Optimization Theory

Lecture 06

Fudan University

luoluo@fudan.edu.cn

Outline

1. Polyak's Heavy Ball Method

2. Nesterov's Acceleration

Outline

1. Polyak's Heavy Ball Method

2. Nesterov's Acceleration

GD for Quadratic Problem

Consider the quadratic problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} Q(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{b}^\top\mathbf{x}, \tag{1}$$

where $\mathbf{A}\in\mathbb{R}^{d\times d}$ is positive definite and $\mathbf{b}\in\mathbb{R}^d$.

The gradient descent method

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

with $\eta\in(0, 2/L)$ holds that

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2 \leq \rho^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2$$

with $\rho = \max\{1-\eta\mu, |1-\eta L|\} < 1$, where $L = \lambda_1(\mathbf{A})$ and $\mu = \lambda_d(\mathbf{A})$.

Polyak's Heavy Ball Method

The iteration of the heavy ball method is

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla Q(\mathbf{x}_t) + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}),$$

where $\mathbf{x}_{-1} = \mathbf{x}_0$, $\eta > 0$ and $\beta \in (0, 1)$.

1. The motion proceeds not in the direction of the force (i.e. negative gradient) because of the presence of inertia.
2. The term $\beta(\mathbf{x}_t - \mathbf{x}_{t-1})$, giving inertia to the motion, will lead to motion along the "essential" direction.

Polyak's Heavy Ball Method

Theorem

Solving problem (1) by Polyak's heavy ball method

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla Q(\mathbf{x}_t) + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}),$$

with $\eta > 0$ *and* $\beta \in (0, 1)$ *such that* $\beta \geq \max\{(1 - \sqrt{\eta L})^2, (1 - \sqrt{\eta \mu})^2\}$*. Then we have*

$$\begin{bmatrix} \mathbf{x}_{t+1} - \mathbf{x}^* \\ \mathbf{x}_t - \mathbf{x}^* \end{bmatrix} = \mathbf{M} \begin{bmatrix} \mathbf{x}_t - \mathbf{x}^* \\ \mathbf{x}_{t-1} - \mathbf{x}^* \end{bmatrix}.$$

all $t \geq 0$ *and some* $\mathbf{M}$ *with spectral radius of* β*.*

Polyak's Heavy Ball Method

We define

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{x}_{t+1} - \mathbf{x}^* \\ \mathbf{x}_t - \mathbf{x}^* \end{bmatrix}$$

For any $\epsilon > 0$, there exist $N^+ \in \mathbb{N}$ such that for all $t > N^+$, we have

$$\left\|\mathbf{z}_t\right\|_2 < (\beta + \epsilon)^t \left\|\mathbf{z}_0\right\|_2 .$$

Let

$$\eta = \left(\frac{2}{\sqrt{L} + \sqrt{\mu}}\right)^2,$$

then we have

$$\beta = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \approx 1 - \frac{2}{\sqrt{\kappa}}.$$

Outline

1. Polyak's Heavy Ball Method

2. Nesterov's Acceleration

Nesterov's Acceleration

We consider the general problem

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

where $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth and μ-strongly-convex.

The iteration of Nesterov's accelerated gradient descent (AGD)

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

where $\mathbf{x}_{-1} = \mathbf{x}_0$, $\eta_t > 0$ and $\beta_t \in (0,1)$.

Nesterov's Acceleration

The iteration of heavy ball method is

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

which is equivalent to

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

Replacing $\nabla f(\mathbf{x}_t)$ by $\nabla f(\mathbf{y}_t)$ leads to

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

Nesterov's Acceleration

Running AGD iteration

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

with

$$\mathbf{x}_{-1} = \mathbf{x}_0, \quad \eta_t = \eta = \frac{1}{L}, \quad \beta_t = \beta = \frac{1-\theta}{1+\theta} \quad \text{and} \quad \theta = \sqrt{\eta\mu},$$

we have

$$f(\mathbf{x}_t) \leq f(\mathbf{x}^*) + \left(1 - \sqrt{\frac{\mu}{L}}\right)^t \left(f(\mathbf{x}_0) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right).$$

Nesterov's Acceleration

We also conduct the more general framework

$$\begin{cases} \dfrac{\theta_t^2}{\eta_t} = \theta_t \mu + (1-\theta_t)\gamma_{t-1} \\ \gamma_t = (1-\theta_t)\gamma_{t-1} + \theta_t \mu \\ \beta_t = \dfrac{(\theta_t - \mu\eta_t)(1-\theta_{t-1})}{\theta_{t-1}(1-\mu\eta_t)} \\ \mathbf{y}_t = \mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1}) \\ \mathbf{x}_{t+1} = \mathbf{y}_t - \eta_t \nabla f(\mathbf{y}_t) \end{cases}$$

with

$$\mathbf{x}_{-1} = \mathbf{x}_0, \quad \theta_0 = 1, \quad \eta \leq \frac{1}{L} \quad \text{and} \quad \gamma_0 \in \left[\mu,\ \frac{1}{\eta_0}\right].$$

Then we have

$$f(\mathbf{x}_t) \leq f(\mathbf{x}^*) + \prod_{s=1}^{t}(1-\theta_s)\left(f(\mathbf{x}_0) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right).$$

Estimate Sequence

Definition

A pair of sequences $\{\phi_t : \mathbb{R}^d \to \mathbb{R}\}_{t=0}^{+\infty}$ *and* $\{\lambda_t \geq 0\}_{t=0}^{+\infty}$ *is called an estimate sequence of function* $f : \mathbb{R}^d \to \mathbb{R}$ *if*

$$\lim_{t \to +\infty} \lambda_t = 0$$

and for any $\mathbf{x} \in \mathbb{R}^d$ *and all* $t \geq 0$ *we have*

$$\phi_t(\mathbf{x}) \leq (1 - \lambda_t) f(\mathbf{x}) + \lambda_t \phi_0(\mathbf{x}).$$

Lemma

If we have $f(\mathbf{x}_t) \leq \min_{\mathbf{x} \in \mathbb{R}^d} \phi_t(\mathbf{x})$*, for all* $t \geq 0$*, then*

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \lambda_t(\phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)),$$

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

Estimate Sequence

Lemma

For L-smooth and μ-strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$, we define

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

and

$$\psi_t(\mathbf{x}; \mathbf{y}_t) = f(\mathbf{x}_{t+1}) - \frac{1}{2\eta} \|\mathbf{x}_{t+1} - \mathbf{y}_t\|_2^2 + \frac{1}{\eta}\langle \mathbf{y}_t - \mathbf{x}_{t+1}, \mathbf{x} - \mathbf{x}_{t+1}\rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}_t\|_2^2$$

for $\eta_t \leq 1/L$. Then it holds

$$\psi_t(\mathbf{x}; \mathbf{y}_t) \leq f(\mathbf{x}).$$

Estimate Sequence

We can define an estimate sequence $\{(\phi_t, \lambda_t)\}_{t=0}^{+\infty}$ recursively as

$$\phi_t(\mathbf{x}) = (1-\theta_t)\phi_{t-1}(\mathbf{x}) + \theta_t\psi_t(\mathbf{x}; \mathbf{y}_t) \quad \text{and} \quad \lambda_t = (1-\theta_t)\lambda_{t-1}$$

with

$$\phi_0(\mathbf{x}) = f(\mathbf{x}_0) + \frac{\gamma_0}{2}\|\mathbf{x}-\mathbf{x}_0\|_2^2 \qquad \text{and} \qquad \lambda_0 = 1$$

for some $\theta_t \in (0,1)$ such that

$$\lambda_t = \lim_{t\to+\infty}\prod_{s=0}^{t}(1-\theta_s) = 0.$$

Then we only needs to find θ_t that guarantees $f(\mathbf{x}_t) \le \min_{\mathbf{x}\in\mathbb{R}^d}\phi_t(\mathbf{x})$.

Estimate Sequence

By checking Nesterov's acceleration, we achieve

$$\mathbf{v}_t = \frac{1}{\theta_t}(\mathbf{x}_{t+1} - (1-\theta_t)\mathbf{x}_t),$$

where

$$\mathbf{v}_t = \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^d} \phi_t(\mathbf{x}).$$

Then we can prove $f(\mathbf{x}_t) \le \min_{\mathbf{x}\in\mathbb{R}^d} \phi_t(\mathbf{x})$ by induction. Thus we have

$$f(\mathbf{x}_t) \le f(\mathbf{x}^*) + \left(1 - \sqrt{\frac{\mu}{L}}\right)^t \left(f(\mathbf{x}_0) - f(\mathbf{x}^*) + \frac{\mu}{2}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2\right)$$

by taking $\eta_t = 1/L$ and $\beta_t = (1-\theta)/(1+\theta)$ with $\theta = \sqrt{\eta\mu}$.

Nesterov's Acceleration

For the general framework

$$\begin{cases} \dfrac{\theta_t^2}{\eta_t} = \theta_t \mu + (1 - \theta_t)\gamma_{t-1} \\ \gamma_t = (1 - \theta_t)\gamma_{t-1} + \theta_t \mu \\ \beta_t = \dfrac{(\theta_t - \mu\eta_t)(1 - \theta_{t-1})}{\theta_{t-1}(1 - \mu\eta_t)} \\ \mathbf{y}_t = \mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1}) \\ \mathbf{x}_{t+1} = \mathbf{y}_t - \eta_t \nabla f(\mathbf{y}_t) \end{cases}$$

we can start with $\theta_0 = 1$ and gradually decrease it until $\theta \to \sqrt{\lambda\eta}$.

Nesterov's Acceleration

For general convex objective, we set $\gamma_0 = 1/\eta$ and $\mu = 0$, then

$$\lambda_t \leq \theta_t^2 \leq \frac{4}{(t+2)^2}$$

that implies

$$f(\mathbf{x}_t) \leq f(\mathbf{x}^*) + \frac{4}{(t+2)^2}\left(f(\mathbf{x}_0) - f(\mathbf{x}^*) + \frac{\gamma_0}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right).$$