# Lecture 09. Convex Optimization Problems

Lecturer: Jie Wang Date: Nov 3, 2021

The major reference of this lecture is [2, 3].

## 1 Introduction

We are given a data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We would like to fit the data by linear models. We have learned how to find the optimal linear model by two different approach. The good news is that the problem admits a closed form solution:

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

which involves computing the inverse matrix. This can be computationally intractable. Thus, we would like to find $\hat{\mathbf{w}}$ by an iterative approach, that is, gradient descent.

To simplify notations, we use $\|\cdot\|$ to denote the Euclidean norm $\|\cdot\|_2$.

## 2 Basic Terminology

**Definition 1.** A general **convex optimization problem** takes the form as follows.

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}) \\ &\text{s.t. } \mathbf{x} \in D, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a proper convex function and $D \subseteq \mathbb{R}^n$ is a nonempty convex set with $D \subseteq \mathbf{dom}\ f$. The set $D$ is the **feasible set**, and each element in $D$ is called a **feasible solution**.

**Definition 2.** A point $\mathbf{x}^* \in \mathbb{R}^n$ is an **optimal point**, or solves the problem (1), if $\mathbf{x}^*$ is a feasible solution, i.e., $\mathbf{x}^* \in D$, and

$$f(\mathbf{x}^*) = f^* = \inf_{\mathbf{x} \in D} f(\mathbf{x}). \tag{2}$$

The value $f^*$ defined in Eq. (2) is the **optimal value**. The set of all optimal points is the **optimal set**, denoted by

$$X^* = \{\mathbf{x}^* : \mathbf{x}^* \in D, f(\mathbf{x}^*) = f^*\}.$$

***Remark*** 1.

- If the problem (1) has an optimal solution, we say the optimal value is *attained* or *achieved*, and the problem is *solvable*. Otherwise ($X^*$ is empty), we say the optimal value is not attained or not achieved.
- A feasible point $\mathbf{x}$ with $f(\mathbf{x}) \leq f^* + \epsilon$ ($\epsilon > 0$) is called *$\epsilon$-suboptimal*, and the set of all $\epsilon$-suboptimal points is called *$\epsilon$-suboptimal set* for the problem (1).

**Proposition 1.** *Suppose that the problem (1) is a convex optimization problem and solvable. Then, the optimal set $X^*$ is convex.*

*Proof.* If there is only one point in $X^*$, we can see that $X^*$ is clearly convex. Thus, we consider the cases where there are multiple points in $X^*$.

Suppose that $\mathbf{x}, \mathbf{y} \in X^*$ and $\mathbf{x} \neq \mathbf{y}$. As $X^* \subseteq D$, the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ belongs to the feasible set $D$ as well. Let $\theta \in (0, 1)$. Then,

$$f^* \leq f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) = f^*,$$

which implies that

$$f^* = f(\theta\mathbf{x} + (1-\theta)\mathbf{y}).$$

Thus, the points on the segment joining $\mathbf{x}$ and $\mathbf{y}$ belong to $X^*$, and thus $X^*$ is convex. This completes the proof. □

**Definition 3.** A feasible point $\mathbf{x}$ is **locally optimal** if there is a $\delta > 0$ such that

$$f(\mathbf{x}) = \inf\{f(\mathbf{z}) : \mathbf{x} \in D, \|\mathbf{z} - \mathbf{x}\| < \delta\}.$$

**Theorem 1.** *Suppose that the problem (1) is a convex optimization problem and solvable. Then, if* $\mathbf{x}$ *is a local optimum, it is also a global optimum.*

*Proof.* Let $\mathbf{y} \in D$ be an arbitrary feasible point other than $\mathbf{x}$. Thus, to show that the claim holds, it suffices to show that,

$$f(\mathbf{x}) \leq f(\mathbf{y}). \tag{3}$$

As $\mathbf{x}$ is a local optimum, we can find a $\delta > 0$ such that

$$f(\mathbf{x}) \leq f(\mathbf{z}), \forall\, \mathbf{z} \in D \cap B := \{\mathbf{z} : \|\mathbf{z} - \mathbf{x}\| < \delta\}.$$

Clearly, if $\mathbf{y} \in B$, the inequality (3) holds. Thus, we only need to consider the case where $\mathbf{y} \notin B$, i.e.,

$$\|\mathbf{y} - \mathbf{x}\| \geq \delta.$$

Due to the convexity of $D$, all the points on the line segment $\ell$ joining $\mathbf{x}$ and $\mathbf{y}$ belong to $D$. Let

$$\theta = 1 - \frac{\delta}{2\|\mathbf{y} - \mathbf{x}\|},$$

and

$$\mathbf{z}_0 = \theta\mathbf{x} + (1-\theta)\mathbf{y}.$$

We can see that $\mathbf{z}_0$ is on the line segment $\ell$ as $\theta \in (0, 1)$, and

$$\|\mathbf{z}_0 - \mathbf{x}\| = \frac{\delta}{2}.$$

This implies that $\mathbf{z}_0 \in B$ and thus

$$f(\mathbf{x}) \leq f(\mathbf{z}_0). \tag{4}$$

Combining with the convexity of $f$, we have

$$f(\mathbf{x}) \leq f(\mathbf{z}_0) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}).$$

By moving $\theta f(\mathbf{x})$ to the LHS, and dividing both sides by $(1-\theta)$, we can see that the inequality (3) holds. This completes the proof. □

***Remark*** 2. Another way—which is much easier—to show Theorem 1 is by noting that, for any $\mathbf{z}$ lying on the line segment joining $\mathbf{x}$ and $\mathbf{y}$, we have

$$f(\mathbf{z}) \leq \max\{f(\mathbf{x}), f(\mathbf{y})\}.$$

**Proposition 2.** *Suppose that the problem (1) is solvable. Then, if $f$ is strictly convex, the problem (1) has a unique global optimum.*

**Proposition 3.** *Consider the problem (1). If $f$ is strongly convex and continuous over its domain, and the feasible set is closed, then the problem (1) is solvable and has a unique global optimum.*

## 3 Optimality Conditions

**Theorem 2.** *Suppose that the problem (1) is solvable. If $f$ is continuously differentiable, then $\mathbf{x}$ is optimal if and only if $\mathbf{x} \in D$ and*

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0, \forall \mathbf{y} \in D. \tag{5}$$

*Proof.*

$\Leftarrow$ Suppose that the inequality (6) holds. Combining the convexity of $f$ leads to

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\ \Rightarrow f(\mathbf{y}) &\geq f(\mathbf{x}), \forall \mathbf{y} \in D. \end{aligned}$$

$\Rightarrow$ Suppose that $\mathbf{x}$ is optimal. Then,

$$\frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} \geq 0, \forall t \in (0, 1].$$

Letting $t$ goes to zero on both sides leading to

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \lim_{t \downarrow 0} \frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} \geq 0.$$

This completes the proof. □

**Corollary 1.** *Suppose that the function $f$ is continuously differentiable in problem (1). If $\mathbf{x}^*$ is an interior point of $D$, then*

$$\mathbf{x}^* \in \underset{\mathbf{x} \in D}{\mathbf{argmin}}\ f(\mathbf{x}) \Leftrightarrow \nabla f(\mathbf{x}^*) = 0.$$

If we do not require the differentiability of $f$, we have the counterparts of Theorem 2 and Corollary 1 as follows.

**Proposition 4.** *Suppose that the problem (1) is solvable. If $\mathbf{x} \in \mathbf{int}\,(\mathbf{dom}\ f)$, then $\mathbf{x}$ is optimal if and only if $\mathbf{x} \in D$ and there exists a $\mathbf{g} \in \partial f(\mathbf{x})$ such that*

$$\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \geq 0, \forall \mathbf{y} \in D. \tag{6}$$

**Corollary 2.** *Suppose that the problem (1) is solvable. If $\mathbf{x}^*$ is an interior point of $D$, then*

$$\mathbf{x}^* \in \underset{\mathbf{x} \in D}{\mathbf{argmin}}\ f(\mathbf{x}) \Leftrightarrow 0 \in \partial f(\mathbf{x}^*).$$

**Example 1.** Let $f(x) = |x|$, where $x \in \mathbb{R}$. Find $x^*$.

**Example 2.** Lasso takes the form of

$$\min_{\mathbf{w} \in \mathbb{R}^n} \frac{1}{n} \|\mathbf{y} - X\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|_1. \tag{7}$$

Suppose that $\hat{\mathbf{w}}$ solves the above problem. Please write down the optimality condition at $\hat{\mathbf{w}}$.

# 4 Problem Setup

We consider the unconstrained optimization problem as follows.

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}). \tag{8}$$

We further assume that

1. $g : \mathbb{R}^n \to \mathbb{R}$ is a continuous convex function, which is possibly nonsmooth;
2. the objective function $f$ is convex and continuously differentiable, and thus
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle, \ \forall\, \mathbf{x}, \mathbf{y}; \tag{9}$$
3. the gradient of function $f$ is Lipschitz continous, i.e.,
$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \tag{10}$$
where $L > 0$ is the so-called Lipschitz constant;
4. the problem in (8) is solvable, i.e., there exists $\mathbf{x}^*$ such that
$$F(\mathbf{x}^*) = F^* = \min F(\mathbf{x}). \tag{11}$$
Notice that, the point $\mathbf{x}^*$ that satisfies Eq. (11) may not be unique.

# 5 The Proximal Gradient Algorithm

We introduce an efficient algorithm to solve the problem (8), called Iterative Shrinkage-Thresholding Algorithm (ISTA) [1], which is a special case of the popular **proximal gradient methods** for solving nonsmooth optimization problems.

## 5.1 The basic approximation model

**Lemma 1.** *Suppose that a function $f$ is continuously differentiable. If the gradient of $f$ is Lipschitz continuous with Lipschitz constant $L$, i.e., the inequality (10) holds, then we have*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{12}$$

*Proof.* Suppose that the inequality (10) holds. Then,

$$\begin{aligned}
f(\mathbf{y}) =& f(\mathbf{x}) + \int_{\mathbf{x}}^{\mathbf{y}} \nabla f(\mathbf{z}) d\mathbf{z} \\
=& f(\mathbf{x}) + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x}\rangle dt \\
=& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle dt \\
\leq& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|\|\mathbf{y} - \mathbf{x}\| dt \\
\leq& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \int_0^1 Lt\|\mathbf{y} - \mathbf{x}\|^2 dt \\
\leq& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2,
\end{aligned}$$

which completes the proof. □

We consider the following quadratic approximation of $F$ at a given point $\mathbf{x}_c$:

$$Q(\mathbf{x};\mathbf{x}_c) = f(\mathbf{x}_c) + \langle \nabla f(\mathbf{x}_c), \mathbf{x}-\mathbf{x}_c\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{x}_c\|^2 + g(\mathbf{x}),$$

which admits a unique minimizer

$$\begin{aligned} p(\mathbf{x}_c) &= \mathbf{argmin}\{Q(\mathbf{x};\mathbf{x}_c) : \mathbf{x}\in\mathbb{R}^n\} \\ &= \mathbf{argmin}_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{L}{2}\left\|\mathbf{x} - \left(\mathbf{x}_c - \frac{1}{L}\nabla f(\mathbf{x}_c)\right)\right\|^2 \right\}. \end{aligned}$$

Thus, the basic step of ISTA to solve the problem (8) is

$$\mathbf{x}_{k+1} = p(\mathbf{x}_k).$$

**Algorithm 1** ISTA

**Input:** An initial point $\mathbf{x}_0$, a Lipschitz constant $L$, and $k=0$.

1: **while** the *termination condition* does not hold **do**
2: $\quad \mathbf{x}_{k+1} \leftarrow p(\mathbf{x}_k)$,
3: $\quad k \leftarrow k+1$,
4: **end while**

**Example 3 (The Shrinkage Operator).** Please find $p(\mathbf{w})$ for the Lasso problem (7).

**Solution:** To simplify notations, let

$$\mathbf{w}^+ := p(\mathbf{w}),$$

and

$$\mathbf{z} = \mathbf{w} - \frac{1}{L}\nabla f(\mathbf{w}) = \mathbf{w} - \frac{2}{Ln}X^\top(X\mathbf{w}-\mathbf{y}).$$

Then,

$$0 \in \partial\lambda\|\mathbf{w}^+\|_1 + \frac{L}{2}\nabla_{\mathbf{w}}\|\mathbf{w}^+ - \mathbf{z}\|^2 \Rightarrow \frac{L}{\lambda}(\mathbf{z}-\mathbf{w}^+) \in \partial\|\mathbf{w}^+\|_1,$$

which leads to

$$w_i^+ = \begin{cases} z_i + \dfrac{\lambda}{L}, & \text{if } z_i < -\dfrac{\lambda}{L}, \\ 0, & \text{if } |z_i| < \dfrac{\lambda}{L}, \\ z_i - \dfrac{\lambda}{L}, & \text{if } z_i > \dfrac{\lambda}{L}. \end{cases} \tag{13}$$

The mapping defined in Eq. (13) is the so-called shrinkage operator—which is a special case of the **proximal operator** for the nonsmooth $\ell_1$ norm—leading to an efficient implementation of ISTA for Lasso. ■

# 6 Convergence Property of ISTA

In this section, we analyze the convergence property of Algorithm 1. We first show that the function values generated by Algorithm 1 monotonically decrease. This is where "descent" in gradient descent comes from. Then, we show that the function values approach $F^*$ with a rate of $O(1/k)$. We will see that convexity and Lipschitz continuity play a central role in deriving the convergence properties.

## 6.1 Convergence in terms of the Function Values

In this section, we show that $F(\mathbf{x}_k) \to F(\mathbf{x}^*)$ with a convergence rate $O(1/k)$. We first show a *descent* lemma as follows.

**Lemma 2.** *For any $\mathbf{x}_c \in \mathbb{R}^n$, one has $\mathbf{x}_c^+ = p(\mathbf{x}_c)$ if and only if there exists $\mathbf{s} \in \partial g(\mathbf{x}_c^+)$, such that*

$$\nabla f(\mathbf{x}_c) + L(\mathbf{x}_c^+ - \mathbf{x}_c) + \mathbf{s} = 0.$$

*Proof.* The claim follows immediately by Corollary 2. □

**Lemma 3.** *Let $\mathbf{x}_c \in \mathbb{R}^n$, $L > 0$, and $\mathbf{x}_c^+ = p(\mathbf{x}_c)$ such that*

$$F(\mathbf{x}_c^+) \le Q(\mathbf{x}_c^+; \mathbf{x}_c). \tag{14}$$

*Then, for any $\mathbf{x} \in \mathbb{R}^n$,*

$$F(\mathbf{x}) - F(\mathbf{x}_c^+) \ge \frac{L}{2}\|\mathbf{x}_c^+ - \mathbf{x}_c\|^2 + L\langle \mathbf{x}_c - \mathbf{x}, \mathbf{x}_c^+ - \mathbf{x}_c\rangle.$$

*Proof.* By (14), we have

$$F(\mathbf{x}) - F(\mathbf{x}_c^+) \ge F(\mathbf{x}) - Q(\mathbf{x}_c^+; \mathbf{x}_c).$$

As $f, g$ are convex, we have

$$\begin{aligned} f(\mathbf{x}) &\ge f(\mathbf{x}_c) + \langle \nabla f(\mathbf{x}_c), \mathbf{x} - \mathbf{x}_c\rangle, \\ g(\mathbf{x}) &\ge g(\mathbf{x}_c^+) + \langle \mathbf{s}, \mathbf{x} - \mathbf{x}_c^+\rangle, \end{aligned}$$

where $\mathbf{s}$ is defined in Lemma 2. Summing the above inequalities yields

$$F(\mathbf{x}) \ge f(\mathbf{x}_c) + \langle \nabla f(\mathbf{x}_c), \mathbf{x} - \mathbf{x}_c\rangle + g(\mathbf{x}_c^+) + \langle \mathbf{s}, \mathbf{x} - \mathbf{x}_c^+\rangle.$$

On the other hand, the definition of $\mathbf{x}_c^+$ leads to

$$Q(\mathbf{x}_c^+; \mathbf{x}_c) = f(\mathbf{x}_c) + \langle \nabla f(\mathbf{x}_c), \mathbf{x}_c^+ - \mathbf{x}_c\rangle + \frac{L}{2}\|\mathbf{x}_c^+ - \mathbf{x}_c\|^2 + g(\mathbf{x}_c^+).$$

All together, we have

$$\begin{aligned} F(\mathbf{x}) - F(\mathbf{x}_c^+) &\ge -\frac{L}{2}\|\mathbf{x}_c^+ - \mathbf{x}_c\|^2 + \langle \nabla f(\mathbf{x}_c) + \mathbf{s}, \mathbf{x} - \mathbf{x}_c^+\rangle \\ &= -\frac{L}{2}\|\mathbf{x}_c^+ - \mathbf{x}_c\|^2 + L\langle \mathbf{x}_c - \mathbf{x}_c^+, \mathbf{x} - \mathbf{x}_c^+\rangle \\ &= \frac{L}{2}\|\mathbf{x}_c^+ - \mathbf{x}_c\|^2 + L\langle \mathbf{x}_c^+ - \mathbf{x}_c, \mathbf{x}_c - \mathbf{x}\rangle, \end{aligned}$$

which completes the proof. □

***Remark*** 3. When we set $\mathbf{x} := \mathbf{x}_c$, Lemma 3 implies that

$$F(\mathbf{x}_c^+) \leq F(\mathbf{x}_c) - \frac{L}{2}\|\mathbf{x}_c^+ - \mathbf{x}_c\|^2.$$

That is, the sequence of function values $F(\mathbf{x}_0), F(\mathbf{x}_1), F(\mathbf{x}_2), \ldots$ generated by Algorithm 1 monotonically decreases as long as $\mathbf{x}_{k+1} \neq \mathbf{x}_k$.

**Theorem 3.** *Let $(\mathbf{x}_k)$ be the sequence generated by Algorithm 1. Then, for any $k \geq 1$*

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2k}, \forall\, \mathbf{x}^* \in X^*.$$

*Proof.* Invoking Lemma 3 with $\mathbf{x} = \mathbf{x}^*$, $\mathbf{x}_c = \mathbf{x}_n$, we have

$$\begin{aligned}\frac{2}{L}(F(\mathbf{x}^*) - F(\mathbf{x}_{n+1})) \geq& \|\mathbf{x}_{n+1} - \mathbf{x}_n\|^2 + 2\langle \mathbf{x}_{n+1} - \mathbf{x}_n, \mathbf{x}_n - \mathbf{x}^*\rangle \\ =& \|\mathbf{x}^* - \mathbf{x}_{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}_n\|^2.\end{aligned}$$

Summing this inequality over $n = 0, \ldots, k-1$ leads to

$$\frac{2}{L}\left(kF(\mathbf{x}^*) - \sum_{n=0}^{k-1} F(\mathbf{x}_{n+1})\right) \geq \|\mathbf{x}^* - \mathbf{x}_k\|^2 - \|\mathbf{x}^* - \mathbf{x}_0\|^2.$$

By noting that $F(\mathbf{x}_k)$ monotonically decreases, we have

$$\frac{2k}{L}\left(F(\mathbf{x}^*) - F(\mathbf{x}_k)\right) \geq \frac{2}{L}\left(kF(\mathbf{x}^*) - \sum_{n=0}^{k-1} F(\mathbf{x}_{n+1})\right) \geq \|\mathbf{x}^* - \mathbf{x}_k\|^2 - \|\mathbf{x}^* - \mathbf{x}_0\|^2,$$

which leads to

$$F(\mathbf{x}^*) - F(\mathbf{x}_k) \leq \frac{L}{2k}\left(\|\mathbf{x}^* - \mathbf{x}_0\|^2 - \|\mathbf{x}^* - \mathbf{x}_k\|^2\right) \leq \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2k}.$$

This completes the proof. □

## References

[1] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Process.*, 18:2419–2434, 2009.

[2] D. Bertsekas. *Convex Optimization Theory.* Athena Scientific, 2009.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.