Proximal Gradient Descent Method

1 Composite Convex Optimization Problem

In this lecture, we consider the following composite convex optimization problem:

$$\min_{x \in \mathbb{R}^d} \phi(x) \qquad \phi(x) = [f(x) + g(x)], \tag{1}$$

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.

The optimization problem (1) is equivalent to optimizing over C:

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

An example is

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

Another related example is

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

Usually $g(x)$ is a regularizer, which is common in machine learning. If $g(x)$ is nonsmooth, one may use smoothing to obtain a $1/\epsilon$-smooth regularizer up to accuracy of ϵ. However, this will slow down the convergence. In this lecture, we consider a different approach called proximal gradient method, which does not suffer from this problem.

2 Proximal Mapping

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

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

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, which is described in Algorithm 1.

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 1 *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 2 *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

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

For example, if we take $C = \{x : \|x\|_\infty \le 1\}$, then

$$\mathrm{prox}_\eta(x) = [\mathrm{prox}_\eta(x_j)]_{j=1,\ldots,d} \qquad \mathrm{prox}_\eta(x_j) = \begin{cases} 1 & x_j \ge 1 \\ x_j & x_j \in (-1, 1) \\ -1 & x_j \le -1 \end{cases}.$$

3 Convergence Analysis

Proposition 1 *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 $\phi(x) \le Q_t(x)$ and

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

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

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

Proof Since

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

we have $\phi(x) = f(x) + g(x) \leq Q_t(x)$.

Moreover, we know that

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

Therefore by definition, the minimizer of $Q_t(x)$ is $x_t = \mathrm{prox}_\eta(x_{t-1} - \eta_t \nabla f(x_{t-1}))$. It implies that $\exists \xi \in \partial Q_t(x)|_{x=x_t}$ such that $\xi^\top (x - x_t) \geq 0$ for all $x \in C$. Since $Q(x)$ is $\eta_t^{-1} + \lambda'$ strongly convex, we have

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

This proves the proposition. ■

Theorem 1 *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 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 $f(x)$.

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. ■

Theorem 2 *Assume that $f(x)$ is L-smooth. Let $\eta_t = \eta \leq 1/L$, then for all $\bar{x} \in C$:*

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

Proof Similar to the proof of Theorem 1, with $\lambda = \lambda' = 0$, we obtain

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

Summing over $t = 1$ to $t = T$, we obtain the desired bound. ■

Note that the results of this section are similar to those of gradient descent without proximal mapping. The results only depend on the smoothness parameter of $f(x)$, but not on the smoothness parameter of $g(x)$. If $g(x)$ is non-smooth, this leads to faster convergence rate.

4 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 the proof of Theorem 1 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 $\eta_t \leq 1/L$. The condition can be rewritten as:

$$\begin{aligned} f(x_t) \leq & f(x_{t-1}) + \nabla f(x_{t-1})^\top (x_t - x_{t-1}) + \frac{1}{2\eta_t}\|x_t - x_{t-1}\|_2^2, \\ x_t = & \mathrm{prox}_{\eta_t}(x_{t-1} - \eta_t \nabla f(x_{t-1})), \end{aligned} \tag{3}$$

which can be regarded as a generalization of the Armijo-Goldstein condition at $c = 0.5$. The larger η_t is, the better convergence rate we will obtain. Therefore backtracking can be performed so that we can find a large η_t that satisfies (3). This leads to Algorithm 2. The convergence follows from the same analysis of Theorem 1.

Theorem 3 *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 $\bar{x} \in C$:*

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

Algorithm 2: Proximal Gradient Descent with Backtracking Line Search

---

Input: $f(\cdot)$, $g(\cdot)$, x_0, and η_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 $x_t = \text{prox}_{\eta_t}(\tilde{x}_t)$
    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

Since the learning rate depends on an estimate of the smoothness of $f(x)$. We may generalize the BB method that employs the following estimate of inverse of the smoothness parameter of $f(x)$:

$$\frac{\|x_t - x_{t-1}\|_2^2}{(x_t - x_{t-1})^\top (\nabla f(x_t) - \nabla f(x_{t-1}))},$$

which leads to Algorithm 3.

Algorithm 3: Proximal Gradient Descent with BB Learning Rate

Input: $f(x)$, x_0, η_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 $x_t = \text{prox}_{\eta_{t-1}}(\tilde{x}_t)$
  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

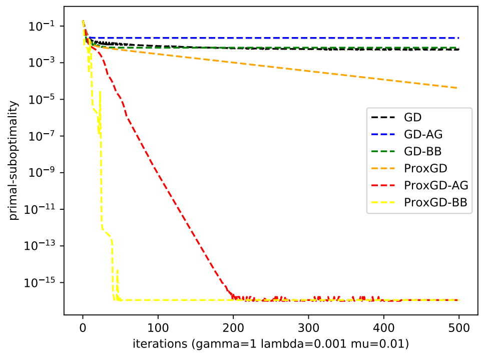
---

5 Empirical Studies

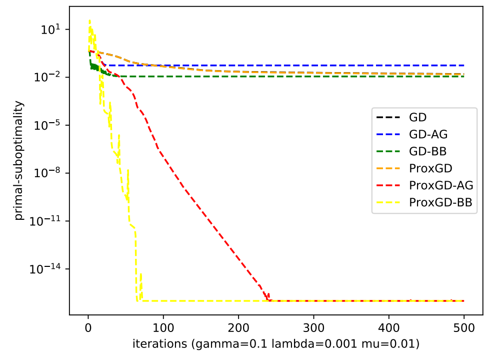
We study the smoothed hinge loss function $\phi_\gamma(z)$ as the last lectures, with L_1 regularization:

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

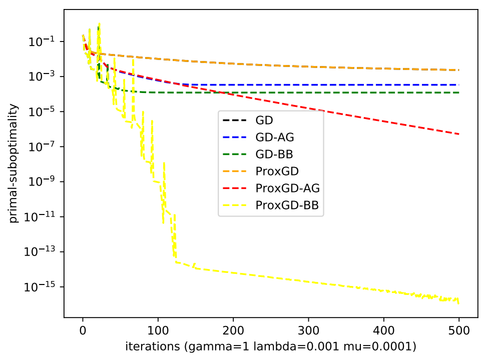
Comparisons are given in Figure 1. We can see that proximal methods work better when $f(x)$ is smoother and the non-smoooth part $g(x)$ is more important (μ is larger). This is consistent with our theoretical results.



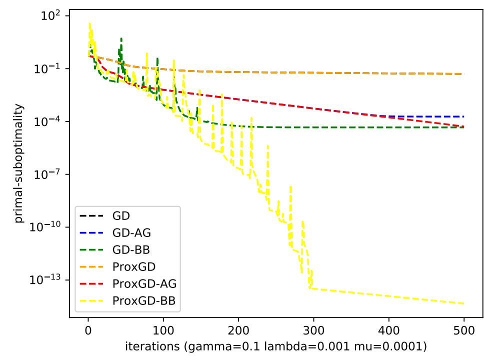
(a) $\gamma = 1$ and $\mu = 10^{-2}$



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



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



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

Figure 1: Convergence Comparisons with Different Smoothing Parameter