Accelerated Proximal Gradient Descent

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, with smoothness parameter L, and $g(x)$ may be nonsmooth convex function.

We have shown that in general, we replace the gradient step

$$y - \eta \nabla f(y)$$

by the proximal gradient step

$$\text{prox}_\eta(y - \eta \nabla f(y)), \tag{2}$$

where

$$\text{prox}_\eta(y) = \arg\min_{z \in C} \left[\frac{1}{2\eta} \|z - y\|_2^2 + g(z) \right]. \tag{3}$$

The smoothness of the system is the smoothness of $f(x)$. The strong convexity of the system is $\lambda + \lambda'$, where $f(x)$ is λ strongly convex and $g(x)$ is λ' strongly convex. In fact, if we define

$$\tilde{f}(x) = f(x) + \frac{\lambda}{2}\|x\|_2^2, \quad \tilde{g}(x) = g(x) - \frac{\lambda}{2}\|x\|_2^2,$$

and define

$$\widetilde{\text{prox}}_{\tilde{\eta}}(y) = \arg\min_{z \in C} \left[\frac{1}{2\tilde{\eta}} \|z - y\|_2^2 + \tilde{g}(z) \right],$$

then

$$\text{prox}_\eta(y - \eta \nabla f(y)) = \widetilde{\text{prox}}_{\tilde{\eta}}(y - \tilde{\eta} \nabla \tilde{f}(y)),$$

where $\tilde{\eta} = \eta/(1 + \eta\lambda')$.

Therefore for an algorithm with composition $f(x) + g(x)$, we can get equivalent algorithm with composition $\tilde{f}(x) + \tilde{g}(x)$, with $\tilde{f}(x)$ being $\lambda + \lambda'$ strongly convex.

2 Convergence Checking

In the standard gradient descent methods (including accelerated gradient descent), for smooth optimization, one may simply check the value of gradient $\nabla f(x)$ for convergence. However, the method fails for composite optimization. This is because $\nabla f(x)$ may not converge to zero.

If proximal gradient method converges, then we have $x_t \to x_*$. From the proximal iteration, we have

$$x_t = \text{prox}_{\eta_t}(x_{t-1} - \eta_t \nabla f(x_{t-1})).$$

It follows that

$$x_* = \text{prox}_{\eta_t}(x_* - \eta_t \nabla f(x_*)).$$

We may define

$$D_\eta \phi(x) = \frac{1}{\eta}\left(x - \text{prox}_\eta(x - \eta \nabla f(x))\right),$$

which can replace the gradient for checking convergence. Note that if $g(x) = 0$, then $D_\eta \phi(x) = \nabla f(x)$ is the gradient. The following result is analogous of the result for gradient descent. We have the following result:

Proposition 1 *Assume $f(x)$ is L-smooth, and $g(x)$ is λ' strongly convex. Let*

$$x^+ = \text{prox}_\eta(x - \eta \nabla f(x)).$$

Given a learning rate $\eta > 0$ such that $\eta(L - \lambda') \le 1$, we have

$$\phi(x^+) \le \phi(x) - \eta(1 + \eta(\lambda' - L)/2)\|D_\eta \phi(x)\|_2^2 \le \phi(x) - 0.5\eta\|D_\eta \phi(x)\|_2^2.$$

Proof Let

$$Q(z) = f(x) + \nabla f(x)^\top (z - x) + \frac{1}{2\eta}\|z - x\|_2^2 + g(z), \tag{4}$$

then x^+ is the solution of $\min_z Q(z)$, and $Q(z)$ is $\eta^{-1} + \lambda'$ strongly convex. This implies that

$$Q(x) - Q(x^+) \ge \frac{\eta^{-1} + \lambda'}{2}\|x - x^+\|_2^2. \tag{5}$$

Moreover, by the smoothness of f, we have

$$\begin{aligned}
\phi(x^+) =& f(x^+) + g(x^+) \le f(x) + \nabla f(x)^\top (x^+ - x) + \frac{L}{2}\|x^+ - x\|_2^2 + g(x^+) \\
=& Q(x^+) + \frac{L - \eta^{-1}}{2}\|x^+ - x\|_2^2 \\
\le& Q(x) + \frac{L - \lambda' - 2\eta^{-1}}{2}\|x^+ - x\|_2^2.
\end{aligned}$$

The first inequality is due to the smoothness of $f(x)$. The second inequality is due to (5). Note that $Q(x) = f(x)$, we obtain the desired bound. ■

Proposition 2 *Assume that $f(x)$ is an L-smooth convex function and $\phi(x)$ is λ_ϕ strongly convex. Let*

$$x^+ = \mathrm{prox}_\eta(x - \eta\nabla f(x)).$$

Given a learning rate $\eta > 0$, we have

$$f(x^+) \leq f(x_*) + \frac{\max(1, \eta L)^2}{2\lambda_\phi}\|D_\eta\phi(x)\|_2^2.$$

Proof From the fact that x^+ is the solution of (4), we obtain the following first order condition: $\exists\xi \in \partial g(x^+)$ such that for all $x_* \in C$:

$$(\nabla f(x) + \xi + \eta^{-1}(x^+ - x))^\top (x_* - x^+) \geq 0.$$

This implies that

$$\begin{aligned}
(\nabla f(x^+) + \xi)^\top (x_* - x^+) =&(\nabla f(x^+) - f(x))^\top (x_* - x^+) + (\nabla f(x) + \xi)^\top (x_* - x^+) \\
\geq&(\nabla f(x^+) - f(x))^\top (x_* - x^+) + \eta^{-1}(x - x^+)^\top (x_* - x^+) \\
=&(\nabla \tilde{f}(x^+) - \tilde{f}(x))^\top (x_* - x^+) \\
\geq& - \max(L, \eta^{-1})\|x^+ - x\|_2\|x^+ - x_*\|_2,
\end{aligned}$$

where $\tilde{f}(z) = f(z) - 0.5\eta^{-1}\|z\|_2^2$, which may not be convex. The last inequality is due to the fact that $\tilde{f}(x)$ is at most $\max(\eta^{-1}, L)$ smooth.

The above inequality implies that

$$- \max(L, \eta^{-1})\|x^+ - x\|_2\|x^+ - x_*\|_2 \leq (\nabla f(x^+) + \xi)^\top (x_* - x^+) \leq \phi(x_*) - \phi(x^+) - \frac{\lambda_\phi}{2}\|x_* - x^+\|_2^2.$$

The second inequality follows from the strong convexity of $\phi(x)$. The above inequality implies that

$$\begin{aligned}
&\phi(x_*) - \phi(x^+) \\
\geq& \inf_z \left[\frac{\lambda + \lambda'}{2}\|z - x^+\|_2^2 - \max(L, \eta^{-1})\|x^+ - x\|_2\|x^+ - z\|_2\right] \\
=&\frac{\max(\eta^{-1}, L)^2}{2\lambda_\phi}\|x^+ - x\|_2^2.
\end{aligned}$$

This proves the desired bound. ∎

These results imply that for problems with smooth $f(x)$ and strongly convex $\phi(x)$, we obtain convergence when $D_\eta\phi(x)$ converges to zero. Therefore this quantity can be used to check convergence. Using the same example as that of the last lecture, the convergence can be shown in Figure 1.

3 Accelerated Proximal Gradient Descent

In this section, we consider a generalization of Nesterov's accelerated gradient descent (Algorithm 3 of Lecture 07) to handle proximal mapping. The general method is presented in Algorithm 1. A version of the resulting method is also known as FISTA [1].

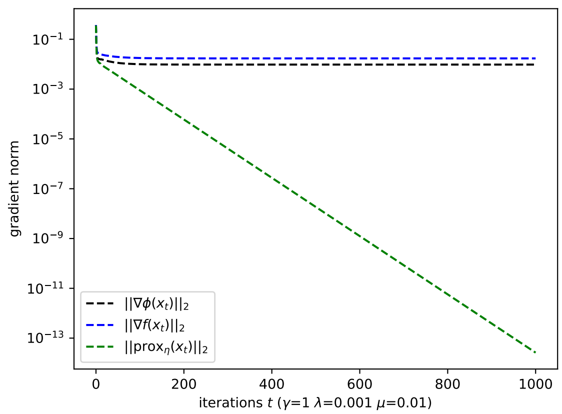


Figure 1: Convergence of Gradients with $L_1 - L_2$ regularized Smoothed Hinge Optimization

---

Algorithm 1: Nesterov's General Accelerated Proximal Gradient Method

---

Input: $f(x)$, x_0, $\{\eta_t\} \leq 1/L$

$\lambda \in [0, 1/L]$ (default is $\lambda = 0$)

$\lambda' \geq 0$ (default is $\lambda' = 0$)

$\gamma_0 \in [\lambda + \lambda', \eta_0^{-1} + \lambda']$ (default is $\gamma_0 = \eta_0^{-1} + \lambda'$)

Output: x_T

Let $x_{-1} = x_0$

Let $\theta_0 = \sqrt{\gamma_0 \eta_0 / (1 + \eta_0 \lambda')}$

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

Solve for θ_t: $\theta_t^2(\eta_t^{-1} + \lambda') = \theta_t(\lambda + \lambda') + (1 - \theta_t)\gamma_{t-1}$

Let $\gamma_t = (1 - \theta_t)\gamma_{t-1} + \theta_t(\lambda + \lambda')$

Let $\beta_t = (\theta_t^{-1} - 1)(\theta_{t-1}^{-1} - 1)\gamma_{t-1}/(\eta_t^{-1} - \lambda)$

Let $y_t = x_{t-1} + \beta_t(x_{t-1} - x_{t-2})$

Let $\tilde{x}_t = y_t - \eta_t \nabla f(y_t)$

Let $x_t = \mathrm{prox}_{\eta_t}(\tilde{x}_t)$

Return: x_T

We will have the following general Theorem.

Theorem 1 *Assume $f(x)$ is L-smooth and λ-strongly convex, and $g(x)$ is λ' strongly convex. Then for all $x_* \in C$, we have*

$$\phi(x_t) \leq \phi(x_*) + \lambda_t \left[\phi(x_0) - \phi(x_*) + \frac{\gamma_0}{2} \|x_* - x_0\|_2^2 \right],$$

where

$$\lambda_t = \prod_{s=1}^{t} (1 - \theta_s).$$

If we let $\eta_t = \eta \leq 1/L$, $\gamma_0 = \lambda + \lambda'$, and $\theta_t = \theta = \sqrt{\eta(\lambda + \lambda')/(1 + \eta\lambda')}$. Then we have

Corollary 1 *Assume that $f(x)$ is L smooth and λ strongly convex, and $g(x)$ is λ' strongly convex. We may take $\eta \leq 1/L$, $\theta = \sqrt{\eta(\lambda+\lambda')/(1+\eta\lambda')}$, and $\beta = (1-\theta)/(1+\theta)$. The following result holds for all $x_* \in C$:*

$$\phi(x_t) \leq \phi(x_*) + (1-\theta)^t \left[\phi(x_0) - \phi(x_*) + \frac{\lambda+\lambda'}{2}\|x_* - x_0\|_2^2\right].$$

In order for the theorem to be valid, η_t only needs to satisfy the following inequality

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

which is required in the proof of Lemma 1.

Similar to the case of Proximal Gradient Descent with backtracking, one may use backtracking to adjust learning rate η for Nesterov's method. We may also use the observed convergence with $D_\eta \phi(y_t)$ to determine β, as in Proposition 1.

This leads to the adaptive version in Algorithm 2. A similar generalization can be obtained using the heavy-ball update, where $\mathrm{prox}_{\alpha_t}(y_t - \alpha_t \nabla f(y_t))$ is replaced by $\mathrm{prox}_{\alpha_t}(y_t - \alpha_t \nabla f(x_{t-1}))$.

Algorithm 2: Adaptive Accelerated Proximal Gradient Method with AG Learning Rate

Input: $f(x)$, x_0, α_0, $\tau = 0.8$, $c = 0.5$
Output: x_T
Let $x_{-1} = x_0$
Let $\gamma = 0$
Let $y_0 = x_0$
for $t = 1, \ldots, T$ **do**
  Let $\beta = \min(1, \exp(\gamma))$
  Let $y_t = x_{t-1} + \beta(x_{t-1} - x_{t-2})$
  Let $\alpha_t = \alpha_{t-1}$
  Let $x_t = \mathrm{prox}_{\alpha_t}(y_t - \alpha_t \nabla f(y_t))$
  Let $\tilde{\eta} = (f(x_t) - f(y_t))/\|(x_t - y_t)/\alpha_t\|_2^2$
  while $\tilde{\eta} \leq c\alpha_t$ *and* $\tilde{\eta} \geq 10^{-4}\alpha_0$ **do**
    Let $\alpha_t = \tau\alpha_t$
    Let $x_t = \mathrm{prox}_{\alpha_t}(y_t - \alpha_t \nabla f(y_t))$
    Let $\tilde{\eta} = (f(y_t) - f(x_t))/\|(x_t - y_t)/\alpha_t\|_2^2$
  if $\tilde{\eta} \geq \tau^{-1} c\alpha_t$ **then**
    Let $\alpha_t = \tau^{-0.5}\alpha_t$
  Let $\gamma = 0.8\gamma + 0.2\ln(\|(x_t - y_t)/\alpha_t\|_2^2 / \|(x_{t-1} - y_{t-1})/\alpha_{t-1}\|_2^2)$
Return: x_T

4 Empirical Studies

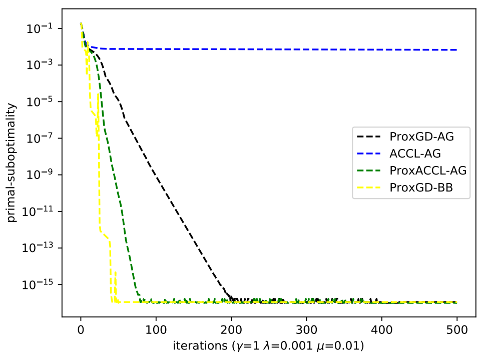
We study the smoothed hinge loss function $\phi_\gamma(z)$ as the last lectures, with L_1 regularization:

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

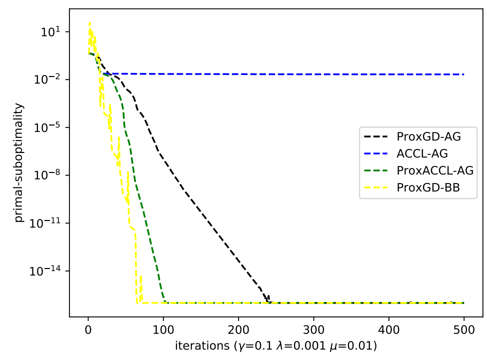
We note that the larger γ is, the smoother $f(x)$ is, and the larger μ is, the more important the non-smooth term $g(w) = \mu\|w\|_1$ is.

Comparisons of Nesterov's accelerated methods and non-accelerated methods are given in Figure 2.

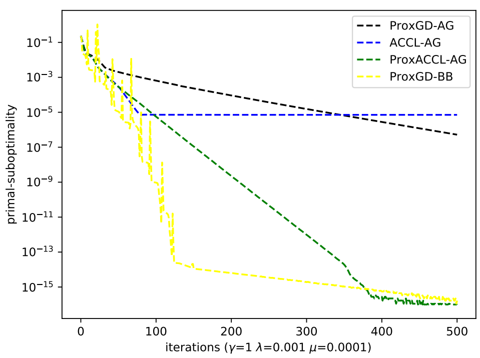
Comparisons of Nesterov's accelerated methods and Heavy-ball methods are given in Figure 3.



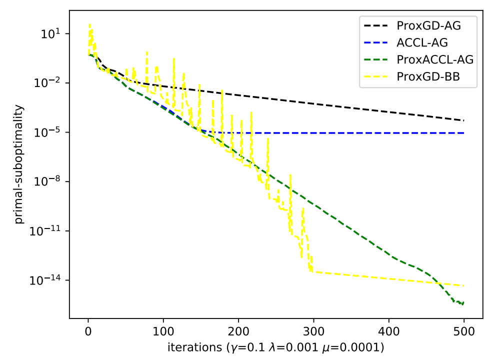
(a) $\gamma = 1$ and $\mu = 10^{-2}$



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



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



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

Figure 2: Convergence Comparisons (Acceleration versus non-Acceleration)

5 Proof Sketch of Theorem 1

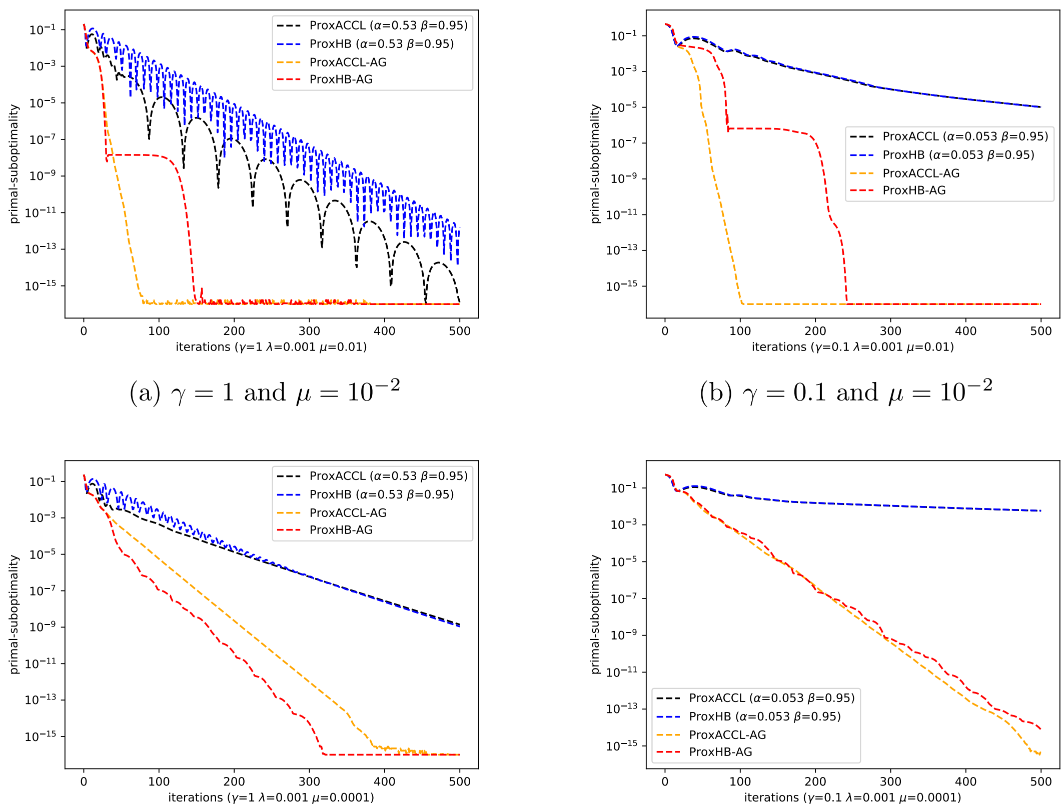
Similar to the analysis of Lecture 7, we can derive the theorem using estimate sequence, which can be constructed as follows.

Lemma 1 *Let* $x^+ = \mathrm{prox}_{\eta_t}(y - \eta_t \nabla f(y))$. *We define*

$$\psi_t(z; y) = \phi(x^+) - \frac{\eta_t^{-1} + \lambda'}{2}\|x^+ - y\|_2^2 + (\eta_t^{-1} + \lambda')(y - x^+)^\top (z - x^+) + \frac{\lambda + \lambda'}{2}\|z - y\|_2^2.$$

Then the following inequality holds:

$$\psi(z; y) \le \phi(z).$$



(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 3: Convergence Comparisons (Acceleration versus Heavy Ball)

Proof We have the first order condition: $\exists \xi \in \partial g(x^+)$ such that for all $z \in C$

$$(\nabla f(y) + \xi + \eta_t^{-1}(x^+ - y))^\top (z - x^+) \geq 0.$$

Therefore

$$\begin{aligned}
\phi(z) =& f(z) + g(z) \\
\geq & f(y) + \nabla f(y)^\top (z - y) + \frac{\lambda}{2}\|z - y\|_2^2 + g(x^+) + \xi^\top (z - x^+) + \frac{\lambda'}{2}\|z - x^+\|_2^2 \\
=& f(y) + \nabla f(y)^\top (x^+ - y) + (\nabla f(y) + \xi)^\top (z - x^+) + \frac{\lambda}{2}\|z - y\|_2^2 \\
& \quad + \frac{\lambda'}{2}[\|z - y\|_2^2 - \|y - x^+\|_2^2 - 2(z - x^+)^\top (x^+ - y)] \\
\geq & f(x^+) - \frac{\eta_t^{-1} + \lambda'}{2}\|x^+ - y\|_2^2 + (\eta_t^{-1} + \lambda')(y - x^+)^\top (z - x^+) + \frac{\lambda + \lambda'}{2}\|z - y\|_2^2 \\
=& \psi_t(z; y).
\end{aligned}$$

The first inequality uses the strong convexity. The second inequality uses the smoothness with $1/\eta_t \geq L$, and the first order condition to $\nabla f(y) + \xi$ by $\eta_t^{-1}(y - x^+)$. ■

Using notations in Lecture 06, we may define an estimate sequence recursively as

$$\phi_t(z) = (1 - \theta_t)\phi_{t-1}(z) + \theta_t \psi_t(z; y_t), \quad \lambda_t = (1 - \theta_t)\lambda_{t-1},$$

with

$$\phi_0(z) = f(x_0) + \frac{\gamma_0}{2}\|z - x_0\|_2^2, \quad \lambda_0 = 1.$$

We prove that for this estimate sequence, the following holds. Results of Lecture 06 then implies the theorem.

Lemma 2 *We have*

$$\phi(x_t) \leq \phi_t(v_t) = \min_z \phi_t(z).$$

Proof The proof is the same as that in Lecture 07, where we replace η_t^{-1} by $\eta_t^{-1} + \lambda'$ and λ by $\lambda + \lambda'$. ■

References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.