New insights in smoothness and strong convexity with improved convergence of gradient descent

Lu Zhang∗ Jiani Wang† Hui Zhang‡

November 1, 2021

Abstract

The starting assumptions to study the convergence and complexity of gradient-type methods may be the smoothness (also called Lipschitz continuity of gradient) and the strong convexity. In this note, we revisit these two basic properties from a new perspective that motivates their definitions and equivalent characterizations, along with an improved linear convergence of the gradient descent method.

Keywords. smoothness, strong convexity, gradient descent, Lipschitz continuity, linear convergence

AMS subject classifications. 90C25, 65K05.

1 Introduction

In the field of optimization, certain classes of functions have to be figured out so that useful optimization theory (including optimality, complexity, and convergence, etc.) can be established. The class of functions with Lipschitz continuous gradient is one of the most important classes, for which some convergence theory of the gradient descent method can be built. The gradient Lipschitz continuity for an objective function $f$, usually called smoothness for brevity, is tantamount to requiring that the gradient of $f$ is continuous in the following sense:

\begin{equation}
\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|,
\end{equation}

where $L > 0$ is called the Lipschitz constant. However, this class of functions is too wide to guarantee convergence even to a local minimum, let alone get global performance. Thus, the strong convexity was imposed as a restriction so that better convergence guarantees are possible. It is defined by the following property:

\begin{equation}
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \mu \| x - y \|^2,
\end{equation}

where $\mu > 0$ is called the strong-convexity constant.

∗Department of Mathematics, National University of Defense Technology, Changsha, Hunan 410073, China
†School of Mathematical Sciences, Chinese Academy of Sciences, Beijing 100049, China
‡Corresponding author. Department of Mathematics, National University of Defense Technology, Changsha, Hunan 410073, China. Email: h.zhang1984@163.com
Up to now, it has become a well-known fact that both smoothness and strong convexity are very fundamental properties for analyzing minimization processes, especially for gradient-type methods. Many existing classic textbooks stated them as basic definitions, along with a couple of equivalent characterizations; see e.g. [3, 4, 7]. In the literature of optimization, they were widely cited as the most natural assumptions. However, it seems unclear where they come from, why they have their own forms (1.1) and (1.2), and how the existing equivalent conditions be constructed. In this short note, we try to answer these questions based on a simple observation in gradient descent, that is, each subproblem in the gradient descent method should be a “good” approximation to the objective function. From this perspective, we define the smoothness and strong convexity as a quantified description of the “good” approximation. In order to construct a series of equivalent characterizations for these two properties, we first reformulate our definitions into convexity of certain functions and then apply a group of equivalent convexity conditions to them.

The remainder of the paper is organized as follows. In Section 2 we introduce the simple observation in gradient descent, and then define the smoothness and strong-convexity. In Section 3, we present a group of characterizations for the smoothness and strong-convexity. In Section 4, we obtain an improved linear convergence by using one of the equivalent characterizations of the convex $L$-smoothness.

**Notation.** We restrict our attention to an arbitrary finite dimensional space $\mathbb{R}^d$ associated with dot product $\langle x, y \rangle := \sum_{i=1}^d x_i y_i$ and its induced norm $\| \cdot \| := \sqrt{\langle \cdot, \cdot \rangle}$. For a fixed function $f$, the Bregman distance is defined as

$$D_f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle,$$

the conjugate function $f^*$ is given by

$$f^*(u) = \sup_{x \in \mathbb{R}^d} \{ \langle x, u \rangle - f(x) \}.$$

### 2 Motivation of definition

We begin our study with the following simple observation in gradient descent, which will motivate us to define more general notions than the smoothness and strong-convexity.

#### 2.1 A simple observation from gradient descent

The gradient descent is a well-known method for solving $\min_{x \in \mathbb{R}^d} f(x)$; for $k \geq 0$ it reads as

$$x_{k+1} = x_k - \frac{1}{t} \nabla f(x_k), \quad \text{(2.1)}$$

where $t > 0$ is the step size. From the perspective of approximation, the gradient descent scheme (2.1) can be obtained by solving a quadratic approximation subproblem as follows:

$$x_{k+1} := \arg \min_{x \in \mathbb{R}^d} \{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{t}{2} \| x - x_k \|^2 \}. \quad \text{(2.2)}$$

In order to force $x_{k+1}$ to be a good approximation to the minimizer of $f$, we expect that the objective function in the subproblem (2.2) is a “good” approximation to the original function $f$, that is

$$f(x) \approx f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{t}{2} \| x - x_k \|^2.$$
Put the terms about $f$ together to yield the following equivalent approximation formulation

$$\frac{t}{2}\|x - x_k\|^2 \approx f(x) - f(x_k) - \langle \nabla f(x_k), x - x_k \rangle.$$  \hspace{1cm} (2.3)

To write the formulation above in a symmetrical way, we let $\tilde{\varphi}(x) = \frac{t}{2}\|x\|^2$. Then, it becomes

$$\tilde{\varphi}(x) - \tilde{\varphi}(x_k) - \langle \nabla \tilde{\varphi}(x_k), x - x_k \rangle \approx f(x) - f(x_k) - \langle \nabla f(x_k), x - x_k \rangle.$$  \hspace{1cm} (2.4)

Since there may exist some more suitable function $\varphi$ than $\tilde{\varphi}$ such that the approximation (2.4) becomes better, we consider the following more general approximation:

$$\varphi(x) - \varphi(x_k) - \langle \nabla \varphi(x_k), x - x_k \rangle \approx f(x) - f(x_k) - \langle \nabla f(x_k), x - x_k \rangle.$$  \hspace{1cm} (2.5)

In terms of the Bregman distance, (2.5) can be simplified into

$$D_{\varphi}(x, x_k) \approx D_f(x, x_k).$$  \hspace{1cm} (2.6)

Due to the arbitrariness of the initial point $x_0$, we actually require that

$$D_{\varphi}(x, y) \approx D_f(x, y), \ \forall \ x, y \in \mathbb{R}^d.$$  \hspace{1cm} (2.7)

There are at least two ways to quantify the approximation (2.7). The first way is to bound the difference between $D_{\varphi}(x, y)$ and $D_f(x, y)$, that is there exists a constant $C < +\infty$ such that

$$\sup_{x, y \in \mathbb{R}^d} |D_{\varphi}(x, y) - D_f(x, y)| = C.$$

A drawback is that the constant $C$ is not scale invariant since for any $\gamma > 0$, we have

$$\sup_{x, y \in \mathbb{R}^d} |D_{\gamma \varphi}(x, y) - D_{\gamma f}(x, y)| = \gamma C.$$

To overcome it, we consider the second way that bound the ratio between them, that is there exists two constants $\mu$ and $L$ such that for any $x, y \in \mathbb{R}^d$ with $x \neq y$ we have

$$\mu \leq \frac{D_f(x, y)}{D_{\varphi}(x, y)} \leq L.$$  \hspace{1cm} (2.8)

Further, we assume the strictly convexity of $\varphi$ so that the above can be equivalently written as

$$\mu D_{\varphi}(x, y) \leq D_f(x, y) \leq LD_{\varphi}(x, y), \forall x, y \in \mathbb{R}^d,$$  \hspace{1cm} (2.9)

which is just the relative smoothness (also called Lipschitz-like/convexity condition in $[11]$) and the relatively strong convexity, proposed recently in $[6]$. Since the purpose of this note is to study the smoothness and the strong-convexity, we restrict our attention to (2.9) with $\varphi = \varphi_0 := \frac{1}{2}\|\cdot\|^2$. The pursuit of more general discussion may be leaved as future work.
2.2 The smoothness and strong-convexity

Now, we are ready to define the smoothness and the strong-convexity in terms of the Bregman distance as follows.

**Definition 2.1.** Let \( f : \mathbb{R}^d \to \mathbb{R} \) be a differentiable function and \( \varphi_0 = \frac{1}{2} \| \cdot \|^2 \). Let \( L \) and \( \mu \) be two given nonnegative constant. We say that \( f \) is \( L \)-smooth if

\[
D_f(x, y) \leq LD_{\varphi_0}(x, y), \forall x, y \in \mathbb{R}^d, \tag{2.10}
\]

and \( \mu \)-strongly-convex if

\[
\mu D_{\varphi_0}(x, y) \leq D_f(x, y), \forall x, y \in \mathbb{R}^d. \tag{2.11}
\]

After some simple calculations, we can separately write the \( L \)-smoothness and the \( \mu \)-strong-convexity in the following equivalent forms

\[
f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \| y - x \|^2, \tag{2.12a}
\]

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \| y - x \|^2. \tag{2.12b}
\]

They are the usual definitions appeared in textbooks. The advantage of Definition 2.1 is that we can easily turn (2.12a)-(2.12b) into convexity of certain functions. Actually, applying the linearity of the Bregman distance to (2.10), we obtain

\[
D_{L\varphi_0 - f}(y, x) \geq 0, \text{ which equivalently means the convexity of } L\varphi_0 - f. \]

Similarly, the \( \mu \)-strong-convexity is equivalent to the convexity of \( f - \mu\varphi_0 \).

We summarize these results in the following lemma.

**Lemma 2.1.** Let \( f : \mathbb{R}^d \to \mathbb{R} \) be a differentiable function and \( \varphi_0 = \frac{1}{2} \| \cdot \|^2 \). Then,

1. \( f \) is \( L \)-smooth if and only if \( L\varphi_0 - f \) is convex;

2. \( f \) is \( \mu \)-strongly-convex if and only if \( f - \mu\varphi_0 \) is convex.

3 Characterization of equivalence

In this section, we will present three groups of equivalent characterizations for the smoothness and strong convexity. The first group is based on Lemma 2.1 and different definitions of convexity. The second group is obtained from a point of dual view. The last group is obtained by using some Fenchel dual properties. To this end, we first introduce two basic results.

**Lemma 3.1** (Proposition 14.2 in [2]). Let \( f : \mathbb{R}^d \to \mathbb{R} \) be a differentiable convex function, \( \varphi_0 = \frac{1}{2} \| \cdot \|^2 \), and \( \gamma > 0 \). Then, \( \gamma\varphi_0 - f \) is convex if and only if \( f^* - \gamma^{-1}\varphi_0 \) is convex.

**Lemma 3.2** (Theorem 23.5 and Corollary 23.5.1 in [8]). Let \( f : \mathbb{R}^d \to \mathbb{R} \) be a convex function. Then, the conditions \( f(u) + f^*(u^*) = \langle u, u^* \rangle \), \( u^* \in \partial f(u) \), and \( u \in \partial f^*(u^*) \) are equivalent. Moreover, \( \partial f^* \) is the inverse of \( \partial f \).

We also need the equivalent definitions of convexity.
Let $$f: \mathbb{R}^d \to \mathbb{R}$$ be a differentiable function. Then, it is convex if and only if one of the following conditions holds:
\[
\begin{align*}
& f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y), \forall x, y \in \mathbb{R}^d, \lambda \in [0, 1], \quad (3.1a) \\
& \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0, \forall x, y \in \mathbb{R}^d, \quad (3.1b) \\
& f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \forall x, y \in \mathbb{R}^d. \quad (3.1c)
\end{align*}
\]

### 3.1 Characterizations of smoothness

Combining the first statement in Lemma 2.1 with the different convexity conditions in Lemma 3.3, we have the following equivalent characterization of smoothness.

**Proposition 3.1.** Let $$f: \mathbb{R}^d \to \mathbb{R}$$ be a differentiable function. Then, it is $$L$$-smooth if and only if one of the following conditions holds:
\[
\begin{align*}
& f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) - \frac{L}{2} \lambda(1 - \lambda)\|x - y\|^2, \forall x, y \in \mathbb{R}^d, \lambda \in [0, 1], \quad (3.2a) \\
& \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|^2, \forall x, y \in \mathbb{R}^d, \quad (3.2b) \\
& f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2, \forall x, y \in \mathbb{R}^d. \quad (3.2c)
\end{align*}
\]

In order to obtain a group of dual characterizations, we need to assume that the differentiable function $$f$$ is also convex so that Lemma 3.1 can be invoked. In other words, the convexity of $$L\varphi_0 - f$$ is equivalent to the convexity of $$f^* - L^{-1}\varphi$$. Before stating the dual characterizations, we denote by $$\Gamma(\mathbb{R}^d)$$ the class of functions with the following properties:

- $$f$$ is convex and differentiable;
- its conjugate $$f^*$$ is differentiable.

Now, together with Lemma 3.3, we immediately have the following result, which seems new to the best of our knowledge. It is obtained at the cost of restricting the objectives into the class of functions $$\Gamma(\mathbb{R}^d)$$.

**Proposition 3.2.** Let $$f \in \Gamma(\mathbb{R}^d)$$. Then, it is $$L$$-smooth if and only if one of the following conditions holds:
\[
\begin{align*}
& f^*(\lambda u + (1 - \lambda)v) \leq \lambda f^*(u) + (1 - \lambda)f^*(v) - \frac{1}{2L} \lambda(1 - \lambda)\|u - v\|^2, \forall u, v \in \mathbb{R}^d, \lambda \in [0, 1], \quad (3.3a) \\
& \langle \nabla f^*(u) - \nabla f^*(v), u - v \rangle \geq \frac{1}{L}\|u - v\|^2, \forall u, v \in \mathbb{R}^d, \quad (3.3b) \\
& f^*(u) \geq f^*(v) + \langle \nabla f^*(v), u - v \rangle + \frac{1}{2L}\|u - v\|^2, \forall u, v \in \mathbb{R}^d. \quad (3.3c)
\end{align*}
\]

**Corollary 3.1.** Let $$f$$ be convex and differentiable. Then, it is $$L$$-smooth if and only if one of the following conditions holds:
\[
\begin{align*}
& \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2, \forall x, y \in \mathbb{R}^d, \quad (3.4a) \\
& f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2, \forall x, y \in \mathbb{R}^d, \quad (3.4b) \\
& \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^d. \quad (3.4c)
\end{align*}
\]
Proof. First of all, we show the equivalence under the assumption of \( f \in \Gamma(\mathbb{R}^d) \) so that Proposition 3.2 can be applied. Actually, the first two conditions follow from (3.3b) and (3.3c) separately by letting \( \nabla f^*(u) = x, \nabla f^*(v) = y \) and using the fact \((\nabla f)^{-1} = \nabla f^*\) and the following relationships in Lemma 3.2:

\[
\begin{align*}
f^*(u) + f(x) &= \langle u, x \rangle = \langle \nabla f(x), x \rangle, \\
f^*(v) + f(y) &= \langle v, y \rangle = \langle \nabla f(y), y \rangle.
\end{align*}
\] (3.5a)

(3.5b)

\( L \)-smoothness implies \((3.4c) \) follows by applying the Cauchy-Schwartz inequality to (3.4a); while \((3.4c) \) \( \Rightarrow \) \( L \)-smoothness follows by applying the Cauchy-Schwartz inequality to \( \langle \nabla f(x) - \nabla f(y), x - y \rangle \) to obtain (3.2b).

Now, we drop the restriction of \( f^* \) being differentiable. To this end, we let \( f_\epsilon := f + \epsilon \varphi_0 \) with \( \epsilon > 0 \), i.e., \( f_\epsilon \) must be strongly convex and hence its conjugate is differentiable [3]. Therefore, \( f_\epsilon \in \Gamma(\mathbb{R}^d) \). Now, we complete the proof by showing the following chain of implication:

\[
(2.12a) \Rightarrow (3.4a) \Rightarrow (3.4a) \Rightarrow (3.4c) \Rightarrow (2.12a).
\]

In fact, (2.12a) says that \( f \) is \( L \)-smooth and hence \( L \varphi_0 - f \) is convex. This further implies that \( (L + \epsilon)\varphi_0 - f \) is convex, i.e., \( f_\epsilon \) is \( (L + \epsilon) \)-smooth. Thus, applying (3.4b) to \( f_\epsilon \), we have

\[
f_\epsilon(y) \geq f_\epsilon(x) + \langle \nabla f_\epsilon(x), y - x \rangle + \frac{1}{2(L + \epsilon)} \| \nabla f_\epsilon(x) - \nabla f_\epsilon(y) \|^2, \forall x, y \in \mathbb{R}^d.
\]

Letting \( \epsilon \) in the above inequality tend to zero, we immediately obtain (3.4b). From (3.4b), we have

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \| \nabla f(x) - \nabla f(y) \|^2,
\]

\[
f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \| \nabla f(x) - \nabla f(y) \|^2.
\]

Adding these two inequalities, we get (3.4a), which implies (3.4c) by invoking the Cauchy-Schwartz inequality. It remains to show (3.4a) \( \Rightarrow (2.12a) \). By the fundamental theorem of calculus, we have

\[
f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt.
\]

Therefore, \( f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt \)

\[
\leq \int_0^1 \| \nabla f(x + t(y - x)) - \nabla f(x) \| \| y - x \| dt \leq \int_0^1 tL \| y - x \|^2 dt = \frac{L}{2} \| y - x \|^2,
\]

where the first inequality follows by the Cauchy-Schwartz inequality and the second one from (3.3c). This completes the proof.

Combining Proposition 3.1 and Corollary 3.1, we conclude that for any convex and differentiable function \( f \), its \( L \)-smoothness can be equivalently characterized by one of the conditions (3.2a)-(3.2c) and (3.4a)-(3.4c). Especial attention should be paid to the conditions (3.4a) and (3.4b), since each of them implies the convexity of \( f \). Thus, we have the following additional result.

**Corollary 3.2.** Let \( f \) be differentiable. Then, it is convex and \( L \)-smooth if and only if one of the conditions (3.4a) and (3.4b) holds.

It should be noted although the equivalence between the \( L \)-smoothness and the conditions (3.4a) and (3.4b) is well-known for convex and differentiable, the current presentation in Corollary 3.2 seems more accurate.
3.2 Characterizations of strong-convexity

The results in this part will be obtained in the similar way to that in the previous subsection. Combining the second statement in Lemma 2.1 with the different convexity conditions in Lemma 3.3, we have the following equivalent characterization of strong-convexity.

Proposition 3.3. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then, it is $\mu$-strongly-convex if and only if one of the following conditions holds:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2} \lambda(1 - \lambda)\|x - y\|^2, \forall x, y \in \mathbb{R}^d, \lambda \in [0, 1],$$

(3.6a)

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2, \forall x, y \in \mathbb{R}^d,$$

(3.6b)

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

(3.6c)

The corresponding dual characterizations are followings:

Proposition 3.4. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then, it is $\mu$-strongly-convex if and only if one of the following conditions holds:

$$f^*(\lambda u + (1 - \lambda)v) \geq \lambda f^*(u) + (1 - \lambda)f^*(v) - \frac{1}{2\mu} \lambda(1 - \lambda)\|u - v\|^2, \forall u, v \in \mathbb{R}^d, \lambda \in [0, 1],$$

(3.7a)

$$\langle \nabla f^*(u) - \nabla f^*(v), u - v \rangle \leq \frac{1}{\mu} \|u - v\|^2, \forall u, v \in \mathbb{R}^d,$$

(3.7b)

$$f^*(u) \leq f^*(v) + \langle \nabla f^*(v), u - v \rangle + \frac{1}{2\mu} \|u - v\|^2, \forall u, v \in \mathbb{R}^d.$$  

(3.7c)

The above clarifies the equivalent conditions between $\mu$-strong-convexity of $f$ and $\mu^{-1}$-smoothness of $f^*$. A classic example is $f(x) = \frac{1}{2}x^TQx$, where $Q$ is a non-singular symmetric positive semi-definite $d \times d$ matrix; the conjugate $f^*(x) = \frac{1}{2}x^TQ^{-1}x$. Obviously, $f$ is $\lambda_{\min}(Q)$-strongly-convex while $f^*$ is $\lambda_{\min}(Q)^{-1}$-smooth, where $\lambda_{\min}(Q)$ represents the smallest eigenvalue of $Q$.

Corollary 3.3. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then, it is $\mu$-strongly-convex if and only if one of the following conditions holds:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \frac{1}{\mu} \|\nabla f(x) - \nabla f(y)\|^2, \forall x, y \in \mathbb{R}^d,$$

(3.8a)

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\mu} \|\nabla f(x) - \nabla f(y)\|^2, \forall x, y \in \mathbb{R}^d,$$

(3.8b)

$$\|\nabla f(x) - \nabla f(y)\| \geq \mu \|x - y\|, \forall x, y \in \mathbb{R}^d.$$  

(3.8c)

The proof is similar to that of Corollary 3.1, we omit the details.

3.3 Characterizations of smoothness and strong-convexity

Now, we want to characterize the class of functions that satisfy (2.12a) and (2.12b), i.e., both $f - \mu \varphi_0$ and $L \varphi_0 - f$ are convex. Let $\tilde{f} := f - \mu \varphi_0$. Then, it is equivalent to the condition that $\tilde{f}$ is convex and $(L - \mu)$-smooth. Recalling the equivalent conditions in Corollary 3.2 we have the following characterizations for smoothness and strong-convexity.
Proposition 3.5. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then, it is $L$-smooth and $\mu$-strongly-convex if and only if one of the following conditions holds:

\begin{align}
\langle \nabla f(x) - \nabla f(y), x - y \rangle &\leq \frac{L\mu}{L + \mu} \|x - y\|^2 + \frac{1}{L + \mu} \|\nabla f(x) - \nabla f(y)\|^2, \forall x, y \in \mathbb{R}^d, \tag{3.9a}\\
f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 + \\
\frac{\mu L}{2(L - \mu)} \|x - y\| - \frac{1}{L} &\|\nabla f(x) - \nabla f(y)\|, \forall u, v \in \mathbb{R}^d. \tag{3.9b}
\end{align}

The equivalent condition (3.9b) was essentially discovered in [10]. As a necessary condition to the $L$-smoothness and $\mu$-strong-convexity, (3.9a) appeared in the textbook [7]. Here, we highlight that it is also sufficient for a differentiable $f$ to be $L$-smooth and $\mu$-strongly-convex.

4 Improved linear convergence for gradient descent

First of all, we recall a well-known linear convergence result for the gradient descent method, stated in [5].

Lemma 4.1. Consider the unconstrained optimization problem $\min_{x \in \mathbb{R}^d} f(x)$, where $f$ is $L$-smooth, has a nonempty solution set $\mathcal{X}^*$, and satisfies the Polyak-Lojasiewicz (PL) inequality

\[ \frac{1}{2} \|\nabla f(x)\|^2 \geq \nu(f(x) - \bar{f}), \forall x \in \mathbb{R}^d, \tag{4.1} \]

where $\bar{f}$ denotes the optimal function value. Then the gradient descent (2.1) with a step-size of $\frac{1}{L}$ has a global linear convergence rate

\[ f(x_{k+1}) - \bar{f} \leq (1 - \frac{\nu}{L}) (f(x_k) - \bar{f}). \tag{4.2} \]

The proof consists of two steps. First, using the $L$-smoothness (2.12a) with $y = x_{k+1}$ and $x = x_k$, we obtain

\[ f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2. \tag{4.3} \]

Second, using the PL inequality with $x = x_k$, we get

\[ \frac{1}{2} \|\nabla f(x_k)\|^2 \geq \nu(f(x_k) - \bar{f}). \tag{4.4} \]

The linear convergence result (4.2) follows directly by combining (4.3) and (4.4). In what follows, we will show that the rate of linear convergence (4.2) can be improved if the objective $f$ is not only $L$-smooth but also convex. From Corollary 3.2, we know that (3.4b) characterizes the $L$-smoothness and the convexity of $f$ at the same time. Now, using (3.4b) with $x = x_{k+1}$ and $y = x_k$, we obtain

\[ f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 - \frac{1}{2L} \|\nabla f(x_{k+1})\|^2. \tag{4.5} \]

Note that the PL inequality $x = x_{k+1}$ implies that

\[ \frac{1}{2} \|\nabla f(x_{k+1})\|^2 \geq \nu(f(x_{k+1}) - \bar{f}). \tag{4.6} \]

Now, combining (4.5), (4.4) and (4.6), we have the following improved linear convergence.
Proposition 4.1. Consider the unconstrained optimization problem \( \min_{x \in \mathbb{R}^d} f(x) \), where \( f \) is convex and \( L \)-smooth, has a nonempty solution set \( X^* \) with \( \bar{f} \) being the optimal function value, and satisfies the PL inequality \((4.1)\). Then the gradient descent \((2.1)\) with a step-size of \( \frac{1}{L} \) has a global linear convergence rate
\[
f(x_{k+1}) - \bar{f} \leq \frac{L - \nu}{L + \nu} (f(x_k) - \bar{f}).
\] \((4.7)\)

Acknowledgements

The third author was supported by the National Science Foundation of China (No.11971480), the Natural Science Fund of Hunan for Excellent Youth (No.2020JJ3038), and the Fund for NUDT Young Innovator Awards (No. 20190105).

References

[1] H. H. Bauschke, J. Bolte, and M. Teboulle, A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications, Math. Oper. Res., 42 (2016), pp. 330–348.

[2] H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer International Publishing, 2011.

[3] A. Beck, First-Order Methods in Optimization, SIAM-Society for Industrial and Applied Mathematics, 2017.

[4] S. Bubeck, Convex optimization: Algorithms and complexity, Foundations and Trends in Machine Learning, 8 (2014), pp. 231–357.

[5] H. Karimi, J. Nutini, and M. Schmidt, Linear convergence of proximal-gradient methods under the Polyak-Lojasiewicz condition., arXiv:1608.04636v1 [cs.LG] 16 Aug 2016., (2016).

[6] H. Lu, R. M. Freund, and Y. Nesterov, Relatively smooth convex optimization by first-order methods, and applications, SIAM J. Optim., 28 (2018), pp. 333–354.

[7] Y. Nesterov, Introductory lectures on convex optimization: A basic course, Kluwer Academic Publishers, 2004.

[8] R. T. Rockafellar, Convex Analysis, Princeton University Press, 2015.

[9] R. T. Rockafellar and R. J. B. Wets, Variational analysis, Springer, 1998.

[10] A. B. Taylor, J. M. Hendrickx, and F. Glineur, Smooth strongly convex interpolation and exact worst-case performance of first-order methods, Math. Program., Ser. A, 161 (2017), pp. 307–345.