Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 11: Proximal Gradient Descent Method

Composite Convex Optimization

In this lecture, we consider the compoiste convex optimization optimization problem:

$$\min_{x \in \mathbb{R}^d} \phi(x), \qquad \phi(x) = f(x) + g(x).$$

where $g(x)$ may be defined on the convex domain $C \subset \mathbb{R}^d$.
That is, $g(x) = +\infty$ when $x \notin C$.
Here we assume that $f(x)$ is a smooth convex function defined on C, and $g(x)$ may be nonsmooth convex function.

Examples

Example 1 (Lasso):

$$g(x) = \mu\|x\|_1, \qquad C = \mathbb{R}^d.$$

and

$$\phi(x) = f(x) + \mu\|x\|_1.$$

Example 2 (Constraint):

$$g(x) = 0, \qquad C = \{x : \|x\|_2 \leq R\}.$$

and the problem is

$$\min_{x \in C} f(x)$$

Proximal Mapping

In proximal gradient method, we assume that the following optimization can be solved efficiently:

$$\mathrm{prox}_\eta(x) = \arg\min_{z \in \mathbb{R}^d} \left[\frac{1}{2\eta} \|z - x\|_2^2 + g(z) \right]. \tag{1}$$

Upper Bound with Proximal Mappoing

Using proximal mapping, we may form an upper bound of $\phi(x)$ as follows:

$$\phi(x) \leq Q(x; y) := f(y) + \nabla f(y)^\top (x - y) + \frac{1}{2\eta}\|x - y\|_2^2 + g(x),$$

where $\eta \leq 1/L$. We note that $Q(x; y) = f(y)$.

Therefore similar to gradient descent, we may minimize the right hand side to obtain y_+ from y so that $\phi(y_+) \leq \phi(y)$. It is easy to check that the solution is

$$\mathrm{prox}_\eta(y - \eta\nabla f(y)).$$

This mapping leads to proximal gradient descent algorithm.

Proximal Gradient Descent

```
Algorithm 1: Proximal Gradient Descent
Input: $f(\cdot)$, $g(\cdot)$, $x_0$, and $\eta_1, \eta_2, \ldots$
Output: $x_T$
for $t = 1, 2, \ldots, T$ do
    Let $\tilde{x}_t = x_{t-1} - \eta_t \nabla f(x_{t-1})$
    Let $x_t = \mathrm{prox}_{\eta_t}(\tilde{x}_t)$
Return: $x_T$
```

Example: Lasso

Example

Consider the following optimization problem

$$\min_{x \in \mathbb{R}^d} \left[f(x) + \mu \|x\|_1 \right].$$

It is easy to check that

$$\mathrm{prox}_\eta(x) = [\mathrm{prox}_\eta(x_j)]_{j=1,\ldots,d} \qquad \mathrm{prox}_\eta(x_j) = \begin{cases} x_j - \eta\mu & x_j > \eta\mu \\ 0 & |x_j| \le \eta\mu \\ x_j + \eta\mu & x_j < -\eta\mu \end{cases}.$$

Example: Constraint

Consider the following optimization problem

$$\min_{x \in C} f(x).$$

We may take

$$g(x) = \begin{cases} 0 & x \in C \\ +\infty & \text{otherwise} \end{cases}.$$

Then

$$\mathrm{prox}_{\eta}(x) = \mathrm{proj}_C(x) = \arg\min_{z \in C} \|z - x\|_2.$$

Convergence

Proposition

If we let

$$Q_t(x) = f(x_{t-1}) + \nabla f(x_{t-1})^\top (x - x_{t-1}) + \frac{1}{2\eta_t}\|x - x_{t-1}\|_2^2 + g(x),$$

then $\phi(x) \leq Q_t(x)$ *and*

$$x_t = \arg\min_x Q_t(x).$$

Moreover, if $g(x)$ *is* λ'*-strongly convex, then* $\forall x \in C$*:*

$$Q(x) - Q(x_t) \geq \frac{\eta_t^{-1} + \lambda'}{2}\|x - x_t\|_2^2.$$

Convergence Theorem

Theorem

Assume that $f(x)$ is an L-smooth convex and λ-strongly convex function, and $g(x)$ is a λ' strongly convex function. Let $\eta_t = \eta \leq 1/L$, then for all $\bar{x} \in C$:

$$\phi(x_t) \leq \phi(\bar{x}) + (1-\theta)^t [\phi(x_0) - \phi(\bar{x})],$$

where $\theta = (\eta\lambda + \eta\lambda')/(\eta\lambda' + 1)$.

Proof (part I)

We have

$$\begin{aligned}\phi(x_t) \leq& Q_t(x_t) \leq Q_t(x) - \frac{\eta_t^{-1} + \lambda'}{2}\|x - x_t\|_2^2 \\ \leq& f(x) - \frac{\lambda}{2}\|x - x_{t-1}\|_2^2 + \frac{1}{2\eta_t}\|x - x_{t-1}\|_2^2 + g(x) - \frac{\eta_t^{-1} + \lambda'}{2}\|x - x_t\|_2^2 \\ =& \phi(x) + \frac{1}{2}\left(\frac{1}{\eta_t} - \lambda\right)\|x - x_{t-1}\|_2^2 - \frac{\eta_t^{-1} + \lambda'}{2}\|x - x_t\|_2^2.\end{aligned}$$

In the above derivation, the first two inequalities are due to Proposition 1. The third inequality is due to the strong convexity of $f(x)$.

Poorf (part II)

Let $x = x_{t-1} + \theta(\bar{x} - x_{t-1})$ for some $\theta \in (0,1)$, we have

$$\begin{aligned}
&(1-\theta)\phi(x_{t-1}) + \theta\phi(\bar{x}) - \phi(x) \\
=&(1-\theta)[\phi(x_{t-1}) - \phi(x) - \nabla\phi(x)^\top(x_{t-1} - x)] + \theta[\phi(\bar{x}) - \phi(x) - \nabla\phi(x)^\top(\bar{x} - x)] \\
\geq&(1-\theta)\frac{\lambda+\lambda'}{2}\|x_{t-1} - x\|_2^2 + \theta\frac{\lambda+\lambda'}{2}\|\bar{x} - x\|_2^2 \\
=&(1-\theta)\theta\frac{\lambda+\lambda'}{2}\|\bar{x} - x_{t-1}\|_2^2.
\end{aligned}$$

The inequality is due to the $\lambda + \lambda'$ strong convexity of $\phi(x)$.
Therefore

$$\phi(x_t) \leq (1-\theta)\phi(x_{t-1}) + \theta\phi(\bar{x}) - \theta(1-\theta)\frac{\lambda+\lambda'}{2}\|\bar{x} - x_{t-1}\|_2^2 + \frac{\theta^2}{2}\left(\frac{1}{\eta_t} - \lambda\right)\|\bar{x} - x_{t-1}\|_2^2.$$

Taking $\eta_t = \eta$ and $\theta = (\lambda + \lambda')/(\lambda' + \eta^{-1})$, we obtain

$$\phi(x_t) \leq (1-\theta)\phi(x_{t-1}) + \theta\phi(\bar{x}).$$

This implies the desired bound.

Convergence: non-strongly convex problem

Theorem

Assume that $f(x)$ is L-smooth. Let $\eta_t = \eta \leq 1/L$, then for all $\bar{x} \in C$:

$$\frac{1}{T}\sum_{t=1}^{T} \phi(x_t) \leq \phi(\bar{x}) + \frac{1}{2\eta T}\|\bar{x} - x_0\|_2^2.$$

Backtracking Line Search

Similar to the case of gradient descent, it is possible to generalize the inexact line search method to deal with proximal mapping.

Observe that in the proof of Theorem 2, the convergence result holds as long as the learning rate satisfies the condition

$$\phi(x_t) \leq Q_t(x_t).$$

Note that this condition holds as long as $\eta_t \leq 1/L$.

Algorithm

Algorithm 2: Proximal Gradient Descent with Line Search

```
Input: $f(\cdot)$, $g(\cdot)$, $x_0$, and $\eta_0$, $\tau \in (0,1)$ (default $= 0.8$)
Output: $x_T$
for $t = 1, 2, \ldots, T$ do
    Let $\eta_t = \eta_{t-1}$
    while true do
        Let $\tilde{x}_t = x_{t-1} - \eta_t \nabla f(x_{t-1})$
        Let $x_t = \mathrm{prox}_{\eta_t}(\tilde{x}_t)$
        if $f(x_t) \leq f(x_{t-1}) + \nabla f(x_{t-1})^\top (x_t - x_{t-1}) + \frac{1}{2\eta_t}\|x_t - x_{t-1}\|_2^2$ then
            break
        Let $\eta_t = \tau \eta_t$
    if $f(x_t) \leq f(x_{t-1}) + \nabla f(x_{t-1})^\top (x_t - x_{t-1}) + \frac{\tau}{2\eta_t}\|x_t - x_{t-1}\|_2^2$ then
        Let $\eta_t = \tau^{-0.5} \eta_t$
Return: $x_T$
```

Theory

Theorem

Assume that $f(x)$ is λ-strongly convex and $g(x)$ is λ' strongly convex. Moreover, $\{\eta_t\}$ are obtained in Algorithm 2. Then for all $\bar{x} \in C$:

$$\phi(x_t) \leq \phi(\bar{x}) + \prod_{t=1}^{T} \left(1 - \frac{\eta_t}{1 + \eta_t \lambda'}(\lambda + \lambda')\right) [\phi(x_0) - \phi(\bar{x})].$$

Proximal-Gradient BB

Algorithm 3: Proximal Gradient Descent with BB Learning Rate

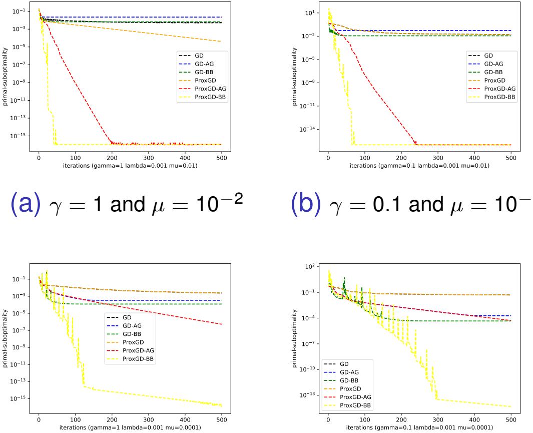
```
Input: $f(x)$, $x_0$, $\eta_0$ , $\tau = 0.8$, $c = 0.5$
Output: $x_T$
Let $g_0 = \nabla f(x_0)$ be a subgradient
for $t = 1, \ldots, T$ do
    Let $\tilde{x}_t = x_{t-1} - \eta_{t-1} g_{t-1}$
    Let $x_t = \mathrm{prox}_{\eta_{t-1}}(\tilde{x}_t)$
    Let $g_t = \nabla f(x_t)$ be a subgradient
    Let $\eta_t = \|x_t - x_{t-1}\|_2^2 / ((x_t - x_{t-1})^\top (g_t - g_{t-1}))$
Return: $x_T$
```

Empirical Studies

We study the smoothed hinge loss function $\phi_\gamma(z)$ as the last lectures, with L_1 regularization:

$$\min_w \left[\frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(w^\top x_i y_i) + \frac{\lambda}{2}\|w\|_2^2 + \mu\|w\|_1 \right].$$

Convergence Comnparisons



(a) $\gamma = 1$ and $\mu = 10^{-2}$

(b) $\gamma = 0.1$ and $\mu = 10^{-2}$

(c) $\gamma = 1$ and $\mu = 10^{-4}$

(d) $\gamma = 0.1$ and $\mu = 10^{-4}$

Figure: Convergence Comparisons with Different Smoothing Parameter

Summary

Composite convex optimization problem

$$\underbrace{f(x)}_{\text{smooth}} + \underbrace{g(x)}_{\text{nonsmooth}}$$

- deal with simple non-smooth function $g(x)$ using proximal mapping
- deal with constraints with projection

Proximal Gradient Descent Method

- smoothness L depends on $f(x)$
- strong convexity depends on both $f(x)$ and $g(x)$
- Learning rate depends on $1/L$
- Line search methods can be generalized