CS 726: Nonsmooth Convex Optimization

Jelena Diakonikolas

Fall 2023

All of the methods we have seen so far were working under the assumption that the objective function f were smooth. However, such an assumption does not hold for many interesting cases, such as piecewise linear functions. As we discussed at the very beginning of this class, *some* assumptions that enforce certain regularity of the objective are necessary. In this lecture, we will learn how to optimize objective functions that are nonsmooth, but are convex and Lipschitz continuous. We will consider possibly constrained settings, assuming we have access to a projection operator for the feasible set.

1 Setup

We begin by describing the setup we will be working with. Recall that our basic optimization problem is

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

For this lecture, we will additionally be assuming that:

- The norm associated with the space $\|\cdot\|$ is arbitrary but fixed and its dual norm is $\|\cdot\|_*$;
- f is M-Lipschitz continuous w.r.t. $\|\cdot\|$, convex, and minimized by some $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$;
- $\mathcal{X}$ is closed and convex. We have access to efficiently computable projections for $\mathcal{X}$.

2 Subdifferentiability

We are not assuming anymore that f is differentiable, so we cannot make use of the gradient of f in the same way as we did for previous methods we wrote down and analyzed. Yet, there is still a notion of supporting hyperplanes for Lipschitz continuous convex functions that allows us to estimate the value of $f(\mathbf{x}^*)$ when we analyze convergence of algorithms. These supporting hyperplanes use a generalization of gradient vectors, called subgradients, that exist for any continuous function, at any point.

Definition 2.1. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is subdifferentiable at $\mathbf{x} \in \operatorname{dom} f(\mathbf{x})$ if there exists a vector $\mathbf{g}_\mathbf{x} \in \mathbb{R}^d$ such that for all $\mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}_\mathbf{x}, \mathbf{y} - \mathbf{x} \rangle . \tag{1}$$

Vector $\mathbf{g}_\mathbf{x}$ is then said to be a subgradient of f at $\mathbf{x}$. The set of all subgradients of f at $\mathbf{x}$ is called the subdifferential set of f at $\mathbf{x}$ and is denoted by $\partial f(\mathbf{x})$.

The subdifferential operator ∂ has the following property: for any two convex functions f and h and any positive constants $a, b : \partial(af + bh)(\mathbf{x}) = a\partial f(\mathbf{x}) + b\partial h(\mathbf{x})$ for $\mathbf{x}$ in the interior of the domain of f and h.

The reason that subdifferentiability is useful comes from the following fact, which we state without a proof.

Fact 2.2. *Every convex lower semicontinuous function is subdifferentiable everywhere on the interior of its domain.*

A specific example that is useful to know allows us to view constrained and unconstrained problems as one and the same.

Example 2.3. Let

$$I_{\mathcal{X}} = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathcal{X} \\ +\infty & \text{if } \mathbf{x} \notin \mathcal{X} \end{cases}$$

be the indicator function of a closed convex set. We saw in one of the first lectures that this function is lower semi-continuous. The subdifferential of this function has a very simple expression: $\forall \mathbf{x} \in \mathcal{X} : \partial I_{\mathcal{X}}(\mathbf{x}) = N_{\mathcal{X}}(\mathbf{x})$, where $N_{\mathcal{X}}(\mathbf{x})$ is the normal cone of $\mathcal{X}$ at $\mathbf{x}$.

This is example is particularly interesting once we go back to the stationarity (or first-order optimality) condition for convex problems with differentiable objectives f :

$$\begin{aligned} & -\nabla f(\mathbf{x}) \in N_{\mathcal{X}}(\mathbf{x}) \\ \Leftrightarrow \quad & \mathbf{0} \in \nabla f(\mathbf{x}) + N_{\mathcal{X}}(\mathbf{x}) \\ \Leftrightarrow \quad & \mathbf{0} \in \partial(f + I_{\mathcal{X}})(\mathbf{x}). \end{aligned}$$

The last condition can be written also for functions that are not differentiable, but only subdifferentiable. You can prove as an exercise that for a convex l.s.c. function f and a closed convex set $\mathcal{X}$, $\mathbf{0} \in \partial(f + I_{\mathcal{X}})(\mathbf{x}^*)$ if and only if $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$.

The following characterization nonsmooth Lipschitz continuous convex functions is particularly useful for analyzing algorithms.

Theorem 2.4. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a convex lower semicontinuous function. Then f is M-Lipschitz continuous w.r.t. a norm $\|\cdot\|$ on the interior of its domain if and only if*

$$(\forall \mathbf{x} \in \operatorname{int} \operatorname{dom}(f))(\forall \mathbf{g}_x \in \partial f(\mathbf{x})) : \quad \|\mathbf{g}_{\mathbf{x}}\|_* \leq M. \tag{2}$$

Proof. Suppose first that f is M-Lipschitz continuous. Fix $\mathbf{x} \in \operatorname{int} \operatorname{dom}(f)$ and $\mathbf{g}_{\mathbf{x}} \in \partial f(\mathbf{x})$ (by Fact 2.2, $\partial f(\mathbf{x})$ is non-empty, so at least one such $\mathbf{g}_{\mathbf{x}}$ exists) and define $\mathbf{y} = \mathbf{x} + \operatorname{argmax}_{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| \leq 1} \langle \mathbf{u}, \mathbf{g}_x \rangle$. Then, by the definition of the dual norm, we have $\langle \mathbf{g}_{\mathbf{x}}, \mathbf{y} - \mathbf{x} \rangle = \max_{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| \leq 1} \langle \mathbf{u}, \mathbf{g}_{\mathbf{x}} \rangle = \|\mathbf{g}_{\mathbf{x}}\|_*$. Further, by the definition of a subgradient,

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{g}_{\mathbf{x}}, \mathbf{y} - \mathbf{x} \rangle = \|\mathbf{g}_{\mathbf{x}}\|_*. \tag{3}$$

On the other hand, as f is M-Lipschitz continuous, we have

$$f(\mathbf{y}) - f(\mathbf{x}) \leq M \|\mathbf{y} - \mathbf{x}\|. \tag{4}$$

Hence, combining (3) and (4), we conclude that $\|\mathbf{g}_{\mathbf{x}}\|_* \leq M$.

For the other direction, now assume that (2) holds. Fix any $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(f)$. By the definition of a subgradient, we have that there exist $\mathbf{g}_{\mathbf{x}} \in \partial f(\mathbf{x})$ and $\mathbf{g}_{\mathbf{y}} \in \partial f(\mathbf{y})$

$$f(\mathbf{y}) - f(\mathbf{x}) \leq \langle \mathbf{g}_{\mathbf{y}}, \mathbf{y} - \mathbf{x} \rangle \leq \|\mathbf{g}_{\mathbf{y}}\|_* \|\mathbf{y} - \mathbf{x}\| \leq M \|\mathbf{y} - \mathbf{x}\|$$

and

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \mathbf{g}_{\mathbf{x}}, \mathbf{x} - \mathbf{y} \rangle \leq \|\mathbf{g}_{\mathbf{x}}\|_* \|\mathbf{y} - \mathbf{x}\| \leq M \|\mathbf{y} - \mathbf{x}\|.$$

The last two inequalities lead to $|f(\mathbf{y}) - f(\mathbf{x})| \leq M \|\mathbf{y} - \mathbf{x}\|$, completing the proof that f is M-Lipschitz continuous, and thus completing the proof of the theorem. □

3 Projected Subgradient Descent

In this section, we assume that f is M-Lipschitz continuous w.r.t. the Euclidean norm $\|\cdot\|_2$ and minimized by some $\mathbf{x}^* \in \mathcal{X}$ on the feasible set $\mathcal{X}$. Since we are not assuming that the function f is smooth (or even differentiable!) anymore, we will not be able to prove a sufficient decrease property. Hence, the method we obtain now will **not** be a descent method anymore. In particular, projected (sub)gradient descent, defined by

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{y} \in \mathcal{X}} \left\{ a_k \langle \mathbf{g}_{\mathbf{x}}, \mathbf{y} \rangle + \frac{1}{2} \|\mathbf{y} - \mathbf{x}_k\|_2^2 \right\}, \tag{PSubGD}$$

where $a_k > 0$ is the step size, is **not** a descent method.

The question that arises is: how are we even able to argue about convergence in this case? Subdifferentiability turns out to be a crucial concept here, as it allows us to still construct a lower bound on $f(\mathbf{x}^*)$, similar to what we had done for all the descent methods (including projected gradient descent) that we have seen so far:

$$f(\mathbf{x}^*) \geq L_k := \frac{1}{A_k} \sum_{i=0}^{k} a_i \big(f(\mathbf{x}_i) + \langle \mathbf{g}_{\mathbf{x}_i}, \mathbf{x}^* - \mathbf{x}_i \rangle \big), \tag{5}$$

where $\mathbf{g}_{\mathbf{x}_i} \in \partial f(\mathbf{x}_i)$.

For the output point $\mathbf{x}_k^{\text{out}}$ and the upper bound on $f(\mathbf{x}_k^{\text{out}})$, in this case we choose

$$\mathbf{x}_k^{\text{out}} = \frac{1}{A_k} \sum_{i=0}^{k} a_i \mathbf{x}_i$$

and

$$U_k := \frac{1}{A_k} \sum_{i=0}^{k} a_i f(\mathbf{x}_i) \tag{6}$$

Observe that U_k defined by (6) is a valid upper bound, as, due to Jensen's inequality, $U_k \geq f(\mathbf{x}_k^{\text{out}})$.

Defining, as before, $G_k := U_k - L_k$, we clearly have

$$f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*) \leq G_k = \frac{1}{A_k} \sum_{i=0}^{k} a_i \left\langle \mathbf{g}_{\mathbf{x}_i}, \mathbf{x}^* - \mathbf{x}_i \right\rangle. \tag{7}$$

Based on the definition of G_k, both the (scaled) initial gap $A_0 G_0$ and the change in the (scaled) gap $A_k G_k - A_{k-1} G_{k-1}$ take the same form: $-a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_k \right\rangle$, so to bound both, it suffices to bound $-a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_k \right\rangle$ for an arbitrary $k \geq 0$.

Let $h_k(\mathbf{y}) = a_k \left\langle \mathbf{g}_{\mathbf{x}}, \mathbf{y} \right\rangle + \frac{1}{2}\|\mathbf{y} - \mathbf{x}_k\|_2^2$, so that $\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{y} \in \mathcal{X}} h_k(\mathbf{x})$. Since h_k is a quadratic function, we have

$$h_k(\mathbf{y}) = h_k(\mathbf{x}) + \langle \nabla h_k(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

Since h_k is minimized by $\mathbf{x}_{k+1}$, for all $\mathbf{y} \in \mathcal{X}$, $\langle \nabla h_k(\mathbf{x}_{k+1}), \mathbf{y} - \mathbf{x}_{k+1} \rangle \geq 0$. Hence,

$$h(\mathbf{x}^*) \geq h(\mathbf{x}_{k+1}) + \frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2.$$

Using the definition of h_k, the last inequality can be written as

$$a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_{k+1} \right\rangle \geq \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 - \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 + \frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2. \tag{8}$$

Plugging (8) into $a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_k \right\rangle = a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_{k+1} \right\rangle + a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}_{k+1} - \mathbf{x}_k \right\rangle$, we get

$$\begin{aligned}
-a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_k \right\rangle &\leq -a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}_{k+1} - \mathbf{x}_k \right\rangle - \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 + \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 - \frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \\
&\leq a_k \|\mathbf{g}_{\mathbf{x}_k}\|_2 \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 - \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 + \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 - \frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \\
&\leq a_k M \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 - \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 + \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 - \frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \\
&\leq -\frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 + \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 + \frac{{a_k}^2 M^2}{2},
\end{aligned}$$

where the second inequality is by Cauchy-Schwarz, the third inequality is by Theorem 2.4, and the final inequality follows from Young's inequality. Summing the last inequality from $k = 0$ to some $K \geq 0$, we now have

$$\begin{aligned}
A_K G_K &= -\sum_{k=0}^{K} a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_k \right\rangle \\
&\leq \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 + \sum_{k=0}^{K} \frac{{a_k}^2 M^2}{2}.
\end{aligned}$$

Therefore, we can conclude that

$$f(\mathbf{x}_K^{\text{out}}) - f(\mathbf{x}^*) \leq \frac{\frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 + \sum_{k=0}^{K} \frac{{a_k}^2 M^2}{2}}{A_K} \tag{9}$$

and to obtain the final convergence bound it remains to choose the sequence $\{a_k\}_{k\geq 0}$.

Consider first choosing $a_k = C$ for some constant C. Then $A_K = (K+1)C$ and the right-hand side of (9) becomes $\frac{\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{2(K+1)C} + \frac{CM^2}{2}$. This expression is minimized for $C = \frac{\|\mathbf{x}_0-\mathbf{x}^*\|_2}{M\sqrt{K+1}}$ (why?), leading to

$$f(\mathbf{x}_K^{\text{out}}) - f(\mathbf{x}^*) \leq \frac{M\|\mathbf{x}^* - \mathbf{x}_0\|_2}{\sqrt{K+1}}. \tag{10}$$

This bound is unimprovable; however, it requires both knowing $\|\mathbf{x}_0 - \mathbf{x}^*\|_2$ and fixing the number of iterations in advance to set the correct step size a_k.

The former issue can be resolved in two ways. In bounded constrained settings, the diameter D of the feasible set is often known, in which case we can choose $a_k = \frac{D}{M\sqrt{K+1}}$, leading to the bound

$$f(\mathbf{x}_K^{\text{out}}) - f(\mathbf{x}^*) \leq \frac{MD}{\sqrt{K+1}}. \tag{11}$$

For the case that we do not want to fix the number of iterations in advance, we can choose $a_k = \frac{D}{M\sqrt{k+1}}$, which leads to a slightly worse bound:

$$f(\mathbf{x}_K^{\text{out}}) - f(\mathbf{x}^*) = O\Big(\frac{MD\log(K+1)}{\sqrt{K+1}}\Big). \tag{12}$$

Finally, if the set $\mathcal{X}$ is unbounded or its diameter is unknown, then we could set $a_k = \frac{1}{\sqrt{k+1}}$, leading to

$$f(\mathbf{x}_K^{\text{out}}) - f(\mathbf{x}^*) = O\Big(\frac{(\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 + M^2)\log(K+1)}{\sqrt{K+1}}\Big). \tag{13}$$

The last bound still reduces at rate $\frac{\log(k)}{k}$; however, the dependence on the problem parameters $\|\mathbf{x}^* - \mathbf{x}_0\|_2$ and M is worse than in previous bounds.

Exercises

1. Let f be an M-Lipschitz continuous μ-strongly convex function and consider running (PSubGD), where, as usual, we assume that $\mathcal{X}$ is closed and convex. Use Eq. (8) to argue that

$$\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 \leq (1 - 2a_k\mu)\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 + {a_k}^2 M^2. \tag{14}$$

Take $a_i = \frac{1}{2\mu(i+1)}$, for $i \geq 0$. Use (14) to argue that, $\forall k \geq 1$,

$$\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 \leq \frac{1}{k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 + \frac{M^2}{4\mu^2} \cdot \frac{\log(k+1)+1}{k+1}. \tag{15}$$

How would the right-hand side of (15) change if instead you chose $a_i = \frac{i/2+1}{2\mu(i+1)}$ for all $i \geq 1$?

How many iterations would you need to take to guarantee that $\|\mathbf{x}^* - \mathbf{x}_k\|_2 \leq \epsilon$, for a given $\epsilon > 0$ in either of the two cases?

2. Let f be an M-Lipschitz continuous convex function and suppose that on a given set $\mathcal{X}$, $\min_{\mathbf{x}} f(\mathbf{x}) = 0$, and, further, f is μ-strongly convex for some $\mu > 0$.

Using (9), propose a step size that for a given number of iterations leads to the following guarantee:

$$f(\mathbf{x}_k^{\text{out}}) \leq \frac{M\sqrt{f(\mathbf{x}_0)}}{\sqrt{\mu}\sqrt{k+1}}.$$

How many iterations are needed to ensure that $f(\mathbf{x}_k^{\text{out}}) \leq \frac{f(\mathbf{x}_0)}{2}$? Propose a restarting algorithm that ensures that on each restart r the function value at the restart output point $\mathbf{x}^r$ satisfies $f(\mathbf{x}^r) \leq \frac{1}{2^r} f(\mathbf{x}_0)$, where $\mathbf{x}_0$ is the initial point for the entire restarted algorithm. Argue that the total number of vector operations that your algorithm takes until it finds a point with function value at most ϵ is (at most)

$$\text{const.} \cdot \frac{M^2}{\mu f(\mathbf{x}_0)} \sum_{i=1}^{\log_2(\frac{f(\mathbf{x}_0)}{\epsilon})} 2^i = O\Big(\frac{M^2}{\mu\epsilon}\Big).$$