Lecture 17: Nonsmooth Optimization

Yudong Chen

All methods we have seen so far work under the assumption that the objective function f is smooth and in particular differentiable. In this lecture, we consider nonsmooth functions.

Examples include the absolute value $f(x) = |x|$ and more generally the ℓ_1 norm $f(x) = \|x\|_1 = \sum_{i=1}^d |x(i)| = \sum_{i=1}^d \max\{x(i), -x(i)\}$,[1] as well as the so-called Rectified Linear Unit (ReLU) $f(x) = \max\{x, 0\}$. In general, the maximum of (finitely many) smooth functions is a nonsmooth function.

1 Nonsmooth optimization

Consider the problem

$$\min_{x \in \mathcal{X}} f(x). \tag{P}$$

Assumptions:

- f is M-Lipschitz continuous for some $M \in (0, \infty)$, i.e.,

  $$|f(x) - f(y)| \le M \|x - y\|, \quad \forall x, y \in \mathrm{dom}(f),$$

  under some norm $\|\cdot\|$, whose dual norm is $\|\cdot\|_*$. Here, $\|\cdot\|$ can be an arbitrary norm. Later when we discuss the projected subgradient descent method, we will restrict to the ℓ_2 norm.

- f is convex and minimized by some $x^* \in \mathrm{argmin}_{x \in \mathcal{X}} f(x)$.

- $\mathcal{X} \subseteq \mathbb{R}^d$ is closed, convex and non-empty, and we can efficiently compute projection onto $\mathcal{X}$.

In this setting, f is not necessarily differentiable. But, it is *subdifferentiable*.

2 Subdifferentiability

Definition 1. We say that a convex function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is subdifferentiable at $x \in \mathrm{dom}(f)$ if there exists $g_x \in \mathbb{R}^d$ such that

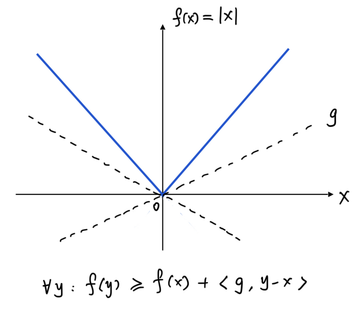
$$\forall y \in \mathbb{R}^d : \quad f(y) \ge f(x) + \langle g_x, y - x \rangle .$$

Such a vector g_x is called a *subgradient* of f at x. The set of all subgradients of f at x is called the *subdifferential* of f at x and denoted by $\partial f(x)$.

[1] In this lecture, $x(i)$ denotes the i-th coordinate of the vector x

Example 1. Let $f(x) = |x|$ be the absolute value function. Then

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



Exercise 1. What is $\partial f(x)$ for the function $f(x) = \max\{x, 0\}$?

It is easy to see that if f is in fact convex and differentiable, then $\partial f(x) = \{\nabla f(x)\}$ is a singleton.

2.1 Optimality condition

For a differentiable convex function f, we know from previous lectures that x^* is a minimizer if and only if $\nabla f(x^*) = 0$. The following theorem provides a generalization to potentially non-differentiable functions.

Theorem 1. *For a convex function f, a point x^* is a minimizer if and only if $0 \in \partial f(x^*)$.*

Proof. Observe that

$$\begin{aligned} & 0 \in \partial f(x^*) \\ \iff & f(y) \geq f(x^*) + \langle 0, y - x^* \rangle, \forall y && \text{by Definition 1} \\ \iff & x^* \text{ is a minimizer} \end{aligned}$$

□

2.2 Properties of subdifferential (optional)

The subdifferential has many important properties. We discuss a few of them below; see Wright-Recht Sections 8.2–8.4 for more.

Fact 1. *Every convex lower semicontinuous function is subdifferentiable everywhere on the interior its domain.*

Example 2. Let $I_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X}, \\ \infty, & x \notin \mathcal{X}, \end{cases}$ be the indicator function of a closed convex nonempty set $\mathcal{X}$. Then for each $x \in \mathcal{X}$, $\partial I_{\mathcal{X}}(x) = N_{\mathcal{X}}(x)$, where $N_{\mathcal{X}}(x)$ is the normal cone at x.

For smooth functions, the gradient has a linearity property: $\nabla(af + bh)(x) = a\nabla f(x) + b\nabla h(x)$. A similar property holds for the subdifferential.

Fact 2 (Linearity). *For any two convex functions f, h and any positive constants a, b, we have*

$$\partial(af + bh)(x) = a\partial f(x) + b\partial(x) = \{ag + bg' : g \in \partial f(x), g' \in \partial h(x)\}$$

for x in the interior of $\mathrm{dom}(f) \cap \mathrm{dom}(g)$.

Exercise 2. What is $\partial f(x)$ for the ℓ_1 norm $f(x) = \|x\|_1 := \sum_{i=1}^d |x_i|$?

With the above facts, we can unify the first-order optimality conditions for constrained and unconstrained problems:

$$\begin{aligned}
& -\nabla f(x) \in N_{\mathcal{X}}(x) & \\
\iff & -\nabla f(x) \in \partial I_{\mathcal{X}}(x) & \text{by Exercise 2} \\
\iff & 0 \in \nabla f(x) + \partial I_{\mathcal{X}}(x) & \\
\iff & 0 \in \partial\left(f + I_{\mathcal{X}}(x)\right). & \text{by Fact 2}
\end{aligned}$$

2.3 Lipschitz continuity

The theorem below relates the subgradients and Lipschitz continuity.

Theorem 2. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a convex function. f is M-Lipschitz-continuous w.r.t a norm $\|\cdot\|$ if and only if*

$$(\forall x \in \mathrm{dom}(f))\,(\forall g_x \in \partial f(x)): \quad \|g_x\|_* \le M.$$

Proof. $\Longrightarrow$ direction. Suppose f is M-Lipschitz. Fix any x and $g_x \in \partial f(x)$. By definition of subgradient and the LIpschitz property, we have

$$\langle g_x, u\rangle \le f(x+u) - f(x) \le M\|u\|, \quad \forall u,$$

hence

$$\begin{aligned}
\|g_x\|_* &= \max_{u:\|u\|=1} \langle g_x, u\rangle & \text{definition of dual norm} \\
&\le \max_{u:\|u\|=1} M\|u\| = M.
\end{aligned}$$

$\Longleftarrow$ direction. Assume that $(\forall x \in \mathrm{dom}(f))\,(\forall g_x \in \partial f(x)) : \|g_x\|_* \le M$. Then for all y:

$$\begin{aligned}
& f(y) \ge f(x) + \langle g_x, y - x\rangle \\
\Longrightarrow\ & f(x) - f(y) \le \langle g_x, x - y\rangle \le \|g_x\|_* \|x - y\| \le M\|x - y\|.
\end{aligned}$$

Switching the roles of x and y gives

$$f(y) - f(x) \le \langle g_y, y - x\rangle \le \|g_y\|_* \|y - x\| \le M\|y - x\|.$$

Combining gives $|f(x) - f(y)| \le M\|x - y\|$. □

3 Projected subgradient descent

For the rest of the lecture, we assume f is M-Lipschitz w.r.t. the *Euclidean ℓ_2 norm* $\|\cdot\|_2$.

We consider the following projected subgradient descent (PSubGD) method:

$$\begin{aligned}
x_{k+1} &= \operatorname*{argmin}_{y \in \mathcal{X}} \left\{ a_k \langle g_{x_k}, y - x_k\rangle + \frac{1}{2}\|y - x_k\|_2^2 \right\} \\
&= P_{\mathcal{X}}\left(x_k - a_k g_{x_k}\right),
\end{aligned}$$

where one may take any subgradient g_{x_k} from the set $\partial f(x_k)$, and $a_k > 0$ is the stepsize.

Without smoothness, we cannot get a descent lemma. In particular, it is not necessarily true that $f(x_{k+1}) \leq f(x_k)$. Nevertheless, we can still argue about convergence for the (weighted) *average of the iterate*s, defined as

$$x_k^{\text{out}} := \frac{1}{A_k} \sum_{i=0}^{k} a_i x_i,$$

where $A_k := \sum_{i=0}^{k} a_i$.

3.1 Convergence rate

We follow the proof strategy introduced in the Frank-Wolfe lecture and restated below.

General strategy:

1. Maintain an upper bound $U_k \geq f(x_k^{\text{out}})$ and a lower bound $L_k \leq f(x^*)$.
2. With $G_k := U_k - L_k \geq f(x_k^{\text{out}}) - f(x^*)$, show that

$$A_k G_k - A_{k-1} G_{k-1} \leq E_k \implies G_k \leq \frac{A_0 G_0 + \sum_{i=1}^{k} E_i}{A_k}.$$

3. Choose $\{a_k\}$ so that the above right hand decays to 0 fast.

By subdifferentiability and convexity, we have the lower bound

$$L_k := \frac{1}{A_k} \sum_{i=0}^{k} a_i \left(f(x_i) + \langle g_{x_i}, x^* - x_i \rangle \right) \leq f(x^*).$$

and the upper bound

$$U_k := \frac{1}{A_k} \sum_{i=0}^{k} a_i f(x_i) \geq f\left(\frac{1}{A_k} \sum_{i=0}^{k} a_i x_i \right) = f(x_k^{\text{out}}).$$

Hence $f(x_k^{\text{out}}) - f(x^*) \leq U_k - L_k =: G_k$. It follows that

$$\begin{aligned} A_k G_k - A_{k-1} G_{k-1} &= -a_k \langle g_{x_k}, x^* - x_k \rangle \\ &= a_k \langle g_{x_k}, x_{k+1} - x^* \rangle + a_k \langle g_{x_k}, x_k - x_{k+1} \rangle . \end{aligned}$$

Recall $x_{k+1} = \text{argmin}_{y \in \mathcal{X}} \left\{ a_k \langle g_{x_k}, y \rangle + \frac{1}{2} \|y - x_k\|_2^2 \right\} = P_{\mathcal{X}}(x_k - a_k g_{x_k})$. By 1st-order optimality condition of x_{k+1} (equivalently, the minimum principle for projection):

$$\langle x_{k+1} - x_k + a_k g_{x_k}, u - x_{k+1} \rangle \geq 0, \quad \forall u \in \mathcal{X}.$$

In particular, for $u = x^*$:

$$\begin{aligned} a_k \langle g_{x_k}, x_{k+1} - x^* \rangle &\leq \langle x_{k+1} - x_k, x^* - x_{k+1} \rangle \\ &= \frac{1}{2} \|x_k - x^*\|_2^2 - \frac{1}{2} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2} \|x_{k+1} - x_k\|_2^2 . \end{aligned}$$

It follows that

$$\begin{aligned}
A_k G_k - A_{k-1} G_{k-1} \leq & \frac{1}{2} \|x_k - x^*\|_2^2 - \frac{1}{2} \|x_{k+1} - x^*\|_2^2 \\
& - \frac{1}{2} \|x_{k+1} - x_k\|_2^2 + a_k \langle g_{x_k}, x_k - x_{k+1} \rangle \\
\leq & \frac{1}{2} \|x_k - x^*\|_2^2 - \frac{1}{2} \|x_{k+1} - x^*\|_2^2 \\
& - \frac{1}{2} \|x_{k+1} - x_k\|_2^2 + a_k M \|x_k - x_{k+1}\|_2 \qquad \text{Cauchy-Schwarz, } \|g_{x_k}\|_2 \leq M \\
\leq & \frac{1}{2} \|x_k - x^*\|_2^2 - \frac{1}{2} \|x_{k+1} - x^*\|_2^2 + \frac{a_k^2 M^2}{2}. \qquad \text{because } -\frac{p^2}{2} + pq \leq \frac{q^2}{2}.
\end{aligned} \tag{1}$$

On the other hand, we also have

$$A_0 G_0 = a_0 \langle g_{x_0}, x_0 - x^* \rangle \leq \frac{a_0^2 M^2}{2} + \frac{1}{2} \|x_0 - x^*\|_2^2 - \frac{1}{2} \|x_1 - x^*\|_2^2.$$

Summing over k and telescoping, we get

$$A_K G_K \leq \frac{1}{2} \|x_0 - x^*\|_2^2 + \sum_{k=0}^{K} \frac{a_K^2 M^2}{2},$$

hence

$$f(x_K^{\text{out}}) - f(x^*) \leq G_K \leq \frac{\|x_0 - x^*\|_2^2}{2A_K} + \frac{M^2 \sum_{k=0}^{K} a_k^2}{2A_K}. \tag{2}$$

It remains to choose the stepsize sequence $\{a_k\}$ to get a good convergence bound. Consider using a constant stepsize $a_k = C, \forall k$, then $A_K = C(K+1)$. Then

$$f(x_K^{\text{out}}) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2C(K+1)} + \frac{M^2 C}{2}.$$

The RHS is minimized when the two RHS terms are balanced:

$$\frac{\|x_0 - x^*\|_2^2}{C(K+1)} = \frac{M^2 C}{2} \qquad \iff \qquad C = \frac{\|x_0 - x^*\|_2}{M\sqrt{K+1}}.$$

We conclude that with the choice $a_k = \frac{\|x_0 - x^*\|_2}{M\sqrt{K+1}}, \forall k$, it holds that

$$f(x_K^{\text{out}}) - f(x^*) \leq \frac{M \|x_0 - x^*\|_2}{\sqrt{K+1}}.$$

This is slower than the $\frac{1}{K}$ rate for minimizing a smooth convex function.

3.2 Other considerations

The above choice of $\{a_k\}$ and the final bound require:

(i) knowing $\|x_0 - x^*\|_2$;

(ii) fixing the total number of iterations K before setting $\{a_k\}$.

To address issue (i) , note that we usually know (an upper bound of) the diameter of $\mathcal{X}$, i.e., $D := \max_{x,y\in\mathcal{X}} \|x-y\|_2$. If D is finite, then $\|x_0 - x^*\| \le D$. In this case we can choose $a_k = \frac{D}{M\sqrt{K+1}}, \forall k$. Plugging into (2), we get

$$f(x_K^{\text{out}}) - f(x^*) \le \frac{D^2 + M^2 \sum_{k=0}^{K} a_k^2}{2A_K} \le \frac{DM}{\sqrt{K+1}}.$$

To address issue (ii), we could instead choose $a_k = \frac{D}{M\sqrt{k+1}}$, which gives a so-called "anytime algorithm" with the slightly worse bound

$$f(x_K^{\text{out}}) - f(x^*) = O\left(\frac{DM\log K}{\sqrt{K+1}}\right).$$

Finally, if D is unknown or unbounded, then we can use $a_k = \frac{1}{\sqrt{k+1}}$. Note that this choice does not require knowledge of the Lipschitz M either. In this case we have

$$f(x_K^{\text{out}}) - f(x^*) = O\left(\frac{\left(\|x_0 - x^*\|_2^2 + M^2\right)\log K}{\sqrt{K+1}}\right).$$

4 Lower bounds (optional)

The $O\left(\frac{1}{\sqrt{K}}\right)$ rate above is order-wise optimal for first-order methods in a sense similar to the optimality of AGD. Consider a first-order method that generates iterates $x_1, x_2, x_3...$ satisfying $x_1 = 0$ and

$$x_{k+1} \in \operatorname{Lin}\{g_1, \dots g_k\}, \quad \forall k \ge 1,$$

where $g_k \in \partial f(x_k)$ is an arbitrary subgradient at x_k. Note that the iterates x_k and x_k^{out} of PSubGD both satisfy this assumption. We have the following lower bound.

Theorem 3. *There exists a convex and M-Lipschitz function f such that for any first-order method satisfying the above assumption, we have*

$$\min_{1\le k\le K} f(x_k) - f(x^*) \ge \frac{M\|x^* - x_1\|_2}{2(1+\sqrt{K})}.$$

Proof. Consider a function $f : \mathbb{R}^K \to \mathbb{R}$ defined as

$$f(x) = \gamma \max_{1\le i\le K} x(i) + \frac{1}{2}\|x\|_2^2,$$

where $\gamma = \frac{M\sqrt{K}}{1+\sqrt{K}}$. Then

$$\partial f(x) = x + \gamma \operatorname{conv}\left\{e_i : i \in \underset{1\le j\le K}{\operatorname{argmax}}\, x(j)\right\},$$

where $e_i \in \mathbb{R}^K$ is the ith standard basis vector and $\operatorname{conv}\{\cdot\}$ denotes the convex hull.

A minimizer of f is x^* with $x^*(i) = -\frac{\gamma}{K}, \forall i$, because $0 \in \partial f(x^*)$ (Theorem 1). Hence

$$\|x^* - x_1\|_2 = \|x^*\|_2 = \frac{\gamma}{\sqrt{K}} = \frac{M}{1+\sqrt{K}} \tag{3}$$

and the optimal value is

$$f(x^*) = -\frac{\gamma^2}{K} + \frac{1}{2}\frac{\gamma^2}{K} = -\frac{M^2}{2(1+\sqrt{K})^2}.$$

Note that if $\|x\|_2 \leq \frac{\gamma}{\sqrt{K}}$, then $\|g\|_2 \leq \frac{\gamma}{\sqrt{K}} + \gamma = M, \forall g \in \partial f(x)$. By Theorem 2 we know that f is M-Lipschitz on the ball $\left\{x : \|x\|_2 \leq \frac{\gamma}{\sqrt{K}}\right\}$.

Under our assumption for first-order methods, it is easy to see that

$$x_k \in \text{Lin}\left\{g_1, \dots g_{k-1}\right\} \subseteq \text{Lin}\left\{e_1, \dots, e_{k-1}\right\}.$$

Therefore, for all $k \leq K$, we have $x_k(K) = 0$ and thus $f(x_k) \geq 0$. It follows that the optimality gap is lower bounded as

$$f(x_k) - f(x^*) \geq 0 - \frac{M^2}{2(1+\sqrt{K})^2} = \frac{M\left\|x^* - x_1\right\|_2}{2(1+\sqrt{K})},$$

where the last step follows from (3). $\square$