Optimization Theory

Lecture 05

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

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. Suppose that $\nabla^2 f(\cdot)$ is continuous in an open neighborhood of $\mathbf{x}^* \in \mathbb{R}^d$.

1. *If $\mathbf{x}^*$ is a local minimizer of $f(\cdot)$, then it holds that*

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}^*) \succeq \mathbf{0}.$$

2. *If it holds that*

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}^*) \succ \mathbf{0},$$

then the point $\mathbf{x}^*$ is a strict local minimizer of $f(\cdot)$.

Outline

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

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

and

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

Applications in Matrix Approximation

Given a symmetric positive-definite matrix $\mathbf{K} \in \mathbb{R}^{d\times d}$ and we sample a subset of columns $\mathbf{C} \in \mathbb{R}^{d\times m}$, where $m < d$.

We want to establish the estimator of $\mathbf{K}$ by the formulation

$$\min_{\mathbf{U}\in\mathbb{R}^{m\times m},\, \delta\in\mathbb{R}} f(\mathbf{U},\delta) \triangleq \left\| \mathbf{K} - (\mathbf{C}\mathbf{U}\mathbf{C}^\top + \delta \mathbf{I}_d) \right\|_F^2 .$$

It has global solution

$$\mathbf{U}^{\mathrm{ss}} = \mathbf{C}^\dagger \mathbf{K} (\mathbf{C}^\dagger)^\top - \delta^{\mathrm{ss}} (\mathbf{C}^\top \mathbf{C})^\dagger$$

and

$$\delta^{\mathrm{ss}} = \frac{1}{d-m} \Big(\mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{C}^\dagger \mathbf{K} \mathbf{C}) \Big) .$$

Applications in Matrix Approximation

We can show that

$$\mathbf{C}\mathbf{U}^{\text{SS}}\mathbf{C}^\top + \delta^{\text{SS}}\mathbf{I}_d \succ \mathbf{0}$$

and

$$\begin{aligned}&(\mathbf{Q}\mathbf{U}^{\text{SS}}\mathbf{Q}^\top + \delta^{\text{SS}}\mathbf{I}_d)^{-1}\\ =&(\delta^{\text{SS}})^{-1}\mathbf{I}_n - (\delta^{\text{SS}})^{-2}\mathbf{Q}(\mathbf{I}_m + (\delta^{\text{SS}})^{-1}\mathbf{U}^{\text{SS}})^{-1}\mathbf{U}^{\text{SS}}\mathbf{Q}^\top.\end{aligned}$$

is well-defined, where $\mathbf{Q} \in \mathbb{R}^{d\times m}$ is the orthogonal bias of $\mathbf{C} \in \mathbb{R}^{d\times m}$.

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

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

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 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 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.