Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 8: Non-Smooth Convex Optimization

Convex Optimization

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.
To obtain theoretical results, we assume that

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

Example

Example

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, it is Lipschitz during the optimization process, where we consider the region

$$\{w : f(w) \leq f(w_0)\}.$$

Subgradient Method

This is the counterpart of gradient.

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

Learning Rate

Example

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.

Convergence

Theorem

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

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

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.

Interpretation of Convergence

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

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

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

Convergence Interpretation

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

Corollary

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

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

Smoothing

We show that it is possible to achieve better convergence rate than subgradient method of Algorithm 1. The idea is to solve a smoothed problem.

Definition

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

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

Approximate Optimization

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

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

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

Smoothing

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

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 of Smoothness

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 $\phi(z, x)$ is convex in x.
Let $\phi(x) = \sup_z \phi(z, x)$ is convex in x, it follows that

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

This implies the result.

Proof of Approximation

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

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

Given x, let z be the optimal solution of infimal convolution. Therefore

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

Example

Example

Consider $f(x) = |x|$, and let

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

Then

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

Convergence

Using smoothing, we can find an ϵ-approximate sub-optimality solution using an $L = G^2/2\epsilon$-smooth function $\tilde{f}(x)$.

We can apply gradient descent and Nesterov's acceleration.

- With gradient descent, we achieve the same convergence as that of subgradient on the original nonsmooth problem.
- With acceleration method, we achieve faster convergence.

Algorithm 2: Nesterov's Acceleration Method (Non Strongly Convex)

Input: $\tilde{f}(x)$, x_0, $\eta \leq 2\epsilon/G^2$

Output: x_T

Let $x_{-1} = x_0$

Let $\theta_0 = 1$

for $t = 1, \ldots, T$ **do**

- Solve for θ_t: $\theta_t^2 = (1 - \theta_t)\theta_{t-1}^2$
- Let $\beta_t = (\theta_{t-1}^{-1} - 1)\theta_t$
- Let $y_t = x_{t-1} + \beta_t(x_{t-1} - x_{t-2})$
- Let $x_t = y_t - \eta \nabla f(y_t)$

Return: x_T

Convergence With Nesterov's Acceleration

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

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

Summary

We have studied nonsmooth optimization

- We introduced subgradient method.
- To achieve ϵ suboptimality, it requires $O(1/\epsilon^2)$ iterations
- We introduced smoothing, which achieves ϵ-approximation of the non-smooth function with a $1/\epsilon$ smooth function.
- With Nesterov acceleration, it requires $O(1/\epsilon)$ iterations to achieve ϵ suboptimality