EE3001 Machine Learning Fall 2024

Lecture 08. Convex Optimization Problems

Lecturer: Xiaojun Chang

Date: Oct 30, 2025
The major reference of this lecture is [2, 3].

1 Introduction

We are given a data set $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ 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:

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

which involves computing the inverse matrix. This can be computationally intractable. Thus, we would like to find $\hat{w}$ 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.

$$\min_x f(x) \quad \text{s.t. } x \in D, \tag{1}$$

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 $D \subseteq \operatorname{dom} f$. The set D is the feasible set, and each element in D is called a feasible solution.

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

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

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

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

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 x with $f(x) \leq f^* + \varepsilon$ ($\varepsilon > 0$) is called ε-suboptimal, and the set of all ε-suboptimal points is called the ε-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 $x, y \in X^*$ and $x \neq y$. As $X^* \subseteq D$, the line segment connecting x and y belongs to the feasible set D as well. Let $\theta \in (0, 1)$. Then,

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

leading to $f^* = f(\theta x + (1-\theta)y)$. This implies that the points on the segment joining x and y belong to X^*, and thus X^* is convex. □

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

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

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

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

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

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

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

Clearly, if $y \in B$, the inequality (3) holds. Thus, we only need to consider the case where $y \notin B$, i.e. $\|y - x\| \geq \delta$. Due to the convexity of D, all the points on the line segment ℓ joining x and y belong to D. Let

$$\theta = 1 - \frac{\delta}{2\|y - x\|}, \qquad z_0 = \theta x + (1-\theta)y.$$

We can see that z_0 is on the line segment ℓ as $\theta \in (0, 1)$, and $\|z_0 - x\| = \delta/2$. This implies that $z_0 \in B$ and thus

$$f(x) \leq f(z_0). \tag{4}$$

Combining with the convexity of f, we have

$$f(x) \leq f(z_0) \leq \theta f(x) + (1-\theta) f(y).$$

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

Remark 2. We can show Theorem 1 by contradiction. This approach is inspired by the epigraph. Suppose that x is not the global optimum, that is, we can find a feasible solution y such that

$$f(y) < f(x).$$

Then, for any $\theta \in [0, 1]$,

$$f((1-\theta)x + \theta y) = f(x + \theta(y - x)) \leq f(x) + \theta(f(y) - f(x)) < f(x),$$

which implies that we cannot find a neighborhood $B_\varepsilon(x) = \{z : \|z - x\| < \varepsilon\}$ of x such that $f(x) \leq f(z)$ for all $z \in B_\varepsilon(x)$. This contradicts the fact that x is a local optimum.

Remark 3. Another way—which is much easier—to show Theorem 1 is by noting that, for any z lying on the line segment joining x and y, we have

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

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 x is optimal if and only if $x \in D$ and*

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

Proof. ($\Leftarrow$) Suppose that the inequality (5) holds. Combining the convexity of f leads to

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x\rangle \Rightarrow f(y) \geq f(x), \;\; \forall y \in D.$$

($\Rightarrow$) Suppose that x is optimal. Then,

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

Letting t go to zero on both sides leading to

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

This completes the proof. □

Question 1. Recall that, for closed convex set $C \subseteq \mathbb{R}^n$ and a point $x \in \mathbb{R}^n$ with $x \notin C$, a point $z \in C$ is the projection of x on C if and only if

$$\langle x - z,\, y - z\rangle \leq 0, \;\; \forall y \in C.$$

The inequality is the so-called variational inequality. Can you show that the above variational inequality holds by Theorem 2?

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

$$x^* \in \underset{x \in D}{\arg\min}\, f(x) \iff \nabla f(x^*) = 0.$$

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 $x \in \operatorname{int}(\operatorname{dom} f)$, then x is optimal if and only if $x \in D$ and there exists a $g \in \partial f(x)$ such that*

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

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

$$x^* \in \underset{x \in D}{\arg\min}\, f(x) \iff 0 \in \partial f(x^*).$$

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

Solution: By Corollary 2, we can see that $0 \in \partial f(x^*)$. As we have seen that

$$\partial f(x) = \begin{cases} \{1\}, & x > 0, \\ [-1, 1], & x = 0, \\ \{-1\}, & x < 0, \end{cases}$$

we can conclude that $x^* = 0$. □

Example 2. Lasso takes the form of

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

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

4 Problem Setup

We consider the unconstrained optimization problem as follows.

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

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
$$f(y) \geq f(x) + \langle \nabla f(x), \, y - x \rangle \, , \; \forall x, y; \tag{9}$$
3. the gradient of function f is Lipschitz continous, i.e.,
$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \tag{10}$$
where $L > 0$ is the so-called Lipschitz constant;
4. the problem in (8) is solvable, i.e., there exists x^* such that
$$F(x^*) = F^* = \min F(x). \tag{11}$$
Notice that the point 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)* [**?**], 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*

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

Proof. Suppose that the inequality (10) holds. We have

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

which completes the proof. □

Consider the following quadratic approximation of F at a given point x_c:

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

In view of Lemma 1, we can see that the function $F(x)$ is upper bounded by $Q(x;x_c)$, that is,

$$F(x) \leq Q(x;x_c),\ \forall x. \tag{13}$$

When $x = x_c$, the inequality becomes equality. Moreover, we note that $Q(x;x_c)$ admits a unique minimizer (why?)

$$p(x_c) := \arg\min\{Q(x;x_c) : x \in \mathbb{R}^n\} = \arg\min_x \left(g(x) + \frac{L}{2}\left\| x - \left(x_c - \frac{1}{L}\nabla f(x_c)\right)\right\|^2 \right). \tag{14}$$

Combining (13) and (14) leads to

$$F(p(x_c)) \leq Q(p(x_c);x_c) \leq Q(x_c;x_c) = F(x_c).$$

This implies that we can improve the function value $F(x_k)$ by minimizing its quadratic approximation $Q(x;x_k)$ based on its current solution x_k. Repeating this procedure, we may expect that the sequence (x_k) converges to the optimal solution. Indeed, this is the idea of ISTA, which is described in Algorithm 1.

Algorithm 1 ISTA

```
1. Input: An initial point $x_0$, a Lipschitz constant $L$, and $k = 0$.
2. while the termination condition does not hold do
    2.1. $x_{k+1} \leftarrow p(x_k)$,
    2.2. $k \leftarrow k + 1$,
3. end while
```

In view of Algorithm 1, we can see that the key is how to find $p(x_k)$. This can be highly nontrivial. Fortunately, for many popular $g(\cdot)$, we do have closed-form solutions, leading to highly efficient implementations of ISTA.

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

Solution: Let $f(w) = \frac{1}{n}\|y - Xw\|^2$ and $g(w) = \lambda\|w\|_1$. To simplify notations, let

$$w^+ := p(w) = \arg\min_z \left(g(z) + \frac{L}{2}\left\| z - \left(w - \frac{1}{L}\nabla f(w)\right)\right\|^2 \right),$$

and

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

Then, by Corollary 2, we can see that

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

leading to

$$w_i^+ = \begin{cases} u_i - \dfrac{\lambda}{L}, & u_i > \dfrac{\lambda}{L}, \\ 0, & |u_i| \leq \dfrac{\lambda}{L}, \\ u_i + \dfrac{\lambda}{L}, & u_i < -\dfrac{\lambda}{L}. \end{cases} \tag{15}$$

The mapping defined in Eq. (15) is the so-called *shrinkage operator*—which is a special case of the proximal operator for the nonsmooth ℓ_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 the convexity and the Lipschitz continuity play a central role in analyzing the convergence behaviors.

6.1 Convergence in terms of the Function Values

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

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

$$\nabla f(x_c) + L(x_c^+ - x_c) + s = 0.$$

Lemma 3. *Let $x_c \in \mathbb{R}^n$, $L > 0$, and $x_c^+ = p(x_c)$. Then, for any $x \in \mathbb{R}^n$, we have*

$$F(x) - F(x_c^+) \geq \frac{L}{2}\|x_c^+ - x_c\|^2 + L\left\langle x_c - x,\, x_c^+ - x_c\right\rangle.$$

Remark 4. When we set $x := x_c$, Lemma 3 implies that

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

That is, the sequence of function values $F(x_0), F(x_1), F(x_2), \ldots$ generated by Algorithm 1 monotonically decreases as long as $x_{k+1} \neq x_k$.

Theorem 3. *Let (x_k) be the sequence generated by Algorithm 1. Then, for any $k \geq 1$*

$$F(x_k) - F(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k}, \ \forall x^* \in X^*.$$

Proof. Invoking Lemma 3 with $x = x^*$, $x_c = x_n$, and $x_c^+ = x_{n+1}$, we have

$$\frac{2}{L}(F(x^*) - F(x_{n+1})) \geq \|x_{n+1} - x_n\|^2 + 2\langle x_{n+1} - x_n,\, x_n - x^*\rangle = \|x^* - x_{n+1}\|^2 - \|x^* - x_n\|^2.$$

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

$$\frac{2}{L}\left(kF(x^*) - \sum_{n=0}^{k-1} F(x_{n+1})\right) \geq \|x^* - x_k\|^2 - \|x^* - x_0\|^2.$$

By noting that $F(x_k)$ monotonically decreases, we have

$$\frac{2k}{L}(F(x^*) - F(x_k)) \geq \frac{2}{L}\left(kF(x^*) - \sum_{n=0}^{k-1} F(x_{n+1})\right) \geq \|x^* - x_k\|^2 - \|x^* - x_0\|^2,$$

which leads to

$$F(x_k) - F(x^*) \leq \frac{L}{2k}\left(\|x^* - x_0\|^2 - \|x^* - x_k\|^2\right) \leq \frac{L\|x_0 - x^*\|^2}{2k}.$$

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.