Almost Sure Convergence Rates of Stochastic Zeroth-order Gradient Descent for Łojasiewicz Functions

Tianyu Wang*

Abstract

We prove *almost sure convergence rates* of Zeroth-order Gradient Descent (SZGD) algorithms for Łojasiewicz functions. The SZGD algorithm iterates as

$$x_{t+1} = x_t - \eta_t \widehat{\nabla} f(x_t), \qquad t = 0, 1, 2, 3, \cdots,$$

where f is the objective function that satisfies the Łojasiewicz inequality with Łojasiewicz exponent θ, η_t is the step size (learning rate), and $\widehat{\nabla} f(x_t)$ is the approximate gradient estimated using zeroth-order information. We show that, for smooth Łojasiewicz functions, the sequence $\{x_t\}_{t\in\mathbb{N}}$ generated by SZGD converges to a bounded point x_∞ almost surely, and x_∞ is a critical point of f. If $\theta \in (0, \frac{1}{2}]$, $f(x_t) - f(x_\infty)$, $\sum_{s=t}^{\infty} \|x_{s+1} - x_s\|^2$ and $\|x_t - x_\infty\|$ ($\|\cdot\|$ is the Euclidean norm) converge to zero *linearly almost surely*. If $\theta \in (\frac{1}{2}, 1)$, then $f(x_t) - f(x_\infty)$ (and $\sum_{s=t}^{\infty} \|x_{s+1} - x_s\|^2$) converges to zero at rate $o\left(t^{\frac{1}{1-2\theta}} \log t\right)$ almost surely; $\|x_t - x_\infty\|$ converges to zero at rate $o\left(t^{\frac{1-\theta}{1-2\theta}} \log t\right)$ almost surely. To the best of our knowledge, this paper provides the first *almost sure convergence rate* guarantee for stochastic zeroth order algorithms for Łojasiewicz functions.

1 Introduction

Zeroth order optimization is a central topic in mathematical programming and related fields. Algorithms for zeroth order optimization find important real-world applications, since often times in practice, we cannot directly access the derivatives of the objective function. To optimize the function in such scenarios, one can estimate the gradient/Hessian first and deploy first/second order algorithms with the estimated derivatives. Previously, many authors have considered this problem. Yet stochastic zeroth order methods for Łojasiewicz functions have not been carefully investigated (See Section 2 for more discussion). In this paper, we study the performance of gradient descent with estimated gradient for (smooth) Łojasiewicz functions.

In particular, we study algorithms governed by the following rule

$$x_{t+1} = x_t - \eta_t \widehat{\nabla} f_k^{\delta_t}(x_t), \qquad t = 0, 1, 2, \cdots, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the unknown objective function, $\eta_t > 0$ is the step size (learning rate), and $\widehat{\nabla} f_k^{\delta_t}(x_t)$ is the estimator of ∇f at x_t defined as follows.

$$\widehat{\nabla} f_k^{\delta_t}(x) := \frac{n}{2\delta_t k} \sum_{i=1}^{k} \left(f(x + \delta_t v_i) - f(x - \delta_t v_i)\right) v_i, \quad \forall x \in \mathbb{R}^n, \tag{2}$$

where $[v_1, v_2, \cdots, v_k]$ is uniformly sampled from the Stiefel's manifold $\mathrm{St}(n, k) := \{V \in \mathbb{R}^{n\times k} : V^\top V = I_k\}$, and $\delta_t := 2^{-t}$ is the finite difference granularity. Previously, the statistical properties of (2) have been investigated by [11] (See Section 3) and [16] have consider stochastic zeroth order optimization algorithms for Polyak-Łojasiewicz functions (See Section 2 for detailed discussions). Throughout, we use Stochastic Zeroth-order Gradient Descent (SZGD) to refer to the above update rule (1) with the estimator (2).

*wangtianyu@fudan.edu.cn

Łojasiewicz functions are real-valued functions that satisfy the Łojasiewicz inequality [21]. The Łojasiewicz inequality generalizes the Polyak–Łojasiewicz inequality [29], and is a special case of the Kurdyka–Łojasiewicz (KL) inequality [17, 18]. Such functions may give rise to spiral gradient flow even if smoothness and convexity are assumed [9]. Also, Łojasiewicz functions may not be convex. The compatibility with nonconvexity has gained them increasing amount of attention, due to the surge in nonconvex objectives from machine learning and deep learning. Indeed, the Łojasiewicz inequality can well capture the local landscape of neural network losses, since some good local approximators for neural network losses, including polymonials and semialgebraic functions, locally satisfy the Łojasiewicz inequality.

Previously, the understanding of Łojasiewicz functions have been advanced by many researchers [29, 21, 17, 18, 19, 5, 1, 23]. In their seminal work, [1] proved the-state-of-the-art convergence rate for $\{\|x_t - x_\infty\|\}_t$, where $\{x_t\}_t$ is the sequence generated by the proximal algorithm, x_∞ is the limit of $\{x_t\}_t$ that is also a critical point of the objective f, and $\|\cdot\|$ denotes the Euclidean norm.

While [1] focused on the convergence rate of $\{\|x_t - x_\infty\|\}_t$, we focus on the convergence rate of $\{f(x_t)\}_t$, where x_t is governed by the SZGD algorithm. As discussed above, the reason is twofold: 1. the gradient flow of Łojasiewicz functions may inevitably be spiral [9]; 2. recent motivations from deep learning put more emphasis on the convergence rate of $\{f(x_t)\}_t$. In particular, we prove the following results for SZGD. Let the objective function f satisfy the Łojasiewicz inequality with exponent θ (Definition 2). Then under the conditions in [1], plus an L-smoothness condition (Definition 1), the followings are true.

- The sequence $\{x_t\}_t$ converges to a limit x_∞ almost surely. In addition, x_∞ is a critical point of f.
- If $\theta \in (0, \frac{1}{2}]$, then there exists $Q > 1$ such that $\{Q^t(f(x_t) - f(x_\infty))\}_t$ converges to zero almost surely.
- If $\theta \in (\frac{1}{2}, 1)$, then $\{\frac{t^{\frac{1}{2\theta-1}}}{\log t}(f(x_t) - f(x_\infty))\}_t$ converges to zero almost surely.

These results imply that (1) when $\theta \in (0, \frac{1}{2}]$, $\{f(x_t) - f(x_\infty)\}_t$ converges to zero linearly almost surely, (2) when $\theta \in (\frac{1}{2}, 1)$, $\{f(x_t) - f(x_\infty)\}_t$ converges to zero at rate $o\left(t^{\frac{1}{1-2\theta}} \log t\right)$ almost surely. To the best of our knowledge, these results are the first almost sure convergence rates for stochastic zeroth order algorithms on Łojasiewicz functions.

Although we have argued that the convergence results of $\{x_t\}_t$ are perhaps not as important nowadays, we prove the following convergence rate for $\{x_t\}_t$.

- If $\theta \in (0, \frac{1}{2}]$, then there exists $Q > 1$ such that $\{Q^t \sum_{s=t}^\infty \|x_{s+1} - x_s\|^2\}_t$ and $\{Q^t \|x_t - x_\infty\|\}_t$ converges to zero almost surely.
- If $\theta \in (\frac{1}{2}, 1)$, then $\{\frac{t^{\frac{1}{2\theta-1}}}{\log t} \sum_{s=t}^\infty \|x_{s+1} - x_s\|^2\}_t$ and $\{\frac{t^{\frac{1-\theta}{2\theta-1}}}{\log t} \|x_t - x_\infty\|\}_t$ converges to zero almost surely.

Our results suggest that, when $\{x_t\}_t$ is generated by SZGD, $\{f(x_t)\}_t$ tends to converge faster than $\{\|x_t - x_\infty\|\}_t$. This observation resonates with the recent example in [9].

In Section 5, we discuss and relate our results to the classic work [1]. In particular, we show that $\{f(x_t)\}_t$ ($\{x_t\}_t$ governed by the proximal algorithm) converges at rate $O(t^{\frac{1}{1-2\theta}})$ if $\theta \in (\frac{1}{2}, 1)$. Together with the seminal results [1], it suggests that, when $\{x_t\}_t$ is governed by the proximal algorithm, $\{f(x_t)\}_t$ converges faster than $\{\|x_t - x_\infty\|\}_t$. This observation is similar to that for SZGD, again resonating with the recent example in [9].

2 Related Works

Zeroth order optimization is a central scheme in many fields (e.g., [25, 8, 32]). Among many zeroth order optimization mechanisms, a classic and prosperous line of works focuses on estimating gradient/Hessian using zeroth order information and use the estimated gradient/Hessian for downstream optimization algorithms.

A classic line of related works is the Robbins–Monro–Kiefer–Wolfowitz-type algorithms [30, 15] from stochastic approximation. See (e.g., [20, 3]) for exposition. The Robbins–Monro and Kiefer–Wolfowitz scheme has been used in stochastic optimization and related fields (e.g., [26, 35]). In particular, [4] have shown that stochastic gradient descent algorithm either converges to a stationary point or goes to infinity, almost surely. While the results of [4] is quite general, no almost sure convergence *rate* is given. As an example of recent development, [35] showed that convergence results for Robbins–Monro when the objective is nonconvex. While the study of Robbins–Monro and Kiefer–Wolfowitz has spanned 70 years, stochastic zeroth order optimization has not come to its modern form until early this century.

In recent decades, due to lack of direct access to gradients in real-world applications, zeroth order optimization has attracted the attention of many researchers. In particular, [12] introduced the single-point gradient estimator for the purpose of bandit learning. Afterwards, many modern gradient/Hessian estimators have been introduced and subsequent zeroth order optimization algorithms have been studied. To name a few, [10, 28] have studied zeroth order optimization algorithm for convex objective and established in expectation convergence rates. [2] used the Stein's identity for Hessian estimators and combined this estimator with cubic regularized Newton's method [27]. [34, 33] provided refined analysis of Hessian/gradient estimators over Riemannian manifolds. [24] studied zeroth order optimization over Riemannian manifolds and proved in expectation convergence rates. The above mentioned stochastic zeroth order optimization works focus *in expectation* convergence rates. Probabilistically stronger results have also been established for stochastic optimization methods recently. For example, [22] provide high probability convergence rates for composite optimization problems. Yet the almost sure convergence rates for stochastic zeroth order optimization have not been established until [16], which we will soon discuss with more details.

The study of Łojasiewicz functions forms an important cluster of related works. Łojasiewicz functions satisfies the Łojasiewicz inequality with Łojasiewicz exponent θ [21]. An important special case of the Łojasiewicz inequality is the Polyak–Łojasiewicz inequality [29], which corresponds to the Łojasiewicz inequality with $\theta = \frac{1}{2}$. In [17, 18], the Łojasiewicz inequality was generalized to the Kurdyka–Łojasiewicz inequality. Subsequently, the geometric properties has been intensively studied, along with convergence studies of optimization algorithms on Kurdyka–Łojasiewicz -type functions [29, 21, 17, 18, 19, 5, 1, 23]. Yet no prior works focus on stochastic zeroth order methods for Łojasiewicz functions.

Perhaps the single most related work is the recent beautiful work by [16]. In [16], almost sure convergence rates are established for convex objectives, and *in expectation* convergence rates are proved for Polyak–Łojasiewicz functions. Compared to [16]'s study of Polyak–Łojasiewicz functions, our results are stronger since we provide *almost sure convergence rates* for Łojasiewicz functions, not to mention that Łojasiewicz functions are more general than Polyak–Łojasiewicz functions. To the best of our knowledge, we establish the first *almost sure convergence rates* for Łojasiewicz functions.

3 Preliminaries

3.1 Gradient Estimation

Consider gradient estimation tasks in $\mathbb{R}^n$. The gradient estimator we use is [11]:

$$\widehat{\nabla} f_k^{\delta}(x) := \frac{n}{2\delta k} \sum_{i=1}^{k} \left(f(x + \delta v_i) - f(x - \delta v_i) \right) v_i, \quad \forall x \in \mathbb{R}^n, \tag{3}$$

where $[v_1, v_2, \cdots, v_k] =: V$ is uniformly sampled from the Stiefel's manifold $\mathrm{St}(n,k) = \{X \in \mathbb{R}^{n\times k} : X^\top X = I_k\}$, and δ is the finite difference granularity. In practice, one can firstly generate a random matrix $U \in \mathbb{R}^{n\times k}$ of *i.i.d.* standard Gaussian ensemble. Then apply the Gram-Schmit process on U to obtain the matrix V.

Definition 1. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called L-smooth if it is continuously differentiable, and*

$$\|\nabla f(x) - \nabla f(x')\| \le L\|x - x'\|, \quad \forall x, x' \in \mathbb{R}^n.$$

More generally, a function $f : \mathbb{R}^n \to \mathbb{R}$ is called (p, L')-smooth ($p \geq 2$, $L' > 0$) if it is p-times continuously differentiable, and

$$\|\partial^{p-1} f(x) - \partial^{p-1} f(x')\| \leq L' \|x - x'\|, \quad \forall x, x' \in \mathbb{R}^n,$$

where $\partial^{p-1} f(x)$ is the $(p-1)$-th total derivative of f at x, and $\|\cdot\|$ is the spectral norm when applied to symmetric multi-linear forms and the Euclidean norm when applied to vectors. The spectral norm of a symmetric multi-linear form F or order m is $\|F\| := \sup_{v \in \mathbb{R}^n, \|v\|=1} F[\underbrace{v, v, v, \cdots, v}_{p \text{ times}}]$.

With the above description of smoothness, we can state theorem on statistical properties of the estimator (3). These properties are in Theorems 1 and 2.

Theorem 1 ([12, 11])**.** *The gradient estimator $\widehat{\nabla} f_k^\delta$ satisfies*

(a) If f is L-smooth, then for all $x \in \mathbb{R}^n$, $\left\| \mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right] - \nabla f(x) \right\| \leq \frac{Ln\delta}{n+1}$.

(b) If f is $(4, L_4)$-smooth, then for all $x \in \mathbb{R}^n$, the bias of gradient estimator satisfies

$$\left\| \mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right] - \nabla f(x) \right\| \leq \frac{\delta^2}{2n} \sqrt{\sum_{i=1}^n \left(\sum_{j=1}^n F_{jji} \right)^2} + \frac{\delta^3 L_4 n}{24},$$

where F_{ijl} denotes the (i, j, l)-component of $\partial^3 f(x)$.

Theorem 2 ([11])**.** *If f is $(3, L_3)$-smooth, the variance of the gradient estimator for f (Eq. 3) satisfies*

$$\begin{aligned} &\mathbb{E}\left[\left\| \widehat{\nabla} f_k^\delta(x) - \mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right] \right\|^2 \right] \\ \leq &\left(\frac{n}{k} - 1\right) \|\nabla f(x)\|^2 + \frac{L_3 \delta^2}{3} \left(\frac{n^2}{k} - n\right) \|\nabla f(x)\| + \frac{L_3^2 n^2 \delta^4}{36k}, \quad \forall x \in \mathbb{R}^n. \end{aligned}$$

While Theorem 2 provides a variance bound for $(3, L_3)$-functions, a similar result holds true for L-smooth functions (Theorem 3).

Theorem 3. *If f is L-smooth, then variance of the gradient estimator for f (Eq. 3) satisfies*

$$\begin{aligned} &\mathbb{E}\left[\left\| \widehat{\nabla} f_k^\delta(x) - \mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right] \right\|^2 \right] \\ \leq &\left(\frac{n}{k} - 1\right) \|\nabla f(0)\|^2 + \frac{4L\delta}{\sqrt{3}} \left(\frac{n^2}{k} - n\right) \|\nabla f(0)\| + \frac{4L^2 n^2 \delta^2}{3k}, \end{aligned}$$

for all $x \in \mathbb{R}^n$.

Compared to Theorem 2, Theorem 3 provides a bound that is order-of-magnitude looser in terms of the granularity δ. Although not as tight, Theorem 3 is sufficient for our purpose. Throughout, we will use Theorem 3, since it makes milder assume on the function. The proof of Theorem 3 can be found in Section 7.

Before ending this subsection, we also recall a basic property for L-smooth functions.

Proposition 1. *If f is L-smooth, then it holds that*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|^2, \qquad x, y \in \mathbb{R}^n.$$

The above proposition is known as the quadratic upper bound. This fact can be found in expository works in the field of optimization (e.g., [6]).

3.2 Łojasiewicz Functions

Łojasiewicz functions are functions that satisfies the Łojasiewicz inequality. We start with the differentiable Łojasiewicz functions (Definition 2). A more general version of the Łojasiewicz inequality [21], where gradient is replaced by subgradients, is discussed in Section 5.

Definition 2. *A differentiable function is said to be a (differentiable) Łojasiewicz function with Łojasiewicz exponent $\theta \in (0,1)$ if for any x^* with $\nabla f(x^*) = 0$, there exist constants $\kappa, \mu > 0$ such that*

$$|f(x) - f(x^*)|^\theta \leq \kappa \|\nabla f(x)\|, \quad \forall x \text{ with } \|x - x^*\| \leq \mu. \tag{4}$$

We call (4) Łojasiewicz inequality. Without loss of generality, we let $\kappa = 1$ to avoid clutter.

An important special case of the Łojasiewicz inequality is the Polyak–Łojasiewicz inequality [29], which corresponds to a case of the Łojasiewicz exponential θ being $\frac{1}{2}$. Also, Łojasiewicz inequality is an important special case of the Kurdyka–Łojasiewicz (KL) inequality [17, 18]. Roughly speaking, the KL inequality does not assume differentiability, and replaces the left-hand-side of (4) with a more general function (in $f(x)$). In its general form, the Łojasiewicz inequality does not require the objective function to be differentiable. In such cases, gradient on the right-hand-side of (4) is replaced with subgradient. We will discuss the subgradient version in Section 5.

3.3 Conventions and Notations

Before proceeding to the main results, we put forward several conventions.

- We use C to denote non-stochastic constants that is independent of t, not necessarily referring to the same value at each occurrence.
- We use $\mathcal{F}_t$ to denote the σ-algebra generated by all randomness after arriving at x_t, but before obtaining the estimator $\widehat{\nabla} f_k^{\delta_t}(x_t)$. We use $\mathbb{E}_t[\cdot]$ to denote $\mathbb{E}[\cdot|\mathcal{F}_t]$.

For Section 4, the objective function f satisfies Assumption 1.

Assumption 1. *Throughout Section 4, the objective function f satisfies:*

(i) *f is L-smooth for some constant $L > 0$. (See Definition 1).*

(ii) *f is a (differentiable) Łojasiewicz function with Łojasiewicz exponent θ. (See Definition 2)*

(iii) $\inf_{x \in \mathbb{R}^n} f(x) > -\infty$.

(iv) *Let $\{x_t\}_t$ be the sequence generated by the SZGD algorithm. We assume $\{x_t\}_t$ is bounded with probability 1.*

Note that compared to the classic work [1], the only additional requirement is item (i) in Assumption 1. Items (ii), (iii), and (iv) are all assumed in [1]. This (only) additional requirement might be inevitable for stochastic zeroth order optimization tasks (See Figure 3 and related discussions in Section 6).

Also, we want to point out that the last two items in Assumption 1 are common and mild. Item (iii) is typical for (unconstrained) minimization problems. Item (iv) is also mild since in practice one can easily check whether $\{x_t\}_t$ stays bounded. In addition, if item (iv) is violated, then functions as simple as $x \mapsto \frac{1}{1+e^{-x}}$ (it satisfies items (i), (ii) and (iii) in Assumption 1) can vandalize an innocent algorithm.

4 Convergence Analysis for SZGD

This section presents convergence analysis for the SZGD algorithm. Before proceeding, we first summarize SZGD in Algorithm 1.

Algorithm 1 Stochastic Zeroth-order Gradient Descent (SZGD)

1: **Input:** Dimension n, Number of orthogonal Directions for the Estimators k. /* function f is L-smooth. */
2: Pick $x_0 \in \mathbb{R}^n$ and $\delta_0 = 1$. /* or $\delta_0 \in (0,1)$. */
3: **for** $t = 0, 1, 2, \cdots$ **do**
4: $\quad x_{t+1} = x_t - \eta_t \widehat{\nabla} f_k^{\delta_t}(x_t)$.
5: $\quad \delta_{t+1} = \delta_t / 2$.
6: **end for**

Remark 1. *Similar to most (if not all) convergence analysis, some requirements need to be imposed on the step size η_t. Same as [1], we require that there exist $0 < \eta_- < \eta_+ < \infty$ such that $\eta_- \leq \eta_t \leq \eta_+$ for all t. To avoid notational clutter in the analysis, we simply let $\eta_t = \eta \in (0, \infty)$ for all t henceforth. For theoretical purpose, setting η_t to a positive constant does not sacrifice generality, compared to setting η_t to be in a compact interval on the positive half line.*

The main almost sure guarantee for SZGD is in Theorem 4.

Theorem 4. *Instate Assumption 1 and Remark 1. Let x_t be a sequence generated by SZGD (Algorithm 1). Then $\{x_t\}$ converges to a critical point x_∞ almost surely. In addition, it holds that*

(a) if $\theta \in (0, \frac{1}{2}]$ and η satisfies that $-1 \leq \left(\frac{Ln\eta^2}{2k} - \eta\right) < 0$, there exists a constant $Q > 1$ such that $\{Q^t (f(x_t) - f(x_\infty))\}_t$ converges to 0 almost surely.

(b) if $\theta \in (\frac{1}{2}, 1)$ and $\eta \in \left(0, \frac{2k}{Ln}\right)$, then it holds that $\{\frac{t^{\frac{1}{2\theta-1}}}{\log t}(f(x_t) - f(x_\infty))\}_t$ converges to 0 almost surely.

This theorem gives almost sure convergence rate of $\{f(x_t)\}_{t\in\mathbb{N}}$. It states that, with probability 1, (*a*) $\{f(x_t)\}_{t\in\mathbb{N}}$ converges linearly if $\theta \in (0, \frac{1}{2}]$, and (*b*) $\{f(x_t)\}_{t\in\mathbb{N}}$ converges at rate $o\left(t^{\frac{1}{1-2\theta}} \log t\right)$ if $\theta \in (\frac{1}{2}, 1)$. Similar results for convergence guarantees for $\{x_t\}_t$ can be found in Section 4.3.

The rest of this section is organized as follows. In Section 4.1, we show that Algorithm 1 converges almost surely. In particular, we show that the sequences $\{x_t\}_{t\in\mathbb{N}}$ generated by SZGD converges to a critical point almost surely. We also prove that with probability 1, $\{f(x_t) > f(x_{t+1})\}$ occurs finitely many times. There properties not only reveal behavior patterns of SZGD, but also serve as step stones for proving almost sure convergence rates. A major reason for proving asymptotic convergence first is that the Łojasiewicz inequality only describes local function landscape near critical points. Thus we first need to show that $\{x_t\}_t$ gets close to a critical point almost surely, before we can apply the Łojasiewicz inequality to prove convergence rates. In Section 4.2, we establish almost sure convergence rates for $\{f(x_t)\}_t$ associated with Algorithm 1. Then in Section 4.2, we establish almost sure convergence rates for $\{x_t\}_t$.

4.1 Asymptotic Convergence

In Lemma 1, we show that $\{\|\nabla f(x_t)\|\}_{t\in\mathbb{N}}$ converges to zero almost surely.

Lemma 1. *Instate Assumption 1 and Remark 1. Let $\{x_t\}_t$ be the sequence governed by Algorithm 1. Let step size $\eta \in \left(0, \frac{2k}{Ln}\right)$. Then $\{\|\nabla f(x_t)\|\}_{t\in\mathbb{N}}$ converges to zero almost surely.*

Proof. By Theorems 1 and 3, we have

$$\mathbb{E}_t\left[\widehat{\nabla} f_k^{\delta_t}(x_t)\right] = \nabla f(x_t) + O\left(\frac{n\delta_t}{n+1}\mathbf{1}\right), \tag{5}$$

and

$$\mathbb{E}_t\left[\|\widehat{\nabla} f_k^{\delta_t}(x_t)\|^2\right] \leq \frac{n}{k}\|\nabla f(x_t)\|^2 + O\left(\left(\frac{n^2}{k} - n\right)\|\nabla f(x_t)\|\delta_t\right). \tag{6}$$

By L-smoothness and Proposition 1, we have

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2,$$

which gives

$$f(x_{t+1}) \leq f(x_t) - \eta \nabla f(x_t)^\top \widehat{\nabla} f_k^{\delta_t}(x_t) + \frac{L\eta^2}{2}\|\widehat{\nabla} f_k^{\delta_t}(x_t)\|^2.$$

By (5) and (6), taking conditional expectation on both sides of the above inequality gives

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1})\right] &\leq f(x_t) - \eta \nabla f(x_t)^\top \mathbb{E}_t\left[\widehat{\nabla} f_k^{\delta_t}(x_t)\right] + \frac{L\eta^2}{2}\mathbb{E}_t\left[\|\widehat{\nabla} f_k^{\delta_t}(x_t)\|^2\right] \\
&\leq f(x_t) - \eta \nabla f(x_t)^\top \left(\nabla f(x_t) + O\left(\frac{n\delta_t}{n+1}\mathbf{1}\right)\right) \\
&\quad + \frac{L\eta^2}{2}\left(\frac{n}{k}\|\nabla f(x_t)\|^2 + O\left(\left(\frac{n^2}{k} - n\right)\|\nabla f(x_t)\|\delta_t\right)\right) \\
&= f(x_t) + \left(\frac{L\eta^2 n}{2k} - \eta\right)\|\nabla f(x_t)\|^2 + O\left(\frac{n^2}{k}\|\nabla f(x_t)\|\delta_t\right).
\end{aligned}$$

The above implies that, there exists a constant C such that, when t is sufficiently large,

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1})\right] &\leq f(x_t) + \left(\frac{L\eta^2 n}{2k} - \eta\right)\|\nabla f(x_t)\|^2 + C \cdot \left(\frac{n^2}{k}\|\nabla f(x_t)\|\delta_t\right) \\
&\leq f(x_t) + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right)\|\nabla f(x_t)\|^2 + C\delta_t. \qquad (7)
\end{aligned}$$

Suppose, in order to get a contradiction, that there exists $\alpha > 0$ such that $\|\nabla f(x_t)\|^2 > \alpha$ infinitely often. Then let T_0 be a constant such that $\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t < \frac{1}{2}\left(\frac{L\eta^2 n}{2k} - \eta\right) < 0$ for all $t \geq T_0$. Then if $\|\nabla f(x_t)\|^2 > \alpha$ infinitely often, (7) gives

$$\mathbb{E}\left[f(x_t)\right] \leq \mathbb{E}\left[f(x_{T_0})\right] + \frac{1}{2}M_t\left(\frac{L\eta^2 n}{2k} - \eta\right)\alpha + C, \quad t \geq T_0$$

for some $\{M_t\}_t \subseteq \mathbb{N}$ such that $\lim_{t\to\infty} M_t = \infty$.

Taking limits on both sides of the above inequality gives

$$\liminf_{t\to\infty} \mathbb{E}\left[f(x_t)\right] \leq \mathbb{E}\left[f(x_{T_0})\right] + \liminf_{t\to\infty} \frac{1}{2}M_t\left(\frac{L\eta^2 n}{2k} - \eta\right)\alpha + C = -\infty,$$

which leads to a contradiction to $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$ (Assumption 1).

By the above proof-by-contradiction argument, we have shown

$$\mathbb{P}\left(\|\nabla f(x_t)\|^2 > \alpha \text{ infinitely often}\right) = 0, \qquad \forall \alpha > 0.$$

Note that

$$\begin{aligned}
\left\{\lim_{t\to\infty}\|\nabla f(x_t)\|^2 = 0\right\}^c &= \bigcup_{\alpha>0}\{\|\nabla f(x_t)\|^2 > \alpha \text{ infinitely often}\} \\
&= \bigcup_{\alpha\in\mathbb{Q}_+}\{\|\nabla f(x_t)\|^2 > \alpha \text{ infinitely often}\},
\end{aligned}$$

where we replace the union over an uncountable set with a union over a countable set.

Thus it holds that

$$\begin{aligned}
1 - \mathbb{P}\left(\lim_{t\to\infty}\|\nabla f(x_t)\|^2 = 0\right) &= \mathbb{P}\left(\bigcup_{\alpha\in\mathbb{Q}_+}\{\|\nabla f(x_t)\|^2 > \alpha \text{ infinitely often}\}\right) \\
&\leq \sum_{b\in\mathbb{Q}_+}\mathbb{P}\left(\|\nabla f(x_t)\|^2 > \alpha \text{ infinitely often}\right) \\
&= 0,
\end{aligned}$$

which concludes the proof.

□

As consequences of Lemma 1, we know that $\{x_t\}_{t\in\mathbb{N}}$ and $\{f(x_t)\}_{t\in\mathbb{N}}$ converge almost surely.

Proposition 2. *Instate Assumption 1 and Remark 1. Let $\{x_t\}_{t\in\mathbb{N}}$ be the sequence of Algorithm 1. Let step size $\eta \in \left(0, \frac{2k}{Ln}\right)$. Then $\{x_t\}_t$ converges to a bounded critical point of f almost surely. Let x_∞ be the almost sure limit of $\{x_t\}_{t\in\mathbb{N}}$. It holds that $\lim_{t\to\infty} f(x_t) = f(x_\infty)$ almost surely.*

Proof of Proposition 2. By Lemma 1, continuity of f, and boundedness of $\{x_t\}_{t\in\mathbb{N}}$, the sequence $\{x_t\}_{t\in\mathbb{N}}$ must converges to a critical point of f almost surely. Otherwise there will be a contradiction.

□

Next we show that, with probability 1, the sequence $\{f(x_t)\}_t$ is non-increasing when t is sufficiently large. This property is desirable for most minimization tasks.

Lemma 2. *Instate Assumption 1 and Remark 1. Let $\{x_t\}_{t\in\mathbb{N}}$ be the sequence of Algorithm 1. Let step size $\eta \in \left(0, \frac{2k}{Ln}\right)$. Then it holds that*

$$\mathbb{P}\left(f(x_{t+1}) > f(x_t) \text{ infinitely often}\right) = 0.$$

In other words, there exists a constant T_0, such that $f(x_{t+1}) \le f(x_t)$ for all $t \ge T_0$ with probability 1.

Proof. Consider the event

$$\left\{f(x_{t+1}) > f(x_t) + C\delta_t \text{ for arbitrarily large } t\right\}.$$

When $f(x_{t+1}) > f(x_t) + C\delta_t$, taking total expectation on both sides of (7) gives

$$\begin{aligned}\mathbb{E}\left[f(x_t)\right] + C\delta_t &< \mathbb{E}\left[f(x_{t+1})\right] \\ &\le \mathbb{E}\left[f(x_t)\right] + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right)\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + C\delta_t,\end{aligned}$$

which is absurd for large t, since $\left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right)\mathbb{E}\left[\|\nabla f(x_t)\|^2\right]$ is nonpositive for large t ($\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t \le 0$ for large t since $\eta \in (0, \frac{2k}{Ln})$). Note that we can pick C so that the C's on both sides of the above inequality take the same value.

Thus it holds that, for any subsequence $\{m_t\}_{t\in\mathbb{N}}$ of $\mathbb{N}$,

$$\mathbb{P}\left(f(x_{m_t+1}) > f(x_{m_t}) + C\delta_{m_t} \text{ for arbitrarily large } t\right) = 0 \tag{8}$$

Note that the following events are equivalent

$$\begin{aligned}&\bigcup_{\alpha\in\mathbb{Q}_+}\left\{f(x_{t+1}) > f(x_t) + \alpha \text{ infinitely often}\right\} \\ =& \bigcup_{\alpha\in\mathbb{Q}_+} \bigcup_{\substack{\{m_t\}_{t\in\mathbb{N}} \text{ is a} \\ \text{subsequence of } \mathbb{N}}} \bigcap_{t\in\mathbb{N}} \left\{f(x_{m_t+1}) > f(x_{m_t}) + \alpha\right\}.\end{aligned}$$

For any $\alpha \in \mathbb{Q}_+$, there exists T_0' ($T_0' \ge \frac{\log C - \log\alpha}{\log 2}$ suffices)

$$\begin{aligned}&\bigcap_{t\in\mathbb{N}}\left\{f(x_{m_t+1}) > f(x_{m_t}) + \alpha\right\} \\ \subseteq& \bigcap_{t\ge T_0'}\left\{f(x_{m_t+1}) > f(x_{m_t}) + C\delta_{m_t}\right\} \\ \subseteq& \left\{f(x_{m_t+1}) > f(x_{m_t}) + C\delta_{m_t} \text{ for arbitrarily large } t\right\}.\end{aligned}$$

Thus it holds that

$$
\begin{aligned}
&\mathbb{P}\left(f(x_{t+1}) > f(x_t) \text{ infinitely often}\right) \\
&= \mathbb{P}\left(\bigcup_{\alpha\in\mathbb{Q}_+} \{f(x_{t+1}) > f(x_t) + \alpha \text{ infinitely often}\}\right) \\
&= \mathbb{P}\left(\bigcup_{\alpha\in\mathbb{Q}_+} \bigcup_{\substack{\{m_t\}_{t\in\mathbb{N}} \text{ is a} \\ \text{subsequence of } \mathbb{N}}} \bigcap_{t\in\mathbb{N}} \{f(x_{m_t+1}) > f(x_{m_t}) + \alpha\}\right) \\
&\le \mathbb{P}\left(\bigcup_{\alpha\in\mathbb{Q}_+} \bigcup_{\substack{\{m_t\}_{t\in\mathbb{N}} \text{ is a} \\ \text{subsequence} \\ \text{of } \mathbb{N}}} \{f(x_{m_t+1}) > f(x_{m_t}) + C\delta_{m_t} \text{ for arbitrarily large } t\ \}\right) \\
&\le \sum_{\alpha\in\mathbb{Q}_+} \sum_{\substack{\{m_t\}_{t\in\mathbb{N}} \text{ is a} \\ \text{subsequence of } \mathbb{N}}} \mathbb{P}\left(f(x_{m_t+1}) > f(x_{m_t}) + C\delta_{m_t} \text{ for arbitrarily large } t\right) \\
&= 0,
\end{aligned}
$$

where the second last line uses (8).

□

4.2 Convergence Rate of $\{f(x_t)\}_{t\in\mathbb{N}}$

In the previous subsection, we have proved asymptotic convergence results for the SZGD algorithm. This section is devoted to convergence rate analysis of $\{f(x_t)\}_{t\in\mathbb{N}}$. Since the SZGD algorithm converges to a critical point almost surely, x_t stays close to a critical point when t is large, and the Łojasiewicz inequality can be used. We first state Proposition 3 that holds true for large t.

Proposition 3. *Instate Assumption 1 and Remark 1. Let $\{x_t\}_{t\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Let $\eta \in \left(0, \frac{2k}{Ln}\right)$. Let x_∞ be the almost sure limit of $\{x_t\}_{t\in\mathbb{N}}$ (Proposition 2). Then there exists constants T_0 and C_0 such that the following holds with probability 1:*

(i) *x_∞ is a critical point of f (Proposition 2) and $\|x_t - x_\infty\| \le \mu$ for all $t \ge T_0$. $\|\nabla f(x_t)\| < 1$ for all $t \ge T_0$. (Lemma 1)*

(ii) *$f(x_{t+1}) \ge f(x_t)$ for all $t \ge T_0$ (Lemma 2).*

(iii) *$0 \le f(x_t) - f(x_\infty) \le 1$ for all $t \ge T_0$. (Proposition 2, Lemma 2)*

(iv) *$\mathbb{E}\left[f(x_t) - f(x_\infty)\right] \in \left(0, \left(\frac{1}{-2\theta\left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right)}\right)^{\frac{1}{2\theta-1}}\right)$ for all $t \ge T_0$. (Proposition 2, Lemma 2)*

(v) *$\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t < \frac{1}{2}\left(\frac{L\eta^2 n}{2k} - \eta\right)$ for all $t \ge T_0$;*

In addition, given fixed $\theta \in (\frac{1}{2}, 1)$,

(vi) *$C_0 > 1$ and $\frac{C_0}{2\theta-1}t^{\frac{2\theta}{1-2\theta}} + C_0^{2\theta}\left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right)t^{\frac{2\theta}{1-2\theta}} + C\delta_t \le 0$ for all $t \ge T_0$;*

(vii) *$\mathbb{E}\left[f(x_{T_0}) - f(x_\infty)\right] \le C_0 T_0^{\frac{1}{1-2\theta}}$.*

Proof. Items (i), (ii), (iii) and (iv) immediately follow from Proposition 2 and Lemma 2. Item (v) is a standard fact, since δ_t converges to zero (very fast). Note that items (i), (ii), (iii), (iv) and (v) does not involve C_0.

Since items (vi) and (vii) involve both C_0 and T_0, we need to be slightly more careful. For item (vi), since $C\delta_t t^{\frac{2\theta}{2\theta-1}}$ converges to zero (really fast), we can first pick T_0 so that $\delta_t t^{\frac{2\theta}{2\theta-1}}$ is small. Let's pick T_0 so that

$$C\delta_t t^{\frac{2\theta}{2\theta-1}} \leq \frac{1}{2\theta-1}, \quad t \geq T_0. \tag{9}$$

Consider a T_0 so that item (v) and the above inequality are satisfied. Since $2\theta > 1$, we can pick C_0 large enough so that, for $t \geq T_0$,

$$\begin{aligned}
&\frac{C_0}{2\theta-1} + C_0^{2\theta}\left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right) + C\delta_t t^{\frac{2\theta}{2\theta-1}} \\
\leq &\frac{C_0+1}{2\theta-1} + \frac{1}{2}C_0^{2\theta}\left(\frac{L\eta^2 n}{2k} - \eta\right) \leq 0,
\end{aligned}$$

where the second inequality uses (9) and item (v). Thus (vi) follows by multiplying both sides of the above inequality by $t^{\frac{2\theta}{1-2\theta}}$.

Finally, we can find C_0 satisfying item (vii), since $\mathbb{E}\left[f(x_{T_0}) - f(x_\infty)\right]$ is absolutely bounded for any given T_0. □

Remark 2. *Let T_0 be a number so that Proposition 3 holds, and recall the definition of x_∞ from Proposition 3. By items (i), (ii) and (iii) in Proposition 3, we know that, for $t \geq T_0$, x_t is close to a critical point, x_∞ is a critical point, and $f(x_t) \geq f(x_\infty)$. Thus the Łojasiewicz inequality implies that*

$$(f(x_t) - f(x_\infty))^{2\theta} \leq \|\nabla f(x_t)\|, \quad \textit{almost surely}$$

for all $t \geq T_0$.

By Proposition 3 and Remark 2, (7) implies that, for $t \geq T_0$,

$$\begin{aligned}
&\mathbb{E}_t\left[f(x_{t+1})\right] - f(x_\infty) \\
\leq &f(x_t) - f(x_\infty) + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right)\|\nabla f(x_t)\|^2 + C\delta_t \\
\leq &f(x_t) - f(x_\infty) + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right)(f(x_t) - f(x_\infty))^{2\theta} + C\delta_t.
\end{aligned} \tag{10}$$

The above inequality is a stochastic relation for the stochastic sequence $\{f(x_t) - f(x_\infty)\}_{t\in\mathbb{N}}$. In what follows, we will study the convergence behavior of this sequence.

Remark 3. *Since $f(x_\infty)$ is a constant, assuming $f(x_\infty) = 0$ does not affect the convergence analysis for (10). To avoid clutter, we assume $f(x_\infty) = 0$ henceforth. This assumption does not sacrifice generality.*

With Remark 3, we can write (10) as

$$\mathbb{E}_t\left[f(x_{t+1})\right] \leq f(x_t) + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right) f(x_t)^{2\theta} + C\delta_t. \tag{11}$$

Now we are ready to prove Theorem 4.

Proof of Theorem 4(a). Let C_0 and T_0 be two constants so that Proposition 3 is true.

By (11) and the Łojasiewicz inequality in Remark 2, with probability 1 it holds that, for all $t \geq T_0$,

$$\mathbb{E}_t\left[f(x_{t+1})\right] \leq f(x_t) + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right) f(x_t)^{2\theta} + C\delta_t \tag{12}$$

$$\leq \left(1 + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right)\right) f(x_t) + C\delta_t, \tag{13}$$

where the last inequality uses item (iii) in Proposition 3 and that $\theta \in (0, \frac{1}{2}]$.

For simplicity, write

$$\alpha_t := 1 + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t \right),$$

which is a decreasing sequence since δ_t is decreasing. Since $-1 \leq \left(\frac{Ln\eta^2}{2k} - \eta \right) < 0$, it holds that $\alpha_t \in (0, 1)$ for all $t \geq T_0$.

Taking total expectation on both sides of (13) gives

$$\mathbb{E}\left[f(x_{t+1})\right] \leq \alpha_t \mathbb{E}\left[f(x_t)\right] + C\delta_t$$

Since α_t decreases with t, there exists constants C and $A \in (0, 1)$ such that

$$\mathbb{E}\left[f(x_t)\right] \leq A^t f(x_0) + C2^{-t}. \tag{14}$$

Let $Q \in (1, \min(A^{-1}, 2))$ and let $Z_t := Q^t f(x_t)$.

Then applying the Markov inequality to Z_t gives, for an arbitrary positive number b,

$$\mathbb{P}\left(Z_t \geq b\right) \leq \frac{Q^t \mathbb{E}\left[f(x_t)\right]}{b} \leq \frac{(QA)^t f(x_0) + C(Q/2)^t}{b}.$$

Thus we have,

$$\mathbb{P}\left(Z_t > b \text{ infinitely often}\right) \leq \limsup_{t\to\infty} \frac{(QA)^t f(x_0) + C(Q/2)^t}{b} = 0.$$

Also, we have

$$\left\{ \lim_{t\to\infty} Z_t = 0 \right\}^c = \bigcup_{b>0} \{Z_t > b \text{ infinitely often}\} = \bigcup_{b\in\mathbb{Q}_+} \{Z_t > b \text{ infinitely often}\},$$

where we replace the union over an uncountable set with a union over a countable set.

Thus it holds that

$$\begin{aligned} 1 - \mathbb{P}\left(\lim_{t\to\infty} Z_t = 0 \right) &= \mathbb{P}\left(\bigcup_{b\in\mathbb{Q}_+} \{Z_t > b \text{ infinitely often}\} \right) \\ &\leq \sum_{b\in\mathbb{Q}_+} \mathbb{P}\left(Z_t > b \text{ infinitely often}\right) \\ &= 0, \end{aligned}$$

which shows that Z_t converges to zero with probability 1.

□

Proof of Theorem 4(b). Let C_0 and T_0 be two constants so that Proposition 3 holds true.

Taking total expectation on both sides of (11) gives, for sufficiently large t,

$$\begin{aligned} \mathbb{E}\left[f(x_{t+1})\right] &\leq \mathbb{E}\left[f(x_t)\right] + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t \right) \mathbb{E}\left[f(x_t)^{2\theta}\right] + C\delta_t \\ &\leq \mathbb{E}\left[f(x_t)\right] + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t \right) \mathbb{E}\left[f(x_t)\right]^{2\theta} + C\delta_t, \end{aligned}$$

where the last inequality uses Jensen's inequality.

For simplicity, write $y_t := \mathbb{E}\left[f(x_t)\right]$ for all t.

Next, we use induction to show that

$$y_t \leq C_0 t^{\frac{1}{1-2\theta}} \tag{15}$$

for all $t \geq T_0$.

Suppose that $y_t \leq C_0 t^{\frac{1}{1-2\theta}}$, which is true when $t = T_0$ (item (vii) in Proposition 3). Then for y_{t+1} we have

$$\begin{aligned} y_{t+1} &\leq y_t + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right) y_t^{2\theta} + C\delta_t \\ &\leq C_0 t^{\frac{1}{1-2\theta}} + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right) C_0^{2\theta} t^{\frac{2\theta}{1-2\theta}} + C\delta_t, \end{aligned} \tag{16}$$

where second inequality uses item (iv) in Proposition 3 and that the function

$$z \mapsto z + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right) z^{2\theta}$$

is strictly increasing when $z \in \left(0, \left(\frac{1}{-2\theta\left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right)}\right)^{\frac{1}{2\theta-1}}\right)$.

By applying Taylor's theorem (mean value theorem) to the function $h(w) = (t+w)^{\frac{1}{1-2\theta}}$, we have $h(1) = h(0) + h'(z)$ for some $z \in [0,1]$. This gives

$$(t+1)^{\frac{1}{1-2\theta}} = t^{\frac{1}{1-2\theta}} + \frac{1}{1-2\theta}(t+z)^{\frac{2\theta}{1-2\theta}} \geq t^{\frac{1}{1-2\theta}} + \frac{1}{1-2\theta} t^{\frac{2\theta}{1-2\theta}}, \tag{17}$$

since $z \in [0,1]$. Thus by (16), we have

$$\begin{aligned} y_{t+1} &\leq C_0 t^{\frac{1}{1-2\theta}} + \left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right) C_0^{2\theta} t^{\frac{2\theta}{1-2\theta}} + C\delta_t \\ &\leq C_0 (t+1)^{\frac{1}{1-2\theta}} \\ &\quad + C_0 \frac{1}{2\theta - 1} t^{\frac{2\theta}{1-2\theta}} + C_0^{2\theta}\left(\frac{L\eta^2 n}{2k} - \eta + C\frac{n^4}{k^2}\delta_t\right) t^{\frac{2\theta}{1-2\theta}} + C\delta_t \\ &\leq C_0 (t+1)^{\frac{1}{1-2\theta}}, \end{aligned}$$

where the second inequality uses (17), and the last inequality uses (vi) in Proposition 3.

By Markov inequality, we have, for an arbitrary positive number b,

$$\mathbb{P}\left(\frac{t^{\frac{1}{2\theta-1}}}{\log t} f(x_t) \geq b\right) \leq \frac{\frac{t^{\frac{1}{2\theta-1}}}{\log t}\mathbb{E}\left[f(x_t)\right]}{b} \leq \frac{C_0}{b \log t}.$$

Thus it holds that

$$\mathbb{P}\left(\frac{t^{\frac{1}{2\theta-1}}}{\log t} f(x_t) \geq b \text{ infinitely often}\right) \leq \limsup_{t\to\infty} \frac{C_0}{b\log t} = 0.$$

We conclude the proof by noting that

$$\begin{aligned} 1 - \mathbb{P}\left(\lim_{t\to\infty} \frac{t^{\frac{1}{2\theta-1}}}{\log t} f(x_t) = 0\right) &= \mathbb{P}\left(\bigcup_{b\in\mathbb{Q}_+}\left\{\frac{t^{\frac{1}{2\theta-1}}}{\log t} f(x_t) \geq b \text{ infinitely often}\right\}\right) \\ &\leq \sum_{b\in\mathbb{Q}_+} \mathbb{P}\left(\frac{t^{\frac{1}{2\theta-1}}}{\log t} f(x_t) \geq b \text{ infinitely often}\right) \\ &= 0. \end{aligned}$$

□

4.3 Convergence Rate of $\{x_t\}_{t\in\mathbb{N}}$

In the previous subsection, we have proved convergence rate of $\{f(x_t)\}_{t\in\mathbb{N}}$. In this section, we prove convergence rates for $\{\|x_t - x_\infty\|\}_{t\in\mathbb{N}}$ and $\{\sum_{s=t}^{\infty}\|x_s - x_{s+1}\|^2\}_{t\in\mathbb{N}}$.

Theorem 5. *Instate Assumption 1 and Remarks 1, 2 and 3. Let x_t be a sequence generated by the SZGD algorithm, and let x_∞ be the almost sure limit of $\{x_t\}_t$. Then it holds that*

(a) if $\theta \in (0, \frac{1}{2}]$ and η satisfies that $-1 \le \left(\frac{Ln\eta^2}{2k} - \eta\right) < 0$, there exists a constant $Q > 1$ such that $\{Q^t \sum_{s=t}^{\infty}\|x_{s+1} - x_s\|^2\}_t$ converges to 0 almost surely.

(b) if $\theta \in (\frac{1}{2}, 1)$ and $\eta \in \left(0, \frac{2k}{Ln}\right)$, then it holds that $\left\{\frac{t^{\frac{1}{2\theta-1}}}{\log}\sum_{s=t}^{\infty}\|x_{s+1} - x_s\|^2\right\}_t$ converges to 0 almost surely.

Proof. It holds that

$$\begin{aligned}
&\frac{k}{n}\left(\eta - \frac{L\eta^2 n}{2k}\right)\mathbb{E}_t\left[\|\widehat{\nabla} f_k^{\delta_t}(x_t)\|^2\right] \\
\le\ &\left(-\frac{L\eta^2 n}{2k} + \eta + C\frac{n^4}{k^2}\delta_t\right)\|\nabla f(x_t)\|^2 + C\cdot((n-k)\|\nabla f(x_t)\|\delta_t) \\
\le\ &f(x_t) - \mathbb{E}_t\left[f(x_{t+1})\right] + C\delta_t + C(n-k)\|\nabla f(x_t)\|\delta_t,
\end{aligned}$$

where the first inequality uses (6) and the second inequality uses (7).

Since $x_{t+1} = x_t - \eta\widehat{\nabla} f_k^{\delta_t}(x_t)$, the above inequality implies that, for all t

$$\begin{aligned}
&\frac{k}{n}\left(\frac{1}{\eta} - \frac{Ln}{2k}\right)\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right] \\
\le\ &f(x_t) - \mathbb{E}_t\left[f(x_{t+1})\right] + C\delta_t + C(n-k)\|\nabla f(x_t)\|\delta_t.
\end{aligned} \tag{18}$$

Taking total expectation on both sides of the above inequality gives

$$\begin{aligned}
&\frac{k}{n}\left(\frac{1}{\eta} - \frac{Ln}{2k}\right)\mathbb{E}\left[\|x_{s+1} - x_s\|^2\right] \\
\le\ &\mathbb{E}\left[f(x_t)\right] - \mathbb{E}\left[f(x_{t+1})\right] + C\delta_t + C(n-k)\,\mathbb{E}\left[\|\nabla f(x_t)\|\right]\delta_t.
\end{aligned}$$

By item (i) in Proposition 3, summing up the above inequality gives

$$\frac{k}{n}\left(\frac{1}{\eta} - \frac{Ln}{2k}\right)\mathbb{E}\left[\sum_{s=t}^{\infty}\|x_{s+1} - x_s\|^2\right] \le \mathbb{E}\left[f(x_t)\right] + C(n-k)\delta_t, \quad \forall t \ge T_0.$$

Since we have proved the convergence rate for $\{\mathbb{E}\left[f(x_t)\right]\}_t$, we can conclude the proof by probabilistic arguments similar to those for Theorem 4.

□

Next we provide almost sure convergence rate guarantee for $\{\|x_t - x_\infty\|\}_t$ in Theorem 6.

Theorem 6. *Instate Assumption 1 and Remarks 1, 2 and 3. Let x_t be a sequence generated by the SZGD algorithm, and let x_∞ be the almost sure limit of $\{x_t\}_t$. Then it holds that*

(a) if $\theta \in (0, \frac{1}{2}]$ and η satisfies that $-1 \le \left(\frac{Ln\eta^2}{2k} - \eta\right) < 0$, there exists a constant $Q > 1$ such that $\{Q^t\|x_t - x_\infty\|\}_t$ converges to 0 almost surely.

(b) if $\theta \in (\frac{1}{2}, 1)$ and $\eta \in \left(0, \frac{2k}{Ln}\right)$, then it holds that $\left\{\frac{t^{\frac{1-\theta}{2\theta-1}}}{\log t}\|x_t - x_\infty\|\right\}_t$ converges to 0 almost surely.

Proof. In [1], Attouch and Bolte uses properties of the function $x \mapsto -x^{1-\theta}$ ($x > 0$, $\theta \in (0,1)$) to study the convergence of $\{\|x_t - x_\infty\|\}_t$. Here we follow a similar path, but in a probabilistic manner. By convexity of the function $x \mapsto -x^{1-\theta}$ ($x > 0$, $\theta \in (0,1)$), it holds that

$$f(x_t)^{1-\theta} - f(x_{t+1})^\theta \geq (1-\theta) f(x_t)^{-\theta} \left(f(x_t) - f(x_{t+1})\right).$$

Taking conditional expectation (conditioning on $\mathcal{F}_t$) on both sides of the above equation gives

$$\begin{aligned}
&f(x_t)^{1-\theta} - \mathbb{E}_t\left[f(x_{t+1})^{1-\theta}\right] \\
&\geq (1-\theta) f(x_t)^{-\theta} \left(f(x_t) - \mathbb{E}_t\left[f(x_{t+1})\right]\right) \\
&\geq (1-\theta) f(x_t)^{-\theta} \\
&\quad \cdot \left(\frac{k}{n}\left(\frac{1}{\eta} - \frac{Ln}{2k}\right)\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right] - C\delta_t - C\left(n-k\right)\|\nabla f(x_t)\|\delta_t\right) \\
&\geq (1-\theta) f(x_t)^{-\theta} \left(\frac{k}{n}\left(\frac{1}{\eta} - \frac{Ln}{2k}\right)\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right] - C\delta_t\right),
\end{aligned} \tag{19}$$

where the second last line uses (18), and the last line holds with probability 1 since $\|\nabla f(x_t)\|$ is bounded (Lemma 1).

Also, it holds that, for sufficiently large t,

$$\begin{aligned}
0 < f(x_t)^\theta &\overset{①}{\leq} \|\nabla f(x_t)\| \overset{②}{\leq} \left\|\mathbb{E}_t\left[\widehat{\nabla} f_k^\delta(x_t)\right] + O\left(\frac{n\delta_t}{n+1}\mathbf{1}\right)\right\| \\
&\leq \left\|\mathbb{E}_t\left[\widehat{\nabla} f_k^\delta(x_t)\right]\right\| + C\delta_t = \frac{1}{\eta}\|\mathbb{E}_t\left[x_{t+1} - x_t\right]\| + C\delta_t \\
&\overset{③}{\leq} \frac{1}{\eta}\sqrt{\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right]} + C\delta_t \overset{④}{\leq} \frac{1}{\eta}\sqrt{\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right]} + \sqrt{C\delta_t},
\end{aligned} \tag{20}$$

where ① uses the Łojasiewicz inequality, ② uses Theorem 1, ③ uses the Jensen's inequality, and ④ uses that $C\delta_t \leq 1$ for large t ($t \geq \log_2 C$ suffices).

Note that (20) gives $0 < f(x_t)^\theta \leq \frac{1}{\eta}\sqrt{\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right]} + C\delta_t$. Combining this with (19) gives

$$\begin{aligned}
&f(x_t)^{1-\theta} - \mathbb{E}_t\left[f(x_{t+1})^{1-\theta}\right] \\
&\geq \frac{(1-\theta)\left(\frac{k}{n}\left(1 - \frac{Ln\eta}{2k}\right)\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right] - C\delta_t\right)}{\sqrt{\mathbb{E}_t\left[\|x_{t-1} - x_t\|^2\right]} + \sqrt{C\delta_t}} \\
&= (1-\theta)\frac{k}{n}\left(1 - \frac{Ln\eta}{2k}\right)\frac{\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right] - C\delta_t}{\sqrt{\mathbb{E}_t\left[\|x_{t-1} - x_t\|^2\right]} + \sqrt{C\delta_t}}.
\end{aligned} \tag{21}$$

For sufficiently large t such that $C\delta_t < 1$, we have

$$\begin{aligned}
&\sqrt{\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right]} - \left(C\delta_t\right)^{1/2} \\
&\leq \frac{\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right] - C\delta_t}{\sqrt{\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right]} + \sqrt{C\delta_t}} \\
&\leq \frac{n}{(1-\theta)k\left(1 - \frac{Ln\eta}{2k}\right)}\left(f(x_t)^{1-\theta} - \mathbb{E}_t\left[f(x_{t+1})^{1-\theta}\right]\right)
\end{aligned}$$

where the last inequality uses (21).

Combining Jensen's inequality and the above inequality gives

$$\begin{aligned}
&\mathbb{E}_t\left[\|x_{t+1} - x_t\|\right] - \left(C\delta_t\right)^{1/2} \\
&\leq \sqrt{\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right]} - \left(C\delta_t\right)^{1/2} \\
&\leq \frac{n}{(1-\theta)k\left(1 - \frac{Ln\eta}{2k}\right)}\left(f(x_t)^{1-\theta} - \mathbb{E}_t\left[f(x_{t+1})^{1-\theta}\right]\right).
\end{aligned}$$

Taking total expectation on both sides of the above inequality and summing over times gives

$$\begin{aligned}\sum_{s=t}^{\infty}\mathbb{E}\left[\|x_{s+1}-x_s\|\right] &\leq \frac{n}{(1-\theta)k\left(1-\frac{Ln\eta}{2k}\right)}\mathbb{E}\left[f(x_t)^{1-\theta}\right]+C\sum_{s=t}^{\infty}\delta_s^{1/2}\\ &= \frac{n}{(1-\theta)k\left(1-\frac{Ln\eta}{2k}\right)}\mathbb{E}\left[f(x_t)^{1-\theta}\right]+C\delta_t^{1/2}\end{aligned}$$

The above implies that

$$\mathbb{E}\left[\|x_t-x_\infty\|\right] \leq \frac{n}{(1-\theta)k\left(1-\frac{Ln\eta}{2k}\right)}\mathbb{E}\left[f(x_t)^{1-\theta}\right]+C\delta_t^{1/2}.$$

In Theorem 4, we have proved convergence rate for $\{f(x_t)\}_t$ and $\{\mathbb{E}[f(x_t)]\}_t$. Combined with the above inequality, the convergence rate in Theorem 4 implies convergence rate of $\mathbb{E}[\|x_t-x_\infty\|]$. Once the convergence rate of $\mathbb{E}[\|x_t-x_\infty\|]$ is obtained, we can apply the probabilistic argument in the proof for Theorem 4 to obtain the almost sure convergence rate.

□

4.4 Implications on the Non-stochastic Case

The almost sure convergence rate for SZGD implies sure convergence rate of the gradient descent algorithm. An intriguing gap between SZGD and its non-stochastic counterpart happens when $\theta \in (\frac{1}{2}, 1)$. In this section, we will first display convergence rate results for the gradient descent algorithm on Łojasiewicz functions. At the end of this section, we discuss the gap between the stochastic case and the non-stochastic case in Remark 4. Note that the classic work [1] provides convergence rate for the proximal algorithm, not the gradient descent algorithm. Also [14] provides analysis for the gradient descent on the Polyak-Łojasiewicz functions, not the Łojasiewicz functions. Thus this convergence rate of $\{f(x_t)\}_t$ governed by gradient descent for smooth Łojasiewicz functions is one of our contributions, although it may not be as important as the results in previous sections.

Recall the gradient descent algorithm iterates as

$$x_{t+1} = x_t - \eta_t \nabla f(x_t). \qquad \text{(Gradient Descent (GD))}$$

Compare to GD, the SZGD algorithm does not require one to have access to first-order information of the objective. In this sense, SZGD algorithm makes weaker assumptions about the environment.

Corollary 1. *Instate Assumption 1 and Remarks 1, 2 and 3. Let x_t be a sequence generated by the gradient descent algorithm, and let x_∞ be the almost sure limit of $\{x_t\}_t$. Let $\eta_t = \eta$. Then if $\theta \in (0, \frac{1}{2}]$ and η satisfies that $-1 \leq \left(\frac{L\eta^2}{2}-\eta\right) < 0$, there exists a constant $Q > 1$ such that $\{Q^t\|x_t - x_\infty\|\}_t$ converges to 0 and $\{Q^t\left(f(x_t)-f(x_\infty)\right)\}_t$ converges to 0.*

Corollary 1 follow immediately from Theorems 4 and 6. If $\theta \in (0, \frac{1}{2}]$, Corollary 1 provides linear convergence rate of $\{\|x_t - x_\infty\|\}_t$ and $\{f(x_t)-f(x_\infty)\}_t$ with x_t governed by the GD algorithm. When $\theta \in (0, \frac{1}{2}]$, linear convergence rate for $\{\sum_{s=t}^{\infty}\|x_{s+1}-x_s\|\}_t$ can be similarly obtained.

If $\theta \in (\frac{1}{2}, 1)$, then the deterministic case (GD) converges faster than the stochastic case (SZGD), as shown in Theorem 7. To prove Theorem 7, we first need the following proposition.

Proposition 4. *Instate Assumption 1 and Remarks 1, 2 and 3. Let $\{x_t\}_{t\in\mathbb{N}}$ be the sequence governed by the gradient descent algorithm. Let learning rate $\eta_t = \eta \in \left(0, \frac{2}{L}\right)$. Let $\{x_t\}_{t\in\mathbb{N}}$ be bounded and let x_∞ be the limit of $\{x_t\}_{t\in\mathbb{N}}$ (Proposition 2). Then there exists a constant T_0 such that the following holds:*

(i) x_∞ is a critical point of f (Proposition 2) and $\|x_t - x_\infty\| \leq \mu$ for all $t \geq T_0$. $\|\nabla f(x_t)\| < 1$ for all $t \geq T_0$. (Lemma 1)

(ii) $f(x_{t+1}) \geq f(x_t)$ for all $t \geq T_0$ (Lemma 2).

(iii) $0 \le f(x_t) - f(x_\infty) \le 1$ *for all* $t \ge T_0$. *(Proposition 2, Lemma 2)*

(iv) $\mathbb{E}\left[f(x_t) - f(x_\infty)\right] \in \left(0, \left(\frac{1}{-2\theta\left(\frac{L\eta^2 n}{2k} - \eta\right)}\right)^{\frac{1}{2\theta-1}}\right)$ *for all* $t \ge T_0$. *(Proposition 2, Lemma 2)*

In addition, given T_0 *and* $\theta \in (\frac{1}{2}, 1)$, *there exists constant* C_0 *such that*

(v) $\frac{1}{2\theta-1} \le C_0^{2\theta}\left(\eta - \frac{L\eta^2}{2}\right)$.

(vi) $f(x_{T_0}) - f(x_\infty) \le C_0 T_0^{\frac{1}{1-2\theta}}$.

Proof. The first six items can be easily verified. The last two items can be obtained via similar ways as those for Proposition 3. □

With the above proposition, we are ready to prove Theorem 7.

Theorem 7. *Instate Assumption 1 and Remarks 1, 2 and 3. Let* x_t *be a sequence generated by the gradient descent algorithm, and let* x_∞ *be the almost sure limit of* $\{x_t\}_t$. *Let* $\eta_t = \eta$. *If* $\theta \in (\frac{1}{2}, 1)$ *and* $\eta \in \left(0, \frac{2}{L}\right)$, *then* $f(x_t) \le O\left(t^{\frac{1}{1-2\theta}}\right)$.

Proof. Let T_0 and C_0 be the constants so that Proposition 4 holds true.

By L-smoothness by f, it holds that

$$\begin{aligned} f(x_{t+1}) &\le f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_t - x_{t+1}\|^2 \\ &= f(x_t) + \left(\frac{L\eta^2}{2} - \eta\right)\|\nabla f(x_t)\|^2. \end{aligned}$$

Since x_t is close to a critical point for all $t \ge T_0$ (item (i) in Proposition 4), combining the above inequality and the Łojasiewicz inequality gives

$$f(x_{t+1}) \le f(x_t) + \left(\frac{L\eta^2}{2} - \eta\right) f(x_t)^{2\theta}. \tag{22}$$

For a given $t \ge T_0$, consider the function $h(x) = (t+x)^{\frac{1}{1-2\theta}}$. Applying Taylor's theorem to h gives

$$(t+1)^{\frac{1}{1-2\theta}} = t^{\frac{1}{1-2\theta}} + \frac{1}{1-2\theta}(t+z)^{\frac{2\theta}{1-2\theta}}$$

for some $z \in [0, 1]$, which implies

$$(t+1)^{\frac{1}{1-2\theta}} \ge t^{\frac{1}{1-2\theta}} + \frac{1}{1-2\theta} t^{\frac{2\theta}{1-2\theta}}. \tag{23}$$

We will use induction to finish the proof. Item (vi) in Proposition 4 states $f(x_{T_0}) \le C_0 T_0^{\frac{1}{1-2\theta}}$. Inductively, if $f(x_t) \le C_0 t^{\frac{1}{1-2\theta}}$ $(t \ge T_0)$, then (22) implies that

$$\begin{aligned} f(x_{t+1}) &\le C_0 t^{\frac{1}{1-2\theta}} + C_0^{2\theta}\left(\frac{L\eta^2}{2} - \eta\right) t^{\frac{2\theta}{1-2\theta}} \\ &\le C_0 (t+1)^{\frac{1}{1-2\theta}} + C_0^{2\theta}\left(\frac{L\eta^2}{2} - \eta\right) t^{\frac{2\theta}{1-2\theta}} + \frac{1}{2\theta-1} t^{\frac{2\theta}{1-2\theta}} \\ &\le C_0 (t+1)^{\frac{1}{1-2\theta}}, \end{aligned}$$

where the first line uses (22), the induction hypothesis, and that the function $x \mapsto x + (\frac{L\eta^2}{2} - \eta)x^{2\theta}$ is strictly increasing on $\left(0, \left(\frac{1}{-2\theta\left(\frac{L\eta^2 n}{2k} - \eta\right)}\right)^{\frac{1}{2\theta-1}}\right)$, the second line uses (23), and the last line uses item (v) in Proposition 4. □

Remark 4. *There is a small gap between the convergence rate of SZGD and GD when $\theta \in (\frac{1}{2}, 1)$. Theorem 4 states that $\{f(x_t)\}_t$ generated by SZGD converges almost surely at rate $o(t^{\frac{1}{1-2\theta}} \log t)$. Theorem 7 states that $\{f(x_t)\}_t$ generated by GD converges at rate $O(t^{\frac{1}{1-2\theta}})$. The difference between the two rates is small. The reason for this logarithmic gap is the invoking of the Markov inequality in proof of Theorem 4. While one can further reduce this gap to order $O(\log \log t)$ (or a gap of even smaller order, e.g., $\log \log \log t$), it is unlikely that such gaps can be fully closed.*

5 The Proximal Algorithm for Nonsmooth Łojasiewicz Functions

Let f be a function that is continuous but possibly nonsmooth. In such cases, we use the (subgradient) proximal algorithm to solve this optimization problem. This section serves to provide a convergence rate analysis for the proximal algorithm for nonsmooth Łojasiewicz functions with exponent $\theta \in (0, \frac{1}{2})$. Previously, [1] showed that, when the Łojasiewicz exponent $\theta \in (\frac{1}{2}, 1)$, the sequence $\{x_t\}$ governed by the proximal algorithm (25) converges at rate $\|x_t - x_\infty\| \leq O\left(t^{\frac{1-\theta}{1-2\theta}}\right)$. In this section, we show that the proximal algorithm satisfies $f(x_t) \leq O\left(t^{\frac{1}{1-2\theta}}\right)$. When $\theta \in (\frac{1}{2}, 1)$ the function value $\{f(x_t)\}_t$ tends to converge at a faster rate than the point sequence $\{\|x_t - x_\infty\|\}_t$. This phenomenon for the proximal algorithm again suggests that the convergence rate of $\{f(x_t)\}_t$ may be more important and informative than the convergence rate of $\{\|x_t - x_\infty\|\}_t$, since the trajectory of $\{x_t\}_t$ may be inevitably spiral [9].

5.1 Preliminaries for Nonsmooth Analysis and Proximal Algorithm

Before proceeding, we first review some preliminaries for nonsmooth analysis and the proximal algorithm. We begin by the concept of subdifferential and subgradient in nonsmooth analysis.

Definition 3 ([31])**.** *Consider a proper lower semicontinuous function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. The effective domain (or simply domain) of f (written $\mathrm{dom} f$) is $\mathrm{dom} f := \{x \in \mathbb{R}^n : -\infty < f(x) < +\infty\}$. For each $x \in \mathrm{dom} f$, the Fréchet subdifferential of f at x, written $\hat{\partial} f(x)$, is the set of vectors $g^* \in \mathbb{R}^n$ such that*

$$\liminf_{\substack{y \neq x \\ y \to x}} \frac{f(y) - f(x) - \langle g^*, y - x \rangle}{\|x - y\|} \geq 0.$$

If $x \notin \mathrm{dom} f$, by convention $\hat{\partial} f(x) = \emptyset$.

The limiting subdifferential of f at x, written $\partial f(x)$, is

$$\partial f(x) := \{g \in \mathbb{R}^n : \exists x_n \to x, f(x_n) \to f(x), g_n^* \in \hat{\partial} f(x_n) \to g\}.$$

A element $g \in \partial f(x)$ is call a subgradient of f at x.

Same as [1], we will make the following assumptions on f:

Assumption 2 ([1])**.** *The function f satisfies*

1. *f is continuous on $\mathrm{dom} f$;*
2. *For any $x_* \in \mathbb{R}^n$ with $\partial f \ni 0$, it holds that: there exists $\kappa, \mu > 0$, and $\theta \in [0, 1)$, such that*

$$|f(x) - f(x_*)|^\theta \leq \kappa \|g\|, \quad \forall x \in B(x_*, \mu), \forall g \in \partial f(x), \tag{24}$$

   where $B(x_*, \mu)$ is the ball of radius μ centered at x_*. Without loss of generality, we let $\kappa = 1$ to avoid clutter.

3. *$\inf_{x \in \mathbb{R}^n} f(x) > -\infty$;*
4. *Let $\{x_t\}_t$ be the sequence generated by the GD algorithm. We assume $\{x_t\}_t$ is bounded.*

In the above assumption, (24) is the Łojasiewicz inequality. Compared to the one in Definition 2, gradient is replaced by subgradient.

Next we review the elements of the proximal algorithm. The proximal algorithm is described by the following inclusion recursion:

$$x_{t+1} \in \arg\min \left\{ f(z) + \frac{1}{2\eta_t}\|z - x_t\|^2 \right\}, \tag{25}$$

with a given $x_0 \in \mathbb{R}^n$.

In each iteration, the proximal algorithm solves an optimization problem whose solution set is compact and nonempty [1]. By the optimality condition (Theorem 10.1, [31]), we know

$$0 \in \partial \left(f(x_{t+1}) + \frac{1}{2\eta_t}\|x_{t+1} - x_t\|^2 \right).$$

By the subadditivity property of subdifferential (e.g., Exercise 10.10, [31]), we have

$$0 \in \partial \left(f(x_{t+1}) + \frac{1}{2\eta_t}\|x_{t+1} - x_t\|^2 \right) \subseteq \{\frac{1}{\eta}(x_{t+1} - x_t)\} + \partial f(x_{t+1}),$$

where $+$ is the Minkowski sum when two summands are sets. Thus it holds that

$$x_{t+1} = x_t - \eta_t g_{t+1},$$

for some $g_{t+1} \in \partial f(x_{t+1})$.

Previous results we will reply on are below in Theorem 8 and 9 [1].

Theorem 8 ([1])**.** *Let f satisfy Assumption 2 and let $\{x_t\}_{t\in\mathbb{N}}$ be generated by the proximal algorithm. If $\{x_t\}_{t\in\mathbb{N}}$ is bounded then it converges to a critical point of f.*

Theorem 9 ([1])**.** *Let f satisfy Assumption 2 and let $\{x_t\}_{t\in\mathbb{N}}$ be generated by the proximal algorithm. Let f satisfy the Łojasiewicz inequality with Łojasiewicz exponent $\theta \in (\frac{1}{2}, 1)$. If $\{x_t\}_{t\in\mathbb{N}}$ is bounded, then it holds that*

$$\|x_t - x_\infty\| \le O(t^{\frac{2\theta}{1-2\theta}}) \quad \textit{and} \quad \|x_t - x_{t+1}\| \le O(t^{\frac{2\theta}{1-2\theta}}),$$

where x_∞ is the limit of $\{x_t\}_t$.

5.2 Convergence of the Proximal Algorithm

Similar to the stochastic case, we focus on the convergence analysis for $\{f(x_t)\}_t$. We start with the following proposition.

Proposition 5. *Let $\{x_t\}_{t\in\mathbb{N}}$ be the bounded sequence generated by the proximal algorithm. Then there exists a sequence $\{w_t\}_t \subseteq [0, \infty)$ such that*

- $\lim_{t\to\infty} \frac{w_t}{\|x_t - x_{t+1}\|} = 0$;
- $f(x_t) \ge f(x_{t+1}) + g_{t+1}^\top (x_t - x_{t+1}) - w_t, \quad \forall t \in \mathbb{N}.$

Proof. By Defintion 3, we have

$$f(x_t) - f(x_{t+1}) - \langle g_{t+1}, x_t - x_{t+1}\rangle \ge -o\left(\|x_t - x_{t+1}\|\right), \quad \forall t,$$

which concludes the proof.

□

Next, we state below some numerical properties when t is large.

Proposition 6. *Fix any learning rate η. For any $\theta \in (\frac{1}{2}, 1)$, there exists constants T_0 and C_0 such that*

(i) $\frac{2}{\eta} \leq (2\theta-1)C_0^{2\theta-1}$;

(ii) *For all* $t \geq T_0$, *it holds that*

$$\|x_t - x_\infty\| \leq \mu,$$

where μ *is defined as in Definition 2.*

(iii) *For all* $t \geq T_0$, *it holds that*

$$\frac{C_0}{(2\theta-1)} t^{\frac{2\theta}{1-2\theta}} + w_t \leq \eta C_0^{2\theta}(t+1)^{\frac{2\theta}{1-2\theta}},$$

where w_t *is a sequence satisfying Proposition 5.*

(iv) $f(x_{T_0}) \leq C_0 T_0^{\frac{1}{1-2\theta}}$.

Proof. Clearly we can find a constant C_0' such that $2\eta \leq (2\theta-1)C^{2\theta-1}$ for all $C \geq C_0'$, and this C_0' does not depend on T_0. Thus item (i) can be easily satisfied. By Theorem 8, we can find T_0 so that item (ii) is true. By Theorem 9, we have $\|x_t - x_{t+1}\| \leq O\left(t^{\frac{2\theta}{1-2\theta}}\right)$ and thus

$$\lim_{t\to\infty} w_t t^{\frac{2\theta}{2\theta-1}} = 0. \tag{26}$$

By item (i) and (26), for any $C \geq C_0'$, it holds that

$$\lim_{t\to\infty} \frac{\frac{C}{2\theta-1} t^{\frac{2\theta}{1-2\theta}} + w_t}{\eta C^{2\theta}(t+1)^{\frac{2\theta}{1-2\theta}}} = \frac{\frac{C}{\eta(2\theta-1)}}{C^{2\theta}} \leq \frac{1}{2}.$$

Thus we can find T_0 and C_0' that satisfies item (iii). For item (iv), given a T_0, we can find C_0'' such that $f(x_{T_0}) \leq C_0'' T_0^{\frac{1}{1-2\theta}}$, since the sequence $\{f(x_t)\}_{t\in\mathbb{N}}$ is absolutely bounded (Theorem 8).

Letting $C_0 = \max\{C_0', C_0''\}$ concludes the proof. □

Theorem 10. *Instate Assumption 2 and Remarks 1, 2 and 3. Let* $\{x_t\}_{t\in\mathbb{N}}$ *generated by the proximal algorithm. If* $\{x_t\}_{t\in\mathbb{N}}$ *is bounded then it converges to a critical point of* f. *Let* x_∞ *be the limit of* $\{x_t\}_{t\in\mathbb{N}}$. *If* $\theta \in (\frac{1}{2}, 1)$, *then it holds that*

$$f(x_t) \leq O\left(t^{\frac{1}{1-2\theta}}\right).$$

Proof. Let $\{w_t\}_t$ be a sequence satisfying Proposition 5. Thus we have

$$\begin{aligned} f(x_t) &\geq f(x_{t+1}) + g_{t+1}^\top (x_t - x_{t+1}) - w_t \\ &= f(x_{t+1}) + \eta\|g_{t+1}\|^2 - w_t. \end{aligned}$$

Since $\{x_t\}$ converges (Theorem 8) and the Łojasiewicz inequality holds with exponent θ, there exists t_0, such that for all $t \geq t_0$,

$$f(x_t) \geq f(x_{t+1}) + \eta\|g_{t+1}\|^2 - w_t \geq f(x_{t+1}) + \eta\left(f(x_{t+1}) - f(x_\infty)\right)^{2\theta} - w_t,$$

which gives

$$f(x_t) - f(x_\infty) \geq f(x_{t+1}) - f(x_\infty) + \eta\left(f(x_{t+1}) - f(x_\infty)\right)^{2\theta} - w_t.$$

To avoid notational clutter, we assume $f(x_\infty) = 0$. Thus all $f(x_t)$ are nonnegative since the proximal algorithm always decreases function value.

For any positive integer t, define

$$h(s) = (t+s)^{\frac{1}{1-2\theta}}.$$

By mean value theorem, one has $h(1) = h(0) + h'(z)$ for some $z \in [0, 1]$. Thus it holds that

$$(t+1)^{\frac{1}{1-2\theta}} = t^{\frac{1}{1-2\theta}} + \frac{1}{1-2\theta}(t+z)^{\frac{2\theta}{1-2\theta}} \geq t^{\frac{1}{1-2\theta}} + \frac{1}{1-2\theta}t^{\frac{2\theta}{1-2\theta}}. \tag{27}$$

Next we use induction to prove the convergence rate. By item (iv) in Proposition 6, we can find C_0 and T_0 such that $f(x_{T_0}) \leq C_0 T_0^{\frac{1}{1-2\theta}}$. Inductively, if $f(x_t) \leq C_0 t^{\frac{1}{1-2\theta}}$ $(t \geq T_0)$, then it holds that

$$\begin{aligned} f(x_{t+1}) + \eta f(x_{t+1})^{2\theta} &\leq f(x_t) + w_t \\ &\leq C_0 t^{\frac{1}{1-2\theta}} + w_t \\ &\leq C_0 (t+1)^{\frac{1}{1-2\theta}} + C_0 \frac{1}{2\theta - 1} t^{\frac{2\theta}{1-2\theta}} + w_t \\ &\leq C_0 (t+1)^{\frac{1}{1-2\theta}} + \eta C_0^{2\theta} (t+1)^{\frac{2\theta}{1-2\theta}}, \end{aligned} \tag{28}$$

where the second last inequality uses (27) and the last inequality uses (iii) in Proposition 6.

Since the function $x \mapsto x + \eta x^{2\theta}$ is monotonic and strictly increasing on $(0, \infty)$, the above inequality (28) implies

$$f(x_{t+1}) \leq C_0 (t+1)^{\frac{1}{1-2\theta}},$$

which concludes the proof. □

6 Empirical Studies

In this section, we empirically study the performance of the SZGD algorithm. All experiments are carried out on the following test functions F_1 and F_2 defined over $\mathbb{R}^{30}$

$$F_1(x) = \left(x^\top Q x\right)^{3/4} \quad \text{and} \quad F_2(x) = \left(x^\top Q x\right)^{1/4},$$

where $Q \in \mathbb{R}^{30 \times 30}$ is a PSD matrix with eigenvalues following exponential distribution with *p.d.f.* $f_{exp}(x) = \frac{1}{5}\exp\left(-\frac{x}{5}\right)\mathbb{I}_{[x \geq 0]}$ and eigenvectors independently sampled from the unit sphere. The results are summarized in Figures 2, 3 and 4. To avoid numerical instability, we set $\delta_t = \max\left(0.1 \times 2^{-t}, 0.00001\right)$ in all numerical experiments.

At first glance, Figure 3 shows seemingly implausible results. However, these results are natural after we investigate the landscape of the test function F_1. While fully visualizing F_1 is impossible, we can consider a 1-d version of F_1: $f_1(x) = \sqrt{|x|}$, whose graph is shown below in Figure 1. This nonconvex function has the following properties, all highly aligned with empirical observations.

- When x is far from the origin, the function's surface is flat. In this case, randomness in the gradient, which may increase the magnitude of gradient, can speedup the convergence. In Figure 3, we observe that SZGD algorithms converges faster than GD at the beginning, which agrees with the landscape of F_1.
- At zero, the function is not differentiable, and thus not L-smooth. This means that the gradient estimator may not be accurate near zero. In Figure 3(a), we observe that GD converges to a more optimal value at the end, which agrees with the non-differentiability of F_1 at zero.
- Near zero, the gradient of the function is not continuous. Therefore, the trajectory of the GD algorithm oscillates near zero. For the 1-d example in Figure 1, the trajectory of GD will jump back and forth around zero, but hits zero with little chance. In Figure 3(a), we observe that the $\|x\|$ value of DG is fuzzy as it goes to zero, which agrees with the shape of F_1 near zero.

Other observations from Figures 2 and 3 are summarized below:

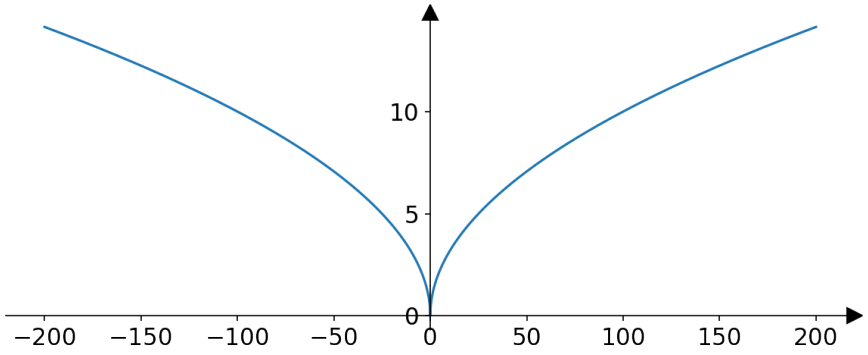


Figure 1: The plot of function $f(x) = \sqrt{|x|}$.

- Figure 2 shows that, on test function F_1, with fixed step size η, SZGD with larger values of k in general converges faster given a fixed number of iterations. However, the gap between different choices of k is not significant.
- In general, $\{f(x_t)\}_t$ converges faster than $\{\|x_t - x_\infty\|\}_t$, which agrees with our theoretical results.

We also empirically study the convergence rates versus the number of function evaluations. Note that one iteration may require more than one function evaluations: k random orthogonal directions are sampled and $2k$ function evaluations are needed to obtain the gradient estimator defined in (3). The convergence results versus number of function evaluations are shown in Figure 4. Some observations from Figure 4 include

- Given the same η and the same number of function evaluations, the sequence $\{f(x_t)\}_t$ converges faster when k is smaller.
- Given the same η and the same number of function evaluations, the sequence $\{\|x_t - x_\infty\|\}_t$ converges fastest when $k = 10$ or $k = 1$ for both F_1 and F_2. This is an intriguing observation, and suggests that there might be some fundamental relation between number of function evaluations needed and convergence of $\{\|x_t - x_\infty\|\}_t$.
- When measured against number of function evaluations, $\{f(x_t)\}_t$ converges faster than $\{\|x_t - x_\infty\|\}_t$. This is similar to the observations of Figures 2 and 3.

7 Proof of Theorem 3

This section is devoted to proving Theorem 3. The proof mimics that for Theorem 2 [11]. To start with, we need the following facts in Propositions 7 and 8.

Proposition 7 ([7]). *Let $V := [v_1, v_2, \cdots, v_k] \in \mathbb{R}^{n\times k}$ be uniformly sampled from the Stiefel's manifold $St(n, k)$. Then the marginal distribution for any v_i is uniform over the unit sphere $\mathbb{S}^{n-1}$.*

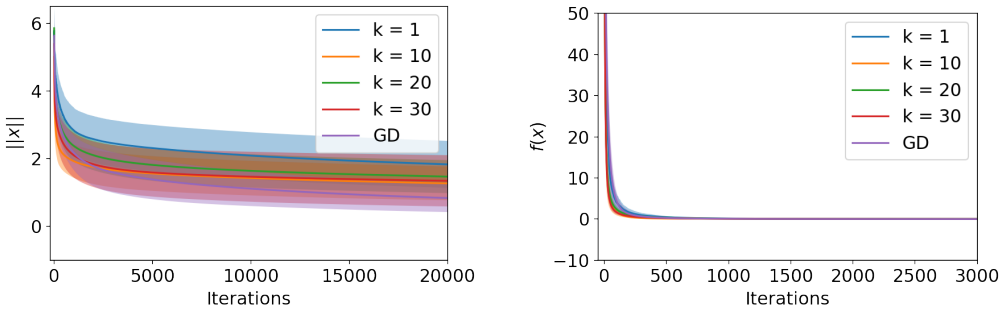
In particular, the uniform measure over the Stiefel's manifold $\mathrm{St}(n, k)$ (the Hausdorff measure with respect to the Frobenius inner product of proper dimension, which is rotation-invariant) can be decomposed into a wedge product of the spherical measure over $\mathbb{S}^{n-1}$ and the uniform measure over $\mathrm{St}(n-1, k-1)$ [7].

Another useful computational fact is the following proposition.

Proposition 8. *Let v be a vector uniformly randomly sampled from the unit sphere $\mathbb{S}^{n-1}$. Then it holds that*

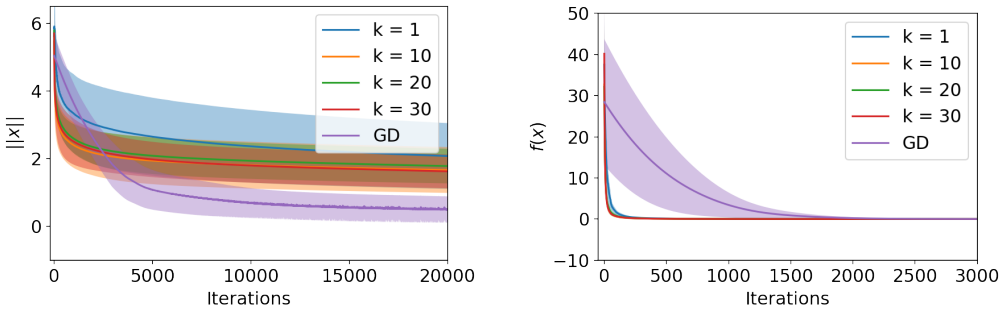
$$\mathbb{E}\left[vv^\top\right] = \frac{1}{n} I.$$

where I is the identity matrix (of size $n \times n$).



(a) (b)

Figure 2: Results of SZGD and GD on test function F_1. The lines labeled with $k = 1$ (resp. $k = 10$, etc.) show results of SZGD with $k = 1$ (resp. $k = 10, etc.$). The line labeled GD plots results of the gradient descent algorithm. Subfigure (a) shows observed convergence rate results for $\{\|x_t - x_\infty\|\}_t$ ($x_\infty = 0$); Subfigure (b) shows observed convergence rate results for $\{\|F_1(x_t) - F_1(x_\infty)\|\}_t$ ($F_1(x_\infty) = 0$). For all experiments, the step size η is set to 0.005. The solid lines show average results over 10 runs. The shaded areas below and above the solid lines indicate 1 standard deviation around the average. Since the starting point x_0 is randomly sampled, the trajectory of GD is also random.



(a) (b)

Figure 3: Results of SZGD and GD on test function F_2. The lines labeled with $k = 1$ (resp. $k = 10$, etc.) show results of SZGD with $k = 1$ (resp. $k = 10, etc.$). The line labeled GD plots results of the gradient descent algorithm. Subfigure (a) shows observed convergence rate results for $\{\|x_t - x_\infty\|\}_t$ ($x_\infty = 0$); Subfigure (b) shows observed convergence rate results for $\{\|F_2(x_t) - F_2(x_\infty)\|\}_t$ ($F_2(x_\infty) = 0$). For all experiments, the step size η is set to 0.005. The solid lines show average results over 10 runs. The shaded areas below and above the solid lines indicate 1 standard deviation around the average. Since the starting point x_0 is randomly sampled, the trajectory of GD is also random.

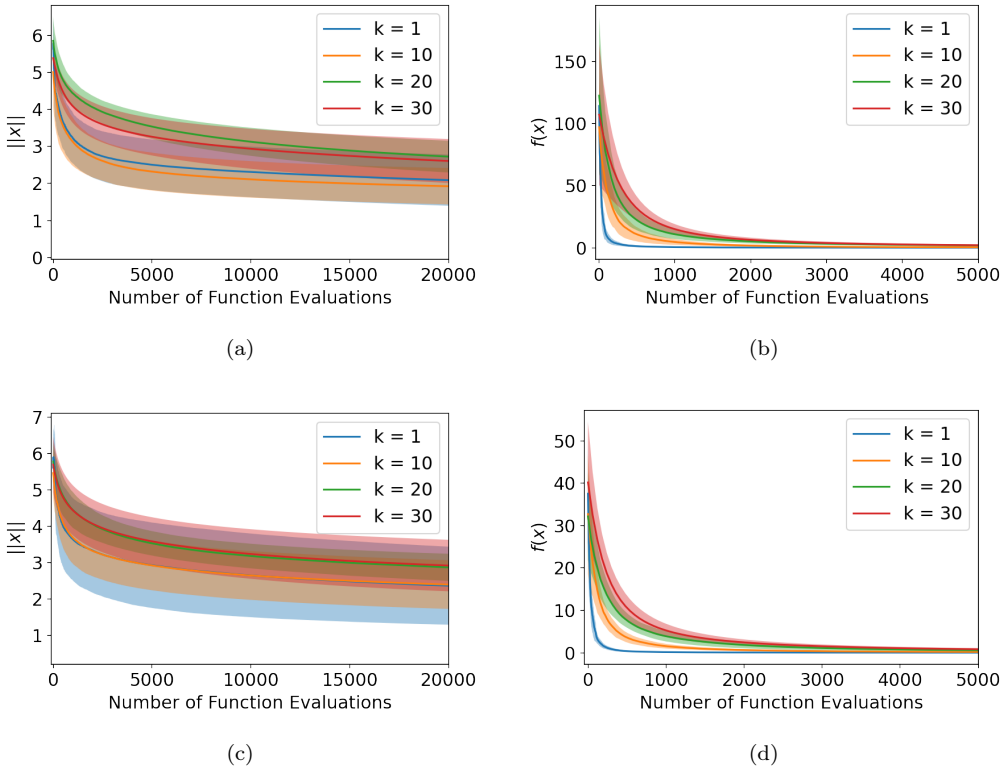


Figure 4: Results of SZGD and GD on test functions F_1 (subfigures (a) and (b)) and F_2 (subfigures (c) and (d)). The solid lines show average results over 10 runs. The lines labeled with $k = 1$ (resp. $k = 10$, etc.) show results of SZGD with $k = 1$ (resp. $k = 10, etc.$). Subfigures (a) and (c) show observed convergence rate results for $\{\|x_t - x_\infty\|\}_t$ ($x_\infty = 0$); Subfigures (b) and (d) show observed convergence rate results for $\{\|F_2(x_t) - F_2(x_\infty)\|\}_t$ ($F_2(x_\infty) = 0$). The step size η is set to 0.005. The shaded areas below and above the solid lines indicate 1 standard deviation around the average. Unlike Figures 2 and 3, this figure plots errors (either $\|x_t - x_\infty\|$ or $f(x_t) - f(x_\infty)$) against number of function evaluations. For example, when $k = 10$, one iteration requires $10 \times 2 = 20$ functions evaluations.

Proof. Let $v[i]$ be the i-th entry of v. For any $a \in [-1,1]$ and $i \neq j$, it holds that $\mathbb{E}[v[i]v[j]] = \mathbb{E}[v[i]|v[j]=a] = 0$. Thus $\mathbb{E}[v[i]v[j]] = 0$ for $i \neq j$. Also, it holds that

$$1 = \mathbb{E}\left[\|v\|^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[v[i]^2\right],$$

which concludes the proof since $\mathbb{E}\left[v[i]^2\right] = \mathbb{E}\left[v[j]^2\right]$ for any $i, j = 1, 2, \cdots, n$ (by symmetry). □

Proof of Theorem 3. Since $f(x)$ is L-smooth ($\nabla f(x)$ is L-Lipschitz), $\nabla^2 f(x)$ (the weak total derivative of $\nabla f(x)$) is integrable. Let $v \in \mathbb{R}^n$ be an arbitrary unit vector. When restricted to any line along direction $v \in \mathbb{R}^n$, it holds that $v^\top \nabla^2 f(x) v$ (the weak derivative of $v^\top \nabla(x)$ along direction v) has bounded L_∞-norm. This is due to the fact that Lipschitz functions on any closed inteval $[a,b]$ forms the Sobolev space $W^{1,\infty}[a,b]$.

In other words, for L-smooth functions, we have $\|\nabla^2 f(x)\| \leq L$ for all x, where $\nabla^2 f(x)$ is the weak total derivative of $\nabla f(x)$.

Next we look at the variance bound for the estimator. Without loss of generality, we let $x = 0$. Bounds for other values of x can be similarly obtained.

Taylor's expansion of f with integral form gives

$$f(\delta v_i) = f(0) + \delta v_i^\top \nabla f(0) + \int_0^\delta (\delta - t) v_i^\top \nabla^2 f(t v_i) v_i \, dt$$

Thus for any $v_i \in \mathbb{S}^{n-1}$ and small δ,

$$\begin{aligned}
&\frac{1}{2}\big(f(\delta v_i) - f(-\delta v_i)\big) \\
&= \delta v_i^\top \nabla f(0) + \int_0^\delta (\delta - t) v_i^\top \nabla^2 f(t v_i) v_i \, dt \\
&\quad - \int_0^{-\delta} (-\delta - t) v_i^\top \nabla^2 f(t v_i) v_i \, dt.
\end{aligned}$$

For simplicity, let $R_i = \int_0^\delta (\delta - t) v_i^\top \nabla^2 f(t v_i) v_i \, dt - \int_0^{-\delta} (-\delta - t) v_i^\top \nabla^2 f(t v_i) v_i \, dt$, and Cauchy-Schwarz inequality gives

$$\begin{aligned}
|R_i| \leq & \left(\int_0^\delta (\delta - t)^2 \, dt\right)^{1/2} \left(\int_0^\delta \left(v_i^\top \nabla^2 f(t v_i) v_i\right)^2 \, dt\right)^{1/2} \\
& - \left(\int_{-\delta}^0 (-\delta - t)^2 \, dt\right)^{1/2} \left(\int_{-\delta}^0 \left(v_i^\top \nabla^2 f(t v_i) v_i\right)^2 \, dt\right)^{1/2} \\
\leq & \frac{2L}{\sqrt{3}} \delta^2
\end{aligned}$$

for all $i = 1, 2, \cdots, k$.

For any i, k, n, it holds that

$$\begin{aligned}
& \mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right\|^2\right] - \left\|\mathbb{E}\left[\frac{\sqrt{k}}{2}\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right]\right\|^2 \\
= & \, \mathbb{E}\left[\left\|\left(\delta v_i^\top \nabla f(0) + R_i\right) v_i\right\|^2\right] - k \left\|\mathbb{E}\left[\left(\delta v_i^\top \nabla f(0) + R_i\right) v_i\right]\right\|^2 \\
\overset{①}{=} & \, \mathbb{E}\left[\delta^2 \nabla f(0)^\top v_i v_i^\top \nabla f(0) + 2\delta \nabla f(0)^\top v_i R_i + R_i^2\right] \\
& - k \left\|\mathbb{E}\left[\delta v_i v_i^\top \nabla f(0) + R_i v_i\right]\right\|^2.
\end{aligned}$$

Since $\mathbb{E}\left[v_i v_i^\top\right] = \frac{1}{n} I$ (Propositions 7 and 8), ① gives

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right\|^2\right] - \left\|\mathbb{E}\left[\frac{\sqrt{k}}{2}\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right]\right\|^2 \\
&= \frac{\delta^2}{n}\|\nabla f(0)\|^2 + 2\delta \nabla f(0)^\top \mathbb{E}\left[R_i v_i\right] + \mathbb{E}\left[R_i^2\right] - k\left\|\mathbb{E}\left[\frac{\delta}{n}\nabla f(0) + R_i v_i\right]\right\|^2 \\
&= \left(\frac{\delta^2}{n} - \frac{\delta^2 k}{n^2}\right)\|\nabla f(0)\|^2 + \left(2\delta - \frac{2\delta k}{n}\right)\nabla f(0)^\top \mathbb{E}\left[R_i v_i\right] \\
&\quad + \left(\mathbb{E}\left[R_i^2\right] - k\mathbb{E}\left[R_i v_i\right]^\top \mathbb{E}\left[R_i v_i\right]\right) \\
&\overset{②}{\leq} \left(\frac{\delta^2}{n} - \frac{\delta^2 k}{n^2}\right)\|\nabla f(0)\|^2 + \frac{4L\delta^3}{\sqrt{3}}\left(1 - \frac{k}{n}\right)\|\nabla f(0)\| + \frac{4L^2\delta^4}{3}.
\end{aligned}
$$

For the variance of the gradient estimator, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla} f_k^\delta(0) - \mathbb{E}\left[\widehat{\nabla} f_k^\delta(0)\right]\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\widehat{\nabla} f_k^\delta(0)\right\|^2\right] - \left\|\mathbb{E}\left[\widehat{\nabla} f_k^\delta(0)\right]\right\|^2 \\
&= \mathbb{E}\left[\left\|\frac{n}{2\delta k}\sum_{i=1}^k \left(f(\delta v_i) - f(\delta v_i)\right) v_i\right\|^2\right] - \left\|\frac{n}{2\delta k}\sum_{i=1}^k \mathbb{E}\left[\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right]\right\|^2 \\
&\overset{③}{=} \frac{n^2}{\delta^2 k^2}\sum_{i=1}^k \mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i) - f(\delta v_i)\right) v_i\right\|^2\right] \\
&\quad - \frac{n^2}{4\delta^2 k^2}\sum_{i,j=1}^k \mathbb{E}\left[\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right]^\top \mathbb{E}\left[\left(f(\delta v_j) - f(-\delta v_j)\right) v_j\right],
\end{aligned}
$$

where the last equation follows from the orthonormality of $\{v_1, v_2 \cdots, v_k\}$. By Proposition 7, we know that $\mathbb{E}\left[\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right] = \mathbb{E}\left[\left(f(\delta v_j) - f(-\delta v_j)\right) v_j\right]$ for all $i, j = 1, 2, \cdots, k$. Thus ③ gives

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla} f_k^\delta(0) - \mathbb{E}\left[\widehat{\nabla} f_k^\delta(0)\right]\right\|^2\right] \\
&\overset{④}{=} \frac{n^2}{\delta^2 k^2}\sum_{i=1}^k \left(\mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right\|^2\right] - \left\|\mathbb{E}\left[\frac{\sqrt{k}}{2}\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right]\right\|^2\right).
\end{aligned}
$$

Combining ② and ④ gives

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla} f_k^\delta(0) - \mathbb{E}\left[\widehat{\nabla} f_k^\delta(0)\right]\right\|^2\right] \\
&= \frac{n^2}{\delta^2 k^2}\sum_{i=1}^k \left(\mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i) - f(\delta v_i)\right) v_i\right\|^2\right] - \left\|\mathbb{E}\left[\frac{\sqrt{k}}{2}\left(f(\delta v_i) - f(-\delta v_i)\right) v_i\right]\right\|^2\right) \\
&\leq \left(\frac{n}{k} - 1\right)\|\nabla f(0)\|^2 + \frac{4L\delta}{\sqrt{3}}\left(\frac{n^2}{k} - n\right)\|\nabla f(0)\| + \frac{4L^2 n^2 \delta^2}{3k}.
\end{aligned}
$$

□

8 Conclusion

This paper studies the SZGD algorithm and its performance on Łojasiewicz functions. In particular, we establish almost sure convergence rates for SZGD algorithms on Łojasiewicz functions. Our

results show that access to noiseless zeroth-order oracle is sufficient for optimizing Łojasiewicz functions. We show that, almost surely, SZGD exhibits convergence behavior similar to its non-stochastic counterpart. Our results suggests $\{f(x_t)\}_t$ tend to converge faster than $\{\|x_t - x_\infty\|\}_t$. We also observe some intriguing facts in the empirical studies. In particular, there might be an optimal choice of k for SZGD to achieve a good convergence rate for $\{\|x_t - x_\infty\|\}_t$.

Acknowledgement

Tianyu Wang thanks Kai Du and Bin Gao for helpful discussions, and thanks Bin Gao for pointers to some important related works.

References

[1] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, 2009.

[2] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42, 2021.

[3] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

[4] Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[5] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] Yasuko Chikuse. *Statistics on Special Manifolds*. Springer New York, NY, 2003.

[8] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.

[9] Aris Daniilidis, Mounir Haddou, and Olivier Ley. A convex function satisfying the ł ojasiewicz inequality but failing the gradient conjecture both at zero and infinity. *Bulletin of the London Mathematical Society*, 54(2):590–608, 2022.

[10] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[11] Yasong Feng and Tianyu Wang. Stochastic Zeroth Order Gradient and Hessian Estimators: Variance Reduction and Refined Bias Bounds. *arXiv preprint arXiv:2205.14737*, 2022.

[12] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

[13] David E Goldberg and John Henry Holland. Genetic algorithms and machine learning. 1988.

[14] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak–ł ojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

[15] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[16] David Kozak, Cesare Molinari, Lorenzo Rosasco, Luis Tenorio, and Silvia Villa. Zeroth-order optimization with orthogonal random directions. *Mathematical Programming*, pages 1–41, 2022.

[17] K. Kurdyka, S. Ł ojasiewicz, and M.A Zurro. Stratifications distinguées comme outil en géométrie semi-analytique. *Manuscripta Math*, 86:81–102, 1995.

[18] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48:769–783, 1998.

[19] Krzysztof Kurdyka, Tadeusz Mostowski, and Adam Parusinski. Proof of the gradient conjecture of r. thom. *Annals of Mathematics*, pages 763–792, 2000.

[20] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[21] S. Ł ojasiewicz. A topological property of real analytic subsets (in french). *Coll. du CNRS, Les équations aux dérivées partielles*, pages 87–89, 1963.

[22] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[23] Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

[24] Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Stochastic zeroth-order riemannian derivative estimation and optimization. *Mathematics of Operations Research*, 2022.

[25] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[26] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[27] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[28] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[29] S. T. Polyak. Gradient methods for minimizing functionals (in russian). *Zh. Vychisl. Mat. Mat. Fiz.*, pages 643–653, 1963.

[30] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[31] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[32] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[33] Tianyu Wang. On Sharp Stochastic Zeroth Order Hessian Estimators over Riemannian Manifolds. *arXiv preprint arXiv:2201.10780*, 2022.

[34] Tianyu Wang, Yifeng Huang, and Didiong Li. From the Greene–Wu Convolution to Gradient Estimation over Riemannian Manifolds. *arXiv preprint arXiv:2108.07406*, 2021.

[35] Xiao Wang, Shiqian Ma, Donald Goldfarb, and Wei Liu. Stochastic quasi-newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017.