Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 12: Accelerated Proximal Gradient Descent

Composite Convex Optimization

In this lecture, we consider the composite 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 $x \notin C$.

- $f(x)$ is smooth
- $g(x)$ may be nonsmooth

Proximal Gradient

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

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

by the proximal gradient step

$$\mathrm{prox}_\eta(y - \eta \nabla f(y)), \tag{1}$$

where

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

Interchangeable Strong Convexity

Assume $f(x)$ is λ strongly convex, $g(x)$ is λ' strongly convex. 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

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

then

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

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

Convergence Checking

In composite optimization, $\nabla f(x)$ may not go to zero. Therefore we may not use it to check convergence. $g(x)$ may not be smooth, we may not include it in gradient calculation.

We define

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

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.

Suboptimality Lower bound: Smoothness

Proposition

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

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

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

$$\phi(x^+) \leq \phi(x) - 0.5\eta \|D_\eta \phi(x)\|_2^2.$$

This proposition can be used to determine learning rate.

Suboptimality Upper Bound: Strong Convexity

Proposition

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

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

Illustration

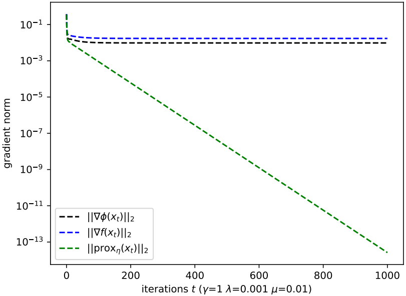


Figure: Convergence with $L_1 - L_2$ regularized Smoothed Hinge Optimization

$$g(x) = \mu\|x\|_1$$

Nesterov's Method

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 θ_0 = sqrt(γ_0 η_0 / (1 + η_0 λ'))
for t = 1, ..., T do
    Solve for θ_t: θ_t^2 (η_t^{-1} + λ') = θ_t (λ + λ') + (1 − θ_t) γ_{t−1}
    Let γ_t = (1 − θ_t) γ_{t−1} + θ_t (λ + λ')
    Let β_t = (θ_t^{-1} − 1)(θ_{t−1}^{-1} − 1) γ_{t−1} / (η_t^{-1} − λ)
    Let y_t = x_{t−1} + β_t (x_{t−1} − x_{t−2})
    Let x̃_t = y_t − η_t ∇f(y_t)
    Let x_t = prox_{η_t}(x̃_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

Convergence

We will have the following general Theorem.

Theorem

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

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

Convergence: Constant α and β

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

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 $x_* \in C$:

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

Algorithm: constant α and β

Algorithm 2: Nesterov's Accelerated Proximal Gradient Method

Input: $f(x)$, x_0, $\eta \leq 1/L$, λ and λ'
Output: x_T

Let $x_{-1} = x_0$
Let $\theta = \sqrt{\eta(\lambda + \lambda')/(1 + \eta\lambda')}$
Let $\beta = (1 - \theta)/(1 + \theta)$
for $t = 1, \ldots, T$ **do**
    Let $y_t = x_{t-1} + \beta(x_{t-1} - x_{t-2})$
    Let $\tilde{x}_t = y_t - \eta\nabla f(y_t)$
    Let $x_t = \mathrm{prox}_{\eta_t}(\tilde{x}_t)$
Return: x_T

Backtracking Line Search

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.

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

Algorithm 3: Adaptive Accelerated Proximal Gradient Method

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 $\tilde{\eta} \le c\alpha_t$ *and* $\tilde{\eta} \ge 10^{-4}\alpha_0$ **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 $\tilde{\eta} \ge \tau^{-1} c\alpha_t$ **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

Heavy-Ball Version

Algorithm 4: Adaptive Heavy-Ball Proximal Gradient Method

Input: $f(x)$, x_0, α_0, $\tau = 0.8$, $c = 0.5$
Output: x_T

```
Let x_{-1} = x_0
Let γ = 0
Let y_0 = x_0
for t = 1, ..., T do
    Let β = min(1, exp(γ))
    Let y_t = x_{t-1} + β(x_{t-1} − x_{t-2})
    Let α_t = α_{t-1}
    Let x_t = prox_{α_t}(y_t − α_t ∇f(x_{t-1}))
    Let η̃ = (f(x_t) − f(y_t)) / ||(x_t − y_t)/α_t||_2^2
    while η̃ ≤ cα_t and η̃ ≥ 10^{-4} α_0 do
        Let α_t = τα_t
        Let x_t = prox_{α_t}(y_t − α_t ∇f(x_{t-1}))
        Let η̃ = (f(y_t) − f(x_t)) / ||(x_t − y_t)/α_t||_2^2
    if η̃ ≥ τ^{-1} cα_t then
        Let α_t = τ^{-0.5} α_t
    Let γ = 0.8γ + 0.2 ln(||(x_t − y_t)/α_t||_2^2 / ||(x_{t-1} − y_{t-1})/α_{t-1}||_2^2)
```

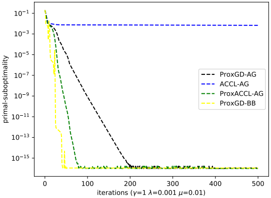
Return: x_T

Empirical Studies

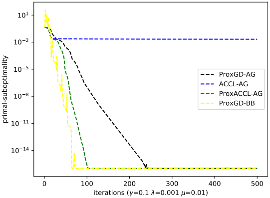
We study the smoothed hinge loss function $\phi_\gamma(z)$ as the last lectures, with L_1 regularization:

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

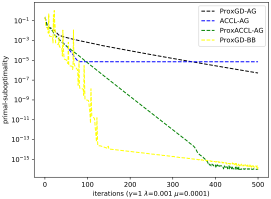
We note that the larger γ is, the smoother $f(x)$ is, and the larger μ is, the more important the non-smooth term $g(w) = \mu\|w\|_1$ is.



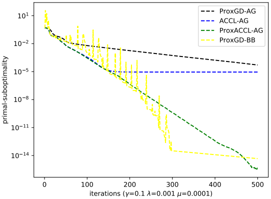
(a) $\gamma = 1$ and $\mu = 10^{-2}$



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

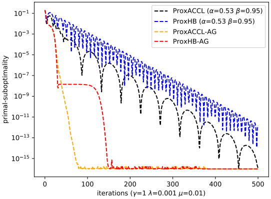


(c) $\gamma = 1$ and $\mu = 10^{-4}$

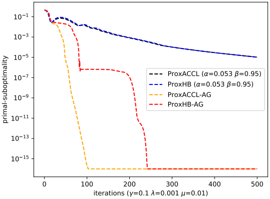


(d) $\gamma = 0.1$ and $\mu = 10^{-4}$

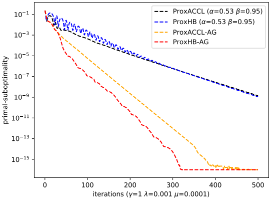
Comparisons (acceleration versus heavy-ball)



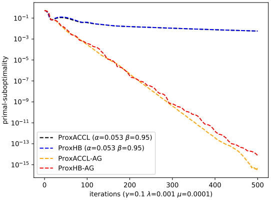
(a) $\gamma = 1$ and $\mu = 10^{-2}$



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



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



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

Summary

Composite convex optimization problem

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

- deal with simple non-smooth function $g(x)$ using proximal mapping
- deal with constraints with projection
- replace gradient by $D_\eta \phi(x)$

Accelerated Proximal Gradient Descent Method

- Straight-forward generalization of non-proximal version
- Adaptive methods work fine
- Can derive heavy-ball version with similar empirical results