Optimization Theory

Lecture 05

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

Polyak–Łojasiewicz Condition

The linear convergence of gradient descent depends on PL condition

$$f(\mathbf{x}) - f^* \leq \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2,$$

where $f^* = \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$. In fact, it does not require strong convexity.

Consider the function

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is nonzero positive semi-definite (possibly not positive definite).

PL condition holds for (1) with the parameter with $\mu = \lambda_k(\mathbf{A})$, where $\lambda_k(\mathbf{A})$ is the smallest nonzero eigenvalue of $\mathbf{A}$.

Gradient descent still has linear convergence rate!

Polyak–Łojasiewicz Condition

Polyak–Łojasiewicz condition and strong convexity:

1. The μ-strong convexity leads to PL condition with parameter μ.
2. PL condition may not lead to (μ-strong) convexity.

Theorem

Let $g : \mathbb{R}^m \to \mathbb{R}$ be smooth and μ-strongly convex and $\mathbf{A} \in \mathbb{R}^{m\times d}$ is nonzero. Define the function $f : \mathbb{R}^d \to \mathbb{R}$ as $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$, then it satisfies PL condition.

Examples

1. Linear regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2+\frac{\lambda}{2}\|\mathbf{x}\|_2^2,$$

where $\mathbf{A}\in\mathbb{R}^{n\times d}$, $\mathbf{b}\in\mathbb{R}^n$ and $\lambda\geq 0$.

2. Logistic regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\ln(1+\exp(-b_i\mathbf{a}_i^\top\mathbf{x}))+\frac{\lambda}{2}\|\mathbf{x}\|_2^2,$$

where $\mathbf{a}_i\in\mathbb{R}^d$, $b_i\in\{1,-1\}$ and $\lambda\geq 0$.

Outline

Step Size (Learning Rate)

For gradient descent method

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

we have showed its convergence with $\eta_t = 1/L$.

1. It is not easy to evaluate the smoothness parameter L.
2. Directly using $\eta = 1/L$ may not performs well in practice.

Line Search Methods

A line search method computes a search direction $\mathbf{p}_k$ and then decides how far to move along that direction.

The iteration is given by

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t \mathbf{p}_t,$$

where the positive scalar α_t is called step size, step length or learning rate.

We typically require $\mathbf{p}_t$ to be a descent direction that satisfies

$$\langle \mathbf{p}_t, \nabla f(\mathbf{x}_k) \rangle < 0.$$

For example

1. $\mathbf{p}_t = -\nabla f(\mathbf{x}_t)$
2. $\mathbf{p}_t = -\mathbf{G}_t^{-1} \nabla f(\mathbf{x}_t)$ with some positive definite $\mathbf{G}_t \in \mathbb{R}^{d \times d}$

Line Search Methods

The ideal choice for α is based on

$$\min_{\alpha>0} \phi(\alpha) \triangleq f(\mathbf{x}_t + \alpha \mathbf{p}_t),$$

but it is not practical.

We want to efficiently select α_t that leads to sufficient reduction in f.

The simple decrease condition

$$f(\mathbf{x}_t + \alpha_t \mathbf{p}_t) < f(\mathbf{x}_t)$$

is not enough.

Wolfe Conditions

We require

$$
\begin{aligned}
f(\mathbf{x}_t + \alpha_t \mathbf{p}_t) \leq & f(\mathbf{x}_t) + c_1 \alpha_t \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle, \\
\langle \nabla f(\mathbf{x}_t + \alpha_t \mathbf{p}_t), \mathbf{p}_t \rangle \geq & c_2 \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle
\end{aligned}
\tag{2}
$$

for some $c_1 \in (0, 1)$ and $c_2 \in (c_1, 1)$, that is Wolfe conditions.

Theorem

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable. Let $\mathbf{p}_t$ be a descent direction at $\mathbf{x}_t$ and assume $\phi(\alpha) = f(\mathbf{x}_t + \alpha \mathbf{p}_t)$ is bounded below on $\alpha \in (0, +\infty)$. Then there exist intervals of step lengths satisfying the conditions (2) with $0 < c_1 < c_2 < 1$.

Wolfe Conditions

We still consider Wolfe conditions

$$
\begin{aligned}
f(\mathbf{x}_t + \alpha_t \mathbf{p}_t) \leq & f(\mathbf{x}_t) + c_1 \alpha_t \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle, \\
\langle \nabla f(\mathbf{x}_t + \alpha_t \mathbf{p}_t), \mathbf{p}_t \rangle \geq & c_2 \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle
\end{aligned}
\tag{3}
$$

for some $c_1 \in (0, 1)$ and $c_2 \in (c_1, 1)$, that is Wolfe condition.

Theorem

Let $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t \mathbf{p}_t$, where $\mathbf{p}_t$ is a descent direction and α_k satisfies the Wolfe conditions. Suppose that continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and lower bounded on $\mathbb{R}^d$ and continuously differentiable. Then

$$
\sum_{t=0}^{+\infty} (\cos\theta_t)^2 \|\nabla f(\mathbf{x}_t)\|_2^2 < +\infty, \quad \text{where } \cos\theta_t = \frac{-\langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle}{\|\nabla f(\mathbf{x}_t)\|_2 \, \|\mathbf{p}_t\|_2}.
$$

Backtracking Line Search

If the algorithm chooses candidate step lengths appropriately, we can use just the sufficient decrease condition.

Algorithm 1 Backtracking Line Search Method

1: **Input:** $\mathbf{x}_t, \mathbf{p}_t \in \mathbb{R}^d$, $\hat{\alpha} > 0$, $\tau, c_1 \in (0, 1)$
2: $\alpha = \hat{\alpha}$
3: **while** $f(\mathbf{x}_t + \alpha \mathbf{p}_t) > f(\mathbf{x}_t) + c_1 \alpha \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle$ **do**
4: $\quad \alpha \leftarrow \tau \alpha$
5: **Output:** $\alpha_t = \alpha$

Barzilai-Borwein Step Size

Gradient descent methods with Barzilai-Borwein step size has the forms of

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

where

$$\alpha_t = \frac{\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2}{\langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \mathbf{x}_t - \mathbf{x}_{t-1} \rangle}$$

or

$$\alpha_t = \frac{\langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \mathbf{x}_t - \mathbf{x}_{t-1} \rangle}{\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|_2^2}.$$

Outline

GD for Quadratic Problem

Consider the quadratic problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} Q(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}, \tag{4}$$

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 $\eta > 0$ 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 (4) 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

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

when $\kappa \gg 1$.