Université Grenoble Alpes
MSIAM 1st year

Numerical Optimization
Tuto 4: Proximal methods

L. Desbat & F. Iutzeler

A. the Proximity operator

In non-smooth optimization, that is when the objective function is not differentiable, the gradient may not be defined at each point. Instead, for any point $x \in \mathbb{R}$ and any convex function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, one can define a subdifferential $\partial g(x) \subset \mathbb{R}^n$ as

$$\partial g(x) = \{u \in \mathbb{R}^n |\ g(z) \geq g(x) + \langle u; z - x\rangle \text{ for all } z \in \mathbb{R}^n\}.$$

The optimality conditions and computation rules roughly translate.

However, the sub-gradient algorithm $x_{k+1} = x_k - \gamma_k g_k$ where $g_k \in\ \partial g(x_k)$ rely on a vanishing stepsize γ_k and is thus very slow in practice. In order to mend this case, a more evolved operator was introduced: its *proximity operator* is defined for some positive constant $\gamma > 0$ as

$$x = \mathbf{prox}_{\gamma g}(y) = \arg\min_{w\in\mathbb{R}^n} \left\{ \gamma g(w) + \frac{1}{2}\|w - y\|^2 \right\}. \tag{A.1}$$

Exercise 1 (First Properties)**.**

a. Justify that for a proper convex function g, this definition as an arg min indeed leads to a unique point. Would it still be the case if g was not convex?
b. This operation is sometimes called *implicit gradient*. Find an explanation why.
   Hint: Use First order optimality conditions.
c. Let $x = \mathbf{prox}_{\gamma g}(y)$ and $x' = \mathbf{prox}_{\gamma g}(y')$, show that
   $$\|x - x'\|^2 \leq \langle x' - y'; x - y\rangle.$$
   Hint: if $g_x \in \partial g(x)$ and $g_{x'} \in \partial g(x')$, the convexity of g gives $\langle x - x'; g_x - g_{x'}\rangle \geq 0$.
d. Deduce that
   $$\|x - x'\|^2 \leq \|y - y'\|^2 - \|(x - y) - (x' - y')\|^2$$
   and investigate the similarities with the gradient of a smooth function.

We showed that the proximity operator of a convex function has the same contraction properties of a gradient operation with step $1/L$ on an L-smooth convex function. Let us now investigate the related algorithm.

Exercise 2 (Proximal point algorithm)**.** The proximal point algorithm is simply obtained by successively applying the proximity operator of a function:

$$x_{k+1} = \mathbf{prox}_{\gamma g}(x_k)$$

a. Let $x^\star$ be a *fixed point* of g (we will suppose that such a point exists), that is $x^\star = \mathbf{prox}_{\gamma g}(x^\star)$. Show that $x^\star$ is a minimizer of g.
   Hint: Use First order optimality conditions.
b. Show that if $x = \mathbf{prox}_{\gamma g}(y)$, then $g(x) \leq g(y) - \frac{1}{2\gamma}\|x - y\|^2$.
   Hint: Use that for f μ-strongly convex and $x^\star$ the minimizer of f, then $f(x^\star) \leq f(y) - \frac{\mu}{2}\|x^\star - y\|^2$.
c. Conclude that the *Proximal Point Algorithm* converge to a minimizer of g.

Now that we have seen the optimization-wise interest of the proximity operator, let us compute it explicitly on some functions.

Exercise 3 (Proximity Operators of basic functions)**.** Compute the proximity operators of the following functions:

a. $g_1(x) = \|x\|_2^2$.
b. $g_2(x) = \iota_C(x)$ with $\iota_C(x) = 0$ if x belongs to convex set C and $+\infty$ elsewhere.
c. $g_3(x) = \|x\|_1$.
d. $g_4(x) = \|x\|_2$.

Unfortunately, in general, no explicit formulation can be found but i) the sub-optimization problems are now strongly convex and thus easier to solve; and more interestingly ii) proximity operator can be merged with other algorithms in order to minimize general functions. These algorithms are called *proximal algorithms* of which the most popular is the proximal gradient algorithm which mixes gradient and proximity operations.

B. the Proximal Gradient algorithm

Let us consider the *composite* optimization problem

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + g(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and convex; and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex. The *proximal gradient algorithm* writes

$$x_{k+1} = \mathbf{prox}_{\gamma g}\left(x_k - \gamma \nabla f(x_k)\right).$$

Exercise 4 (Analysis)**.**

a. Show that the fixed points of the iteration above are minimizers of F.
b. Connect the proximal gradient with the projected gradient algorithm.
c. Show that

$$F(x_{k+1}) \leq F(x_k) - \frac{(2 - \gamma L)}{2\gamma} \|x_{k+1} - x_k\|^2.$$

Hint: Use the descent lemmas for the gradient on smooth functions and the proximal point algorithm.

d. Give a range of stepsizes for which the sequence $F(x_k)$ converges as soon as minimizer exists.

Exercise 5 (Application)**.** The *lasso* problem is a regularized linear regression problem that writes as

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1$$

where A is a full rank $m \times n$ matrix and b is a size m vector.

a. Write the iterations for a proximal gradient algorithm. Which stepsize can be used?
b. The regularization $\lambda\|x\|_1$ is said to be *sparsity enforcing*, guess why.