Optimization Theory

Lecture 07

Fudan University

luoluo@fudan.edu.cn

Outline

1. Black Box Model
2. Gradient Descent Methods
3. Polyak–Łojasiewicz Condition
4. Line Search Methods
5. Barzilai-Borwein Step Size

Outline

1. Black Box Model
2. Gradient Descent Methods
3. Polyak–Łojasiewicz Condition
4. Line Search Methods
5. Barzilai-Borwein Step Size

Convergence Criteria

For the unconstrained convex optimization problem

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

the convergence of an algorithm can be measured by the following in metrics:

1. Convergence in parameter (suppose there exists optimal solution $\mathbf{x}^*$), where we measure the distance

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2 .$$

2. Convergence of objective value, measured by objective suboptimality

$$f(\mathbf{x}_t) - \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}).$$

3. Convergence of gradient

$$\|\nabla f(\mathbf{x}_t)\|_2 .$$

Convergence Criteria

If $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and convex and has an optimal solution $\mathbf{x}^*$, then

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

and

$$\|\nabla f(\mathbf{x}_t)\|_2 = \|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}^*)\|_2 \leq L \|\mathbf{x}_t - \mathbf{x}^*\|_2 ,$$

which implies convergence in parameter implies convergence in objective value and gradient.

The reverse directions may not hold if the objective is not strongly-convex.

Black Box Model

Local black box:

1. The only information available for the numerical scheme is the answer of the oracle.
2. The oracle is local.

Different types of oracles:

1. Zero-order oracle: returns the function value $f(\mathbf{x})$.
2. First-order oracle: returns the function value $f(\mathbf{x})$ and the gradient $\nabla f(\mathbf{x})$.
3. Second-order oracle: returns $f(\mathbf{x})$, $\nabla f(\mathbf{x})$, and the Hessian $\nabla^2 f(\mathbf{x})$.

Black Box Model

There are two participants in the black box model: a learner and an oracle.

1. The learner has
   - infinite computational power,
   - knowledge of the function class to which f belongs,
   - knowledge of the domain.
2. The oracle has specific knowledge of the function.

Black Box Model

The key question:

How many queries to the oracles are necessary and sufficient to find an ϵ-approximate solution?

We will study this question from two perspectives:

1. Upper bound: Designing algorithms.
2. Lower bound: Information theoretic reasoning.

Black Box Model

The strength of the black-box model:

1. It will allow us to derive a complete theory of optimization.
2. We will obtain matching upper and lower bounds on the oracle complexity for various sub-classes of interesting functions.

The weakness of the black-box model:

1. It does not limit our computational resources.
2. The side information of the algorithm is ignored.

Outline

1. Black Box Model
2. Gradient Descent Methods
3. Polyak–Łojasiewicz Condition
4. Line Search Methods
5. Barzilai-Borwein Step Size

Gradient Descent Methods

We consider the optimization problem

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

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

The gradient descent method

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

with $\eta_t = \eta \leq 1/L$ leads to

$$\frac{1}{T}\sum_{t=1}^{t} f(\mathbf{x}_t) \leq f(\hat{\mathbf{x}}) + \frac{L\left\|\mathbf{x}_0 - \hat{\mathbf{x}}\right\|_2^2}{2T}$$

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

Minimizing Convex Function

The gradient descent method

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

with $\eta_t = \eta \leq 1/L$ leads to

$$\frac{1}{T}\sum_{t=1}^{t} f(\mathbf{x}_t) \leq f(\hat{\mathbf{x}}) + \frac{L\left\|\mathbf{x}_0 - \hat{\mathbf{x}}\right\|_2^2}{2T}$$

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

Suppose $f(\cdot)$ has a minimizer $\mathbf{x}^*$ and let $\bar{\mathbf{x}} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{x}_t$, then we need

$$T \geq \left\lceil \frac{L\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{2} \cdot \frac{1}{\epsilon} \right\rceil$$

to guarantee $f(\bar{\mathbf{x}}) - f(\mathbf{x}^*) \leq \epsilon$.

Last-Iterate Convergence

It is also possible the establish the last-iterate convergence

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{2L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{t+4},$$

which is sublinear.

The proof depends on the results

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

and

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\eta}{2} \|\nabla f(\mathbf{x}_t)\|_2^2 .$$

Nonconvex Optimization

The following inequality

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\eta}{2} \|\nabla f(\mathbf{x}_t)\|_2^2 .$$

does not depend on the convexity.

We uniformly sample $\hat{\mathbf{x}}$ from $\{\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}\}$, then

$$\mathbb{E} \|\nabla f(\hat{\mathbf{x}})\|_2^2 \leq \frac{2L(f(\mathbf{x}_0) - f^*)}{T},$$

where we suppose $f^* = \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) > -\infty$

We require

$$T \geq \left\lceil \frac{2L(f(\mathbf{x}_0) - f^*)}{\epsilon^2} \right\rceil$$

to find an ϵ-stationary point of f in expectation.

Minimizing Strongly Convex Function

We consider using gradient descent method

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

with $\eta_t = \eta \leq 1/L$ to solve the optimization problem

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

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

It holds linear convergence rate

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

We require

$$T \geq \left\lceil \kappa \ln \left(\frac{f(\mathbf{x}_T) - f(\mathbf{x}^*)}{\epsilon} \right) \right\rceil$$

to guarantee $f(\mathbf{x}_0) - f(\mathbf{x}^*) \leq \epsilon$, where $\kappa \triangleq L/\mu$ is the condition number.

Example: Regularized Generalized Linear Model

For regularized linear regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \frac{\beta}{2}\|\mathbf{x}\|_2^2$$

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

We have

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^\top\mathbf{A} + \beta\mathbf{I} \qquad \text{and} \qquad \kappa = \frac{\lambda_1(\mathbf{A}^\top\mathbf{A})+\beta}{\lambda_d(\mathbf{A}^\top\mathbf{A})+\beta}.$$

Example: Quadratic Problem

We consider using gradient descent method to solve quadratic problem

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

where $\mathbf{A}$ is positive definite.

Then we have

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \left(1 - \frac{\lambda_d(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^*)),$$

where $\lambda_1(\mathbf{A})$ and $\lambda_d(\mathbf{A})$ are the largest and the smallest eigenvalues of $\mathbf{A}$.

For positive semi-definite $\mathbf{A}$, what about the convergence rate?

Outline

1. Black Box Model
2. Gradient Descent Methods
3. Polyak–Łojasiewicz Condition
4. Line Search Methods
5. Barzilai-Borwein Step Size

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

1. Black Box Model
2. Gradient Descent Methods
3. Polyak–Łojasiewicz Condition
4. Line Search Methods
5. Barzilai-Borwein Step Size

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 and lower bounded. Let $\mathbf{p}_t$ be a descent direction at $\mathbf{x}_t$, 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

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

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:** $\alpha_t = \alpha$

Outline

1. Black Box Model
2. Gradient Descent Methods
3. Polyak–Łojasiewicz Condition
4. Line Search Methods
5. Barzilai-Borwein Step Size

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

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