Optimization Theory and Algorithm	**Lecture 16 - 09/15/2021**

Lecture 16

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Junbo Hao

1 Review

Optimization is a special field that is built on the three intertwined pillars:

- **Model**: gives rise to optimization problems.
- **Algorithm**: solves optimization problems.
- **Theory**: supports algorithms and models.

1.1 Review of Modeling

General form of optimization modeling:

Suppose that $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ is a well-defined function. Then

$$\min_{x} \ f(\mathbf{x}), \tag{1}$$

$$\text{s.t. } \mathbf{x} \in \mathcal{X}, \tag{2}$$

where f is called a s an *objective function*, $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top \in \mathcal{X}$ is a *decision variable*, and $\mathcal{X}$ is the so-called *feasible set*. For the feasible set $\mathcal{X}$, it is commonly denoted as

$$\mathcal{X} = \{\mathbf{x} : f_i(\mathbf{x}) \leqslant 0, i = 1, \dots, m \text{ and } g_j(\mathbf{x}) = 0, j = 1, \dots, l\},$$

where $f_i(\mathbf{x}) \leqslant 0, i = 1, \dots, m$ are m *inequality constrains*, and $g_j(\mathbf{x}) = 0, j = 1, \dots, l$ are l *equality constrains*.

Definition 1.1. (Global Minimum)

Point $\mathbf{x}^* \in \mathcal{X}$ is the global minimum of (1) if for any $\mathbf{x} \in \mathcal{X}$, $f(\mathbf{x}) \geqslant f(\mathbf{x}^*) = f^*$.

Definition 1.2. (Local Minimum)

Point $\mathbf{x}^* \in \mathcal{X}$ is a local minimum of (1) if there exists a neighborhood of $\mathbf{x}^*$, $N(\mathbf{x}^*, \epsilon) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\| \leqslant \epsilon\}$, such that for any $\mathbf{x} \in N(\mathbf{x}^*, \epsilon)$, $f(\mathbf{x}) \geqslant f(\mathbf{x}^*)$.

1.2 Review of Algorithm

Optimization algorithms are to design for finding the local and global minimums for the optimization problem.

Algorithms:

- Closed form: e.g., LS problem, $\min \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2$, with the closed solution, $\mathbf{x}^* = (A^\top A)^{-1}A^\top \mathbf{b}$.
- Iterative method: e.g., gradient descent algorithm, $\mathbf{x}^{t+1} = \mathbf{x}^t - s_t \nabla f(\mathbf{x}^t)$.
- Others

1.3 Review of Theory

Optimality Conditions:

- Necessary: $\nabla f(\mathbf{x}^*) = 0$.
- Necessary: $\nabla f(\mathbf{x}^*) = 0$ and $\nabla^2 f(\mathbf{x}^*) \succeq 0$
- Sufficient: $\nabla f(\mathbf{x}^*) = 0$ and $\nabla^2 f(\mathbf{x}^*) > 0$

Table 1: Convergence Theory

	β-smooth	+ Convex	$+\alpha$-strong Convex
$\min\limits_{1\leqslant t\leqslant T} \|\nabla f(\mathbf{x}^t)\|$	$O(1/\sqrt{T})$	$O(1/T)$	NA
$f(\mathbf{x}^T) - f(\mathbf{x}^*)$	NA	$O(1/T)$	$\frac{\beta}{2}\exp(-\frac{\alpha}{\beta}T)\|\mathbf{x}^0 - \mathbf{x}^*\|^2$
$\|\mathbf{x}^T - \mathbf{x}^*\|^2$	NA	NA	$\exp(-\frac{\alpha}{\beta}T)\|\mathbf{x}^0 - \mathbf{x}^*\|^2$

1.4 Summary

- In the previous semester, we have done for $\mathcal{X} = \mathbb{R}^n$ in Eq.(1) which is the unconstrained optimization.
- we have considered the objective function $f \in C^0$, $f \in C^1$ and $f \in C^2$. Convex or Nonconvex?

Algorithm for $f \in C^0$:

- Sub-gradient descent

- Proximal gradient descent

Algorithm for $f \in C^1$:

- Gradient Descent (linear search, β-smooth, α-strong convex)
- Accelerated Gradient Descent

Algorithm for $f \in C^2$:

- General Newton Method
- SR1, BFGS, DFP

Theory = Optimilaty Condition + Convergence Theory.

What is new? From optimization modeling framework, we consider $\mathcal{X} \neq \mathbb{R}^n$ in this semester. How to do?

1.5 Optimization with Linear Equality Constrains

Let us consider a special case which is called "quadratic programing".

Example 1.3.

$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top P\mathbf{x} + q^\top \mathbf{x} + r, \tag{3}$$

$$s.t.\ A\mathbf{x} = \mathbf{b}, \tag{4}$$

where $P \succ 0$. If we disregard the quality constrain, the optimality condition of unconstrained optimization says: $\nabla f(\mathbf{x}^*) = P\mathbf{x}^* + q = 0$, that is $\mathbf{x}^* = -P^{-1}q$. Thus, a natural question should be asked that what optimality conditions of Eq.(3).

To this end, the optimality conditions of general convex optimization formulation (1) are provided via the following theorem.

Theorem 1.4. $\mathbf{x}^*$ *is optimal of the convex optimization problem* (1) *if and only if*

$$\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geqslant 0, \text{ for all } \mathbf{y} \in \mathcal{X}. \tag{5}$$

Proof. (i) If $\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geqslant 0$ for all $\mathbf{y} \in \mathcal{X}$, then we have $f(\mathbf{y}) \geqslant f(\mathbf{x}^*)$ due to the convexity of f, namely

$$f(\mathbf{y}) \geqslant f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle . \tag{6}$$

(ii) Suppose that $\mathbf{x}^*$ is optimal, but the condition (**??**) does not hold, i.e., there exists $\mathbf{y} \in \mathcal{X}$ such that

$$\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle < 0.$$

Let $\mathbf{z} = \lambda \mathbf{y} + (1-\lambda)\mathbf{x}^*$, then

$$\begin{aligned} \frac{\partial f(\mathbf{z})}{\partial \lambda}|_{\lambda=0} &= \langle \nabla f(\lambda \mathbf{y} + (1-\lambda)\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \, |_{\lambda=0} \\ &= \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle < 0. \end{aligned}$$

This implies that $f(\mathbf{z}) < f(\mathbf{x}^*)$. Contradiction! ■

Remark 1.5.

- *Theorem 1.4 shows that $-\nabla f(\mathbf{x}^*)$ defines a supporting hyperplane to the feasible set at $\mathbf{x}^*$.*

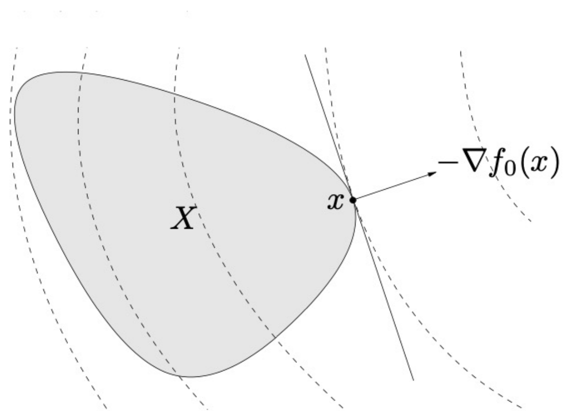

Figure 1: Geometric Interpretation of Optimality Condition

- *If $\mathcal{X} = \mathbb{R}^n$, then the condition (5) reduces to the unconstrained optimality condition,* $\nabla f(\mathbf{x}^*) = 0$.

Example 1.6. Let us consider the following general convex optimization with linear equality constrains.

$$\min_{\mathbf{x}} f(\mathbf{x}), \tag{7}$$

$$s.t. \; A\mathbf{x} = \mathbf{b}. \tag{8}$$

We will write down the optimality condition of (7) according to Theorem 1.4.

First, Theorem 1.4 shows that

$$\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geqslant 0, A\mathbf{x}^* = 0, A\mathbf{y} = 0.$$

So, $A(\mathbf{x}^* - \mathbf{y}) = 0$ and $\mathbf{y} - \mathbf{x}^* \in \mathcal{N}(A)$. Let $\mathbf{v} = \mathbf{y} - \mathbf{x}^*$, then $\mathbf{v}^\top \nabla f(\mathbf{x}^*) \geqslant 0$. However, $\mathcal{N}(A)$ is a linear space, we thus have $\mathbf{y}'$ such that $\mathbf{y}' - \mathbf{x} = -\mathbf{v}$, then $\mathbf{v}^\top \nabla f(\mathbf{x}^*) \leqslant 0$. Finally, we have $\mathbf{v}^\top \nabla f(\mathbf{x}^*) = 0$ and $\nabla f(\mathbf{x}^*) \perp \mathcal{N}(A)$. Thus, $\nabla f(\mathbf{x}^*) \in \mathcal{C}(A^\top)$, there exists $\lambda \in \mathbb{R}^n$ such that

$$\nabla f(\mathbf{x}^*) + A^\top \lambda = 0 \text{ (Optimality Condition)}.$$

To obtain the optimal point, we have to solve the following equations.

$$(*) = \begin{cases} Ax^* = b, \\ \nabla f(\mathbf{x}^*) + A^\top \lambda = 0. \end{cases}$$

For Example 1.3, it becomes a linear equation system:

$$\begin{cases} Ax^* = b, \\ \nabla P\mathbf{x}^* + q + A^\top \lambda = 0. \end{cases}$$

Actually, variable λ is called **dual variable** which will be denoted in the next section.

Q: How to solve the general equation system (*)?