Optimization Theory

Lecture 04

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

Second-Order Characterization

Theorem (Smoothness and Convexity)

Let $f(\cdot)$ be a twice differentiable function defined on $\mathbb{R}^d$

1. *It is L-smooth if and only if $-L\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$ for all $\mathbf{x} \in \mathbb{R}^d$.*
2. *It is convex if and only if $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$ for all $\mathbf{x} \in \mathbb{R}^d$.*
3. *It is μ-strongly-convex if and only if $\nabla^2 f(\mathbf{x}) \succeq \mu\mathbf{I}$ for all $\mathbf{x} \in \mathbb{R}^d$.*

Sometimes, we say $f(\cdot)$ is ℓ-weakly convex if the function

$$g(\mathbf{x}) = f(\mathbf{x}) + \frac{\ell}{2}\|\mathbf{x}\|_2^2$$

is convex for some $\ell > 0$.

Second-Order Characterization

Some examples:

1. For unconstrained 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 $\mathbf{A}\in\mathbb{R}^{d\times d}$. We have

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}.$$

2. For regularized generalized linear model

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^n \phi_i(\mathbf{a}^\top\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2.$$

   where $\phi_i:\mathbb{R}^d\to\mathbb{R}$ is twice differentiable. We have

$$\nabla f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \phi_i'(\mathbf{a}_i^\top\mathbf{x})\mathbf{a} + \lambda\mathbf{x} \quad \text{and} \quad \nabla^2 f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \phi_i''(\mathbf{a}_i^\top\mathbf{x})\mathbf{a}_i\mathbf{a}_i^\top + \lambda\mathbf{I}.$$

Outline

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

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

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

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

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 oracle:

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 convex optimization.
2. We will obtain matching upper and lower bounds on the oracle complexity for various subclasses 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

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

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

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

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

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

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

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

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

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

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 $\{\mathbf{x}_0, \dots, \mathbf{x}_{T-1}\}$, 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 $2L(f(\mathbf{x}_0) - f^*)\epsilon^{-2}$ first-order oracle complexity 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

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

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

It holds linear convergence rate

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

The value $\kappa \triangleq L/\mu \geq 1$ is called condition number of the function f.

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 $L = \lambda_1(\mathbf{A})$, $\mu = \lambda_d(\mathbf{A})$ and

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

where $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$.

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

Outline

Polyak–Łojasiewicz Condition

The linear convergence of gradient descent depends on PL condition

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

which does not require strong convexity.

Consider the function

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

where $\mathbf{A} \neq \mathbf{0}$ is positive semi-definite but not positive definite.

PL condition holds with the parameter of the smallest non-zero eigenvalue of $\mathbf{A}$, which leads to linear convergence rate of gradient descent method.