Lecture 3. Gradient Descent

Lecturer: Jie Wang Date: March 12, 2020

The major reference of this lecture is [1, 2].

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

We consider the problem as follows.

$$\begin{aligned} \min_{\mathbf{x}} \ & f(\mathbf{x}) \\ \text{s.t. } & g_i(\mathbf{x}) \le 0,\ i = 1, \ldots, m, \\ & h_i(\mathbf{x}) = 0,\ i = 1, \ldots, p, \\ & \mathbf{x} \in X, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, $h_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, p$, and $X \subseteq \mathbb{R}^n$. To simplify notations, let $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ be a vector function whose i^{th} component is g_i, and $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$ be a vector function whose i^{th} component is h_i. Then, the problem in (1) becomes

$$\begin{aligned} \min_{\mathbf{x}} \ & f(\mathbf{x}) \\ \text{s.t. } & \mathbf{g}(\mathbf{x}) \le 0, \\ & \mathbf{h}(\mathbf{x}) = 0, \\ & \mathbf{x} \in X. \end{aligned} \tag{2}$$

We call the problem in (2) as the primal problem.

Definition 1.

- The *feasible set* is

$$D = \{\mathbf{x} : \mathbf{g}(\mathbf{x}) \le 0, \mathbf{h}(\mathbf{x}) = 0, \mathbf{x} \in X\}. \tag{3}$$

- Each element in D is called a *feasible solution*.
- The optimal function value is defined by

$$f^* = \inf_{\mathbf{x} \in D} f(\mathbf{x}). \tag{4}$$

Assumption 1. Feasibility and Boundedness *The feasible set is nonempty and the objective function is bounded from below, that is,*

$$-\infty < f^* = \inf_{\mathbf{x}\in D} f(\mathbf{x}) < \infty.$$

Definition 2. We say $\mathbf{x}^*$ is an *optimal point*, or solves the problem (2), if $\mathbf{x}^*$ is feasible and $f(\mathbf{x}^*) = f^*$. 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 problem (2) 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 $f(\mathbf{x}) \le f^* + \epsilon$ ($\epsilon > 0$) is called *ϵ-suboptimal*, and the set of all ϵ-suboptimal points is called *ϵ-suboptimal set* for the problem (2).

Definition 3. Consider the problem (2). Suppose that the functions f, g_i, $i = 1, \dots, m$ are convex, h_i, $i = 1, \dots, p$ are affine, and the set X is convex. Then, we say that the problem (2) is a *convex optimization problem*.

Remark 2. A general convex optimization problem takes the form of

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}) \\ &\text{s.t. } \mathbf{x} \in \mathbf{dom}\ f, \\ &\quad \mathbf{x} \in X, \end{aligned}$$

where f is convex and the feasible set

$$D = \mathbf{dom}\ f \cap X$$

is convex. Throughout this lecture, we assume that D is nonempty.

Proposition 1. *Suppose that the problem (2) 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,

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

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

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

Proposition 2. *Suppose that the problem (2) 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,

$$f(\mathbf{x}) \leq f(\mathbf{y}). \tag{5}$$

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 (5) 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 ℓ 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 ℓ as $\theta \in (0, 1)$, and

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

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

$$f(\mathbf{x}) \leq f(\mathbf{z}_0). \tag{6}$$

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 (5) holds. This completes the proof. □

Proposition 3. *Suppose that the problem (2) is a convex optimization problem and solvable. Then, if* f *is strictly convex, the problem (2) has a unique global optimum.*

Proposition 4. *Suppose that the problem (2) is a convex optimization problem. If* f *is strongly convex and continuous over its domain, and the feasible set is closed, then the problem (2) is solvable and has a unique global optimum.*

Definition 5. If $\mathbf{x}$ is feasible, and $g_i(\mathbf{x}) = 0$, we say the i^{th} inequality constraint $g_i(\mathbf{x}) \leq 0$ is *active* at $\mathbf{x}$; otherwise ($g_i(\mathbf{x}) < 0$), we say the constraint $g_i(\mathbf{x}) \leq 0$ is *inactive* at $\mathbf{x}$.

3 Optimality Conditions

Proposition 5. *Suppose that the problem (2) is a convex optimization problem and solvable. If f is continuously differentiable, then $\mathbf{x}$ is optimal if and only if $\mathbf{x} \in D$ and*

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0, \forall \mathbf{y} \in D. \tag{7}$$

Proof.

$\Leftarrow$ Suppose that the inequality (7) 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 problem (2) is a convex optimization problem, and f is continuously differentiable. If $\mathbf{x}^*$ is an interior point of D, then*

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

4 Problem Setup

For the problem (2), letting $m = p = 0$ and $X = \mathbb{R}^n$ leads to the following unconstrained optimization problem:

$$\min_{\mathbf{x}} f(\mathbf{x}). \tag{8}$$

We further assume that

1. the objective function f is convex and continuously differentiable, and thus

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{x}, \mathbf{y}; \tag{9}$$

2. 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;

3. the function f attains its minimum at $\mathbf{x}^*$, i.e.,

$$f(\mathbf{x}^*) = f^* = \min f(\mathbf{x}). \tag{11}$$

   Notice that, the point $\mathbf{x}^*$ that satisfies Eq. (11) may not be unique.

5 The Gradient Descent Algorithm

We propose to solve Problem (1) by gradient descent in Algorithm 1 as follows.

Algorithm 1 Gradient Descent

Input: An initial point $\mathbf{x}_0$, a constant $\alpha \in (0, 2/L)$, and $k = 0$.

1: **while** the *termination condition* does not hold **do**

2: $k \leftarrow k + 1$,

3:

$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k). \tag{12}$$

4: **end while**

How to measure the performance of the above iterative algorithms?

Definition 6. Suppose that the sequence (a_k) converges to a number L.

- The sequence is said to converge linearly to L, if there exists a number $\mu \in (0, 1)$ such that

$$\lim_{k\to\infty} \frac{|a_{k+1} - L|}{|a_k - L|} = \mu,$$

  where the number μ is called the rate of convergence.

- The sequence is said to converge superlinearly to L, if

$$\lim_{k\to\infty} \frac{|a_{k+1} - L|}{|a_k - L|} = 0.$$

- The sequence is said to converge sublinearly to L, if

$$\lim_{k\to\infty} \frac{|a_{k+1} - L|}{|a_k - L|} = 1.$$

- The sequence is said to converge to L with order q, if there exists a number M such that

$$\lim_{k\to\infty} \frac{|a_{k+1} - L|}{|a_k - L|^q} < M.$$

  If $q = 2$ ($q = 3$), the sequence is quadratic (cubic) convergence.

6 Convergence property

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.

Before we derive the monotonicity of the function values, we first show a useful lemma that is due to the Lipschitz continuity.

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*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{13}$$

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. □

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. *Suppose a function f is continuously differentiable and its gradient is Lipschitz continuous with constant $L > 0$. Then, for the sequence generated by Algorithm 1, we have*

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \alpha \left(1 - \frac{L}{2}\alpha\right) \|\nabla f(\mathbf{x}_k)\|^2, \, \forall \, k = 0, 1, 2 \ldots, \tag{14}$$

i.e., the sequence of function values $\{f(\mathbf{x}_k)\}_k$ is monotonically decreasing.

Proof. In view of Lemma 1 and the update rule in (12), we have

$$\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= f(\mathbf{x}_k) - \alpha \|\nabla f(\mathbf{x}_k)\|^2 + \frac{L}{2}\alpha^2 \|\nabla f(\mathbf{x}_k)\|^2 \\
&= f(\mathbf{x}_k) - \alpha \left(1 - \frac{L}{2}\alpha\right) \|\nabla f(\mathbf{x}_k)\|^2,
\end{aligned} \tag{15}$$

which completes the proof. □

The next result tells us that the sequence of gradients tends to be zero.

Lemma 3. *Suppose that the conditions in Lemma 2 hold. Then,*

$$\nabla f(\mathbf{x}_k) \to 0. \tag{16}$$

Proof. The inequality (14) leads to

$$\begin{aligned}
&\|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})}{\alpha\left(1 - \frac{L}{2}\alpha\right)} \\
\Rightarrow &\sum_{i=0}^{k} \|\nabla f(\mathbf{x}_i)\|^2 \leq \frac{f(\mathbf{x}_0) - f(\mathbf{x}_{k+1})}{\alpha\left(1 - \frac{L}{2}\alpha\right)} \\
\Rightarrow &\sum_{i=0}^{\infty} \|\nabla f(\mathbf{x}_i)\|^2 \leq \frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{\alpha\left(1 - \frac{L}{2}\alpha\right)}.
\end{aligned}$$

The claim follows immediately. □

Remark 3. In both Lemmas 2 and 3, we do not assume convexity of f. Moreover, Lemma 3 implies that, if the sequence $\{\mathbf{x}_k\}$ converges to a point $\bar{x}$, we would have $\nabla f(\bar{x}) = 0$.

We are now ready to state the following theorem.

Theorem 1. *Consider the problem in Section 4 and the sequence generated by Algorithm 1. Then, the sequence of function values $f(\mathbf{x}_k)$ tends to the optimal function value $f(\mathbf{x}^*)$ in a rate of $O(1/k)$. Specifically,*

1. *if $\alpha \in (0, 1/L]$, we have*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{k}\left(\frac{1}{2\alpha}\|\mathbf{x}_0 - \mathbf{x}^*\|^2\right); \tag{17}$$

2. *if $\alpha \in (1/L, 2/L)$, we have*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{k}\left(\frac{1}{2\alpha}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{L\alpha - 1}{2 - L\alpha}(f(\mathbf{x}_0) - f(\mathbf{x}^*))\right). \tag{18}$$

Proof. Combining the inequality (13) and (12) leads to

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \tag{19}$$

$$\Leftrightarrow f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq f(\mathbf{x}_k) - f(\mathbf{x}^*) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2. \tag{20}$$

Due to the convexity of f, we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle. \tag{21}$$

Combining (20) and (21) leads to

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) &\leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= -\frac{1}{\alpha}\langle \mathbf{x}_{k+1} - \mathbf{x}_k, \mathbf{x}_k - \mathbf{x}^* \rangle - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= -\frac{1}{2\alpha}\left(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 - \|\mathbf{x}_k - \mathbf{x}^*\|^2\right) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= \frac{1}{2\alpha}\left(\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right) - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.
\end{aligned}$$

By summing up the above inequality for $i = 0, 1, \ldots, k-1$, we have

$$k(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \leq \sum_{i=0}^{k-1} f(x_{i+1}) - f(\mathbf{x}^*) \tag{22}$$
$$\leq \frac{1}{2\alpha}(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_k - \mathbf{x}^*\|^2) - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\sum_{i=0}^{k-1}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2.$$

By noting (12) and a similar argument in Lemma 3, we have

$$\sum_{i=0}^{\infty}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 \leq \frac{2\alpha}{2 - L\alpha}(f(\mathbf{x}_0) - f(\mathbf{x}^*)). \tag{23}$$

Thus, for $\alpha \in (0, 1/L]$, the inequality in (22) implies that

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{2\alpha k}\|\mathbf{x}_0 - \mathbf{x}^*\|^2. \tag{24}$$

If $\alpha \in (1/L, 2/L)$, by (23), we have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{k}\left(\frac{1}{2\alpha}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{L\alpha - 1}{2 - L\alpha}(f(\mathbf{x}_0) - f(\mathbf{x}^*))\right). \tag{25}$$

This completes the proof that $f(\mathbf{x}_k) \to f(\mathbf{x}^*)$ with a convergence rate $O(1/k)$. □

Remark 4. Why do not show that $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \to 0$? Indeed, this can be wrong, as there might exist multiple optimal solutions, and we do not know in advance which optimal solution the sequence $\mathbf{x}_k$ will converge to.

7 GD for Strongly Convex Optimization Problems

In this section, we analyze the convergence property of Algorithm 1 for problem (1) when the objective function is strongly convex. We first introduce the concept of strongly convex.

Definition 7. A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called strongly convex if there exists a constant $\mu > 0$ such that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2, \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{26}$$

The constant μ is called the convexity parameter of function f.

To avoid any confusion, we explicitly state the problem we shall analyze as follows.

$$\min_{\mathbf{x}} f(\mathbf{x}), \tag{27}$$

where f is strongly convex with convexity parameter $\mu > 0$ and its gradient is Lipschitz continuous with constant $L > 0$.

Remark 5. Problem (27) always admits a unique solution. Why?

We first show a useful result as follows.

Lemma 4. *Suppose f is strongly convex with convexity parameter $\mu > 0$ and continuously differentiable. Then,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2, \forall \mathbf{x}, \mathbf{y}. \tag{28}$$

Proof. Suppose that $\mathbf{x}$ is fixed. Then, the right hand side of (26) is a function of $\mathbf{y}$, which is denoted by

$$q(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{29}$$

Then,

$$f(\mathbf{y}) \geq q(\mathbf{y}) \geq \min_{\mathbf{z}} q(\mathbf{z}) = f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2, \tag{30}$$

which completes the proof. □

We are now ready to analyze the convergence property of Algorithm 1 on problem (27).

Theorem 2. *Consider the problem (27) and the sequence generated by Algorithm 1. Then,*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq (1 - \mu\alpha(2 - L\alpha))^k (f(\mathbf{x}_0) - f(\mathbf{x}^*)). \tag{31}$$

Proof. In view of the inequality (20), we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq & f(\mathbf{x}_k) - f(\mathbf{x}^*) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\alpha^2\|\nabla f(\mathbf{x}_k)\|^2 \\ \leq & f(\mathbf{x}_k) - f(\mathbf{x}^*) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\alpha^2 2\mu(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \quad \text{by Lemma 4} \\ = & (1 - \mu\alpha(2 - L\alpha))^k (f(\mathbf{x}_0) - f(\mathbf{x}^*)), \end{aligned}$$

which completes the proof. □

Remark 6. When $\alpha \in (0, 2/L)$, the coefficient $1 - \mu\alpha(2 - L\alpha)$ in (31) is in $[1 - \mu/L, 1)$, which means that the function values converge linearly to its optimal value (what if $\mu = L$?).

References

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

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