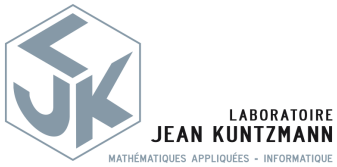

Lecture 8 : Nonsmooth Optimization - Subgradient Descent

This lecture focus on non-smooth optimization, i.e., minimizing (or maximizing) a non-differentiable function. The main challenge of nonsmooth optimization is that one no longer has access to the gradient and higher-order derivatives, thus, renders all algorithms that we studied from the beginning useless. There are, in general, two main approaches to tackle non-smoothness: subgradient and proximal operators. The former will be introduced in this lecture, while the latter will be the subject of the next class.

1 Non-smooth optimization - examples and challenges

In many real-world problems, our objective function is happened not to be smooth. Here are two examples:

Example 1.1 (Max of two functions)**.** A common situation in optimization related to manufacturing or scheduling tasks is:

$$\underset{x\in\mathbb{R}^d}{\text{Minimize}} \quad f(x) := \max(h_1(x), h_2(x)),$$

where $h_1, h_2 : \mathbb{R}^d \to \mathbb{R}$. Even if h_1 and h_2 are (infinitely) differentiable, their maximum is not necessarily differentiable. Take $h_1(x), h_2(x) = -x$ as an example.

Example 1.2. Assume that one has a dataset $\{(x_i, y_i) \mid i = 1, \dots, n\}, (x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ and $d \gg n$. We want to learn a vector $\theta \in \mathbb{R}^d$ such that $\theta^\top x_i \approx y_i$ and we want the vector to be sparse, i.e., it has many zero coefficients. The LASSO provides a convex formulation to perform such a task, by optimizing the following function:

$$\underset{\theta}{\text{Minimize}} \quad f(x) := \sum_{i=1}^{n} (\theta^\top x_i - y_i)^2 + \lambda \|\theta\|_1$$

where $\|\theta\|_1 = \sum_{\ell=1}^{d} |\theta_\ell|$ is the ℓ_1 norm of θ. The function f is not differentiable because $\|\cdot\|_1$ is not.

What happens when the function is not smooth? Perhaps the simplest non-smooth (but convex) functions that we can think of is $f(x) = |x|$ can tell us a lot about non-smooth optimization. In fact, f is differentiable everywhere except $x = 0$ because:

$$|x| = \begin{cases} -x, & \text{if } x < 0 \\ 0, & \text{if } x = 0 \\ x, & \text{otherwise} \end{cases} \quad \Longrightarrow \quad \nabla f(x) = \begin{cases} -1, & \text{if } x < 0 \\ \text{not defined}, & \text{if } x = 0 \\ 1, & \text{otherwise} \end{cases}.$$

Therefore, we can still perform "gradient descent", as long as our current iteration does not sit at 0. With fixed step-size $\alpha > 0$, our update rule becomes:

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

Typically, the dynamic generated by the above equation is: first, x_k will move towards 0 - the minimizer of f. Then, the sequence $\{x_k\}_{k\in\mathbb{N}}$ will oscillate around 0, and typically form a cycle of two points. Therefore, different from the smooth case where we already analysis, the iterates and the ojective function do not even converge to $x^\star$ or $f^\star = \inf f$. This suggests that a different approach, or more broadly, mathemtical framework is required to tackle this new situation. For that reason, we will study the notion of *subgradient*, a counterpart of gradient for non-smooth function.

2 Subdifferentiability and subgradient of a convex function

2.1 Definition and properties

Definition 2.1. Let $f : \mathbb{R}^n \to \bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. Define $\texttt{dom}\, f = \{x \in \mathbb{R}^n \mid f(x) < \infty\}$. An element $g \in \mathbb{R}^d$ is called a subgradient of f at $x \in \texttt{dom}\, f$ if:

$$f(y) \geq f(x) + \langle g, y - x\rangle, \forall y \in \mathbb{R}^n.$$

The set of subgradient of f at a point x is denoted as $\partial f(x) \subseteq \mathbb{R}^d$.

Example 2.2. Take $f(x) = |x|$. We have:

$$\partial f(x) = \begin{cases} -1, & \text{if } x < 0 \\ [-1, 1], & \text{if } x = 0 \\ 1, & \text{otherwise} \end{cases}.$$

Remark that in Example 2.2, $\partial f(x) = \{\nabla f(x)\}$ whereever f is differentiable. This is not a coincidence, due to the following result:

Theorem 2.3 (Subgradient at a differentiable point)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. If f differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.*

Proof. To prove that $\{\nabla f(x)\} \subseteq \partial f(x)$, we use a very similar argument as when we proved that for a C^1 function, one have:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x).$$

This will be left as an exercise. The rest of this proof will be dedicated to the other direction.

Assume that $g \in \partial f(x)$. By definition, we have:

$$f(x + d) - f(x) \leq g^\top d, \forall d \in \mathbb{R}^n.$$

Since the function f is differentiable at x, we also have:

$$\lim_{d\to 0} \frac{f(x + d) - f(x) - \nabla f(x)^\top d}{\|d\|} = 0.$$

This equation implies that for any $\epsilon > 0$, there exists $\delta > 0$ such that $\forall d, \|d\| \leq \delta$, we have:

$$\left\| \frac{f(x + d) - f(x) - \nabla f(x)^\top d}{\|d\|} \right\| \leq \epsilon,$$

or more precisely,

$$f(x+d) - f(x) \leq \nabla f(x)^\top d + \epsilon\|d\|, \forall d \in \mathbb{R}^n, \|d\| \leq \delta.$$

As a consequence, we obtain:

$$(\nabla f(x) - g)^\top d \leq \epsilon\|d\|, \forall d \in \mathbb{R}^n, \|d\| \leq \delta.$$

If $\nabla f(x) - g \neq 0$, one can choose $\epsilon < \|\nabla f(x) - g\|$, and $d = \delta\frac{\nabla f(x)-g}{\|\nabla f(x)-g\|}$, and obtain the contradiction. □

Therefore, if our function is differentiable, its subgradient reduces to its gradient. Note that the converse of Theorem 2.3 is also true, albeit the proof is omitted due to its complication.

Theorem 2.4 (Converse theorem of Theorem 2.3)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. If $\partial f(x) = \{\nabla f(x)\}$, then f differentiable at x.*

Next, we investigate the optimality condition for a non-smooth (but convex) function.

Theorem 2.5 (Optimality condition)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. A solution $x \in \mathbb{R}^d$ is a minimizer of f if and only if $0 \in \partial f(x)$.*

Proof. If x is a minimizer of f, then $0 \in \partial f(x)$ by definition. For the converse direction, it is immediately from the definition of subgradient, which is:

$$f(y) \geq f(x) + 0^\top(y - x) = f(x), \forall y \in \mathbb{R}^d.$$

□

2.2 Calculus with subgradient

Subgradient calculus share many gradient calculus, except for the chain rule. We consider the subgradient of f under certain operators in the following:

Theorem 2.6 (Subgradient calculus)**.** *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ and $f_1, \ldots, f_n : \mathbb{R}^n \to \mathbb{R}$ be convex functions. We have:*

$$\partial(\alpha f)(x) = \alpha\partial f(x)$$
$$\partial(f + g)(x) = \partial f(x) + \partial g(x)$$
$$\partial(f \circ (\mathbf{A}x + b)) = \mathbf{A}^\top\partial f(\mathbf{A}x + b)$$

These results (except the first) one are quite complicated to prove. We accept them without actually proving them.

3 Subgradient descent algorithm and its theoretical guarantees

In this section, we investigate the subgradient descent method and its theoretical guarantees. Similar to gradient descent, subgradient descent update the current iteration by following the opposite direction of its subgradient, i.e.,

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k). \tag{1}$$

If f is differentiable, we recover the gradient descent method (see Theorem 2.4). The main theoretical guarantee for subgradient descent is the following:

Theorem 3.1 (Theoretical guarantee for subgradient descent)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex. Assume that f admit at least one minimzer $x^\star$. We have:*

$$\min_{k=0,\dots,K-1} f(x_k) - f(x^\star) \leq \frac{1}{2} \cdot \frac{\|x_0 - x^\star\|_2^2 + \sum_{i=1}^2 \alpha_k^2 \|g_k\|^2}{\sum_{k=0}^{K-1} \alpha_k}.$$

Proof. We have:

$$\begin{aligned}
\|x_{k+1} - x^\star\|^2 &= \|x_k - x^\star - \alpha_k g_k\|^2 \\
&= \|x_k - x^\star\|^2 - 2\alpha_k g_k^\top (x_k - x^\star) + \alpha_k^2 \|g_k\|^2 \\
&\leq \|x_k - x^\star\|^2 - 2\alpha_k (f(x_k) - f(x^\star)) + \alpha_k^2 \|g_k\|^2.
\end{aligned}$$

Telescoping these inequalities with $k = 0, \dots, K-1$, we got:

$$2 \sum_{k=0}^{K-1} \alpha_k (f(x_k) - f(x^\star)) \leq \|x_0 - x^\star\|^2 - \|x_K - x^\star\|^2 + \sum_{k=0}^{K-1} \alpha_k^2 \|g_k\|^2.$$

Since:

$$2 \sum_{k=0}^{K-1} \alpha_k (f(x_k) - f(x^\star)) \geq 2 \left(\sum_{k=0}^{K-1} \alpha_k \right) \min_{k=0,\dots,K-1} (f(x_k - f(x^\star))),$$

divding both sides with $2 \left(\sum_{k=0}^{K-1} \alpha_k \right)$ yields the result. □

Corollary 3.2 (Guarantees for bounded subgradient function)**.** *Assume the same hypothesis as in Theorem 3.1, and that there exists a constant $G >$ such that $\|g\| \leq G, \forall g \in \partial f(x), \forall x$, then:*

$$\min_{k=0,\dots,K-1} f(x_k) - f(x^\star) \leq \frac{1}{2} \cdot \frac{\|x_0 - x^\star\|_2^2 + G^2 \sum_{k=0}^{K-1} \alpha_k^2}{\sum_{k=0}^{K-1} \alpha_k}.$$

We investigate several cases:

1. If $\alpha_k = \alpha > 0$, then we have:

$$\min_{k=0,\dots,K-1} f(x_k) - f(x^\star) \leq \frac{1}{2} \cdot \left(\frac{\|x_0 - x^\star\|_2^2}{K\alpha} + \alpha G^2 \right).$$

2. If $\alpha_k = \frac{1}{k}$, then we have:

$$\min_{k=0,\dots,K-1} f(x_k) - f(x^\star) \leq \frac{1}{2} \cdot \frac{\|x_0 - x^\star\|_2 + G^2 \pi / 4}{\ln k}.$$

3. If $\alpha = \frac{1}{\sqrt{k}}$, then we have:

$$\min_{k=0,\dots,K-1} f(x_k) - f(x^\star) \leq \frac{1}{2} \cdot \frac{\|x_0 - x^\star\|_2 + G^2 \ln k}{\sqrt{k}}.$$

4. Lower-bound of the upper-bound: Note that

$$\frac{\|x_0 - x^\star\|^2 + G^2 \sum_{k=0}^{K-1} \alpha_k^2}{\sum_{k=0}^{K-1} \alpha_k} \geq \frac{\|x_0 - x^\star\|^2}{\sum_{k=0}^{K-1} \alpha_k} + \frac{1}{k} \left(\sum_{k=0}^{K-1} \alpha_k \right) \geq 2 \cdot \frac{\|x_0 - x^\star\|^2}{\sqrt{k}} = O\left(\frac{1}{\sqrt{k}} \right).$$

Therefore, $O(1/\sqrt{k})$ is the best one can get if we follow this analysis.

Polyak's step-sizes With which step-size strategies can we achieve the optimal rate $O(1/\sqrt{k})$? One of them is known as Polyak's step-sizes. It will be described in the following:

Assume that we know the optimal value $f^\star = f(x^\star)$, we choose:

$$\alpha_k = \frac{f(x_k) - f(x^\star)}{\|g_k\|^2}.$$

Theorem 3.3 (Theoretical guarantee for Polyak's step-sizes)**.** *Assume the same hypothesis as in Theorem 3.1, and that there exists a constant $G >$ such that $\|g\| \leq G, \forall g \in \partial f(x), \forall x$, then:*

$$\min_{k=0,\dots,K-1} f(x_k) - f(x^\star) \leq \frac{G\|x_k - x^\star\|}{\sqrt{k}}.$$

Proof. We consider:

$$\begin{aligned}
\|x_{k+1} - x^\star\|^2 &= \|x_k - x^\star - \alpha_k g_k\|^2 \\
&= \|x_k - x^\star\|^2 - 2\alpha_k g_k^\top (x_k - x^\star) + \alpha_k^2 \|g_k\|^2 \\
&\leq \|x_k - x^\star\|^2 - 2\alpha_k (f(x_k) - f(x^\star)) + \alpha_k^2 \|g_k\|^2 \\
&= \|x_k - x^\star\|^2 - 2\alpha_k (f(x_k) - f(x^\star)) + \alpha_k^2 \|g_k\|^2 \\
&= \|x_k - x^\star\|^2 - \frac{(f(x_k) - f(x^\star))^2}{\|g_k\|^2} \\
&\leq \|x_k - x^\star\|^2 - \frac{(f(x_k) - f(x^\star))^2}{G^2}.
\end{aligned}$$

Therefore, telescoping and using the fact that $f(x_k) - f(x^\star) \geq \min_{\ell=0,\dots,K-1} f(x_\ell) - f(x^\star), \forall K > k$, we get:

$$\min_{k=0,\dots,K-1} f(x_k) - f(x^\star) \leq \frac{G\|x_k - x^\star\|}{\sqrt{k}}.$$

□