

Optimization Theory

Lecture 08

Fudan University

luoluo@fudan.edu.cn

- 1 Subgradient Descent Method
- 2 Smoothing Technique
- 3 Proximal Gradient Methods

1 Subgradient Descent Method

2 Smoothing Technique

3 Proximal Gradient Methods

Subgradient Descent Method

We consider optimization with a nonsmooth objective function

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

Here we assume that $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is G -Lipschitz and convex defined on a convex and closed set $\mathcal{C} \subseteq \mathbb{R}^d$, but not necessarily smooth.

We have introduced the subgradient method

$$\begin{cases} \tilde{\mathbf{x}}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}_t, \\ \mathbf{x}_{t+1} = \text{proj}_{\mathcal{C}}(\tilde{\mathbf{x}}_{t+1}). \end{cases}$$

Let $R = \sup_{\mathbf{x} \in \mathcal{C}} \|\mathbf{x} - \mathbf{x}_0\|_2$.

- ① For convex case, it requires $\mathcal{O}(G^2 R^2 \epsilon^{-2})$ iterations.
- ② For μ -strongly-convex, it requires $\mathcal{O}(G^2 \mu^{-1} \epsilon^{-1})$ iterations.

Optimality of Subgradient Descent Method

Theorem

Given $G > 0$, $\mu > 0$, $d > t \geq 1$ and $\epsilon > 0$, there exists a G -Lipschitz and μ -strongly convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ on

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq \frac{G}{2\mu} \right\},$$

such that a first order optimization algorithm with initial point $\mathbf{x}_0 = 0$ can only produce solutions that satisfy

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \geq \frac{G^2}{8\mu(t+1)} - \epsilon,$$

where \mathbf{x}^ is the solution of $\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x})$.*

Optimality of Subgradient Descent Method

Let $\mathbf{x} = [x_1, \dots, x_d] \in \mathbb{R}^d$ and define

$$f(\mathbf{x}) = \frac{G}{2} \max_{i=1,2,\dots,t+1} \left(x_i + \frac{\epsilon}{i} \right) + \frac{\mu}{2} \|\mathbf{x}\|_2^2.$$

- ① The function f is G -Lipschitz continuous.
- ② Any subgradient $\mathbf{g} \in \partial f(\mathbf{x})$ satisfies $\mathbf{g} = \lambda \mathbf{x} + 0.5G\mathbf{y}$, where

$$\mathbf{y} \in \text{conv} \left\{ \mathbf{e}_i : x_i = \max_{k=1,2,\dots,t+1} x_k \right\}.$$

We use $\text{conv}(\mathcal{S})$ to present the convex hull of \mathcal{S} , which is defined as

$$\text{conv}(\mathcal{S}) = \left\{ \sum_{i=1}^m \alpha_i \mathbf{x}_i : \mathbf{x}_i \in \mathcal{S}, \alpha_i \geq 0, \sum_{i=1}^m \alpha_i = 1 \right\}.$$

- ③ Check the zero-chain property.

Optimality of Subgradient Descent Method

For convex case, we can show the optimality by considering

$$f(\mathbf{x}) = \frac{G}{2} \max_{i=1,2,\dots,t+1} \left(x_i + \frac{\epsilon}{i} \right) + \frac{\mu}{2} \|\mathbf{x}\|_2^2.$$

with

$$\mu = \frac{G}{2R\sqrt{t+1}}.$$

Outline

1 Subgradient Descent Method

2 Smoothing Technique

3 Proximal Gradient Methods

Smoothing Technique

Definition

We say the function $\tilde{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ is an (L, ϵ) -smooth approximation of function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ if \tilde{f} is L -smooth and we have

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

for all $\mathbf{x} \in \mathbb{R}^d$.

We can find approximate solution $\tilde{\mathbf{x}}$ for $\min_{\mathbf{x} \in \mathbb{R}^d} \tilde{f}(\mathbf{x})$ such that

$$\tilde{f}(\tilde{\mathbf{x}}) \leq \inf_{\mathbf{x} \in \mathbb{R}^d} \tilde{f}(\mathbf{x}) + \tilde{\epsilon},$$

then

$$\begin{aligned} f(\tilde{\mathbf{x}}) &\leq \tilde{f}(\tilde{\mathbf{x}}) + \epsilon \leq \inf_{\mathbf{x} \in \mathbb{R}^d} \tilde{f}(\mathbf{x}) + \tilde{\epsilon} + \epsilon \\ &\leq \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) + \tilde{\epsilon} + \epsilon. \end{aligned}$$

Theorem

If $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and G -Lipschitz continuous, then

$$\tilde{f}(\mathbf{x}) = \min_{\mathbf{z} \in \mathbb{R}^d} \left(f(\mathbf{z}) + \frac{L}{2} \|\mathbf{z} - \mathbf{x}\|_2^2 \right).$$

\tilde{f} is convex and it is a $(L, G^2/(2L))$ -smooth approximation of $f(\mathbf{x})$.

Applying AGD to minimize $\tilde{f}(\mathbf{x})$ with $L = \mathcal{O}(G^2/\epsilon)$ can find $\mathcal{O}(\epsilon)$ suboptimal solution of $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$.

- 1 For convex $f(\mathbf{x})$, we require $\tilde{\mathcal{O}}(G/\epsilon)$ iterations.
- 2 For μ -strongly convex $f(\mathbf{x})$, we require $\tilde{\mathcal{O}}(G/\sqrt{\mu\epsilon})$ iterations.

Outline

- 1 Subgradient Descent Method
- 2 Smoothing Technique
- 3 Proximal Gradient Methods

Composite Convex Optimization Problem

We consider the problem of the form

$$\min_{\mathbf{x} \in \mathbb{R}^d} \phi(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}),$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a smooth convex function and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex but possibly nonsmooth.

- ① We focus on the case that g has some simple form.
- ② Subgradient method leads to slow convergence.
- ③ How to obtain the convergence rate like (accelerated) gradient descent?

Proximal Operator

We introduce the proximal operator as follows

$$\text{prox}_h(\mathbf{x}) = \arg \min_{\mathbf{z} \in \mathbb{R}^d} \left(\frac{1}{2} \|\mathbf{z} - \mathbf{x}\|_2^2 + h(\mathbf{x}) \right),$$

where $h : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex but possibly nonsmooth.

Proximal Operator

Recall that optimizing smooth convex function $f(\mathbf{x})$ by gradient descent

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t)$$

is based on minimizing RHS of

$$f(\mathbf{y}) \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

Proximal Gradient Descent

For composite problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} \phi(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}),$$

we can minimize RHS of

$$\phi(\mathbf{y}) = f(\mathbf{y}) + g(\mathbf{y}) \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}_t\|_2^2 + g(\mathbf{y}).$$

That is

$$\mathbf{x}_{t+1} = \text{prox}_{\eta g}(\mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)) \quad \text{with} \quad \eta = 1/L.$$

Proximal Gradient Descent

It can be computed efficiently for some simple $g(\cdot)$. For example:

- ① Let $g(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$, then

$$\text{prox}_{\eta g}(\mathbf{x}) = \begin{bmatrix} \text{sign}(x_1) \max\{|x_1| - \eta\lambda, 0\} \\ \text{sign}(x_2) \max\{|x_2| - \eta\lambda, 0\} \\ \vdots \\ \text{sign}(x_d) \max\{|x_d| - \eta\lambda, 0\} \end{bmatrix},$$

which can be computed efficiently.

- ② Let $g(\mathbf{x}) = \mathbb{1}_{\mathcal{C}}(\mathbf{x})$ for some closed convex \mathcal{C} , then

$$\text{prox}_{\eta g}(\mathbf{x}) = \text{proj}_{\mathcal{C}}(\mathbf{x}),$$

which leads to

$$\mathbf{x}_{t+1} = \text{proj}_{\mathcal{C}}(\mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)).$$

Gradient Mapping

For function $\phi = f + g$ with convex functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $g : \mathbb{R}^d \rightarrow \mathbb{R}$ and $\eta > 0$, we define the gradient mapping as follows

$$\mathcal{G}_{\eta g, f}(\mathbf{x}) = \frac{1}{\eta}(\mathbf{x} - \text{prox}_{\eta g}(\mathbf{x} - \eta \nabla f(\mathbf{x}))),$$

which is a generalization of gradient operator $\nabla f(\mathbf{x})$.

The proximal gradient method

$$\mathbf{x}_{t+1} = \text{prox}_{\eta g}(\mathbf{x}_t - \eta \nabla f(\mathbf{x}_t))$$

is equivalent to

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathcal{G}_{\eta g, f}(\mathbf{x}_t).$$

Gradient Mapping

We consider the composite convex problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} \phi(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}),$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth and convex and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex but possibly nonsmooth.

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

- ① The point \mathbf{x}^* is an optimal solution if and only if $\mathcal{G}_{\eta g, f}(\mathbf{x}^*) = \mathbf{0}$.
- ② Suppose g is μ_g -strongly convex and $\eta < 2/(L - \mu)$, then

$$\|\mathcal{G}_{\eta g, f}(\mathbf{x})\|_2^2 \leq \frac{2/\eta}{2 - \eta(L - \mu_g)} (\phi(\mathbf{x}) - \phi(\mathbf{x}^+)).$$

- ③ Suppose ϕ is μ_ϕ -strongly convex and $\eta < 1/L$, then

$$\phi(\mathbf{x}^+) \leq \phi(\mathbf{x}^*) + \frac{1}{2\mu_\phi} \|\mathcal{G}_{\eta g, f}(\mathbf{x})\|_2^2.$$

Convergence Analysis (Convex)

We consider the composite convex problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} \phi(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}),$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth and convex and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex.

The proximal gradient method with $\eta = 1/L$ holds that

$$\phi(\mathbf{x}_T) \leq \phi(\mathbf{x}^*) + \frac{L}{2T} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2.$$

Additionally suppose ϕ is μ_ϕ -strongly convex leads to

$$\phi(\mathbf{x}_T) - \phi(\mathbf{x}^*) \leq \left(1 - \frac{\mu_\phi}{L + \mu_\phi}\right)^T (\phi(\mathbf{x}_0) - \phi(\mathbf{x}^*)).$$

Convergence Analysis (Nonconvex)

If we only suppose g is convex but allow f be nonconvex, then

$$\mathbb{E} \|\mathcal{G}_{\eta g, f}(\hat{\mathbf{x}})\|_2^2 \leq \frac{2L(\phi(\mathbf{x}_0) - \phi^*)}{T},$$

where $\phi^* = \inf_{\mathbf{x} \in \mathbb{R}^d} \phi(\mathbf{x}) > -\infty$ and $\hat{\mathbf{x}}$ is uniformly sampled from

$$\{\mathbf{x}_0, \dots, \mathbf{x}_{T-1}\}.$$

Here, we say $\hat{\mathbf{x}}$ is an ϵ -stationary point of $\phi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ if

$$\|\mathcal{G}_{\eta g, f}(\hat{\mathbf{x}})\|_2 \leq \epsilon.$$

Accelerated Proximal Gradient Descent

We can also apply Nesterov's acceleration to proximal gradient methods

$$\begin{aligned}\mathbf{y}_t &= \mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1}), \\ \mathbf{x}_{t+1} &= \text{prox}_{\eta g}(\mathbf{y}_t - \eta_t \nabla f(\mathbf{y}_t)).\end{aligned}$$

For convex case, it holds

$$\phi(\mathbf{x}_T) - \phi(\mathbf{x}^*) \leq \mathcal{O}\left(\frac{L}{T^2}\right).$$

For strongly-convex case, it holds

$$\phi(\mathbf{x}_T) - \phi(\mathbf{x}^*) \leq \mathcal{O}\left(\left(1 - \sqrt{\frac{\mu\phi}{L}}\right)^T\right).$$

Subgradient Method vs. Proximal Gradient Method

Solving the composite convex problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} \phi(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}),$$

by subgradient method are based on

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t(\nabla f(\mathbf{x}_t) + \xi_t),$$

where $\xi_t \in \partial g(\mathbf{x}_t)$.

The proximal gradient method is more progressive, since it holds that

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t(\nabla f(\mathbf{x}_t) + \xi_{t+1}),$$

where $\xi_{t+1} \in \partial g(\mathbf{x}_{t+1})$.