Optimization Theory

Lecture 09

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

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

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

$$\mathbf{x}_{-1} = \mathbf{x}_0, \qquad \eta_t = \frac{1}{L} \qquad \text{and} \qquad \beta_t = \frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1}$$

we have

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

Nesterov's Acceleration

For general convex case, 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

$$\mathbf{x}_{-1} = \mathbf{x}_0, \quad \eta_t = \frac{1}{L} \quad \text{and} \quad \beta_t = \frac{1+\lambda_{t-1}}{\lambda_t} \quad \text{where} \quad \lambda_t = \begin{cases} 0, & t = 0, \\ \frac{1+\sqrt{1+4\lambda_{t-1}}}{2}, & t \geq 1, \end{cases}$$

we have

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