Optimization Theory

Lecture 07

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

Nesterov's Acceleration

Nesterov's acceleration:

$$\begin{cases} \mathbf{y}_t = \mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1}), \\ \mathbf{x}_{t+1} = \mathbf{y}_t - \eta_t \nabla f(\mathbf{y}_t). \end{cases}$$

Can we further accelerate Nesterov's acceleration?

First-Order Methods

Let us check our possibilities in minimizing L-smooth convex functions by first-order methods.

Assumption

An iterative method $\mathcal{M}$ generates a sequence of test points $\{\mathbf{x}_t\}$ such that

$$\mathbf{x}_t \in \mathbf{x}_0 + \mathrm{span}\{\nabla f(\mathbf{x}_0), \dots, \nabla f(\mathbf{x}_{t-1})\}.$$

"Worst Functions" and Zero-Chain

Consider the following functions

$$f_t(\mathbf{x}) = \frac{L}{4}\left(\frac{1}{2}\left(x_1^2 + \sum_{i=1}^{t-1}(x_i - x_{i+1})^2 + x_t^2\right) - x_1\right),$$

for $t = 1, \dots, d$, where $\mathbf{x} = [x_1, \dots, x_d]^\top$.

Let $\mathbb{R}^{t,d} = \{\mathbf{x} \in \mathbb{R}^d : x_{t+1} = \cdots = x_d = 0\}$, that is the subspace of $\mathbb{R}^d$, in which only the first t components of the point can differ from zero.

Lemma

Let $\mathbf{x}_0 = \mathbf{0}$. Then for any sequence $\{\mathbf{x}_1, \dots, \mathbf{x}_t\}$ satisfying the condition

$$\mathbf{x}_t \in \mathcal{L}_t = \mathrm{span}\{\nabla f_t(\mathbf{x}_0), \dots, \nabla f_t(\mathbf{x}_{t-1})\},$$

we have $\mathcal{L}_t \subseteq \mathbb{R}^{t,d}$.

"Worst Functions"

Consider the following functions

$$f_t(\mathbf{x}) = \frac{L}{4}\left(\frac{1}{2}\left(x_1^2 + \sum_{i=1}^{t-1}(x_i - x_{i+1})^2 + x_t^2\right) - x_1\right),$$

for $t = 1, \dots, d$, where $\mathbf{x} = [x_1, \dots, x_d]^\top$. They are L-smooth and convex.

We can verify $\nabla^2 f(\mathbf{x}) = \frac{L}{4}\mathbf{A}_t$ with

$$\mathbf{A}_t = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \cdots & -1 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

"Worst Functions"

The equation $\nabla f(\mathbf{x}^*) = \mathbf{0}$ leads to

$$x_i^* = \begin{cases} 1 - \dfrac{i}{t+1}, & i = 1, \dots, t, \\ 0, & i = t+1 \dots, d. \end{cases}$$

Then we have

$$f(\mathbf{x}^*) = -\frac{L}{8}\left(1 - \frac{i}{t+1}\right).$$

and

$$\|\mathbf{x}^*\|_2^2 \leq \frac{t+1}{3}.$$

"Worst Functions"

Lemma

For all $\mathbf{x} \in \mathbb{R}^{t,d}$, *we have* $f_t(\mathbf{x}) = f_p(\mathbf{x})$ *for* $p = t, t+1, \dots, d$.

Corollary

For any $\{\mathbf{x}_t\}_{t=1}^{p}$ *with* $\mathbf{x}_0 = \mathbf{0}$ *and* $\mathbf{x}_t \in \mathcal{L}_t$, *we have* $\mathbf{x}_t \in \mathbb{R}^{t,d}$ *and*

$$f_p(\mathbf{x}_t) = f_t(\mathbf{x}_t) \geq f_t^*$$

for any $p = t, t+1, \dots, d$.

Lower Complexity Bound (Convex)

Theorem

For any t such that $t \in [1, (d-1)/2]$ and any $\mathbf{x}_0 \in \mathbb{R}^d$, there exists an L-smooth and convex function $f : \mathbb{R}^d \to \mathbb{R}$ such that for any first-order algorithm $\mathcal{M}$ with

$$\mathbf{x}_t \in \mathbf{x}_0 + \mathrm{span}\{\nabla f(\mathbf{x}_0), \dots, \nabla f(\mathbf{x}_{t-1})\},$$

we have

$$f(\mathbf{x}_t) - f^* \geq \frac{3L \left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{8(t+1)^2},$$

where $\mathbf{x}^*$ is the minimizer of f and $f^* = f(\mathbf{x}^*)$.

Lower Complexity Bound

1. The above theorem is valid when the iteration number is not too large as compared with the dimension the variables.
2. Without a direct use of finite dimensional arguments, we cannot justify a better complexity of the corresponding numerical scheme.
3. Nesterov's acceleration is optimal for minimizing smooth and convex function by first-order methods.

Lower Complexity Bound (Strongly Convex)

Consider the d-dimensional regularized "worst functions"

$$f(\mathbf{x}) = \frac{L-\mu}{4}\left(\frac{1}{2}\left(x_1^2 + \sum_{i=1}^{d-1}(x_i - x_{i+1})^2 + \left(1 - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)x_d^2\right) - \beta x_1\right) + \frac{\mu}{2}\|\mathbf{x}\|_2^2,$$

for $t = 1, \dots, d$, where $\mathbf{x} = [x_1, \dots, x_d]^\top$, $\beta > 0$ and $\kappa = L/\mu$.

1. The functions are L-smooth and μ-strongly convex.
2. The zero-chain property still holds.
3. The minimizer is $\mathbf{x}^* = [q, q^2, \dots, q^d]$ with $q = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$.

Lower Complexity Bound (Strongly Convex)

Theorem

For any t and d such that $t \leq d/2$, $d \geq 2$ and any $\mathbf{x}_0 \in \mathbb{R}^d$, there exists an L-smooth and μ-strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$ such that for any first-order algorithm $\mathcal{M}$ with

$$\mathbf{x}_t \in \mathbf{x}_0 + \mathrm{span}\{\nabla f(\mathbf{x}_0), \dots, \nabla f(\mathbf{x}_{t-1})\},$$

we have

$$f(\mathbf{x}_t) - f^* \geq \frac{\mu}{2} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2t}.$$

where $\mathbf{x}^*$ is the minimizer of f and $f^* = f(\mathbf{x}^*)$.

Lower Complexity Bound (Nonconvex)

The lower complexity bound for finding an ϵ-stationary point of L-smooth nonconvex function are based on

$$f(\mathbf{x}) = \frac{\sqrt{\mu}}{2}(x_1 - 1)^2 + \frac{1}{2}\sum_{i=1}^{t}(x_{i+1} - x_i)^2 + \mu\sum_{i=1}^{T}\Gamma_r(x_i),$$

where $\mathbf{x} = [x_1, \dots, x_{T+1}]^\top \in \mathbb{R}^{T+1}$ and $\Gamma_r(x) = 120\int_1^x \frac{t^2(t-1)}{1+(t/r)^2}\,\mathrm{d}t$.

1. Let $r \geq 1$ and $\mu \leq 1$. For any $\mathbf{x} \in \mathbb{R}^{T+1}$ such that $x_T = x_{T+1} = 0$, we have $\|\nabla f(\mathbf{x})\|_2 \geq \mu^{3/4}/4$.
2. The lower complexity bound is $\Omega(L\epsilon^{-2}(f(\mathbf{x}_0) - f^*))$. See paper:
    - Yair Carmon, John C. Duchi, Oliver Hinder, Aaron Sidford. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*. 185(1):315–355, 2021.

Outline

Nonsmooth Convex Optimization

We consider optimization with a nonsmooth objective function

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

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

For constrained optimization, we assume the projection operator

$$\mathrm{proj}_{\mathcal{C}}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathcal{C}} \|\mathbf{y}-\mathbf{x}\|_2^2$$

can be efficiently computed for all $\mathbf{y}\in\mathbb{R}^d$.

Subgradient Descent Method

Suppose $\mathcal{C} = \mathbb{R}^d$, we have introduced gradient descent

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

converges by taking $\eta_t = \eta > 0$.

For nonsmooth case, we can replace the gradient by the subgradient and introduce projection step

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

for $t = 0, 1 \dots,$ where $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ and

$$\lim_{t \to +\infty} \eta_t = 0.$$

Convergence Analysis (Convex)

Theorem

A convex function f is G-Lipschitz continuous on $\operatorname{dom} f$ if

$$\max_{\mathbf{g}\in\partial f(\mathbf{x})} \{\|\mathbf{g}\|_2\} \leq G$$

for all $\mathbf{x} \in \operatorname{dom} f$.

Theorem

Let $\mathbf{z} = \operatorname{proj}_{\mathcal{C}}(\mathbf{y})$ for some convex and closed $\mathcal{C} \subseteq \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^d$, then

$$\|\mathbf{z} - \mathbf{x}\|_2^2 \leq \|\mathbf{y} - \mathbf{x}\|_2^2$$

for any $\mathbf{x} \in \mathcal{C}$.

Convergence Analysis (Convex)

We assume the convex function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\max_{\mathbf{g} \in \partial f(\mathbf{x})} \{\|\mathbf{g}\|_2\} \leq G$$

on domain $\mathcal{C}$. Then for all $\hat{\mathbf{x}} \in \mathcal{C}$, we have

$$\frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \eta_t f(\mathbf{x}_t) \leq f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \sum_{t=0}^{T-1} \eta_t^2 G^2}{2 \sum_{t=0}^{T-1} \eta_t}.$$

1. Taking $\eta_t = \eta_0/\sqrt{T}$ leads to

$$\frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) \leq f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \eta_0^2 G^2}{2\eta_0 \sqrt{T}}.$$

2. Taking $\eta_t = \eta_0/(\sqrt{t+1} + \sqrt{t}\,)$ leads to

$$\sum_{t=0}^{T-1} \frac{f(\mathbf{x}_t)}{\sqrt{T(t+1)} + \sqrt{Tt}} \leq f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \eta_0^2(\ln(2T-1) + 2)G^2/2}{2\eta_0 \sqrt{T}}.$$

Convergence Analysis (Strongly Convex)

If we additionally suppose f is μ-strongly convex and set

$$\eta_t = \frac{2}{\mu(t+1)},$$

then

$$\sum_{t=0}^{T-1} \frac{t}{T(T-1)} f(\mathbf{x}_t) \leq f(\hat{\mathbf{x}}) + \frac{2G^2}{\mu(T-1)}.$$