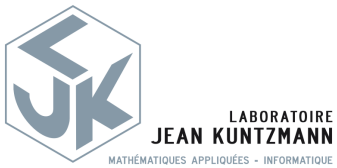

Lecture 9 : Nonsmooth Optimization - Proximal algorithms

This is the second lecture on non-smooth optimization, i.e., minimizing (or maximizing) a non-differentiable function. In particular, we will focus on proximal operators, a class of operations that allows one to work with "non-smooth objective functions as if they are smooth". After the introduction of their definition, we will proceed to two proximal algorithms:

1 Proximal operators - definitions and examples

Definition 1.1 (Lower semi-continuous function)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous if for any sequence $\{x_k\}_{k\in\mathbb{N}}$ that converges to a point $x \in \mathbb{R}^d$, we have:

$$\liminf_{k\to\infty} f(x_k) := \lim_{k\to\infty} \inf_{\ell\geq k} f(x_k) \geq f(x).$$

Example 1.2. Any continuous function is lower semicontinuous. Indicator functions of closed set is also lower semicontinuous because:

$$\mathbf{1}_A = \begin{cases} 0, & \text{if } x \in A, \\ +\infty, & \text{otherwise} \end{cases}.$$

For an example of a non lower-semicontinuous function, one can take the indicator function of an open ball.

Definition 1.3 (Proper function)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper if its domain $\texttt{dom}\, f := \{x \in \mathbb{R} \mid f(x) < \infty\}$ is non-empty.

Now, we can define well the proximal operator of a function f as:

Definition 1.4 (Proximal operator)**.** Let f be a convex, lower-semicontinuous and proper function. The proximal operator of $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ evaluated at x is defined as:

$$\text{prox}_f(x) = \text{argmin}_{y\in\mathbb{R}^d}\, g_x(y) := f(y) + \frac{1}{2}\|x - y\|^2.$$

For a function f, usually, we consider the proximal operators with αf with $\alpha > 0$. We write:

$$\text{prox}_{\alpha f}(x) = \text{argmin}_{y\in\mathbb{R}^d} \alpha f(y) + \frac{1}{2}\|x - y\|^2 = \text{argmin}_{y\in\mathbb{R}^d} f(y) + \frac{1}{2\alpha}\|x - y\|^2.$$

We first argue why proximal operator is well-defined.

Proposition 1.5 (Proximal operator is well-defined)**.** *If $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a convex, lower-semicontinuous and proper, then* $\text{prox}_f(x)$ *is well-defined, i.e., it is an application from* $\mathbb{R}^d \to \mathbb{R}^d$.

Proof. Since the function f is proper, there exists at least one point $z \in \mathbb{R}^d$ such that $f(z) < +\infty$. Therefore, $\inf_y g_x(y) < +\infty$.

Moreover, since f is convex, $f(\cdot) + \frac{1}{2}\|\cdot - x\|^2$ is 1-strongly convex. We accept without prove the following result (see tutorial for its detailed proof): if a function g is μ-strongly convex, then it is coercive, i.e., $\lim_{\|y\|\to+\infty} g(y) = +\infty$.

Therefore, the sequence $\{y_k\}_{k\in\mathbb{N}}$ that approximates the infimum of g_x is bounded, i.e., contained in a compact set $\mathcal{B}(0, M)$ for a constant $M > 0$ sufficiently large. Thus, one can assume that y_k converges to $\bar{y}$ (or we extract a converging sequence of y_k). By lower-semicontinuity of $g_x(y)$, we have:

$$\inf_y g_x(y) = \liminf_{k\to\infty} g_x(y_k) \geq g_x(\bar{y}).$$

Hence, the infimum is attained at $\bar{y}$, i.e., $g_x(\bar{y}) = \inf_y g_x(y)$. Thus, $\text{argmin}_y g_x \neq \emptyset$.

Finally, g_x is 1-strongly convex. Therefore, it cannot have more than one minimizer because otherwise:

$$g_x(tz + (1-t)y) \leq tg_x(z) + (1-t)g_x(y) - \frac{\mu}{2}t(1-t)\|z-y\|^2 < \min g_x(\cdot),$$

a contradiction. Hence, the proximal operator is well-defined. □

Example 1.6 (Examples of proximal operators)**.** Here are several famous examples:

1. If $f(x) = \mathbf{1}_S$, then $\text{prox}_f(x) = \text{proj}_S(x)$.

2. if $f(x) = -\log x$, then $\text{prox}_f(x)$ can be calculated as:

$$x \in \text{argmin}\, f(y) + \frac{1}{2}\|x-y\|^2 \implies -\frac{1}{y} + (y-x) = 0 \implies \text{prox}_f(x) = \frac{x + \sqrt{x^2+4}}{2}.$$

3. if $f(x) = \alpha|x|$, then:

$$\text{prox}_f(x) = \begin{cases} x + \alpha, & \text{if } x < -\alpha \\ 0, & \text{if } |x| \leq \alpha \\ x - \alpha, & \text{otherwise} \end{cases}.$$

Theorem 1.7 (Fixed points of proximal operators)**.** *Consider a lower-semicontinuous, proper and convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. A point $x^\star$ is a minimizer of f if and only if* $\text{prox}_{\alpha f}(x^\star) = x^\star, \forall \alpha > 0$.

Proof. Assume that $x^\star \in \text{argmin} f$. We have:

$$f(x^\star) \leq f(y) \leq f(y) + \frac{1}{2\alpha}\|y - x^\star\|^2 \implies x^\star = \text{prox}_{\alpha f}(x^\star).$$

If $x^\star = \text{prox}_{\alpha f}(x^\star)$, we have:

$$0 \in \partial\left(f(\cdot) + \frac{1}{2\alpha}\|\cdot - x^\star\|^2\right)(x^\star) = \partial f(x^\star) + \frac{1}{\alpha}(x^\star - x^\star) = \partial f(x^\star).$$

Notet that what we use is a bit different from the result stated for the subgradients of the sum of two functions (see tutorial exercise). Therefore, $x^\star$ is a minimizer of f. □

Remark 1.8. In the previous example, we can see that proximal operator can be seen as a generalized operation of the projection.

There is another interpretation: if we assume that f is C^2, then given a point $x \in \mathbb{R}^d$, compute the proximal operator $\text{prox}_{\alpha f}(x)$ is equivalent to solve the following equation:

$$h(y, \alpha) = y + \alpha \nabla f(y) = x.$$

Note that for the C^1 function $h = 0$ (because f is C^2) defines an implicit function $y(\alpha)$. In particular, with $\alpha = 0$, we have: $y(0) = x$ and $\text{Jac}_y h(x, 0) = \mathbf{I}, \text{Jac}_\alpha h(x, 0) = \nabla f(x)$. By the implicit function theorem,

$$\text{Jac}_\alpha y(0) = -\text{Jac}_y h(x, 0)^{-1} \text{Jac}_\alpha h(x, 0) = \nabla f(x).$$

Therefore, we have: $y(\alpha) = x - \alpha \nabla f(x) + o(\alpha)$, which is an approximation of gradient descent step. Thus, for α sufficiently small, $\text{prox}_{\alpha f}(x)$ is an approximation of gradient descent.

2 Proximal algorithms

We consider two proximal algorithms: *proximal point algorithm* (PPA) and *proximal gradient descent algorithm* (PGD).

2.1 Proximal point algorithm

PPA updates their iterations as:

$$x_{k+1} = \text{prox}_{\alpha_k f}(x_k). \tag{PPA}$$

Note that α_k behaves like the step-size in other algorithms that we already saw in previous lectures. The algorithm is queer in the sense that we minimze f plus a quadratic function, which is usually as challenging as f. However, this algorithm will serve as a naive algorithm to showcase several properties of proximal operators. In the next section, we will see a much more practical algorithm.

3 Proximal gradient descent method

The PGD algorithm assume that the objective function f satisfies:

$$f(x) = g(x) + h(x),$$

where g, h are convex functions, but g is smooth and h is prox-friendly, i.e., $\text{prox}_{\alpha h}$ can be calculated efficiently. PGD updates their iterations as:

$$x_{k+1} = \text{prox}_{\alpha_k h}(x - \alpha_k \nabla g(x_k)) \tag{PGD}$$

Note that this algorithm recovers several other algorithms that we already saw/knew:

1. If $g = 0$, we recover (PPA) for $f(x) = h(x)$.
2. If $h = 0$, $\text{prox}_{\alpha h}(x) = x, \forall \alpha$. Therefore, (PGD) reduces to gradient descent for $f(x) = g(x)$.
3. If $h = \mathbf{1}_{\mathcal{F}}$ where $\mathcal{F}$ is a convex feasible set, then (PPA) reduces to a projected gradient descent of the constrained optimization problem: $\min_{x \in \mathcal{F}} f(x)$.

We have the following theoretical guarantee on (PGD).

Theorem 3.1 (Theoretical guarantees for (PGD))**.** *Let $f = g + h$ where g is L-smooth and convex while h is only convex. We assume that f admits at least one minimzer $x^\star$. Consider the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by* (PGD) *with fixed $0 < \alpha_k = \alpha = \frac{1}{L}$, we have:*

$$f(x_k) - f(x^\star) \leq \frac{L}{2k}\|x_0 - x^\star\|^2, \forall k \in \mathbb{N}.$$

Proof. For step-size $\alpha > 0$, we define:

$$G_\alpha(x) := \frac{1}{\alpha}(x - \mathrm{prox}_{\alpha h}(x - \alpha\nabla g(x))).$$

Note that $G_\alpha(x)$ becomes $\nabla g(x)$ if $h(x)$ is a constant. It is the generalization of gradient in non-smooth decomposition settings because:

$$x_{k+1} = x_k - \alpha G_\alpha(x_k).$$

It is sufficient to prove the following:

$$f(x_{k+1}) \leq f(z) + \langle G_\alpha(x_k), x_k - z\rangle - \frac{\alpha}{2}\|G_\alpha(x_k)\|^2. \tag{1}$$

Indeed, assume that (1) holds, then for $z = x_k$, we have:

$$f(x_{k+1}) \leq f(x_k) - \frac{\alpha}{2}\|G_\alpha(x_k)\|^2 \leq f(x_k).$$

In particular, $\{f(x_k)\}_{k\in\mathbb{N}}$ is a non-increasing sequence. Moreover, apply (1) with $z = x^\star$, we have:

$$\begin{aligned} f(x_{k+1}) &\leq f(x^\star) + \langle G_\alpha(x_k), x_k - x^\star\rangle - \frac{\alpha}{2}\|G_\alpha(x_k)\|^2 \\ &\leq f(x^\star) + \frac{1}{2\alpha}\left(\|x_k - x^\star\|^2 - \|\underbrace{x_k - \alpha G_\alpha(x_k)}_{x_{k+1}} - x^\star\|^2\right). \end{aligned}$$

Telescoping the previous inequality yields:

$$\sum_{k=1}^{K}(f(x_k) - f(x^\star)) \leq \frac{1}{2\alpha}\left(\|x_0 - x^\star\|^2 - \|x_{K+1} - x^\star\|^2\right) \leq \frac{1}{2\alpha}\|x_0 - x^\star\|^2.$$

This yields the proof if one chooses $\alpha = 1/L$ and exploits the fact that $f(x_k)$ is a non-increasing sequence.

To obtain (1), we combine convexity with L-smoothness assumption. More specifically:

$$\begin{aligned} g(x_{k+1}) &\leq g(x_k) + \langle\nabla g(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \quad (L\text{-smoothness}) \\ &\leq g(z) - \langle\nabla g(x_k), z - x_k\rangle + \langle\nabla g(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \,(\text{convexity}) \\ &= g(z) + \langle\nabla g(x_k), x_{k+1} - z\rangle + \frac{L\alpha^2}{2}\|G_\alpha(x_k)\|^2. \end{aligned}$$

To continue, we exploit the property of proximal operator. Since $x_{k+1} = \mathrm{prox}_{\alpha h}(x_k - \alpha\nabla g(x_k))$, it implies:

$$\frac{1}{\alpha}\left((x_k - \alpha\nabla g(x_k)) - x_{k+1}\right) \in \partial h(x_{k+1}) \implies G_\alpha(x_k) = \delta + \nabla g(x_k), \delta \in \partial h(x_{k+1}).$$

Applying this inequality yields:

$$\begin{aligned} g(x_{k+1}) + \underbrace{h(z) + \langle \delta, x_{k+1} - z \rangle}_{\geq h(x_{k+1})} &\leq g(z) + h(z) + \langle \delta + \nabla g(x_k), x_{k+1} - z \rangle + \frac{L\alpha^2}{2} \|G_\alpha(x_k)\|^2 \\ &= f(z) + \langle G_\alpha(x_k), x_k - \alpha G_\alpha(x_k) - z \rangle + \frac{L\alpha^2}{2} \|G_\alpha(x_k)\|^2 \\ &= f(z) + \langle G_\alpha(x_k), x_k - z \rangle + \left(\frac{L\alpha^2}{2} - \alpha \right) \|G_\alpha(x_k)\|^2 \\ &\leq f(z) + \langle G_\alpha(x_k), x_k - z \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|^2. \end{aligned}$$

where the last line is due to our constraint on $\alpha \leq 1/L$. □

Corollary 3.2 (Theoretical guarantees for (PPA))**.** *Assume that the function f is convex and admits at least one minimizer $x^\star$, the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by* (PPA) *with fixed $\alpha_k = \alpha$ satisfies:*

$$f(x_k) - f(x^\star) \leq \frac{1}{2k\alpha} \|x_0 - x^\star\|^2.$$

Corollary 3.3 (Theoretical guarantees for projected gradient descent)**.** *Assume that the function f is convex, L-smooth and admits at least one minimizer $x^\star \in \mathcal{F}$ a convex feasible set, the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by projected gradient descent with fixed $\alpha_k = \alpha = \frac{1}{L}$ satisfies:*

$$f(x_k) - f(x^\star) \leq \frac{L}{2k} \|x_0 - x^\star\|^2.$$