Non-Smooth Convex Optimization

1 Convex Optimization Problem

In this lecture, we consider the general unconstrained convex optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x).$$

Here we assume that $f(x)$ is Lipschitz convex function, but not necessarily smooth. We will discuss the convergence of gradient descent method, and then acceleration with smoothing technique.

To obtain theoretical results, we assume that

$$\|\nabla f(x)\|_2 \leq G.$$

Example 1 *We consider the SVM formulation*

$$\min_w f(w) := \left[\frac{1}{n}\sum_{i=1}^n (1 - w^\top x_i y_i)_+ + \frac{\lambda}{2}\|w\|_2^2\right]$$

This is non-smooth. The function $f(w)$ is not Lipschitz globally over $\mathbb{R}^d$. However, if we start with $w_0 = 0$ and consider the region matters for optimization:

$$\{w : f(w) \leq f(0) = 1\} \subset \{w : \|w\|_2 \leq \sqrt{2/\lambda}\}.$$

Then the function is Lipschitz in this region

$$\|\nabla f(w)\|_2 \leq \frac{1}{n}\sum_{i=1}^n \|x_i\|_2 + \sqrt{2\lambda}.$$

2 Subgradient Method

In the smooth case, we have shown that gradient descent will converge with a constant step size. If the function is not smooth, then gradient can be replaced by a subgradient, as in Algorithm 1, but the step size has to be reduced in order to achieve convergence.

Algorithm 1: Subgradient Descent Method

Input: $f(x)$, x_0, $\eta_1, \eta_2, \ldots$
Output: x_T
for $t = 1, \ldots, T$ **do**
  Let $x_t = x_{t-1} - \eta_t g_t$, where $g_t \in \partial f(x_{t-1})$ is a subgradient
Return: x_T

Example 2 *Consider the function $f(x) = |x|$. The optimal solution is $x = 0$. For any constant learning rate $\eta_t = \eta$, if we take $x_0 = \eta/2$, then we have*

$$x_1 = -\eta/2, \quad x_2 = \eta/2, \cdots$$

Therefore the algorithm does not converge with a constant step size. However, with a smaller stepsize, one can obtain a solution closer to the optimal solution.

For the subgradient descent method in Algorithm 1, we may obtain a general convergence result as in the following theorem, which compares the iterate from the algorithm to an arbitrary solution x.

Theorem 1 *Assume that $f(x)$ is G-Lipschitz, then we have*

$$\frac{1}{\sum_{t=1}^T \eta_t} \sum_{t=1}^T \eta_t f(x_{t-1}) \leq f(x) + \frac{\|x_0 - x\|_2^2 + \sum_{t=1}^T \eta_t^2 G^2}{2\sum_{t=1}^T \eta_t}.$$

Proof Given any x, we have

$$\begin{aligned}
\|x_t - x\|_2^2 =& \|(x_t - x_{t-1}) + (x_{t-1} - x)\|_2^2 \\
=& \|x_t - x_{t-1}\|_2^2 + 2(x_t - x_{t-1})^\top (x_{t-1} - x) + \|x_{t-1} - x\|_2^2 \\
=& \eta_t^2 \|g_t\|_2^2 - 2\eta_t g_t^\top (x_{t-1} - x) + \|x_{t-1} - x\|_2^2 \\
\leq& \|x_{t-1} - x\|_2^2 + 2\eta_t g_t^\top (x - x_{t-1}) + \eta_t^2 G^2 \\
\leq& \|x_{t-1} - x\|_2^2 + 2\eta_t [f(x) - f(x_{t-1})] + \eta_t^2 G^2.
\end{aligned}$$

The first inequality is due to the condition $\|g_t\|_2 \leq G$, and the second inequality is due to the definition of subgradient g_t at x_{t-1}.

We can now sum over $t = 1, \ldots, T$, and obtain

$$2\sum_{t=1}^T \eta_t f(x_{t-1}) \leq 2\sum_{t=1}^T \eta_t f(x) + \sum_{t=1}^T \eta_t^2 G^2 + \|x_0 - x\|_2^2.$$

This leads to the result stated in the theorem. ■

To understand the convergence result, we consider the case that we know T in advance, and choose a small constant learning rate $\eta_0/\sqrt{T}$. In this case, we obtain the following result.

Corollary 1 *If we take $\eta_t = \eta = \eta_0/\sqrt{T}$, then*

$$\frac{1}{T} \sum_{t=1}^T f(x_{t-1}) \leq f(x) + \frac{\|x_0 - x\|_2^2 + \eta_0^2 G^2}{2\eta_0 \sqrt{T}}.$$

The result shows that the averaged primal sub-optimality of subgradient iterates will be $O(1/\sqrt{T})$ if we take a small learning rate of $O(1/\sqrt{T})$. This can be compared to the smooth case, where we can achieve a rate of $O(1/T)$ for the last iterate with a constant learning rate. Note it is also possible to derive a convergence result for the last iterate as well, but we do not discuss it here. For a given comparison point x, we may optimize the bound in Corollary 1 over η_0, and obtain a result below when $\eta_0 = \|x - x_0\|_2/G$:

$$\frac{1}{T}\sum_{t=1}^{T} f(x_{t-1}) \leq f(x) + \frac{\|x_0 - x\|G}{\sqrt{T}}.$$

This means choosing stepsize is important, and the larger G is, the smaller η should be.

If we do not know T a priori, we may choose a decaying learning rate schedule and obtain the following result.

Corollary 2 *If we take $\eta_t = \eta_0/(\sqrt{t} + \sqrt{t-1})$, then*

$$\sum_{t=1}^{T} \frac{1}{\sqrt{Tt} + \sqrt{T(t-1)}}[f(x_{t-1}) - f(x)] \leq \frac{\|x_0 - x\|_2^2 + 0.5\eta_0^2(\ln T + 1)G^2}{2\eta_0\sqrt{T}}.$$

Proof We note that $\sum_{t=1}^{T} \eta_t = \eta_0/\sqrt{T}$, and $\sum_{t=1}^{T} \eta_t^2 \leq 0.5\eta_0^2(\sum_{t=1}^{T} 1/t) = 0.5\eta_0^2(\ln T + 1)$. ■

This bound has an additional $\log T$ factor, but matches that of Corollary 1 otherwise.

The bound implies that to achieve ϵ primal sub-optimality, we need

$$O(1/\sqrt{\epsilon})$$

number of iterations.

3 Smoothing

In this section, we show that it is possible to achieve better convergence rate than subgradient method of Algorithm 1. by using smoothing technique introduced by Nesterov. This is because after smoothing, we can apply Nesterov's acceleration method.

Definition 1 *If $\tilde{f}(x)$ is an (L, ϵ)-smooth approximation of $f(x)$ if*

$$\tilde{f}(x) \leq f(x) \leq \tilde{f}(x) + \epsilon.$$

The following result shows that instead of optimizing with $f(x)$, we can obtain an approximation solution by optimizing with its smoothed version $\tilde{f}(x)$.

Theorem 2 *Assume $\tilde{f}(x)$ is an (L, ϵ)-smooth approximation of $f(x)$. Let $\tilde{x}$ be an $\tilde{\epsilon}$-approximate solution of the minimization problem with respect to $\tilde{f}(x)$:*

$$\tilde{f}(\tilde{x}) \leq \min_x \tilde{f}(x) + \tilde{\epsilon},$$

then

$$f(\tilde{x}) \leq \min_x f(x) + \epsilon + \tilde{\epsilon}.$$

Proof Given any x_*, we have

$$f(\tilde{x}) \leq \tilde{f}(\tilde{x}) + \epsilon \leq \tilde{f}(x_*) + \tilde{\epsilon} + \epsilon \leq f(x_*) + \epsilon + \tilde{\epsilon}.$$

■

From the above theorem, it follows that we can solve a smoothed version of a nonsmooth optimization problem. In particularly, for Lipschitz functions, there always exist an (L, ϵ)-smooth approximation with $L = G^2/(2\epsilon)$.

Proposition 1 *If $f(x)$ is G-Lipschitz, then*

$$\tilde{f}(x) = \min_z \left[f(z) + \frac{L}{2}\|x - z\|_2^2 \right] \tag{1}$$

is an $(L, G^2/(2L))$-smooth approximation of $f(x)$.

Proof First we prove the smoothness. First, since $\tilde{f}(x)$ is convex because it is minimum over z with respect to the joint convex function $f(z) + \frac{L}{2}\|x - z\|_2^2$ of (z, x). Observe that

$$\tilde{f}(x) = \frac{L}{2}\|x\|_2^2 - \sup_z \phi(z, x),$$

where

$$\phi(z, x) = \left[L \cdot (z^\top x) - \frac{L}{2}\|z\|_2^2 - f(z) \right],$$

and $\phi(z, x)$ is convex in x. Therefore

$$\sup_z \phi(z, x)$$

is a convex function of x (from Lecture 2, we know that sup of convex functions are convex). Denote it by $\phi(x)$. It follows that

$$\tilde{f}(x) + \phi(x) = \frac{L}{2}\|x\|_2^2.$$

This means that the smoothness condition of $\tilde{f}(x)$ is no more than the smoothness condition of $\frac{L}{2}\|x\|_2^2$, which is L.

Next we show that $\tilde{f}(x)$ is an ϵ-approximation. By definition, we have

$$\tilde{f}(x) \leq f(x) + \frac{L}{2}\|x - x\|_2^2 = f(x).$$

Now, given x, let z be the solution of (1), and thus

$$\nabla f(z) + L(z - x) = 0.$$

Then by Lipschitz condition, we have

$$\|z - x\|_2 \leq \frac{G}{L}.$$

Therefore

$$\tilde{f}(x) = f(z) + \frac{L}{2}\|x - z\|_2^2 \geq f(x) - G\|x - z\|_2 + \frac{L}{2}\|x - z\|_2^2 \geq f(x) - \frac{G^2}{2L}.$$

■

The smoothing method (1) is often referred to as the *Moreau envelope* or *Moreau-Yosida regularization*. In general, for two convex functions $f(x)$ and $g(x)$, the function $\inf_z[f(x) + g(x - z)]$, call *infimal convolution*, is also convex. Their properties are well studied in the convex analysis literature.

In the following, we give an example of (1). It shows that in general, it is relatively easy to apply smoothing.

Example 3 *Consider $f(x) = |x|$, and let*

$$\tilde{f}(x) = \min_z \left[f(z) + \frac{1}{2\epsilon}(x - z)^2 \right].$$

Then

$$\tilde{f}(x) = \begin{cases} |x| - \epsilon/2 & |x| \geq \epsilon \\ \frac{1}{2\epsilon}x^2 & otherwise \end{cases}.$$

Using smoothing, we can find an ϵ-approximate sub-optimality solution using an $L = G^2/2\epsilon$-smooth function $\tilde{f}(x)$. Using the general Nesterov's acceleration for non-strongly convex functions, we can obtain with $\lambda_T = O(1/T^2)$:

$$f(x_T) \leq \tilde{f}(x_T) + \epsilon \leq f(x_*) + \epsilon + \lambda_T \left[f(x_0) - f(x_*) + \frac{G^2}{4\epsilon}\|x_* - x_0\|_2^2 \right].$$

By choosing $T = O(1/\epsilon)$, we obtain

$$f(x_T) \leq f(x_*) + O(\epsilon).$$

Using acceleration algorithm, the resulting convergence rate is better than that of the subgradient method. Note that in Corollary 1, we need $O(1/\sqrt{\epsilon})$ iterations.