DSCC/CSC 435 & ECE 412 Optimization for Machine Learning	Lecture 11
Smoothing Techniques	
Lecturer: Jiaming Liang	*October 31, 2023*

1 Smoothing

Recall that the convex nonsmooth optimization complexity is

$$\mathcal{O}\left(\frac{M^2 d_0^2}{\varepsilon^2}\right),$$

which is optimal for a black-box model, i.e., unstructured problems. However, if we know some structure of the problem, the complexity could be improved by taking advantage of this structural information. In this lecture, we explore the smoothable structure in nonsmooth optimization. The goal is to improve the complexity form $\mathcal{O}(\varepsilon^{-2})$ to $\mathcal{O}(\varepsilon^{-1})$.

Consider

$$\min_{x \in \mathbb{R}^n} \{\phi(x) := f(x) + h(x) + \theta(x)\} \tag{1}$$

where f is convex, differentiable everywhere and L-smooth, h is closed, convex and simple, and θ is convex (but not simple) and smoothable.

Definition 1. *A function θ is (C_1, C_2)-smoothable if there exist a scalar $\mu > 0$ and a convex and differentiable function θ_μ such that*

- $\theta_\mu(x) \le \theta(x) \le \theta_\mu(x) + C_2\mu$;
- $\nabla\theta_\mu$ *is* $\frac{C_1}{\mu}$*-Lipschitz continuous.*

For some $\mu > 0$, we consider an auxiliary problem

$$\min_{x \in \mathbb{R}^n} \{\phi_\mu(x) := f(x) + h(x) + \theta_\mu(x)\}.$$

Let $f_\mu(x) = f(x) + \theta_\mu(x)$, then we know f_μ is convex , differentiable everywhere, and ∇f_μ is $\left(L + \frac{C_1}{\mu}\right)$-Lipschitz continuous. Apply the ACG method with FISTA update to solve the auxiliary problem.

Algorithm 1 FISTA

Input: Initial point $x_0 \in \operatorname{dom} h$, $L_\mu = L + \frac{C_1}{\mu}$, set $y_0 = x_0$, $A_0 = 0$.

for $k \geq 0$ **do**

Step 1. Compute

$$a_k = \frac{1 + \sqrt{1 + 4L_k A_k}}{2L_k}, \quad A_{k+1} = A_k + a_k, \quad \tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_{k+1}} \tag{2}$$

Step 2. Compute x_{k+1} and y_{k+1}

$$\begin{aligned} y_{k+1} &= \operatorname{argmin}\left\{ \ell_{f_\mu}(x; \tilde{x}_k) + h(x) + \frac{L_\mu}{2}\|x - \tilde{x}_k\|^2 : x \in \mathbb{R}^n \right\}, \\ x_{k+1} &= \frac{A_{k+1}}{a_k} y_{k+1} - \frac{A_k}{a_k} y_k. \end{aligned}$$

end for

Theorem 1. *If $\mu = \frac{\varepsilon}{2C_2}$, then FISTA finds y_k such that $\phi(y_k) - \phi_* \leq \varepsilon$ in at most*

$$\mathcal{O}\left(\|x_0 - x_*\| \left(\sqrt{\frac{L}{\varepsilon}} + \frac{\sqrt{C_1 C_2}}{\varepsilon} \right) \right)$$

iterations.

Proof. In view of the first condition in Definition 1, we have for every $x \in \operatorname{dom} h$,

$$\phi_\mu(x) \leq \phi(x) \leq \phi_\mu(x) + C_2\mu.$$

Using this inequality and Theorem 1 of Lecture 7, we have

$$\begin{aligned} \phi(y_k) - \phi_* &= \phi(y_k) - \phi_\mu(y_k) + \phi_\mu(y_k) - \phi_\mu(x_*) + \phi_\mu(x_*) - \phi_* \\ &\leq C_2\mu + \phi_\mu(y_k) - \phi_\mu(x_*) + 0 \\ &= \frac{\varepsilon}{2} + \frac{2L_\mu \|x_0 - x_*\|^2}{k^2} \\ &= \frac{\varepsilon}{2} + 2\left(L + \frac{C_1}{\mu}\right) \frac{\|x_0 - x_*\|^2}{k^2}, \end{aligned}$$

where the last identity is due to the definition of L_μ in Algorithm 1. To find ε-solution, the complexity is

$$\mathcal{O}\left(\|x_0 - x_*\| \left(\sqrt{\frac{L}{\varepsilon}} + \frac{\sqrt{C_1 C_2}}{\varepsilon} \right) \right).$$

□

Example

Consider the saddle point problem

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} f(x)+h(x)+\langle Ax,y\rangle - g(y)$$

or

$$\min_{x\in\mathbb{R}^n} f(x)+h(x)+\max_{y\in\mathbb{R}^m}\langle Ax,y\rangle - g(y),$$

where f is convex, differentiable everywhere and L-smooth, h is closed, convex and simple, g is a closed and convex fucntion, and $\operatorname{dom} g$ is bounded.

Define

$$\theta(x)=\max_{y\in\mathbb{R}^m}\langle Ax,y\rangle - g(y) = g^*(Ax),\quad A\in\mathbb{R}^{m\times n}.$$

Then, the problem is in the form of (1) and $\theta(x)$ is convex but not necessarily smooth.

Lemma 1. *Assume $\tilde{g}$ is a closed and μ-strongly convex function, then*

$$\tilde{\theta}(z)=(\tilde{g})^*(z)=\sup_{y\in\mathbb{R}^m}\langle z,y\rangle-\tilde{g}(y)$$

is convex and differentiable everywhere, and $\nabla\tilde{\theta}(z)=y(z)$. Moreover, $\nabla\tilde{\theta}$ is $\frac{1}{\mu}$-Lipschitz continuous.

Proof. See Lecture 5. □

In our setup, we let

$$\tilde{g}(y)=g(y)+\frac{\mu}{2}\|y-y_0\|$$

for some $y_0\in\operatorname{dom} h$ and

$$\tilde{\theta}_\mu(z)=\sup_{y\in\mathbb{R}^m}\left\{\langle z,y\rangle-g(y)-\frac{\mu}{2}\|y-y_0\|^2\right\}.$$

Then,

$$\nabla\tilde{\theta}_\mu(z)=y_\mu(z)$$

and it is $\frac{1}{\mu}$-Lipschitz continuous. Now let

$$\theta_\mu(x)=\tilde{\theta}_\mu(Ax),$$

then θ_μ is $\frac{\|A\|^2}{\mu}$-Lipschitz continuous. So we have $C_1=\|A\|^2$. Moreover, we have for every $x\in\operatorname{dom} h$,

$$\theta_\mu(x)-\theta(x)\le\frac{\mu}{2}\max_{y\in\operatorname{dom} g}\|y-y_0\|^2,$$

so

$$C_2=\frac{1}{2}\operatorname{Diam}(g)^2.$$

Finally, applying Theorem 1, the complexity to find ε-solution is

$$\mathcal{O}\left(\|x_0-x_*\|\left(\sqrt{\frac{L}{\varepsilon}}+\frac{\|A\|\operatorname{Diam}(g)}{\varepsilon}\right)\right).$$