ is Fejér-monotone with respect to $\{x : Ax \ni 0\}$ and that

$$c\sum_{k=0}^{n} \|J_{\tau_k A}x^k - x^k\|^2 + \|x^{n+1} - x\|^2 \leq \|x^0 - x\|^2$$

for all $n \geq 0$, in particular $\|J_{\tau_k A}x^k - x^k\| \to 0$, as well as, by (35), $x^{k+1} - x^k$. Hence we can conclude the proof as the proof of Theorem 3.1. We could also consider (summable) errors. See [2] for variants, [12] for a similar proof with errors.

5.1.3 Forward-Backward splitting

We now consider a mixture of the two previous, namely the "forward-backward" splitting

$$x^{k+1} = (I + \tau A)^{-1}(I - \tau B)x^k \tag{36}$$

where A is maximal monotone and B μ-co-coercive. Then, as before, if $0 < \tau < 2\mu$, the algorithm is the composition of two averaged operator and converges weakly to a fixed point if it exists. We see that

$$(I + \tau A)^{-1}(I - \tau B)x = x \Leftrightarrow x - \tau Bx \in x + \tau Ax \Leftrightarrow Ax + Bx \ni 0.$$

As B is continuous, this is equivalent to $(A + B)x \ni 0$. Hence, if $A + B$ has a zero, this algorithm converges to a zero of $A + B$.

5.1.4 Douglas-Rachford splitting

This method was introduced under the following form in a paper of Lions and Mercier (79):

$$x^{k+1} = J_{\tau A}(2J_{\tau B} - I)x^k + (I - J_{\tau B})x^k \tag{37}$$

Theorem 5.3. *Let $x^0 \in \mathcal{X}$. Then if x^k defined by* (37)*, $x^k \rightharpoonup x$ such that $w = J_{\tau B}x$ is a solution of $Aw + Bw \ni 0$ (if it exists).*

Proof: we use

$$J_{\tau A} = \tfrac{1}{2}I + \tfrac{1}{2}R_{\tau A}, \quad J_{\tau B} = \tfrac{1}{2}I + \tfrac{1}{2}R_{\tau B}.$$

Hence the operator in the algorithm is

$$\tfrac{1}{2}R_{\tau B} + \tfrac{1}{2}R_{\tau A} \circ R_{\tau B} + (\tfrac{1}{2}I - \tfrac{1}{2}R_{\tau B}) = \tfrac{1}{2}I + \tfrac{1}{2}R_{\tau A} \circ R_{\tau B}$$

so that it is (1/2)-averaged (and hence a resolvent). We deduce from Thm 3.1 that the iterations converge to a fixed point, if it exists, of $R_{\tau A} \circ R_{\tau B}$. One has

$$\begin{aligned} R_{\tau A} \circ R_{\tau B}x = x \Leftrightarrow 2J_{\tau A}(2J_{\tau B}x - x) - (2J_{\tau B}x - x) = x \Leftrightarrow J_{\tau A}(2J_{\tau B}x - x) = J_{\tau B}x \\ \Leftrightarrow 2J_{\tau B}x - x \in J_{\tau B}x + \tau A(J_{\tau B}x) \Leftrightarrow J_{\tau B}x \in x + \tau A(J_{\tau B}x). \end{aligned}$$

Letting $w = J_{\tau B}x$, we see that w satisfies

$$w \in w + \tau Bw + \tau Aw$$

hence $Aw + Bw = 0$. Conversely, if w satisfies this equation and $x = w + Bw = w - Aw$, we see that x is a fixed point. We know, then, by Theorem 3.1, that $x^k \rightharpoonup x$. Then $w = J_{\tau B}x$ is a solution of $Aw + Bw \ni 0$. Further conditions on A, B ensuring that $J_{\tau B}x^k$ converges to a solution are found in [19], variants with errors in [12].

The iterations $x^{k+1} = R_{\tau A}R_{\tau B}x^k$ are known as the *Peaceman-Rachford* splitting algorithm and converge under some conditions to the same point.

5.2 Descent algorithms, acceleration, "FISTA"

5.2.1 Forward-Backward descent

In case $A = \partial g$ and $B = \nabla f$, algorithm (36), which aims at finding a point x where $\partial g(x) + \nabla f(x) \ni 0$, or equivalently a minimizer of

$$\min_{x \in \mathcal{X}} F(x) := f(x) + g(x) \tag{38}$$

where g is, a "simple" convex lsc function and f is a convex function with Lipschitz gradient. The basic idea of the Forward-Backward splitting scheme (FBS) is to combine an explicit step of descent in the smooth part f with a implicit step of descent in g. It iterates the operator:

$$\bar{x} \mapsto \hat{x} = T_\tau \bar{x} := \operatorname{prox}_{\tau g}(\bar{x} - \tau \nabla f(\bar{x})) = (I + \tau \partial g)^{-1}(\bar{x} - \tau \nabla f(\bar{x})). \tag{39}$$

Another name found in the literature [23] is the "composite gradient" descent, as one may see here $(\hat{x} - \bar{x})/\tau$ as a generalised gradient for F at $\bar{x}$. The essential reason why all this is reasonable is that clearly, a fixed point $\hat{x} = \bar{x}$ will satisfy the Euler Lagrange equations $\nabla f(\bar{x}) + \partial g(\bar{x}) \ni 0$ of (38). Observe that in the particular case where $g = \delta_C$ is the characteristic function of a closed, convex set C, then $\operatorname{prox}_{\tau g}(x)$ reduces to $\Pi_C(x)$ (the orthogonal projection onto C) and the mapping T_τ defines a projected gradient descent method.

Algorithm 1 Forward-Backward descent with fixed step

Choose $x_0 \in \mathcal{X}$
for all $k \geq 0$ **do**

$$x^{k+1} = T_\tau x^k = \operatorname{prox}_{\tau g}(x^k - \tau \nabla f(x^k)). \tag{40}$$

end for

The theoretical convergence rate of the plain FBS descent is not very good, as one can merely show the same as for the gradient descent:

Theorem 5.4. *Let $x^0 \in \mathcal{X}$ and x^k be recursively defined by* (40)*, with $\tau \leq 1/L$. Then not only x^k converges to a minimiser, but one has the rates*

$$F(x^k) - F(x^*) \leq \tfrac{1}{2\tau k} \|x^* - x^0\|^2 \tag{41}$$

where x^* is any minimiser of f. If in addition f of g are strongly convex with parameters μ_f, μ_g (with $\mu = \mu_f + \mu_g > 0$), one has

$$F(x^k) - F(x^*) + \tfrac{1+\tau\mu_g}{2\tau} \|x^k - x^*\|^2 \leq \omega^k \tfrac{1+\tau\mu_g}{2\tau} \|x^0 - x^*\|^2. \tag{42}$$

where $\omega = (1 - \tau\mu_f)/(1 + \tau\mu_g)$.

However, its behaviour is improved if the objective is smoother than actually known, moreover, it is quite robust to perturbations and can be overrelaxed, see in particular [10].

5.2.2 FISTA

An "optimal" accelerated version is also available for this method, *cf* Section 2.4.3. This is well described in [25], [23], although a somewhat simpler proof is found in [3], where the algorithm, in the cases where $\mu = \mu_f + \mu_g = 0$, is called "FISTA". The general iteration takes the form:

Algorithm 2 FISTA with fixed step

Choose $x^0 = x^{-1} \in \mathcal{X}$ and $t_0 \geq 0$

for all $k \geq 0$ **do**

$$y^k = x^k + \beta_k(x^k - x^{k-1}) \tag{43}$$

$$x^{k+1} = T_\tau y^k = \mathrm{prox}_{\tau g}(y^k - \tau \nabla f(y^k)) \tag{44}$$

where, in case $\mu = 0$,

$$t_{k+1} = \tfrac{1+\sqrt{1+4t_k^2}}{2} \geq \tfrac{k+1}{2}, \tag{45}$$

$$\beta_k = \tfrac{t_k - 1}{t_{k+1}}, \tag{46}$$

and if $\mu = \mu_f + \mu_g > 0$,

$$t_{k+1} = \tfrac{1-qt_k^2+\sqrt{(1-qt_k^2)^2+4t_k^2}}{2}, \tag{47}$$

$$\beta_k = \tfrac{t_k-1}{t_{k+1}} \tfrac{1+\tau\mu_g - t_{k+1}\tau\mu}{1-\tau\mu_f}, \tag{48}$$

where $q = \tau\mu/(1+\tau\mu_g) < 1$.

end for

In the latter case, we assume $L > \mu_f$, otherwise f is quadratic and the problem is trivial. The following result is then true:

Theorem 5.5. *Assume $t_0 = 0$ and let x^k be generated by the algorithm, in either case $\mu = 0$ or $\mu > 0$. Then, one has the decay rate*

$$F(x^k) - F(x^*) \leq \min\left\{(1+\sqrt{q})(1-\sqrt{q})^k, \tfrac{4}{(k+1)^2}\right\} \tfrac{1+\tau\mu_g}{2\tau} \|x^0 - x^*\|^2.$$

It must be mentioned that in the case $\mu = 0$, a classical choice for t_k is also $t_k = (k+1)/2$, which gives essentially the same rate. An important issue is the stability of these rates when the proximal operators can be only evaluated inexactly — the situation here is worse than for the nonaccelerated algorithm, which has been addressed in several papers.

The proof of of both Theorems 5.4 and 5.5 rely on the following essential but straightforward descent rule: let $\hat{x} = T_\tau \bar{x}$, then for all $x \in \mathcal{X}$,

$$F(x) + (1-\tau\mu_f)\frac{\|x-\bar{x}\|^2}{2\tau} \geq \frac{1-\tau L}{\tau}\frac{\|\hat{x}-\bar{x}\|^2}{2} + F(\hat{x}) + (1+\tau\mu_g)\frac{\|x-\hat{x}\|^2}{2\tau}. \tag{49}$$

In particular, if $\tau L \leq 1$,

$$F(x) + (1-\tau\mu_f)\frac{\|x-\bar{x}\|^2}{2\tau} \geq F(\hat{x}) + (1+\tau\mu_g)\frac{\|x-\hat{x}\|^2}{2\tau}. \tag{50}$$

The proof is elementary: by definition, $\hat{x}$ is the minimiser of the $(\mu_g + (1/\tau))$-strongly convex function

$$x \mapsto g(x) + f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{\|x-\bar{x}\|^2}{2\tau}.$$

It follows that for all x (*cf* (18)):

$$\begin{aligned}F(x) + (1-\tau\mu_f)\frac{\|x-\bar{x}\|^2}{2\tau} \\ &\geq g(x) + f(\bar{x}) + \langle \nabla f(\bar{x}), x-\bar{x}\rangle + \frac{\|x-\bar{x}\|^2}{2\tau} \\ &\geq g(\hat{x}) + f(\bar{x}) + \langle \nabla f(\bar{x}), \hat{x}-\bar{x}\rangle + \frac{\|\hat{x}-\bar{x}\|^2}{2\tau} + (1+\tau\mu_g)\frac{\|x-\hat{x}\|^2}{2\tau}.\end{aligned}$$

But since ∇f is L-Lipschitz, $f(\bar{x})+\langle \nabla f(\bar{x}), \hat{x}-\bar{x}\rangle \geq f(\hat{x}) - (L/2)\|\hat{x}-\bar{x}\|^2$ so that equation (49) follows.

Remark 5.6. *One can more precisely deduce from this computation that*

$$F(x) + (1-\tau\mu_f)\frac{\|x-\bar{x}\|^2}{2\tau} \geq F(\hat{x}) + (1+\tau\mu_g)\frac{\|x-\hat{x}\|^2}{2\tau} + \left(\frac{\|\hat{x}-\bar{x}\|^2}{2\tau} - D_f(\hat{x},\bar{x})\right). \tag{51}$$

where $D_f(x,y) := f(x) - f(y) - \langle \nabla f(y), x-y\rangle \leq (L/2)\|x-y\|^2$ *is the "Bregman f-distance" from y to x [5]. In particular,* (50) *holds as soon as*

$$D_f(\hat{x},\bar{x}) \leq \frac{\|\hat{x}-\bar{x}\|^2}{2\tau}$$

which is always true if $\tau \leq 1/L$ but might also occur in other situations, and in particular, be tested "on the fly" during the iterations. This allows to implement efficient backtracking strategies 'à la' Armijo for the algorithms described in this section when the Lipschitz constant of f is not a priori known.

Remark 5.7. *Observe that if $X \subset \mathcal{X}$ is a closed convex set containing the domain of F, and on which the projection Π_X can be computed, then the same inequality* (50) *holds if $\hat{x} = T_\tau \Pi_X \bar{x}$ (requiring only that ∇f is Lipschitz on X). This means that the same rates are valid if one replaces* (43) *with*

$$y^k = \Pi_X(x^k + \beta_k(x^k - x^{k-1}))$$

which is feasible if X is the domain of F.

5.2.3 Convergence rates

Unaccelerated scheme We start with the rates of the unaccelerated FB descent scheme and prove Theorem 5.4.

First, if $\mu_f = \mu_g = 0$: we start from inequality (50), letting, for $k \geq 0$, $\bar{x} = x^k$ and $\hat{x} = x^{k+1}$. It follows that for any x:

$$F(x) + \frac{\|x-x^k\|^2}{2\tau} \geq F(x^{k+1}) + \frac{\|x-x^{k+1}\|^2}{2\tau}.$$

Choosing $x = x^k$ shows that $F(x^k)$ is nonincreasing. Summing then these inequalities from $k=0$ to $n-1$, $n \geq 1$ yields

$$\sum_{k=1}^{n}(F(x^k) - F(x)) + \sum_{k=1}^{n}\tfrac{1}{2\tau}\|x-x^k\|^2 \leq \sum_{k=0}^{n-1}\tfrac{1}{2\tau}\|x-x^k\|^2.$$

After cancellations and using $F(x^k) \geq F(x^n)$ for $k = 0,\dots,n$, it remains just

$$n(F(x^n) - F(x)) + \tfrac{1}{2\tau}\|x-x^n\|^2 \leq \tfrac{1}{2\tau}\|x-x^0\|^2$$

so that, in particular $F(x^n) - F(x^*) \le \|x^* - x^0\|^2/(2n\tau)$.

Now, if $\mu_f > 0$ or $\mu_g > 0$ we can improve this computation: we now have for any x:

$$F(x) + (1-\tau\mu_f)\frac{\|x-x^k\|^2}{2\tau} \ge F(x^{k+1}) + (1+\tau\mu_g)\frac{\|x-x^{k+1}\|^2}{2\tau}.$$

Choosing $x = x^k$ shows that $F(x^k)$ is nonincreasing. Letting

$$\omega = \frac{1-\tau\mu_f}{1+\tau\mu_g} \le 1, \tag{52}$$

and summing these inequalities from $k=0$ to $n-1$, $n \ge 1$, after multiplication by ω^{-k-1}, yields

$$\sum_{k=1}^{n} \omega^{-k}(F(x^k) - F(x)) + \sum_{k=1}^{n} \omega^{-k}\tfrac{1+\tau\mu_g}{2\tau}\|x-x^k\|^2 \le \sum_{k=0}^{n-1} \omega^{-k-1}\tfrac{1-\tau\mu_f}{2\tau}\|x-x^k\|^2.$$

After cancellations and using $F(x^k) \ge F(x^n)$ for $k=0,\dots,n$, we get

$$\omega^{-n}\left(\sum_{k=0}^{n-1}\omega^k\right)(F(x^n) - F(x)) + \omega^{-n}\tfrac{1+\tau\mu_g}{2\tau}\|x-x^n\|^2 \le \tfrac{1+\tau\mu_g}{2\tau}\|x-x^0\|^2.$$

We deduce, in case $\mu = \mu_f + \mu_g > 0$ so that $\omega < 1$,

$$F(x^k) - F(x^*) + \tfrac{1+\tau\mu_g}{2\tau}\|x^k - x^*\|^2 \le \omega^k \tfrac{1+\tau\mu_g}{2\tau}\|x^0 - x^*\|^2. \tag{53}$$

which is a "linear convergence rate" (however we will see that one can do better).

Convergence rates for FISTA Now we show the accelerated convergence rates. The basic idea consists in first choosing in (50) a generic point of the form $((t-1)x^k + x)/t$, $t \ge 1$, which is a convex combination of the iterate x^k and another generic point (in practice a minimizer) x. We find after some calculation (systematically using the strong convexity inequalities when possible)

$$\begin{aligned} t(t-1)(F(x^k) - F(x)) - \mu\frac{t-1}{2}\|x-x^k\|^2 & \\ + (1-\tau\mu_f)\frac{\|(t-1)x^k + x - ty^k\|^2}{2\tau} & \\ \ge t^2(F(x^{k+1}) - F(x)) + (1+\tau\mu_g)&\frac{\|(t-1)x^k + x - tx^{k+1}\|^2}{2\tau}. \end{aligned} \tag{54}$$

Consider first the case where $\mu = \mu_f + \mu_g = 0$. Then we have

$$\begin{aligned} t^2(F(x^{k+1}) - F(x)) + \frac{\|(t-1)x^k + x - tx^{k+1}\|^2}{2\tau} & \\ \le t(t-1)(F(x^k) - F(x)) + &\frac{\|(t-1)x^k + x - ty^k\|^2}{2\tau}. \end{aligned}$$

We see that the term $F(x^k) - F(x)$ is "shrunk" at each step by a factor $(t-1)/t < 1$, while the other term is not. How can we exploit this?

The basic idea in the proof is to use a *variable* parameter $t = t_{k+1}$, and choose y^k to ensure that the term $(t_{k+1}-1)x^k + x - t_{k+1}y^k$ in the right hand side is the same as the term $(t_{k+1}-1)x^k + x - t_{k+1}x^{k+1}$ of the left hand side at the *previous* iterate, that is,

$$(t_{k+1}-1)x^k + x - t_{k+1}y^k = (t_k - 1)x^{k-1} + x - t_k x^k$$

so that if we sum the inequalities for $k = 0,\dots,n$ the norms will cancel. Hence, we choose:

- $t_{k+1}(t_{k+1}-1) = t_k^2$;
- $y^k = x^k + \beta_k(x^k - x^{k-1})$ with $\beta_k = (t_k - 1)/t_{k+1}$;

we obtain the recursion

$$t_{k+1}^2(F(x^{k+1}) - F(x)) + \frac{\|(t_{k+1}-1)x^k + x - t_{k+1}x^{k+1}\|^2}{2\tau} \\ \le t_k^2(F(x^k) - F(x)) + \frac{\|(t_k-1)x^{k-1} + x - t_k x^k\|^2}{2\tau}.$$

which we can sum from $k = 0, \dots, n-1$ to obtain

$$F(x^n) - F(x) + \frac{1}{2t_n^2\tau}\|(t_{k+1}-1)x^k + x - t_{k+1}x^{k+1}\|^2 \le \frac{1}{2t_n^2\tau}\|x^0 - x\|^2.$$

Observe that $t_{k+1}^2 - t_{k+1} - t_k^2 = 0$ yields $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$ (one can choose $t_0 = 0, t_1 = 1$), and in particular $t_{k+1} \ge 1/2 + t_k \ge (k+1)/2$ for $k \ge 1$, by induction. Therefore, choosing $x = x^*$,

$$F(x^n) - F(x^*) \le \frac{2}{2(n+1)^2\tau}\|x^0 - x\|^2. \tag{55}$$

An important remark is that, if one takes $x = x^*$, $F(x^k) - F(x^*) \ge 0$ so that in fact one can get the same inequalities if one only ensures $t_{k+1}(t_{k+1}-1) \le t_k^2$, and not $=$. For instance, the sequence $t_0 = 0$, $t_k = (k+1)/2$ for $k \ge 1$ is admissible and yields the same rate.

It can be interesting to take slightly smaller t_k, such as $(k+1)/\alpha$ for $\alpha > 2$. One can show in particular the convergence of the iterates (x^k) to a solution in this case, while it is still an open problem for $\alpha = 2$. It has been even observed by Charles Dossal (U. Bordeaux) that in that case, one can show that

$$F(x^n) - F(x^*) = o\left(\frac{1}{n^2}\right)$$

which *does not* contradict the lower bound (5).

Convergence rates for FISTA, strongly convex case We start again from (54) but now we assume that $\mu = \mu_f + \mu_g > 0$. Then, we observe that

$$\begin{aligned} -\mu\frac{t-1}{2}&\|x - x^k\|^2 + (1-\tau\mu_f)\frac{\|x - x^k + t(x^k - y^k)\|^2}{2\tau} \\ &= (1 - \tau\mu_f - \mu\tau(t-1))\frac{\|x - x^k\|^2}{2\tau} + \frac{1-\tau\mu_f}{\tau}t\left\langle x - x^k, x^k - y^k\right\rangle + t^2(1-\tau\mu_f)\frac{\|x^k - y^k\|^2}{2\tau} \\ &= \frac{(1+\tau\mu_g - t\mu\tau)}{2\tau}\|x - x^k + t\tfrac{1-\tau\mu_f}{1+\tau\mu_g - t\mu\tau}(x^k - y^k)\|^2 + t^2(1-\tau\mu_f)\left(1 - \tfrac{1-\tau\mu_f}{1+\tau\mu_g - t\mu\tau}\right)\frac{\|x^k - y^k\|^2}{2\tau} \\ &= \frac{(1+\tau\mu_g - t\mu\tau)}{2\tau}\|x - x^k + t\tfrac{1-\tau\mu_f}{1+\tau\mu_g - t\mu\tau}(x^k - y^k)\|^2 - t^2(t-1)\frac{\tau\mu(1-\tau\mu_f)}{1+\tau\mu_g - t\mu\tau}\frac{\|x^k - y^k\|^2}{2\tau}. \end{aligned}$$

It follows that for any $x \in \mathcal{X}$,

$$\begin{aligned} t(t-1)(F(x^k) - F(x)) + (1+\tau\mu_g - t\mu\tau)&\frac{\|x - x^k - t\tfrac{1-\tau\mu_f}{1+\tau\mu_g - t\mu\tau}(y^k - x^k)\|^2}{2\tau} \\ &\ge t^2(F(x^{k+1}) - F(x)) + (1+\tau\mu_g)\frac{\|x - x^{k+1} - (t-1)(x^{k+1} - x^k)\|^2}{2\tau} \\ &\qquad + t^2(t-1)\frac{\tau\mu(1-\tau\mu_f)}{1+\tau\mu_g - t\mu\tau}\frac{\|x^k - y^k\|^2}{2\tau}. \end{aligned} \tag{56}$$

We let $t = t_{k+1}$ above, then we can get a useful recursion if we let

$$\omega_k = \frac{1+\tau\mu_g - t_{k+1}\mu\tau}{1+\tau\mu_g} = 1 - t_{k+1}\frac{\mu\tau}{1+\tau\mu_g} \in [0,1] \tag{57}$$

$$t_{k+1}(t_{k+1}-1) \leq \omega_k t_k^2, \tag{58}$$

$$\beta_k = \frac{t_k - 1}{t_{k+1}} \frac{1+\tau\mu_g - t_{k+1}\mu\tau}{1-\tau\mu_f} = \omega_k \frac{t_k - 1}{t_{k+1}} \frac{1+\tau\mu_g}{1-\tau\mu_f}, \tag{59}$$

$$y^k = x^k + \beta_k(x^k - x^{k-1}) \tag{60}$$

Denoting $\alpha_k = 1/t_k$ and

$$q = \frac{\tau\mu}{1+\tau\mu_g} = \frac{\tau\mu_f + \tau\mu_g}{1+\tau\mu_g} < 1, \tag{61}$$

one finds the same rules as in formula (2.2.9), p. 80 in [25] (with the minor difference that here we may chose $t_0 = 0, t_1 = 1$, and we have shifted the numbering of the sequences $(x^k), (y^k)$). In this case, we find

$$t_{k+1}^2(F(x^{k+1}) - F(x)) + \frac{1+\tau\mu_g}{2\tau}\|x - x^{k+1} - (t_{k+1}-1)(x^{k+1} - x^k)\|^2$$
$$\leq \omega_k \left(t_k^2(F(x^k) - F(x)) + \frac{1+\tau\mu_g}{2\tau}\|x - x^k - (t_k - 1)(x^k - x^{k-1})\|^2 \right)$$

so that

$$t_k^2(F(x^k) - F(x)) + \frac{1+\tau\mu_g}{2\tau}\|x - x^k - (t-1)(x^k - x^{k-1})\|^2$$
$$\leq \left(\prod_{n=0}^{k-1} \omega_n\right) \left[t_0^2(F(x^0) - F(x)) + \frac{1+\tau\mu_g}{2\tau}\|x - x^0\|^2 \right] \tag{62}$$

The update rule for t_k reads

$$t_{k+1}(t_{k+1} - 1) = (1 - qt_{k+1})t_k^2, \tag{63}$$

so that,

$$t_{k+1} = \frac{1 - qt_k^2 + \sqrt{(1-qt_k^2)^2 + 4t_k^2}}{2}. \tag{64}$$

We need to make sure that $qt_{k+1} \leq 1$ so that (57) holds. This is proved exactly as in the proof of Lemma 2.2.4 in [25]. Assuming (as in [25]) that $\sqrt{q}t_k \leq 1$, we observe that (63) yields

$$qt_{k+1}^2 = qt_{k+1} + (1 - qt_{k+1})qt_k^2.$$

If $qt_{k+1} \geq 1$, it yields $qt_{k+1}^2 \leq qt_{k+1}$ hence $qt_{k+1} \leq q < 1$, a contradiction. Hence $qt_{k+1} < 1$ and we obtain that qt_{k+1}^2 is a convex combination of 1 and qt_k^2, so that $\sqrt{q}t_{k+1} \leq 1$. We have shown that as soon as $\sqrt{q}t_0 \leq 1$ (which we will now assume), $\sqrt{q}t_k \leq 1$ for all k. Eventually, we also observe that

$$t_{k+1}^2 = (1 - qt_k^2)t_{k+1} + t_k^2$$

showing that t_k is an increasing sequence. It remains to estimate the factor

$$\theta_k = t_k^{-2} \prod_{n=0}^{k-1} \omega_n \quad (k \geq 1).$$

From (58) (with an equality) we find

$$1 - \frac{1}{t_{k+1}} = \omega_k \frac{t_k^2}{t_{k+1}^2}$$

so that

$$t_0^2 \theta_k = \frac{t_0^2}{t_k^2} \prod_{n=0}^{k-1} \omega_n = \prod_{n=1}^{k} \left(1 - \frac{1}{t_k}\right) \le (1 - \sqrt{q})^k$$

since $1/t_k \ge \sqrt{q}$. If $t_0 \ge 1$ it shows $\theta_k \le (1 - \sqrt{q})^k / t_0^2$. If $t_0 \in [0, 1[$, we rather write

$$\theta_k = \frac{\omega_0}{t_k^2} \prod_{n=1}^{k-1} \omega_n = \frac{\omega_0}{t_1^2} \prod_{n=2}^{k} \left(1 - \frac{1}{t_k}\right)$$

and observe that (64) yields (using $2 - q \ge 1 \ge q$)

$$t_1 = \frac{1 - q t_0^2 + \sqrt{1 + 2(2-q) t_0^2 + q^2 t_0^4}}{2} \ge 1.$$

Also, $\omega_0 \le 1 - q$ (from (57)), so that

$$\theta_k \le (1 + \sqrt{q})(1 - \sqrt{q})^k.$$

The next step is to bound also θ_k by $O1/k^2$. We exactly follow Lemma 2.2.4 in [25]. In our notation, it reads

$$\frac{1}{\sqrt{\theta_{k+1}}} - \frac{1}{\sqrt{\theta_k}} = \frac{\theta_k - \theta_{k+1}}{\sqrt{\theta_k \theta_{k+1}} (\sqrt{\theta_k} + \sqrt{\theta_{k+1}})} \ge \frac{\theta_k (1 - (1 - 1/t_{k+1}))}{2 \theta_k \sqrt{\theta_{k+1}}}$$

since θ_k is nonincreasing. It follows

$$\frac{1}{\sqrt{\theta_{k+1}}} - \frac{1}{\sqrt{\theta_k}} \ge \frac{1}{2 t_{k+1} \sqrt{\theta_{k+1}}} = \frac{1}{2} \frac{1}{\sqrt{\prod_{n=0}^{k} \omega_n}} \ge \frac{1}{2},$$

showing that $1/\sqrt{\theta_k} \ge (k-1)/2 + t_1/\sqrt{\omega_0} \ge (k+1)/2$. Hence, provided $\sqrt{q} t_0 \le 1$, we also find:

$$\theta_k \le \frac{4}{(k+1)^2}. \tag{65}$$

We have shown the following result, due to Nesterov and stated, with a different proof, in [25]:

Theorem 5.8. *If $\sqrt{q} t_0 \le 1$, $t_0 \ge 0$, then the sequence (x^k) produced by iterations $x^k = T_\tau y^k$ with (64), (57), (59), (60) satisfies*

$$F(x^k) - F(x^*) \le \min \left\{ \frac{(1 - \sqrt{q})^k}{t_0^2}, \frac{4}{(k+1)^2} \right\} \left(t_0^2 (F(x^0) - F(x^*)) + \frac{1 + \tau \mu_g}{2\tau} \|x^0 - x^*\|^2 \right) \tag{66}$$

if $t_0 \ge 1$, and

$$F(x^k) - F(x^*) \le \min \left\{ (1 + \sqrt{q})(1 - \sqrt{q})^k, \frac{4}{(k+1)^2} \right\} \left(t_0^2 (F(x^0) - F(x^*)) + \frac{1 + \tau \mu_g}{2\tau} \|x^0 - x^*\|^2 \right) \tag{67}$$

if $t_0 \in [0, 1]$, where x^* is a minimiser of F.

Theorem 5.5 is a particular case of this result, for $t_0 = 0$.

Remark 5.9. *(Constant steps.)* In case $\mu > 0$ (which is $q > 0$), then an admissible choice which satisfies (57),(58), (59), is to take $t = 1/\sqrt{q}$, $\omega = 1 - \sqrt{q}$,

$$\beta = \omega^2 \frac{1 + \tau\mu_g}{1 - \tau\mu_f} = \frac{\sqrt{1 + \tau\mu_g} - \sqrt{\tau\mu}}{\sqrt{1 + \tau\mu_g} + \sqrt{\tau\mu}}.$$

Then, (66) becomes

$$F(x^k) - F(x^*) \leq (1 - \sqrt{q})^k \left(F(x^0) - F(x^*) + \mu \frac{\|x^0 - x^*\|^2}{2} \right).$$

Remark 5.10. *(Monotone "FISTA", monotone algorithms.) The algorithms studied here are not necessarily "monotone" in the sense that the objective F is not always nonincreasing. A workaround implemented in various papers [34, 3] consists in choosing for x^{k+1} any point with $F(x^{k+1}) \leq F(T_\tau y^k)$*[4]*, which will not change much* (54) *except that in the last term, x^{k+1} should be replaced with $T_\tau y^k$. Then, the same computations carry on, and it is enough to replace the update rule* (60) *for y^k with*

$$\begin{aligned} y^k &= x^k + \beta_k(x^k - x^{k-1}) + \omega_k \tfrac{t_k}{t_{k+1}} \tfrac{1+\tau\mu_g}{1-\tau\mu_f} (T_\tau y^{k-1} - x^k) \\ &= x^k + \beta_k \left((x^k - x^{k-1}) + \tfrac{t_k}{t_k - 1} (T_\tau y^{k-1} - x^k) \right) \end{aligned} \tag{60'}$$

to obtain the same rates of convergence. The most sensible choice for x^{k+1} is to take $T_\tau y^k$ if $F(T_\tau y^k) \leq F(x^k)$, and x^k else, in which case one of the two terms ($x^k - x^{k-1}$ or $T_\tau y^{k-1} - x^k$) vanishes in (60').

Conclusion: compare the geometric rate (52) with $\omega = 1 - \sqrt{q}$ for q given by (61), what do we observe?

5.3 ADMM, Douglas-Rachford splitting

We now consider a class of method which solves another kind of problem, namely of the form

$$\min_{Ax+By=\zeta} f(x) + g(y) \tag{68}$$

where in practice one will ask that the convex, lsc functions f, g are "simple" (and even more than this).

Observe that if f^* is continuous at some point A^*p and if g^* is continuous at some point B^*q, *cf* Section. 4.3.2 (or, in finite dimension, if $A^*p \in \operatorname{ri} \operatorname{dom} f^*$, $B^*q \in \operatorname{ri} \operatorname{dom} g^*$), we can define

$$\tilde{f}(\xi) = \min_{Ax=\xi} f(x), \quad \tilde{g}(\eta) = \min_{By=\eta} g(y),$$

moreover the min is reached in both cases.

Then, one has $\tilde{f}^*(p) = f(A^*p)$, $\tilde{g}^*(q) = g(B^*q)$ and the problem reads

$$\min_\xi \tilde{f}(\xi) + \tilde{g}(\zeta - \xi);$$

it can be seen as an inf-convolution problem. Moreover Corollary 4.20 shows that the value of (68) is also

$$\sup_p \langle \zeta, p \rangle - f^*(A^*p) - g^*(B^*p) \tag{69}$$

[4] this makes sense only if the evaluation of F is easy and does not take too much time.

which gives a dual form for (68).

An "augmented Lagrangian" approach for (68) consists in introducing the constraint in the form

$$\min_{x,y}\sup_{z} f(x)+g(y)-\langle z, Ax+By-\zeta\rangle+\frac{\gamma}{2}\|Ax+By-\zeta\|^2$$

which we observe is equivalent (as the sup is $+\infty$ if $Ax+By\neq\zeta$).

If we introduce the function

$$\mathcal{D}(z)=\inf_{x,y} f(x)+g(y)-\langle z, Ax+By-\zeta\rangle+\frac{\gamma}{2}\|Ax+By-\zeta\|^2$$

we find that, denoting $\bar x,\bar y$ the solution of the problem for z and $\bar x_h,\bar y_h$ the solution for $z+h$ (the min is reached, why?),

$$\begin{aligned}\mathcal{D}(z)&=f(\bar x)+g(\bar y)-\langle z+h, A\bar x+B\bar y-\zeta\rangle+\frac{\gamma}{2}\|A\bar x+B\bar y-\zeta\|^2+\langle h, A\bar x+B\bar y-\zeta\rangle\\&\geq f(\bar x_h)+g(\bar y_h)-\langle z+h, A\bar x_h+B\bar y_h-\zeta\rangle+\frac{\gamma}{2}\|A\bar x_h+B\bar y_h-\zeta\|^2\\&\qquad+\frac{\gamma}{2}\|A(\bar x-\bar x_h)+B(\bar y-\bar y_h)\|^2+\langle h, A\bar x+B\bar y-\zeta\rangle\end{aligned}$$

where we have used the strong convexity of the norm with respect to $Ax+By$. We find

$$\mathcal{D}(z)-\langle h, A\bar x+B\bar y-\zeta\rangle\geq\mathcal{D}(z+h)+\frac{\gamma}{2}\|(A\bar x+B\bar y-\zeta)-(A\bar x_h+B\bar y_h-\zeta)\|^2$$

which shows that $\zeta-A\bar x-B\bar y\in\partial^+\mathcal{D}(z)$ (the super-gradient of the concave function $\mathcal{D}$ at z) and that $z\mapsto A\bar x+B\bar y-\zeta$ is γ-co-coercive, and $(1/\gamma)$-Lipschitz. Hence a natural algorithm, known as "augmented Lagrangian", consists in iteratively solving

$$\begin{cases}(x^{k+1},y^{k+1})=\arg\min_{x,y} f(x)+g(y)-\left\langle z^k, Ax+By-\zeta\right\rangle+\frac{\gamma}{2}\|Ax+By-\zeta\|^2\,,\\ z^{k+1}=z^k+\gamma(\zeta-Ax^{k+1}-By^{k+1}):\end{cases}\tag{70}$$

it is precisely a gradient ascent with fixed step for the concave function $\mathcal{D}$, and will converge (it should be shown then that also x^k,y^k converge to a solution).

Unfortunately, this algorithm is usually not implementable, as the joint minimization step cannot in general be performed. This is why it was proposed [17, 16] to perform these minimizations alternatively instead than simultaneously, see Algorithm 3

Algorithm 3 ADMM

Choose $\gamma>0$, y^0, z^0.
for all $k\geq 0$ **do**
  Find x^{k+1} by minimising $x\mapsto f(x)-\left\langle z^k, Ax\right\rangle+\frac{\gamma}{2}\|\zeta-Ax-By^k\|^2$,
  Find y^{k+1} by minimising $y\mapsto g(y)-\left\langle z^k, By\right\rangle+\frac{\gamma}{2}\|\zeta-Ax^{k+1}-By\|^2$,
  Update $z^{k+1}=z^k+\gamma(\zeta-Ax^{k+1}-By^{k+1})$.
end for

We will relate this approach to other known converging algorithms. Then in a next section, we will show how we can derive rates of convergence for this approach. A classical reference for the convergence is [12], see also `http://stanford.edu/~boyd/admm.html`.

Let us observe that in terms of the functions $\tilde{f}, \tilde{g}$, the algorithm computes, letting $\xi^k = Ax^k$, $\eta^k = By^k$:

$$\xi^{k+1} = \arg\min_\xi \tilde{f}(\xi) - \langle z^k, \xi\rangle + \frac{\gamma}{2}\|\zeta - \xi - \eta^k\|^2 = \operatorname{prox}_{\tilde{f}/\gamma}(\zeta + \tfrac{1}{\gamma}z^k - \eta^k); \tag{71}$$

$$\eta^{k+1} = \arg\min_\eta \tilde{g}(\eta) - \langle z^k, \eta\rangle + \frac{\gamma}{2}\|\zeta - \xi^{k+1} - \eta\|^2 = \operatorname{prox}_{\tilde{g}/\gamma}(\zeta + \tfrac{1}{\gamma}z^k - \xi^{k+1}). \tag{72}$$

Thanks to Moreau's identity (26),

$$\operatorname{prox}_{\gamma\tilde{f}^*}(z^k + \gamma(\zeta - \eta^k)) = z^k + \gamma(\zeta - \eta^k) - \gamma\xi^{k+1}, \tag{73}$$

$$\operatorname{prox}_{\gamma\tilde{g}^*}(z^k + \gamma(\zeta - \xi^{k+1})) = z^k + \gamma(\zeta - \xi^{k+1}) - \gamma\eta^{k+1} = z^{k+1}. \tag{74}$$

Letting $\tilde{f}^*_\zeta(p) := \tilde{f}^*(p) - \langle\zeta, p\rangle = f^*(A^*p) - \langle\zeta, p\rangle$, the first line can also be rewritten

$$\gamma(\xi^{k+1} - \zeta) = z^k - \gamma\eta^k - \operatorname{prox}_{\gamma\tilde{f}^*_\zeta}(z^k - \gamma\eta^k). \tag{75}$$

If we let $u^k = z^k - \gamma\eta^k$, $v^{k+1} = z^k + \gamma(\zeta - \xi^{k+1})$, we find that

$$\gamma\eta^{k+1} = z^k + \gamma(\zeta - \xi^{k+1}) - \operatorname{prox}_{\gamma\tilde{g}^*}(z^k + \gamma(\zeta - \xi^{k+1})) = v^{k+1} - \operatorname{prox}_{\gamma\tilde{g}^*}(v^{k+1}).$$

and

$$u^{k+1} = z^{k+1} - \gamma\eta^{k+1} = 2\operatorname{prox}_{\gamma\tilde{g}^*}(v^{k+1}) - v^{k+1}.$$

On the other hand, (75) gives

$$\operatorname{prox}_{\gamma\tilde{f}^*_\zeta}(u^k) + \gamma\eta^k = z^k + \gamma(\zeta - \xi^{k+1}) = v^{k+1}.$$

Hence the iteration reads

$$v^{k+1} = \operatorname{prox}_{\gamma\tilde{f}^*_\zeta}(2\operatorname{prox}_{\gamma\tilde{g}^*}(v^k) - v^k) + v^k - \operatorname{prox}_{\gamma\tilde{g}^*}(v^k),$$

which is precisely a Douglas-Rachford iteration for the problem

$$0 \in \partial\tilde{g}^* + \partial\tilde{f}^*_\zeta$$

which is the equation for (69).

The theory seen up to now shows that $v^k \rightharpoonup v$ a fixed point of the iteration, which is such that $\operatorname{prox}_{\gamma\tilde{g}^*}(v)$ is a solution of the dual problem. In practice, z^k will converge to a Lagrange Multiplier for (70), and x^k, y^k to a solution, as soon as there is enough coercivity (in particular, in finite dimension).

5.4 Other saddle-point algorithms: Primal-dual algorithm

We remark that thanks to (74) and (71), one has

$$\frac{z^k - z^{k-1}}{\gamma} = \zeta - \xi^k - \eta^k$$

hence

$$\xi^{k+1} = \operatorname{prox}_{\tilde{f}/\gamma}(\xi^k + \tfrac{1}{\gamma}(2z^k - z^{k-1}))$$

while as before

$$z^{k+1} = \operatorname{prox}_{\gamma\tilde{g}^*}(z^k - \gamma(\xi^{k+1} - \zeta)).$$

This is the form of a *primal-dual* algorithm (of “Arrow-Hurwicz” type) which aims at solving a fixed point problem of the form (letting $\tau = 1/\gamma$):

$$\xi + \tau\partial\tilde{f}(\xi) \ni \xi + \tau z, \quad z + \gamma\partial\tilde{g}^*(z) \ni z - \gamma(\xi - \zeta).$$

More generally, for a problem in the standard form

$$\min_x f(Kx) + g(x) = \min_x \sup_y \langle Kx, y\rangle + g(x) - f^*(y),$$

one can implement the Algorithm 4 described below.

Algorithm 4 PDHG

Input: initial pair of primal and dual points (x^0, y^0), steps $\tau, \sigma > 0$.
for all $k \geq 0$ **do**
find (x^{k+1}, y^{k+1}) by solving

$$x^{k+1} = \mathrm{prox}_{\tau g}(x^k - \tau K^* y^k) \tag{76}$$

$$y^{k+1} = \mathrm{prox}_{\sigma f^*}(y^k + \sigma K(2x^{k+1} - x^k)). \tag{77}$$

end for

Let us write now the iterates as follows:

$$\begin{cases} \frac{x^{k+1}-x^k}{\tau} + \partial g(x^{k+1}) \ni -K^* y^k = K^*(y^{k+1} - y^k) - K^* y^{k+1} \\ \frac{y^{k+1}-y^k}{\sigma} + \partial f^*(y^{k+1}) \ni K(x^{k+1} - x^k) + Kx^{k+1}, \end{cases}$$

that is

$$\begin{pmatrix} \frac{1}{\tau}I & -K^* \\ -K & \frac{1}{\sigma}I \end{pmatrix} \begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix} + \begin{pmatrix} \partial g(x^{k+1}) \\ \partial f^*(y^{k+1}) \end{pmatrix} + \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix} \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} \ni 0. \tag{78}$$

We find that this algorithm is a proximal point algorithm for the variable $z^k = (x^k, y^k)^T$, the monotone operator which is the sum of the subgradient of the convex function $(x, y) \mapsto (g(x) + f^*(y))$ and the antisymmetric linear operator $\begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix}$, in the metric

$$M_{\tau,\sigma} := \begin{pmatrix} \frac{1}{\tau}I & -K^* \\ -K & \frac{1}{\sigma}I \end{pmatrix}$$

if this metric is positive definite. To see this we observe that if A is a monotone operator and M a symmetric positive definite operator, $M^{-1}A$ defines a monotone operator in the scalar product $\langle\cdot,\cdot\rangle_M = \langle M\cdot,\cdot\rangle$: if $p \in M^{-1}Ax$, $q \in M^{-1}Ay$,

$$\langle p - q, x - y\rangle_M = \langle M(p - q), x - y\rangle \geq 0$$

as $Mp \in Ax$, $Mq \in Ay$. Hence, in this metric, the resolvent J_A^M is given by $y = (I + M^{-1}A)^{-1}x$, which satisfies the equation $y + M^{-1}Ay \ni x$, that is, $M(y - x) + Ay \ni 0$.

When is the matrix $M_{\tau,\sigma}$ positive definite? We have

$$\left\langle M_{\tau,\sigma} \begin{pmatrix} \xi \\ \eta \end{pmatrix}, \begin{pmatrix} \xi \\ \eta \end{pmatrix} \right\rangle = \frac{1}{\tau}\|\xi\|^2 + \frac{1}{\sigma}\|\eta\|^2 - 2\langle K\xi, \eta\rangle$$

which is positive if and only if for any $X, Y \geq 0$

$$\sup_{\|\xi\|\leq X, \|\eta\|\leq Y} 2\langle K\xi, \eta\rangle = 2\|K\|XY < \frac{X^2}{\tau} + \frac{Y^2}{\sigma}$$

if and only if

$$2\|K\| < \min_{X\geq 0, Y\geq 0} \frac{X}{\tau Y} + \frac{Y}{\sigma X} = \frac{2}{\sqrt{\tau\sigma}}$$

if and only if

$$\tau\sigma\|K\|^2 < 1. \tag{79}$$

We deduce:

Theorem 5.11. *If* (79) *is satisfied, then* $z^k = (x^k, y^k)^T$ *defined by Algorithm 4 converges to a fixed point* $(x, y)^T$ *of the operator, that is, a solution of* (31) *(if one exists).*

5.4.1 Rate

To find a rate, we do as follows. Taking the scalar product of (78) with $z^{k+1} - z$ where z is an arbitrary point, we find

$$\left\langle z^{k+1} - z^k, z^{k+1} - z\right\rangle_{M_{\tau,\sigma}} + \left\langle \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix} \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix}, \begin{pmatrix} x^{k+1} - x \\ y^{k+1} - y \end{pmatrix} \right\rangle \\ + g(x^{k+1}) + f^*(y^{k+1}) \leq g(x) + f^*(y)$$

The scalar product is

$$-\left\langle K^* y^{k+1}, x\right\rangle + \left\langle K x^{k+1}, y\right\rangle$$

while

$$\left\langle z^{k+1} - z^k, z^{k+1} - z\right\rangle_{M_{\tau,\sigma}} = \tfrac{1}{2}\|z^{k+1} - z^k\|^2_{M_{\tau,\sigma}} + \tfrac{1}{2}\|z^{k+1} - z\|^2_{M_{\tau,\sigma}} - \tfrac{1}{2}\|z^k - z\|^2_{M_{\tau,\sigma}}.$$

Therefore, introducing the Lagrangian of (29), as

$$\mathcal{L}(x^{k+1}, y) - \mathcal{L}(x, y^{k+1}) = g(x^{k+1}) + \left\langle y, Kx^{k+1}\right\rangle - f^*(y) - g(x) - \left\langle y^{k+1}, Kx\right\rangle + f^*(y^{k+1})$$

we obtain for any $z = (x, y)^T$:

$$\mathcal{L}(x^{k+1}, y) - \mathcal{L}(x, y^{k+1}) + \tfrac{1}{2}\|z^{k+1} - z^k\|^2_{M_{\tau,\sigma}} + \tfrac{1}{2}\|z^{k+1} - z\|^2_{M_{\tau,\sigma}} \leq \tfrac{1}{2}\|z^k - z\|^2_{M_{\tau,\sigma}}.$$

Summing from $k = 0$ to $n - 1$ and using the convexity of $(\xi, \eta)^T \mapsto \mathcal{L}(\xi, y) - \mathcal{L}(x, \eta)$, we find if we let $Z^n = (X^n, Y^n)^T = (\sum_{k=1}^n z^n)/n$ that

$$\mathcal{L}(X^n, y) - \mathcal{L}(x, Y^n) \leq \frac{1}{2n}\|z^0 - z\|^2_{M_{\tau,\sigma}}. \tag{80}$$

This is a weak form of a rate (as it depends on (x, y)), and there is still some work to convert it into a true rate for the energy. The simplest case is when $\operatorname{dom} f^*, \operatorname{dom} g$ are bounded, then one can take the sup in x, y to find that

$$\mathcal{G}(X^n, Y^n) \leq \frac{C}{2n}$$

where $C = \sup\{\|z^0 - z\|^2_{M_{\tau,\sigma}} : z = (x, y), x \in \operatorname{dom} g, y \in \operatorname{dom} f^*\}$.

5.4.2 Extensions

We present here an extension of Algorithm 4 due to Condat and in a generalized form to Vu (referred usually as Condat-Vu's primal-dual algorithm). A first observation (*cf* Vu, Bot) is that one can replace ∂g and ∂f^* with monotone operators, and get similar results.

A second observation, due to Condat, is that one can iterate the operator with an explicit step of a co-coercive operator. Typically, one if h is a convex function with L-Lipschitz gradient, one can replace (78) with

$$\begin{pmatrix} \frac{1}{\tau}I & -K^* \\ -K & \frac{1}{\sigma}I \end{pmatrix} \begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix} + \begin{pmatrix} \partial g(x^{k+1}) \\ \partial f^*(y^{k+1}) \end{pmatrix} + \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix} \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} \ni \begin{pmatrix} -\nabla h(x^k) \\ 0 \end{pmatrix}.$$

This iteration is of the form (36) and will converge if the operator

$$C = M_{\tau,\sigma}^{-1} \begin{pmatrix} -\nabla h(x) \\ 0 \end{pmatrix}$$

is μ-co-coercive with $\mu > 1/2$, in the metric $M_{\tau,\sigma}$. That is, if for all z, z':

$$\langle M_{\tau,\sigma}(z - z'), Cz - Cz' \rangle \geq \mu \|Cz - Cz'\|^2_{M_{\tau,\sigma}}.$$

This can be rewritten

$$\langle x - x', \nabla h(x) - \nabla h(x') \rangle \geq \tau\mu \|\nabla h(x) - \nabla h(x')\|^2$$

and one will be able to find $\mu > 1/2$ such that this holds as soon as $\tau < 2/L$. In this case again we get the convergence of the Vu-Condat algorithm, which reads:

Algorithm 5 PDHG with explicit step

Input: initial pair of primal and dual points (x^0, y^0), steps $\tau, \sigma > 0$.
for all $k \geq 0$ **do**
find (x^{k+1}, y^{k+1}) by solving

$$x^{k+1} = \mathrm{prox}_{\tau g}(x^k - \tau(K^* y^k + \nabla h(x^k))) \tag{81}$$

$$y^{k+1} = \mathrm{prox}_{\sigma f^*}(y^k + \sigma K(2x^{k+1} - x^k)). \tag{82}$$

end for

Exercise: Show that a fixed point of these iterations solves

$$\min_x f(Kx) + g(x) + h(x) = \min_x \sup_y \langle y, Kx \rangle - f^*(y) + g(x) + h(x).$$

5.5 Acceleration of the Primal-Dual algorithm

If times permits...

References

[1] Jean-Bernard Baillon and Georges Haddad. Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel J. Math.*, 26(2):137–150, 1977.

[2] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces.* CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. With a foreword by Hédy Attouch.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[4] Dimitri P. Bertsekas. *Convex Optimization Algorithms.* Athena Scientific, 2015.

[5] L. M. Brègman. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 7:620–631, 1967.

[6] H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert.* North-Holland Publishing Co., Amsterdam, 1973. North-Holland Mathematics Studies, No. 5. Notas de Matemática (50).

[7] Haïm Brézis. *Analyse fonctionnelle.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. Théorie et applications. [Theory and applications].

[8] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ci You Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208, 1995.

[9] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numer.*, 25:161–319, 2016.

[10] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200, 2005.

[11] R. Cominetti, J. A. Soto, and J. Vaisman. On the rate of convergence of Krasnosel'skiĭ-Mann iterations and their connection with sums of Bernoullis. *Israel J. Math.*, 199(2):757–772, 2014.

[12] Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Programming*, 55(3, Ser. A):293–318, 1992.

[13] I. Ekeland and R. Témam. *Convex analysis and variational problems*, volume 28 of *Classics in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english edition, 1999. Translated from the French.

[14] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions.* CRC Press, Boca Raton, FL, 1992.

[15] H. Federer. *Geometric measure theory.* Springer-Verlag New York Inc., New York, 1969.

[16] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17 – 40, 1976.

[17] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér.*, 9(R-2):41–76, 1975.

[18] Roger A. Horn and Charles R. Johnson. *Matrix analysis.* Cambridge University Press, Cambridge, second edition, 2013.

[19] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[20] George J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Math. J.*, 29:341–346, 1962.

[21] José Luis Morales and Jorge Nocedal. Remark on "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization" [mr1671706]. *ACM Trans. Math. Software*, 38(1):Art. 7, 4, 2011.

[22] Arkadi S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization.* John Wiley & Sons Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[23] Yu. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1, Ser. B):125–161, 2013.

[24] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[25] Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization.* Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[26] Jorge Nocedal and Stephen J. Wright. *Numerical optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[27] Zdzisław Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bull. Amer. Math. Soc.*, 73:591–597, 1967.

[28] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *U.S.S.R. Comput. Math. Math. Phys.*, 4(5):1–17, 1964.

[29] Boris T. Polyak. *Introduction to optimization.* Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.

[30] Simeon Reich. Weak convergence theorems for nonexpansive mappings in Banach spaces. *J. Math. Anal. Appl.*, 67(2):274–276, 1979.

[31] R. T. Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific J. Math.*, 17:497–510, 1966.

[32] R. T. Rockafellar. *Convex analysis.* Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

[33] R. Tyrrell Rockafellar. Convex functions, monotone operators and variational inequalities. In *Theory and Applications of Monotone Operators (Proc. NATO Advanced Study Inst., Venice, 1968)*, pages 35–65. Edizioni "Oderisi", Gubbio, 1969.

[34] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008. *Submitted to SIAM J. Optim. / available at* `http://www.csie.ntu.edu.tw/~b97058/tseng/papers/apgm.pdf`.