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

Tianyu Wang* and Yasong Feng

Abstract

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

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \widehat{\nabla} f(\mathbf{x}_t), \qquad t = 0, 1, 2, 3, \cdots ,$$

where f is the objective function that satisfies the Łojasiewicz inequality with Łojasiewicz exponent θ, η_t is the step size (learning rate), and $\widehat{\nabla} f(\mathbf{x}_t)$ is the approximate gradient estimated using zeroth-order information only. We show that, for smooth Łojasiewicz functions, the sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ generated by SZGD converges to a point $\mathbf{x}_\infty$ almost surely, and $\mathbf{x}_\infty$ is a critical point of f. If $\theta \in (0, \frac{1}{2}]$, $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_{t\in\mathbb{N}}$, $\{\sum_{s=t}^\infty \|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\}_{t\in\mathbb{N}}$ and $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_{t\in\mathbb{N}}$ ($\|\cdot\|$ is the Euclidean norm) converge to zero linearly in expectation. If $\theta \in (\frac{1}{2}, 1)$, then $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_{t\in\mathbb{N}}$ (and $\{\sum_{s=t}^\infty \|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\}_{t\in\mathbb{N}}$) converges to zero at rate $O\left(t^{\frac{1}{1-2\theta}}\right)$ in expectation; $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_{t\in\mathbb{N}}$ converges to zero at rate $O\left(t^{\frac{1-\theta}{1-2\theta}}\right)$ in expectation. Our results show that $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_{t\in\mathbb{N}}$ can converge faster than $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_{t\in\mathbb{N}}$.

Also, we show that for Łojasiewicz functions with Łojasiewicz exponent $\theta \in (\frac{1}{2}, 1)$, the sequence $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_{t\in\mathbb{N}}$ generated by the proximal algorithm converges at rate $O\left(t^{\frac{1}{1-2\theta}}\right)$. This rate is faster than the convergence rate of $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_{t\in\mathbb{N}}$ previously obtained by Attouch and Bolte. Our results for the proximal algorithm reiterate that $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_{t\in\mathbb{N}}$ can converge faster than $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_{t\in\mathbb{N}}$, regardless of whether the objective f is smooth or nonsmooth.

1 Introduction

Zeroth order optimization is a central topic in optimization 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).

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 classic work, [1] proved the state-of-the-art convergence rate for $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$, where $\{\mathbf{x}_t\}_t$ is the sequence generated by the proximal algorithm, $\mathbf{x}_\infty$ is the limit of $\{\mathbf{x}_t\}_t$ that is also a critical point of the objective f, and $\|\cdot\|$ denotes the Euclidean norm.

*Correspondence to : wangtianyu@fudan.edu.cn

Convergence Rates of SZGD

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

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \widehat{\nabla} f_{\mathbf{V}_t}^{\delta_t}(\mathbf{x}_t), \qquad t = 0, 1, 2, \cdots, \tag{1}$$

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

$$\widehat{\nabla} f_{\mathbf{V}_t}^{\delta_t}(\mathbf{x}) := \frac{n}{2\delta_t k} \sum_{i=1}^{k} \left(f(\mathbf{x} + \delta_t \mathbf{v}_{t,i}) - f(\mathbf{x} - \delta_t \mathbf{v}_{t,i}) \right) \mathbf{v}_{t,i}, \quad \forall \mathbf{x} \in \mathbb{R}^n, \tag{2}$$

where $\mathbf{V}_t = [\mathbf{v}_{t,1}, \mathbf{v}_{t,2}, \cdots, \mathbf{v}_{t,k}]$ is uniformly sampled from the Stiefel manifold $\mathrm{St}(n,k) := \{\mathbf{X} \in \mathbb{R}^{n\times k} : \mathbf{X}^\top \mathbf{X} = \mathbf{I}_k\}$, and $\delta_t := 2^{-t}$ is the finite difference granularity. Previously, the statistical properties of (2) have been investigated by [11] (See Section 3). Throughout, we use Stochastic Zeroth-order Gradient Descent (SZGD) to refer to the above update rule (1) with the estimator (2).

In this paper, we prove the following results for SZGD. Let the objective function f satisfy the Łojasiewicz inequality with exponent θ (Definition 2). Let $\{\mathbf{x}_t\}_t$ be the sequence generated by SZGD. Then under Assumption 1,

- The sequence $\{\mathbf{x}_t\}_t$ converges to a limit $\mathbf{x}_\infty$ almost surely. In addition, $\mathbf{x}_\infty$ is a critical point of f.
- If $\theta \in (0, \frac{1}{2}]$, then there exists $Q > 1$ such that $\{Q^t(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))\}_t$ converges to zero in expectation. In other words, if $\theta \in (0, \frac{1}{2}]$, $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_t$ converges to zero linearly in expectation.
- If $\theta \in (\frac{1}{2}, 1)$, $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_t$ converges to zero at rate $O\left(t^{\frac{1}{1-2\theta}}\right)$ in expectation. This rate is faster than the convergence rate of $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ previously obtained in [1], where $\{\mathbf{x}_t\}_t$ is generated by the proximal algorithm.

  Also, we prove the following convergence rate for $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ and $\{\sum_{s=t}^{\infty} \|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\}_t$.
- If $\theta \in (0, \frac{1}{2}]$, then there exists $Q > 1$ such that $\{Q^t \sum_{s=t}^{\infty} \|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\}_t$ and $\{Q^t\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ converges to zero in expectation. In other words, if $\theta \in (0, \frac{1}{2}]$, $\{\sum_{s=t}^{\infty} \|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\}_t$ and $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ converges to zero linearly in expectation.
- If $\theta \in (\frac{1}{2}, 1)$, then $\{\sum_{s=t}^{\infty} \|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\}_t$ converges to zero at rate $O\left(t^{\frac{1}{1-2\theta}}\right)$ in expectation, and $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ converges to zero at rate $O\left(t^{\frac{1-\theta}{2\theta-1}}\right)$ in expectation.

For $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$, the convergence rate of SZGD matches the convergence rate of the proximal algorithm [1], for any $\theta \in (\frac{1}{2}, 1)$. This means one does not need a proximal oracle to obtain the state-of-the-art of convergence rate for $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$. More importantly, the above results imply that for any fixed $\theta \in (\frac{1}{2}, 1)$, $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_t$ can converge much faster than $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$. Also, we show that the Gradient Descent (GD) algorithm converges at the same rate as SZGD.

Convergence Rates of the Proximal Algorithm

We also provide convergence rates of $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_t$ for the proximal algorithm. Let $\{\mathbf{x}_t\}_t$ be the sequence generated by the proximal algorithm. In their classic work [1], Attouch and Bolte showed that, when $\theta \in (\frac{1}{2}, 1)$, $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ converges to zero at rate $O\left(t^{\frac{1-\theta}{1-2\theta}}\right)$. In this paper, we also prove the following result.

- Let $\{\mathbf{x}_t\}_t$ be generated by the proximal algorithm. If $\theta \in (\frac{1}{2}, 1)$, then $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_t$ converges to zero at rate $O\left(t^{\frac{1}{1-2\theta}}\right)$. Note that this rate can be much faster than the convergence rate of $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ previously obtained in [1], which is of order $O\left(t^{\frac{1-\theta}{1-2\theta}}\right)$.

Remark 1 (Summary of contributions)**.** *We prove convergence rates of SZGD for Łojasiewicz Functions. Our results generalize the recent work that studies SZGD for Polyak–Łojasiewicz functions [16].*

More importantly, our results show that, when the Łojasiewicz exponent $\theta \in (\frac{1}{2}, 1)$*,* $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_t$ *can converge much faster than* $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$*. This observation is true for both SZGD and the proximal algorithm. In addition, this observation is true regardless of whether the objective function* f *is smooth (*f *is* L*-smooth, see Definition 1) or nonsmooth (*f *is continuous but not necessarily differentiable).*

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 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.

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 work by [16]. In [16], in expectation convergence rates are proved for Polyak–Łojasiewicz functions. Compared to [16]'s study of Polyak–Łojasiewicz functions, our results are more general since we Łojasiewicz functions are more general than Polyak–Łojasiewicz functions.

3 Preliminaries

3.1 Gradient Estimation

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

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

where $[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k] = \mathbf{V}$ is uniformly sampled from the Stiefel manifold $\mathrm{St}(n,k) = \{\mathbf{X} \in \mathbb{R}^{n \times k} : \mathbf{X}^\top \mathbf{X} = \mathbf{I}_k\}$, and δ is the finite difference granularity. In practice, one can firstly generate a random matrix $\mathbf{U} \in \mathbb{R}^{n \times k}$ of *i.i.d.* standard Gaussian ensemble. Then apply the Gram–Schmit process on $\mathbf{U}$ to obtain the matrix $\mathbf{V}$.

Definition 1. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called L-smooth if it is continuously differentiable, and $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \le L \|\mathbf{x} - \mathbf{x}'\|$, for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$.*

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])**.** *If f is L-smooth, then the gradient estimator $\widehat{\nabla} f_{\mathbf{V}}^{\delta}$ satisfies $\left\| \mathbb{E}\left[\widehat{\nabla} f_{\mathbf{V}}^{\delta}(\mathbf{x}) \right] - \nabla f(\mathbf{x}) \right\| \le \frac{Ln\delta}{n+1}$ for all $\mathbf{x} \in \mathbb{R}^n$.*

Theorem 2. *If f is L-smooth, then variance of the gradient estimator for f (Eq. 3) satisfies*

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

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

The proof of Theorem 2 can be found in Section 7.

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 $\mathbf{x}^*$ with $\nabla f(\mathbf{x}^*) = 0$, there exist constants $\kappa, \mu > 0$ such that*

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

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 exponent 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 $f(\mathbf{x})$). 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 lower case bold letters (e.g. $\mathbf{x}_t, \mathbf{u}$) to refer to vectors, upper case bold letters (e.g., $\mathbf{P}_t$, $\mathbf{V}_t$) to refer to matrices.
- We use C to denote non-stochastic constants that is independent of t, not necessarily referring to the same value at each occurrence. Such constants C always appear in front δ_t – We will only write $C\delta_t$, and terms other than δ_t will not be multiplied to C. Since δ_t decreases to zero exponentially fast, we use such C to avoid notational clutter in front of exponentially small terms.
- For any random variable (or collection of random variables) X, we use $\mathbb{E}_X[\cdot]$ to denote the expectation with respect to X. We use $\mathcal{F}_t$ to denote the σ-algebra generated by all randomness after arriving at $\mathbf{x}_t$, but before obtaining the estimator $\widehat{\nabla} f_{\mathbf{V}_t}^{\delta_t}(\mathbf{x}_t)$, and use $\mathbb{E}_t[\cdot]$ to denote the expectation condition on $\mathcal{F}_t$. Also, we use $\mathbb{E}[\cdot]$ to denote the total expectation.

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

(iv) All critical points of f are isolated points (of $\mathbb{R}^n$).

(v) Let $\{\mathbf{x}_t\}_t$ be the sequence generated by the SZGD algorithm. We assume that there exists a critical point x^* such that $\|\mathbf{x}_t - \mathbf{x}^*\| \le \mu$ for all t.

Many items in the above assumptions are assumed in the classic work [1]. See Assumption 2 for the set of assumptions previously employed in [1].

A final remark before proceeding to the main results is that we focus on stochastic zeroth-order optimization with noiseless function evaluations. The algorithm is random, and the environment is noiseless.

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 $\mathbf{x}_0 \in \mathbb{R}^n$ and $\delta_0 = 1$. /* or $\delta_0 \in (0, 1)$. */
3: Pick step size lower and upper bounds $\eta_-, \eta_+ \in (0, \infty)$.
4: **for** $t = 0, 1, 2, \cdots$ **do**
5: Sample $\mathbf{V}_t \sim \text{Unif}(\text{St}(n, k))$, and use the random directions in $\mathbf{V}_t$ to define $\widehat{\nabla} f_k^{\delta_t}(\mathbf{x}_t)$.
6: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \widehat{\nabla} f_{\mathbf{V}_t}^{\delta_t}(\mathbf{x}_t)$. /* The learning rate η_t satisfies $\eta_t \in [\eta_-, \eta_+]$ for all t. */
7: $\delta_{t+1} = \delta_t / 2$.
8: **end for**

The main convergence rate guarantee for SZGD is in Theorem 3.

Theorem 3. *Instate Assumption 1. Let $\mathbf{x}_t$ be a sequence generated by SZGD (Algorithm 1). Pick step sizes η_t so that there exist $\eta_-, \eta_+ \in (0, \infty)$ such that $\eta_- \le \eta_t \le \eta_+$ for all t. Then $\{\mathbf{x}_t\}$ converges to a critical point $\mathbf{x}_\infty$ almost surely. In addition, it holds that* *(a) if* $\theta \in (0, \frac{1}{2}]$, $-1 \le \left(\frac{Ln\eta_-^2}{2k} - \eta_-\right) < 0$ *and* $-1 \le \left(\frac{Ln\eta_+^2}{2k} - \eta_+\right) < 0$, *then there exists a constant* $Q > 1$ *such that* $\{Q^t\left(f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right)\}_t$ *converges to* 0 *in expectation.*

(b) if $\theta \in (\frac{1}{2}, 1)$ *and* $\eta_-, \eta_+ \in \left(0, \frac{2k}{Ln}\right)$, *then it holds that,* $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_t$ *converges to zero at rate* $O\left(t^{\frac{1}{1-2\theta}}\right)$ *in expectation.*

This theorem gives convergence rate of $\{f(\mathbf{x}_t)\}_{t\in\mathbb{N}}$. Similar convergence guarantees for $\{\mathbf{x}_t\}_t$ 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 $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ generated by SZGD converges to a critical point almost surely. In Section 4.2, we establish convergence rates for $\{f(\mathbf{x}_t)\}_t$ associated with Algorithm 1. Then in Section 4.2, we establish convergence rates for $\{\mathbf{x}_t\}_t$.

4.1 Asymptotic Convergence

We will first show that $\{\mathbf{x}_t\}_t$ converges to a critical point almost surely. For simplicity, let

$$-B := \max\left\{\left(\frac{L\eta_-^2 n}{2k} - \eta_-\right), \left(\frac{L\eta_+^2 n}{2k} - \eta_+\right)\right\}. \tag{5}$$

Proposition 1. *Let* $\mathbf{P}_t := \mathbf{V}_t\mathbf{V}_t^\top$. *Then it holds that*

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - B\frac{n}{k}\nabla f(\mathbf{x}_t)^\top \mathbf{P}_t \nabla f(\mathbf{x}_t) + C\delta_t.$$

In addition, we have $\mathbb{E}_t\left[f(\mathbf{x}_{t+1})\right] \le f(\mathbf{x}_t) - B\|\nabla f(\mathbf{x}_t)\|^2 + C\delta_t$.

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

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

Taylor's expansion of f with integral form gives

$$f(\delta\mathbf{v}) = f(\mathbf{0}) + \delta\mathbf{v}^\top\nabla f(\mathbf{0}) + \int_0^\delta (\delta - t)\mathbf{v}^\top\nabla^2 f(t\mathbf{v})\mathbf{v}\, dt$$

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

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

Therefore,

$$\begin{aligned}
\widehat{\nabla} f^{\delta_t}_{\mathbf{V}_t}(\mathbf{x}_t) &= \frac{n}{2k\delta_t}\sum_{i=1}^{k}\left(f(\mathbf{x}_t+\delta_t\mathbf{v}_{t,i}) - f(-\mathbf{x}_t-\delta_t\mathbf{v}_{t,i})\right)\mathbf{v}_{t,i} \\
&= \frac{n}{k}\sum_{i=1}^{k}\mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top\nabla f(\mathbf{x}_t) + \frac{n}{k}\sum_{i=1}^{k}\mathbf{v}_{t,i}\int_0^{\delta_t}(\delta_t - t)\mathbf{v}_{t,i}^\top\nabla^2 f(\mathbf{x}_t+t\mathbf{v}_{t,i})\mathbf{v}_{t,i}\,dt \\
&\quad - \frac{n}{2k\delta_t}\sum_{i=1}^{k}\mathbf{v}_{t,i}\int_0^{-\delta_t}(-\delta_t - t)\mathbf{v}_{t,i}^\top\nabla^2 f(\mathbf{x}_t+t\mathbf{v}_{t,i})\mathbf{v}_{t,i}\,dt \\
&= \frac{n}{k}\sum_{i=1}^{k}\mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top\nabla f(\mathbf{x}_t) + \frac{n}{2k\delta_t}\sum_{i=1}^{k}\mathbf{v}_{t,i}\int_0^{\delta_t}(\delta_t - t)\mathbf{v}_{t,i}^\top\nabla^2 f(\mathbf{x}_t+t\mathbf{v}_{t,i})\mathbf{v}_{t,i}\,dt \\
&\quad - \frac{n}{2k\delta_t}\sum_{i=1}^{k}\mathbf{v}_{t,i}\int_0^{-\delta_t}(-\delta_t - t)\mathbf{v}_{t,i}^\top\nabla^2 f(\mathbf{x}_t+t\mathbf{v}_{t,i})\mathbf{v}_{t,i}\,dt,
\end{aligned}$$

which implies

$$\begin{aligned}
\nabla f(\mathbf{x}_t)^\top\widehat{\nabla} f^{\delta_t}_{\mathbf{V}_t}(\mathbf{x}_t) &= \frac{n}{k}\nabla f(\mathbf{x}_t)^\top\mathbf{P}_t\nabla f(\mathbf{x}_t) + O(Ln\|\nabla f(\mathbf{x}_t)\|\delta_t), \\
\left\|\widehat{\nabla} f^{\delta_t}_{\mathbf{V}_t}(\mathbf{x}_t)\right\|^2 &= \frac{n}{k}\nabla f(\mathbf{x}_t)^\top\mathbf{P}_t\nabla f(\mathbf{x}_t) + O(Ln\left(\|\nabla f(\mathbf{x}_t)\| + \|\nabla f(\mathbf{x}_t)\|^2\right)\delta_t).
\end{aligned}$$

Since $\{\mathbf{x}_t\}_t$ is bounded and $\|\nabla f(x_t)\|$ is continuous (Assumption 2), we know $O(Ln\|\nabla f(\mathbf{x}_t)\|\delta_t) \le C\delta_t$ and $O(Ln\left(\|\nabla f(\mathbf{x}_t)\| + \|\nabla f(\mathbf{x}_t)\|^2\right)\delta_t) \le C\delta_t$.

Thus by L-smoothness of the f, we have

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1}-\mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \\
&= f(\mathbf{x}_t) - \eta_t\nabla f(\mathbf{x}_t)^\top\widehat{\nabla} f^{\delta_t}_{\mathbf{V}_t}(\mathbf{x}_t) + \frac{L\eta_t^2}{2}\|\widehat{\nabla} f^{\delta_t}_{\mathbf{V}_t}(\mathbf{x}_t)\|^2 \\
&\le f(\mathbf{x}_t) - B\frac{n}{k}\nabla f(\mathbf{x}_t)^\top\mathbf{P}_t\nabla f(\mathbf{x}_t) + C\delta_t.
\end{aligned}$$

Since $\mathcal{F}_t$ contains $\mathbf{x}_t$ but not $\mathbf{V}_t$, it holds that

$$\mathbb{E}_t\left[f(\mathbf{x}_{t+1})\right] \le f(\mathbf{x}_t) - B\frac{n}{k}\nabla f(\mathbf{x}_t)^\top\mathbb{E}_t\left[\mathbf{P}_t\right]\nabla f(\mathbf{x}_t) + C\delta_t.$$

By Propositions 7 and 8, we know $\mathbb{E}_t\left[\mathbf{P}_t\right] = \frac{k}{n}\mathbf{I}$. Combining this fact with the above equation finishes the proof. □

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

Lemma 1. *Instate Assumption 1. Let $\{\mathbf{x}_t\}_t$ be the sequence governed by Algorithm 1. Pick step sizes η_t so that there exist $\eta_-, \eta_+ \in (0,\infty)$ such that $\eta_- \le \eta_t \le \eta_+$ for all t. Then $\{\|\nabla f(\mathbf{x}_t)\|\}_{t\in\mathbb{N}}$ converges to zero almost surely.*

Proof. By Proposition 1, we know

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{n}{k}B\nabla f(\mathbf{x}_t)^\top\mathbf{P}_t\nabla f(\mathbf{x}_t) + C\delta_t. \tag{6}$$

Taking conditional expectation on both sides of the above inequality gives

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)\right] \le &\; - B\mathbb{E}\left[\frac{n}{k}\nabla f(\mathbf{x}_t)^\top\mathbb{E}_t\left[\mathbf{P}_t\right]\nabla f(\mathbf{x}_t)\right] + C\delta_t \\
= &\; - B\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + C\delta_t.
\end{aligned}$$

Suppose, in order to get a contradiction, that there exists $\alpha > 0$ such that $\|\nabla f(\mathbf{x}_t)\|^2 > \alpha$ infinitely often. Thus we have

$$\mathbb{E}\left[f(\mathbf{x}_t)\right] \leq \mathbb{E}\left[f(\mathbf{x}_0)\right] - B\mathbb{E}\left[\sum_{s \in \mathcal{X}_t} \|\nabla f(\mathbf{x}_s)\|^2\right] + C\delta_t, \quad t \geq T_0,$$

where $\mathcal{X}_t = \{0 \leq s < t : \|\nabla f(\mathbf{x}_s)\|^2 > \alpha\}$. The above inequality (6) gives

$$\mathbb{E}\left[f(\mathbf{x}_t)\right] \leq \mathbb{E}\left[f(\mathbf{x}_0)\right] - \alpha B N_t + C\delta_t, \quad t \geq T_0$$

for some $\{N_t\}_t \subseteq \mathbb{N}$ such that $\lim\limits_{t \to \infty} N_t = \infty$.

Taking limits on both sides of the above inequality gives

$$\liminf_{t \to \infty} \mathbb{E}\left[f(\mathbf{x}_t)\right] \leq \mathbb{E}\left[f(\mathbf{x}_0)\right] + \liminf_{t \to \infty}(\alpha B N_t) + C\delta_t = -\infty,$$

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

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

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

Note that $\left\{\lim_{t \to \infty} \|\nabla f(\mathbf{x}_t)\|^2 = 0\right\}^c = \bigcup_{\alpha > 0} \{\|\nabla f(\mathbf{x}_t)\|^2 > \alpha \text{ infinitely often}\} = \bigcup_{\alpha \in \mathbb{Q}_+} \{\|\nabla f(\mathbf{x}_t)\|^2 > \alpha \text{ infinitely often}\}$, where we replace the union over an uncountable set with a union over a countable set.

Thus it holds that

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

which concludes the proof.

□

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

Lemma 2. *Instate Assumption 1. Let $\{\mathbf{x}_t\}_{t \in \mathbb{N}}$ be the sequence of Algorithm 1. Pick step sizes η_t so that there exist $\eta_-, \eta_+ \in (0, \infty)$ such that $\eta_- \leq \eta_t \leq \eta_+$ for all t. Then $\{\mathbf{x}_t\}_t$ converges to a bounded critical point of f almost surely. Let $\mathbf{x}_\infty$ be the almost sure limit of $\{\mathbf{x}_t\}_{t \in \mathbb{N}}$. It holds that $\lim\limits_{t \to \infty} f(\mathbf{x}_t) = f(\mathbf{x}_\infty)$ almost surely.*

Proof of Lemma 2. Let $\mathbf{V}_t \in \mathrm{St}(n, k)$ be the random orthogonal frame at t.

> Let R be an arbitrary realization of $\mathbf{V}_1, \mathbf{V}_2, \cdots$ such that $\{\|\nabla f(\mathbf{x}_t)\|\}_{t \in \mathbb{N}}$ converges to zero. **(H0)**

Next we restrict our attention to this realization R. Note that there is no randomness in $\{\mathbf{x}_t\}_{t \in \mathbb{N}}$ once this realization R is fixed.

Let $\mathbf{u} \in \mathbb{R}^n$ be an arbitrary vector, and let $z_t := \mathbf{u}^\top \mathbf{x}_t$ and $\varphi_t := \eta_t \mathbf{u}^\top \widehat{\nabla} f_k^{\delta_t}(\mathbf{x}_t)$ for all t. With this notation, we have $z_{t+1} = z_t - \varphi_t$ for all t. By Lemma 1, we know $\lim_{t \to \infty} (z_{t+1} - z_t) = 0$. Let $K_\mathbf{u}$ be the set of subsequential limits of $\{z_t\}_t$. Then $K_\mathbf{u}$ is closed and bounded (by Assumption 1).

Claim 1. *The set $K_\mathbf{u}$ is connected for any unit vector $\mathbf{u} \in \mathbb{R}^n$.*

Proof of Claim 1. Suppose, in order to get a contradiction, that $K_\mathbf{u}$ is not connected. **(H)**

Since $K_\mathbf{u}$ is closed and bounded, we know that, if $K_\mathbf{u}$ is disconnected, then there exists $a_1, b_1, a_2, b_2 \in (-\infty, \infty)$ such that

- $[a_1, b_1] \cup [a_2, b_2] \subseteq K_\mathbf{u}$ with $b_1 < a_2$;

- $(b_1, a_2) \cap K_{\mathbf{u}} = \emptyset$.

For simplicity, let $m := \frac{b_1+a_2}{2}$ and $\Delta := a_2 - b_1$. Since b_1 and a_2 are limit points, we can find subsequences $\{z_{i_j^1}\}_j$ and $\{z_{i_j^2}\}_j$ such that

1. $z_{i_j^1} \le m - \frac{\Delta}{4}$;
2. $z_{i_j^2} \ge m + \frac{\Delta}{4}$;
3. $i_j^1 < i_j^2$ for all j.

Since $\{\mathbf{x}_{t+1} - \mathbf{x}_t\}_t$ converges to zero, we know that $\{z_{t+1} - z_t\}_t$ converges to zero. By items 1, 2 and 3 above, we know that there exists $\{i_j^3\}_j$ such that $i_j^1 < i_j^3 \le i_j^2$ and $z_{i_j^3} \in (m - \frac{\Delta}{4}, m + \frac{\Delta}{4}]$ for infinitely many j. Otherwise, there will be a contradiction to the fact that $\{z_{t+1} - z_t\}_t$ converges to zero. Then we know that $\{z_t\}_t$ has a limit point in $[m - \frac{\Delta}{4}, m + \frac{\Delta}{4}]$, since $\{z_{i_j^3}\}_j$ has a limit point in $[m - \frac{\Delta}{4}, m + \frac{\Delta}{4}]$. This is a contradiction to **(H)**.

□

By Claim 1, we know that the limit points of $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ is connected. In this realization R, it holds that that any limit point of $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ is a critical point of f. By Assumption 2, critical points of f are isolated. Thus we know that the limit points of $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ must be a singleton. In other words, the sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ converges to a critical point of f.

The above argument holds for an arbitrary realization R that satisfies **(H0)**. By Lemma 1, almost all realizations satisfy **(H0)**. This finishes the proof.

□

Now that we have shown that $\{\mathbf{x}_t\}_t$ converges to a critical point almost surely. We state a more convenient form of the Łojasiewicz inequality in the following proposition.

Proposition 2. *Instate Assumption 1. Pick step sizes η_t so that there exist $\eta_-, \eta_+ \in (0,\infty)$ such that $\eta_- \le \eta_t \le \eta_+$ for all t. Let $\mathbf{x}_\infty$ be the almost sure limit of $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$. Then it holds that, for all $t \ge T_0$,*

$$(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{2\theta} \le \|\nabla f(\mathbf{x}_t)\|, \quad \textit{almost surely.}$$

Proof. Proposition 2 is a direct consequence of *a.s.* convergence of $\{\mathbf{x}_t\}_t$ (Lemma 2) and the Łojasiewicz inequality. □

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

In the previous subsection, we have proved asymptotic convergence results for the SZGD algorithm. This section is devoted to convergence rate analysis of $\{f(\mathbf{x}_t)\}_{t\in\mathbb{N}}$. We first state Proposition 3 that holds true for large t.

Proposition 3. *Instate Assumption 1. Let $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Pick step sizes η_t so that there exist $\eta_-, \eta_+ \in (0,\infty)$ such that $\eta_- \le \eta_t \le \eta_+$ for all t.Let $\mathbf{x}_\infty$ be the almost sure limit of $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ (Lemma 2).*

If $\theta \in (0, \frac{1}{2}]$, there exists $T_0 < \infty$ such that

(i) $f(\mathbf{x}_t) - f(\mathbf{x}_\infty) \le 1$ *for all* $t \ge T_0$.

If $\theta \in (\frac{1}{2}, 1)$, there exist constants $C_0, T_0 < \infty$ such that

(ii) $\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)] \in \left(0, \left(\frac{1}{2\theta B}\right)^{\frac{1}{2\theta-1}}\right)$ *for all* $t \ge T_0$

(iii) $C_0 > 1$ *and* $\frac{C_0}{2\theta-1} t^{\frac{2\theta}{1-2\theta}} - BC_0^{2\theta} t^{\frac{2\theta}{1-2\theta}} + C\delta_t \le 0$ *for all* $t \ge T_0$;

(iv) $\mathbb{E}[f(\mathbf{x}_{T_0}) - f(\mathbf{x}_\infty)] \le C_0 T_0^{\frac{1}{1-2\theta}}$.

Proof. We first prove item (i). When $\theta \in (0, \frac{1}{2}]$, $-z^{2\theta} \overset{①}{\leq} -\min\{1, z\}$ for all $z \in [0, \infty)$.

Let $\mathcal{X} \subseteq \mathbb{N}$ be a set of times where $f(\mathbf{x}_s) - f(\mathbf{x}_\infty) > 1$ for $s \in \mathcal{X}$. Suppose, in order to get a contradiction, that $|\mathcal{X}| = \infty$.

Clearly we have

$$\lim_{t\to\infty} f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_0) - B \sum_{s\in\mathcal{X}} \nabla f(\mathbf{x}_s)^\top \mathbf{P}_s \nabla f(\mathbf{x}_s) + C \sum_{s=0}^{\infty} \delta_s.$$

Note that $\mathcal{X}$ and $\mathbb{E}_{\mathbf{V}_{s:s\in\mathcal{X}}} [\lim_{t\to\infty} f(\mathbf{x}_{t+1})]$ are random variables contained in $\cup_{s=0}^{\infty} \mathcal{F}_s$. Thus taking expectation with respect to $\{\mathbf{V}_s : s \in \mathcal{X}\}$ on both sides of the above inequality gives

$$\begin{aligned}
\mathbb{E}_{\mathbf{V}_{s:s\in\mathcal{X}}} \left[\lim_{t\to\infty} f(\mathbf{x}_{t+1})\right] &\leq f(\mathbf{x}_0) + C \sum_{s=0}^{\infty} \delta_s - B \sum_{s\in\mathcal{X}} \mathbb{E}_{\mathbf{V}_{s:s\in\mathcal{X}}} \left[\nabla f(\mathbf{x}_s)^\top \mathbf{P}_s \nabla f(\mathbf{x}_s)\right] \\
&= f(\mathbf{x}_0) + C \sum_{s=0}^{\infty} \delta_s - B \sum_{s\in\mathcal{X}} \|\nabla f(\mathbf{x}_s)\|^2 \\
&\leq f(\mathbf{x}_0) + C \sum_{s=0}^{\infty} \delta_s - B \sum_{s\in\mathcal{X}} (f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{2\theta} \qquad \text{(by Proposition 2)} \\
&\leq f(\mathbf{x}_0) + C \sum_{s=0}^{\infty} \delta_s - B \sum_{s\in\mathcal{X}} \min\{1, f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\} \qquad \text{(by ①)} \\
&\leq f(\mathbf{x}_0) + C \sum_{s=0}^{\infty} \delta_s - B|\mathcal{X}|.
\end{aligned}$$

If there does not exist a T_0 such that $f(\mathbf{x}_t) - f(\mathbf{x}_\infty) \leq 1$ for all $t \geq T_0$, then the above implies that $f(\mathbf{x}_t)$ goes to negative infinity. This is a contradiction to that $f(\mathbf{x}_t)$ is bounded from below (Assumption 2). This finishes the proof of item (i).

Item (ii) follows from Lemma 2. Next we prove items (iii) and (iv).

Recall $B = -\min\left\{\left(\frac{L\eta_-^2 n}{2k} - \eta_-\right), \left(\frac{L\eta_+^2 n}{2k} - \eta_+\right)\right\}$. Consider a T_0 so that the following is satisfied for all $t \geq T_0$

$$C\delta_t t^{\frac{2\theta}{2\theta-1}} \leq \frac{1}{2\theta - 1}. \tag{7}$$

Next, since $2\theta > 1$, we can pick $C_0' > 1$ so that

$$\frac{C_0' + 1}{2\theta - 1} - B(C_0')^{2\theta} \leq 0. \tag{8}$$

Thus, for $t \geq T_0$, it holds that

$$\frac{C_0'}{2\theta - 1} - B(C_0')^{2\theta} + C\delta_t t^{\frac{2\theta}{2\theta-1}} \leq \frac{C_0' + 1}{2\theta - 1} - B(C_0')^{2\theta} \leq 0. \qquad \text{(by Eqs. 7 and 8)}$$

By multiplying both sides of the above inequality by $t^{\frac{2\theta}{1-2\theta}}$, we find a C_0' satisfying (iii). In fact, any constant larger than this C_0' satisfies this item.

Finally, we can find C_0'' satisfying item (iv), since $\mathbb{E}\left[f(\mathbf{x}_{T_0}) - f(\mathbf{x}_\infty)\right]$ is absolutely bounded for any given T_0. Indeed, any constant larger than this C_0'' satisfies this item. We finish the proof by taking $C_0 = \max\{C_0', C_0''\}$. □

By Propositions 3 and 2, (6) implies that, for $t \geq T_0$,

$$\begin{aligned}
\mathbb{E}_t \left[f(\mathbf{x}_{t+1})\right] - f(\mathbf{x}_\infty) &\leq f(\mathbf{x}_t) - f(\mathbf{x}_\infty) - B\|\nabla f(\mathbf{x}_t)\|^2 + C\delta_t \\
&\leq f(\mathbf{x}_t) - f(\mathbf{x}_\infty) - B\left(f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right)^{2\theta} + C\delta_t.
\end{aligned} \tag{9}$$

The above inequality (9) is a stochastic relation for the stochastic sequence $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_{t\in\mathbb{N}}$. In what follows, we will study the convergence behavior of this sequence.

Now we are ready to prove Theorem 3.

Proof of Theorem 3(a). Let T_0 be the constant so that item (i) in Proposition 3 is true. By Eq. (6), we have

$$\begin{aligned}\mathbb{E}_{\mathbf{V}_t}\left[f(\mathbf{x}_{t+1})\right] - f(\mathbf{x}_\infty) &\leq f(\mathbf{x}_t) - f(\mathbf{x}_\infty) - B\|\nabla f(\mathbf{x}_t)\|^2 + C\delta_t \\ &\leq f(\mathbf{x}_t) - f(\mathbf{x}_\infty) - B\left(f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right)^{2\theta} + C\delta_t \qquad (10)\\ &\leq \left(1-B\right)\left(f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right) + C\delta_t, \qquad (11)\end{aligned}$$

where (10) uses Proposition 2 and the last inequality uses item (i) in Proposition 3. Taking total expectation on both sides of (11) gives

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty)\right] &\leq (1-B)\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right] + C\delta_t \\ &\vdots \\ &\leq (1-B)^{t-T_0}\mathbb{E}\left[\left(f(\mathbf{x}_{T_0}) - f(\mathbf{x}_\infty)\right)\right] + C(1-B)^{t-T_0}. \qquad (12)\end{aligned}$$

Let $Q \in (1, (1-B)^{-1})$ and let $Z_t := Q^t\left(f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right)$. By the Cauchy–Schwarz inequality, (12) gives

$$\mathbb{E}\left[Z_{t+1}\right] \leq Q^t(1-B)^{t-T_0}Z_{T_0} + CQ^t(1-B)^{t-T_0}.$$

We conclude the proof by noticing $\lim_{t\to\infty}\mathbb{E}\left[Z_{t+1}\right] = 0$.

□

Proof of Theorem 3(b). Let C_0 and T_0 be two constants so that items (ii), (iii) and (iv) in Proposition 3 hold true.

Taking expectation on both sides of (9) gives, for all $t \geq T_0$,

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty)\right] &\leq \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right] - B\mathbb{E}\left[\left(f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right)^{2\theta}\right] + C\delta_t \\ &\leq \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right] - B\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right]^{2\theta} + C\delta_t,\end{aligned}$$

where the last inequality uses Jensen's inequality.

For simplicity, write $y_t := \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right]$ for all t. Next, we use induction to show that

$$y_t \leq C_0 t^{\frac{1}{1-2\theta}}, \quad \forall t \geq T_0. \qquad (13)$$

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

$$y_{t+1} \leq y_t - By_t^{2\theta} + C\delta_t \leq C_0 t^{\frac{1}{1-2\theta}} - BC_0^{2\theta}t^{\frac{2\theta}{1-2\theta}} + C\delta_t, \qquad (14)$$

where second inequality uses item (ii) in Proposition 3 and that the function $z \mapsto z - Bz^{2\theta}$ is strictly increasing when $z \in \left(0, \left(\frac{1}{2\theta B}\right)^{\frac{1}{2\theta-1}}\right)$.

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

$$(t+1)^{\frac{1}{1-2\theta}} = t^{\frac{1}{1-2\theta}} + \frac{1}{1-2\theta}(t+z)^{\frac{2\theta}{1-2\theta}} \geq t^{\frac{1}{1-2\theta}} + \frac{1}{1-2\theta}t^{\frac{2\theta}{1-2\theta}}, \qquad (15)$$

since $z \in [0,1]$. Thus by (14), we have

$$\begin{aligned}y_{t+1} &\leq C_0 t^{\frac{1}{1-2\theta}} - BC_0^{2\theta}t^{\frac{2\theta}{1-2\theta}} + C\delta_t \\ &\leq C_0(t+1)^{\frac{1}{1-2\theta}} + C_0\frac{1}{2\theta-1}t^{\frac{2\theta}{1-2\theta}} - BC_0^{2\theta}t^{\frac{2\theta}{1-2\theta}} + C\delta_t \leq C_0(t+1)^{\frac{1}{1-2\theta}},\end{aligned}$$

where the second inequality uses (15), and the last inequality uses (iii) in Proposition 3.

□

4.3 Convergence Rate of $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$

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

Theorem 4. *Instate Assumption 1. Let $\mathbf{x}_t$ be a sequence generated by the SZGD algorithm, and let $\mathbf{x}_\infty$ be the almost sure limit of $\{\mathbf{x}_t\}_t$. Pick step sizes η_t so that there exist $\eta_-, \eta_+ \in (0,\infty)$ such that $\eta_- \le \eta_t \le \eta_+$ for all t. Then it holds that*

(a) if $\theta \in (0, \frac{1}{2}]$, $-1 \le \left(\frac{Ln\eta_-^2}{2k} - \eta_-\right) < 0$ and $-1 \le \left(\frac{Ln\eta_+^2}{2k} - \eta_+\right) < 0$, there exists a constant $Q > 1$ such that $\{Q^t \sum_{s=t}^{\infty}\|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\}_t$ converges to 0 in expectation.

(b) if $\theta \in (\frac{1}{2}, 1)$ and $\eta_-, \eta_+ \in (0, \frac{2k}{Ln})$, then it holds that $\left\{\sum_{s=t}^{\infty}\|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\right\}_t$ converges to 0 at rate $O\left(t^{\frac{1}{1-2\theta}}\right)$ in expectation.

Proof. Note that $\{\|\nabla f(\mathbf{x}_t)\|\}_t$ is absolutely bounded due to boundedness of $\{\mathbf{x}_t\}_t$ and continuity of $\nabla f(\mathbf{x})$. By Theorems 2 and 1 it holds that

$$\mathbb{E}_{\mathbf{V}_t}\left[\|\widehat{\nabla} f_{\mathbf{V}_t}(\mathbf{x}_t)\|^2\right] \le \|\mathbb{E}_{\mathbf{V}_t}\left[\widehat{\nabla} f_{\mathbf{V}_t}(\mathbf{x}_t)\right]\|^2 + \left(\frac{n}{k} - 1\right)\|\nabla f(\mathbf{x}_t)\|^2$$
$$+ \frac{4L\delta}{\sqrt{3}}\left(\frac{n^2}{k} - n\right)\|\nabla f(\mathbf{x}_t)\| + \frac{4L^2 n^2 \delta^2}{3k} \le \frac{n}{k}\|\nabla f(\mathbf{x}_t)\|^2 + C\delta_t.$$

Thus it holds that

$$\frac{k}{n}\left(\eta_t - \frac{L\eta_t^2 n}{2k}\right)\mathbb{E}_t\left[\|\widehat{\nabla} f_{\mathbf{V}_t}^{\delta_t}(\mathbf{x}_t)\|^2\right] \le \left(\eta_t - \frac{L\eta_t^2 n}{2k}\right)\|\nabla f(\mathbf{x}_t)\|^2 + C\delta_t$$
$$\le f(\mathbf{x}_t) - \mathbb{E}_t\left[f(\mathbf{x}_{t+1})\right] + C\delta_t,$$

where the second inequality uses (6).

Since $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t\widehat{\nabla} f_{\mathbf{V}_t}^{\delta_t}(\mathbf{x}_t)$, the above inequality implies that,

$$\frac{k}{n}\left(\frac{1}{\eta_t} - \frac{Ln}{2k}\right)\mathbb{E}_t\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right] \le f(\mathbf{x}_t) - \mathbb{E}_t\left[f(\mathbf{x}_{t+1})\right] + C\delta_t. \tag{16}$$

Taking total expectation on both sides of the above inequality gives

$$\frac{k}{n}\left(\frac{1}{\eta_t} - \frac{Ln}{2k}\right)\mathbb{E}\left[\|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\right] \le \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right] - \mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty)\right] + C\delta_t.$$

Since $\{\mathbf{x}_t\}_t$ converges almost surely and $\eta_t \in [\eta_-, \eta_+]$, summing up the above inequality gives

$$\frac{k}{n}\left(\frac{1}{\eta_+} - \frac{Ln}{2k}\right)\mathbb{E}\left[\sum_{s=t}^{\infty}\|\mathbf{x}_{s+1} - \mathbf{x}_s\|^2\right] \le \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right] + C\delta_t, \quad \forall t \ge T_0.$$

Since we have proved the convergence rate for $\{\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right]\}_t$, we can conclude the proof by the result of Theorem 3.

□

Next we provide convergence rate guarantee for $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ in Theorem 5.

Theorem 5. *Instate Assumption 1. Let $\mathbf{x}_t$ be a sequence generated by the SZGD algorithm, and let $\mathbf{x}_\infty$ be the almost sure limit of $\{\mathbf{x}_t\}_t$. Pick step sizes η_t so that there exist $\eta_-, \eta_+ \in (0,\infty)$ such that $\eta_- \le \eta_t \le \eta_+$ for all t. Then it holds that*

(a) if $\theta \in (0, \frac{1}{2}]$, $-1 \le \left(\frac{Ln\eta_-^2}{2k} - \eta_-\right) < 0$ and $-1 \le \left(\frac{Ln\eta_+^2}{2k} - \eta_+\right) < 0$, there exists a constant $Q > 1$ such that $\{Q^t\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ converges to 0 in expectation.

(b) if $\theta \in (\frac{1}{2}, 1)$ and $\eta_-, \eta_+ \in \left(0, \frac{2k}{Ln}\right)$, then it holds that $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ converges to 0 at rate $O\left(t^{\frac{1-\theta}{1-2\theta}}\right)$ in expectation.

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 $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$. 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

$$z_2^{1-\theta} - z_1^{1-\theta} \geq (1-\theta) z_2^{-\theta}(z_2 - z_1), \qquad \forall z_1, z_2 > 0,\ \ \theta \in (0,1). \tag{17}$$

By letting $z_1 = f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty)$ and $z_2 = f(\mathbf{x}_t) - f(\mathbf{x}_\infty)$, we have

$$(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{1-\theta} - (f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty))^{\theta} \geq (1-\theta)(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{-\theta} \left(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})\right).$$

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

$$\begin{aligned}
&(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{1-\theta} - \mathbb{E}_t\left[(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty))^{1-\theta}\right] \\
\geq\ & (1-\theta)(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{-\theta}\left(f(\mathbf{x}_t) - \mathbb{E}_t\left[f(\mathbf{x}_{t+1})\right]\right) \\
\geq\ & (1-\theta)(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{-\theta}\left(\frac{k}{n}\left(\frac{1}{\eta_t} - \frac{Ln}{2k}\right)\mathbb{E}_t\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right] - C\delta_t\right),
\end{aligned} \tag{18}$$

where the last line uses (16).

Also, it holds that, for sufficiently large t,

$$\begin{aligned}
0 &< (f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{\theta} \overset{①}{\leq} \sqrt{\|\nabla f(\mathbf{x}_t)\|} \overset{②}{\leq} \left(\left\|\mathbb{E}_t\left[\widehat{\nabla} f_{\mathbf{V}_t}^{\delta_t}(\mathbf{x}_t)\right] + O\left(\frac{n\delta_t}{n+1}\mathbf{1}\right)\right\|\right)^{1/2} \\
&\leq \left(\left\|\mathbb{E}_t\left[\widehat{\nabla} f_{\mathbf{V}_t}^{\delta_t}(\mathbf{x}_t)\right]\right\| + C\delta_t\right)^{1/2} = \left(\frac{1}{\eta_t}\|\mathbb{E}_t\left[\mathbf{x}_{t+1} - \mathbf{x}_t\right]\| + C\delta_t\right)^{1/2} \\
&\leq \sqrt{\frac{1}{\eta_t}\mathbb{E}_t\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right]} + \sqrt{C\delta_t},
\end{aligned} \tag{19}$$

where ① uses Proposition 2, and ② uses Theorem 1.

Combining the above result with (18) gives

$$\begin{aligned}
&(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{1-\theta} - \mathbb{E}_t\left[(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty))^{1-\theta}\right] \\
\geq\ & \frac{(1-\theta)\left(\frac{k}{n}\left(1 - \frac{Ln\eta_t}{2k}\right)\sqrt{\eta_t}\mathbb{E}_t\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right] - C\delta_t\right)}{\sqrt{\mathbb{E}_t\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right]} + \sqrt{C\delta_t}} \\
=\ & (1-\theta)\frac{k}{n}\left(1 - \frac{Ln\eta_t}{2k}\right)\sqrt{\eta_t}\frac{\mathbb{E}_t\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right] - C\delta_t}{\sqrt{\mathbb{E}_t\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right]} + \sqrt{C\delta_t}}.
\end{aligned} \tag{20}$$

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

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

where the last inequality uses (20).

Combining Jensen's inequality and the above inequality gives

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

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

$$\sum_{s=t}^{\infty} \mathbb{E}\left[\|\mathbf{x}_{s+1} - \mathbf{x}_s\|\right] \leq \frac{n\sqrt{\eta_t}}{(1-\theta)k\left(1 - \frac{Ln\eta_t}{2k}\right)} \mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{1-\theta}\right] + C\sum_{s=t}^{\infty} \delta_s^{1/2}$$
$$= \frac{n\sqrt{\eta_t}}{(1-\theta)k\left(1 - \frac{Ln\eta_t}{2k}\right)} \mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{1-\theta}\right] + C\delta_t^{1/2}$$

Since $\|\mathbf{x}_t - \mathbf{x}_\infty\| \leq \sum_{s=t}^{\infty} \|\mathbf{x}_{s+1} - \mathbf{x}_s\|$, the above implies that

$$\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_\infty\|\right] \leq \frac{n\sqrt{\eta_t}}{(1-\theta)k\left(1 - \frac{Ln\eta_t}{2k}\right)} \mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{1-\theta}\right] + C\delta_t^{1/2}$$
$$\leq \frac{n\sqrt{\eta_t}}{(1-\theta)k\left(1 - \frac{Ln\eta_t}{2k}\right)} \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right]^{1-\theta} + C\delta_t^{1/2},$$

where the last inequality uses Jensen's inequality. We can conclude the proof by the result of Theorem 3.

□

4.4 Implications on Gradient Descent

The convergence rate for SZGD implies sure convergence rate of the gradient descent algorithm. In this section, we will display convergence rate results for the gradient descent algorithm on Łojasiewicz functions. 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 $\{f(\mathbf{x}_t)\}_t$ 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

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t). \qquad \text{(Gradient Descent (GD))}$$

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. Let* $\mathbf{x}_t$ *be a sequence generated by the gradient descent algorithm, and let* $\mathbf{x}_\infty$ *be the limit of* $\{\mathbf{x}_t\}_t$. *Suppose there exists* $\eta_-, \eta_+ \in (0, \infty)$ *such that* $\eta_- \leq \eta_t \leq \eta_+$ *for all* t. *Then if* $\theta \in (0, \frac{1}{2}]$, $-1 \leq \left(\frac{L\eta_-^2}{2} - \eta_-\right) < 0$ *and* $-1 \leq \left(\frac{L\eta_+^2}{2} - \eta_+\right) < 0$, *there exists a constant* $Q > 1$ *such that* $\{Q^t\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ *converges to* 0 *and* $\{Q^t\left(f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\right)\}_t$ *converges to* 0.

Corollary 1 follows immediately from Theorems 3 and 5. If $\theta \in (0, \frac{1}{2}]$, Corollary 1 provides linear convergence rate of $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ and $\{f(\mathbf{x}_t) - f(\mathbf{x}_\infty)\}_t$ with x_t governed by the GD algorithm. When $\theta \in (0, \frac{1}{2}]$, linear convergence rate for $\{\sum_{s=t}^{\infty} \|\mathbf{x}_{s+1} - \mathbf{x}_s\|\}_t$ 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 6. To prove Theorem 6, we first need the following proposition.

Proposition 4. *Instate Assumption 1. Let* $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ *be the sequence governed by the gradient descent algorithm. Pick step sizes* η_t *so that there exist* $\eta_-, \eta_+ \in (0, \infty)$ *such that* $\eta_- \leq \eta_t \leq \eta_+$ *for all* t. *Let* $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ *be bounded and let* $\mathbf{x}_\infty$ *be the limit of* $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$. *Then there exist constants* C_0 *and* T_0 *such that the following holds:*

(i) $f(\mathbf{x}_t) - f(\mathbf{x}_\infty) \in \left(0, \left(\frac{1}{2\theta B}\right)^{\frac{1}{2\theta-1}}\right)$ *for all* $t \geq T_0$.

(ii) $-C_0^{2\theta}B + \frac{C_0}{2\theta-1} \leq 0$, *where* B *is defined in (5).*

(iii) $f(\mathbf{x}_{T_0}) - f(\mathbf{x}_\infty) \le C_0 T_0^{\frac{1}{1-2\theta}}$.

Proof. Since f is L-smooth, it holds that

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &\le f(\mathbf{x}_t) - B\|\nabla f(\mathbf{x}_t)\|^2, \end{aligned}$$

where B is defined in (5).

Therefore, $\lim_{t\to\infty} \|\nabla f(\mathbf{x}_t)\| = 0$. Otherwise, the above inequality leads to a contradiction to that $\inf_{\mathbf{x}} f(\mathbf{x}) > -\infty$. By Lemma 2, we know that $\{\mathbf{x}_t\}_t$ converges to a critical point of f. Thus by continuity of f, item (i) can be satisfied since $\{f(\mathbf{x}_t)\}_t$ converges. Item (ii) can be satisfied by picking some $C_0 > 1$. Since $\{f(\mathbf{x}_t)\}_t$ is bounded, we can pick C_0 large enough so that item (iii) is satisfied.

□

With the above proposition, we are ready to prove Theorem 6.

Theorem 6. *Instate Assumption 1. Let $\mathbf{x}_t$ be a sequence generated by the gradient descent algorithm, and let $\mathbf{x}_\infty$ be the limit of $\{\mathbf{x}_t\}_t$. Pick step sizes η_t so that there exist $\eta_-, \eta_+ \in (0,\infty)$ such that $\eta_- \le \eta_t \le \eta_+$ for all t. If $\theta \in (\frac{1}{2}, 1)$, then $f(\mathbf{x}_t) - f(\mathbf{x}_\infty) \le O\left(t^{\frac{1}{1-2\theta}}\right)$.*

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

By L-smoothness by f, it holds that

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) + \left(\frac{L\eta_t^2}{2} - \eta_t\right)\|\nabla f(\mathbf{x}_t)\|^2. \end{aligned}$$

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

$$\begin{aligned} f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty) &\le f(\mathbf{x}_t) - f(\mathbf{x}_\infty) + \left(\frac{L\eta_t^2}{2} - \eta_t\right)(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{2\theta} \qquad (21) \\ &\le f(\mathbf{x}_t) - f(\mathbf{x}_\infty) - B(f(\mathbf{x}_t) - f(\mathbf{x}_\infty))^{2\theta}, \end{aligned}$$

where B is defined in (5).

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

$$(t+1)^{\frac{1}{1-2\theta}} \ge t^{\frac{1}{1-2\theta}} + \frac{1}{1-2\theta} t^{\frac{2\theta}{1-2\theta}}. \qquad (22)$$

We will use induction to finish the proof. Item (iii) in Proposition 4 states $f(\mathbf{x}_{T_0}) - f(\mathbf{x}_\infty) \le C_0 T_0^{\frac{1}{1-2\theta}}$. Inductively, if $f(\mathbf{x}_t) - f(\mathbf{x}_\infty) \le C_0 t^{\frac{1}{1-2\theta}}$ $(t \ge T_0)$, then (21) implies that

$$\begin{aligned} & f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty) \le C_0 t^{\frac{1}{1-2\theta}} - BC_0^{2\theta} t^{\frac{2\theta}{1-2\theta}} \\ \le\ & C_0 (t+1)^{\frac{1}{1-2\theta}} - BC_0^{2\theta} t^{\frac{2\theta}{1-2\theta}} + \frac{C_0}{2\theta - 1} t^{\frac{2\theta}{1-2\theta}} \le C_0 (t+1)^{\frac{1}{1-2\theta}}, \end{aligned}$$

where the first line uses (21), the induction hypothesis, item (i) in Proposition 4, and that the function $x \mapsto x - Bx^{2\theta}$ is strictly increasing on $\left(0, \left(\frac{1}{2\theta B}\right)^{\frac{1}{2\theta-1}}\right)$, the second line uses (22), and the last line uses item (ii) in Proposition 4.

□

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 $\{\mathbf{x}_t\}$ governed by the proximal algorithm (23) converges at rate $\|\mathbf{x}_t - \mathbf{x}_\infty\| \le O\left(t^{\frac{1-\theta}{1-2\theta}}\right)$. In this section, we show that the proximal algorithm satisfies $f(\mathbf{x}_t) \le O\left(t^{\frac{1}{1-2\theta}}\right)$. When $\theta \in (\frac{1}{2}, 1)$ the function value $\{f(\mathbf{x}_t)\}_t$ tends to converge at a faster rate than the point sequence $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$. This phenomenon for the proximal algorithm again suggests that the convergence rate of $\{f(\mathbf{x}_t)\}_t$ may be more important and informative than the convergence rate of $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$, since the trajectory of $\{\mathbf{x}_t\}_t$ 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* dom$f$*) is* $\mathrm{dom} f := \{\mathbf{x} \in \mathbb{R}^n : -\infty < f(\mathbf{x}) < +\infty\}$. *For each $\mathbf{x} \in \mathrm{dom} f$, the Fréchet subdifferential of f at $\mathbf{x}$, written $\hat{\partial} f(\mathbf{x})$, is the set of vectors $\mathbf{g}^* \in \mathbb{R}^n$ such that*

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

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

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

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

An element $\mathbf{g} \in \partial f(\mathbf{x})$ is call a subgradient of f at $\mathbf{x}$.

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

$$\mathbf{x}_{t+1} \in \arg\min_{\mathbf{z}} \left\{ f(\mathbf{z}) + \frac{1}{2\eta_t} \|\mathbf{z} - \mathbf{x}_t\|^2 \right\}, \tag{23}$$

with a given $\mathbf{x}_0 \in \mathbb{R}^n$.

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 $0 \in \partial \left(f(\mathbf{x}_{t+1}) + \frac{1}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right)$. By the subadditivity property of subdifferential (e.g., Exercise 10.10, [31]), we have

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

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

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}_{t+1},$$

for some $\mathbf{g}_{t+1} \in \partial f(\mathbf{x}_{t+1})$.

We also need the following two theorems.

Theorem 7 ([1]). *Let f satisfy Assumption 2 and let $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ be generated by the proximal algorithm. If $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ is bounded then it converges to a critical point of f.*

Theorem 8 ([1]). *Let f satisfy Assumption 2 and let $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ be generated by the proximal algorithm. Let f satisfy the Łojasiewicz inequality with Łojasiewicz exponent $\theta \in (\frac{1}{2}, 1)$. If $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ is bounded, then it holds that*

$$\|\mathbf{x}_t - \mathbf{x}_\infty\| \leq O(t^{\frac{2\theta}{1-2\theta}}) \quad \text{and} \quad \|\mathbf{x}_t - \mathbf{x}_{t+1}\| \leq O(t^{\frac{2\theta}{1-2\theta}}),$$

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

Same as [1], we will make the following assumptions on f. Note that all items in Assumption 2 are assumed in [1].

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

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

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

   where $B(\mathbf{x}_*, \mu)$ is the ball of radius μ centered at x_*. Without loss of generality, we let $\kappa = 1$ to avoid clutter.
3. *$\inf_{x\in\mathbb{R}^n} f(\mathbf{x}) > -\infty$;*
4. *Let $\{\mathbf{x}_t\}_t$ be the sequence generated by the GD algorithm. We assume $\{\mathbf{x}_t\}_t$ is bounded.*
5. *There exists $\eta_-, \eta_+ \in (0, \infty)$ such that $\eta_- \leq \eta_t \leq \eta_+$ for all t.*

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

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

- $\lim_{t\to\infty} \frac{w_t}{\|\mathbf{x}_t - \mathbf{x}_{t+1}\|} = 0$;
- $f(\mathbf{x}_t) \geq f(\mathbf{x}_{t+1}) + \mathbf{g}_{t+1}^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) - w_t, \quad \forall t \in \mathbb{N}$.

Proof. By Definition 3, we have

$$f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) - \langle \mathbf{g}_{t+1}, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \geq -o\left(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|\right), \quad \forall t,$$

which concludes the proof.

□

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

Proposition 6. *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 $\|\mathbf{x}_t - \mathbf{x}_\infty\| \leq \mu$, where μ is defined as in Definition 2.

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

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

where w_t is a sequence satisfying Proposition 5.

(iv) $f(\mathbf{x}_{T_0}) - f(\mathbf{x}_\infty) \leq C_0 T_0^{\frac{1}{1-2\theta}}$.

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 7, we can find T_0 so that item *(ii)* is true. By Theorem 8, we have $\|\mathbf{x}_t - \mathbf{x}_{t+1}\| \leq O\left(t^{\frac{2\theta}{1-2\theta}}\right)$ and thus Proposition 5 gives

$$\lim_{t\to\infty} w_t t^{\frac{2\theta}{2\theta-1}} = 0. \tag{25}$$

By item *(i)* and (25), 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 7). Letting $C_0 = \max\{C_0', C_0''\}$ concludes the proof. □

Theorem 9. *Instate Assumption 2. Let $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ generated by the proximal algorithm. If $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ is bounded then it converges to a critical point of f. Let $\mathbf{x}_\infty$ be the limit of $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$. If $\theta \in (\frac{1}{2}, 1)$, then it holds that $f(\mathbf{x}_t) - f(\mathbf{x}_\infty) \leq O\left(t^{\frac{1}{1-2\theta}}\right)$.*

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

$$f(\mathbf{x}_t) \geq f(\mathbf{x}_{t+1}) + \mathbf{g}_{t+1}^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) - w_t = f(\mathbf{x}_{t+1}) + \eta_t \|\mathbf{g}_{t+1}\|^2 - w_t.$$

Since $\{\mathbf{x}_t\}$ converges (Theorem 7) and the Łojasiewicz inequality holds with exponent θ, there exists T_0, such that for all $t \geq T_0$,

$$f(\mathbf{x}_t) \geq f(\mathbf{x}_{t+1}) + \eta_t \|\mathbf{g}_{t+1}\|^2 - w_t \geq f(\mathbf{x}_{t+1}) + \eta_t \left(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty)\right)^{2\theta} - w_t,$$

which gives

$$f(\mathbf{x}_t) - f(\mathbf{x}_\infty) \geq f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty) + \eta_t \left(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty)\right)^{2\theta} - w_t. \tag{26}$$

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 we have

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

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 $f(\mathbf{x}_{T_0}) - f(\mathbf{x}_\infty) \leq C_0 T_0^{\frac{1}{1-2\theta}}$. Inductively, if $f(\mathbf{x}_t) - f(\mathbf{x}_\infty) \leq C_0 t^{\frac{1}{1-2\theta}}$ ($t \geq T_0$), then by (26) it holds that

$$\begin{aligned}
&f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty) + \eta_t (f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty))^{2\theta} \leq f(\mathbf{x}_t) - f(\mathbf{x}_\infty) + w_t \\
&\leq C_0 t^{\frac{1}{1-2\theta}} + w_t \leq C_0 (t+1)^{\frac{1}{1-2\theta}} + C_0 \frac{1}{2\theta-1} t^{\frac{2\theta}{1-2\theta}} + w_t \\
&\leq C_0 (t+1)^{\frac{1}{1-2\theta}} + \eta_t C_0^{2\theta} (t+1)^{\frac{2\theta}{1-2\theta}},
\end{aligned} \tag{28}$$

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

Since the function $x \mapsto x + \eta_t x^{2\theta}$ is monotonic and strictly increasing on $(0,\infty)$, the above inequality (28) implies $f(\mathbf{x}_{t+1}) - f(\mathbf{x}_\infty) \leq C_0 (t+1)^{\frac{1}{1-2\theta}}$, 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 $\mathbb{R}^{30}$

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

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 1 and 2. To avoid numerical instability, we set $\delta_t = \max\left(0.1 \times 2^{-t}, 0.00001\right)$ in all numerical experiments.

Remark 2. *In this experiments, we compare the results of SZGD with GD. Strictly speaking, this is not a fair comparison since SZGD uses only zeroth-order information while GD can use first-order information.*

At first glance, Figure 1d and 1e show seemingly implausible results. However, these results are natural after we investigate the landscape of the test function F_2. While fully visualizing F_2 is impossible, we can consider a 1-d version of F_2: $f(x) = \sqrt{|x|}$, whose graph is shown below in Figure 1a. Note that $f(x) = \sqrt{|x|}$ is indeed a 1-d version of F_2, since $f(x) = \sqrt{|x|} = (x^2)^{1/4}$. This nonconvex function has the following properties, all highly aligned with empirical observations.

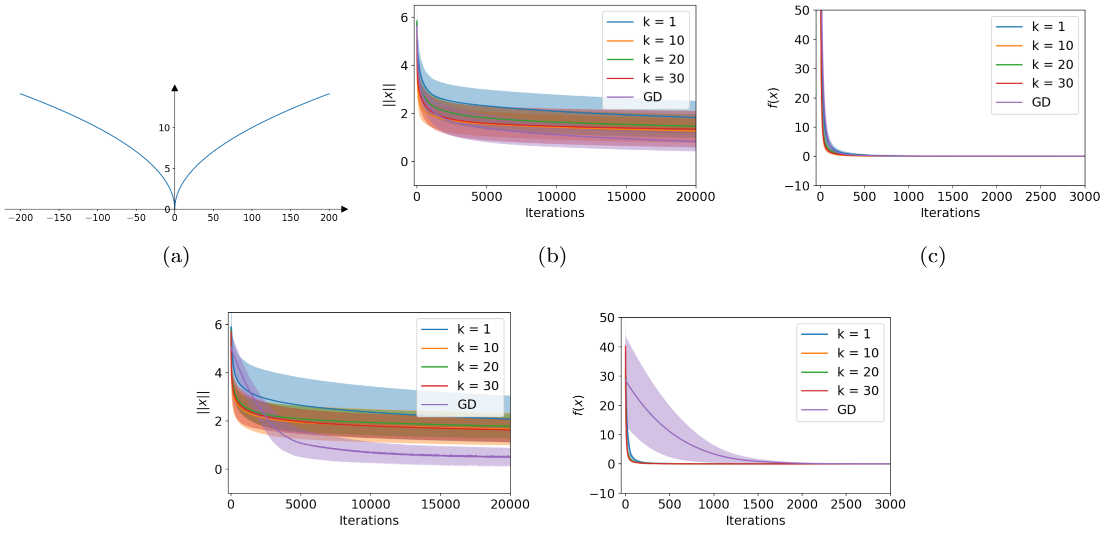
- When $\mathbf{x}$ is far from the origin, the function's surface is flat. In this case, randomness in the gradient, which may increase the magnitude of gradient, can speedup the convergence. In Figure 1d, we observe that SZGD algorithms converge faster than GD at the beginning, which agrees with the landscape of F_2.
- 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 1d, we observe that GD converges to a more optimal value at the end, which agrees with the non-differentiability of F_2 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 1a, the trajectory of GD will jump back and forth around zero, but hits zero with little chance. In Figure 1d, we observe that the $\|\mathbf{x}\|$ value of DG is fuzzy as it goes to zero, which agrees with the shape of F_2 near zero.

Other observations from Figure 1 are summarized below:

- Figure 1b and 1c show 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, $\{f(\mathbf{x}_t)\}_t$ converges faster than $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$, which agrees with our theoretical results.

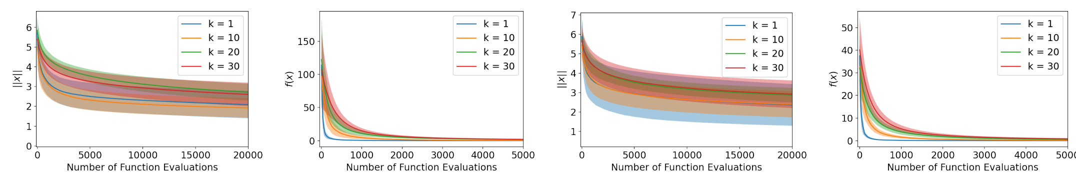
We also empirically study the convergence rates versus the number of function evaluations. Note that one iteration may require more than one function evaluations: k random orthogonal directions are sampled and $2k$ function evaluations are needed to obtain the gradient estimator defined in (3). The convergence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

- Given the same learning rate $\eta_t = \eta$ and the same number of function evaluations, in terms of *function evaluations*, the sequence $\{f(\mathbf{x}_t)\}_t$ converges faster when k is smaller.
- Given the same learning rate $\eta_t = \eta$ and the same number of function evaluations, the sequence $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ 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 $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$.
- When measured against number of function evaluations, $\{f(x_t)\}_t$ converges faster than $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$. This is similar to the observations in Figure 1.



(a) (b) (c) (d) (e)

Figure 1: Subfigure (a) plots the function $f(x) = \sqrt{|x|}$. Subfigures (b) and (c) (resp. (d) and (e)) plot results of SZGD and GD on test function F_1 (resp. 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. For all experiments, the step size η_t is set to 0.005 for all t. Subfigures (b) and (d) show observed convergence rate results for $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ ($x_\infty = 0$); Subfigure (c) (resp. (e)) shows observed convergence rate results for $\{\|F_1(x_t) - F_1(x_\infty)\|\}_t$ (resp. $\{\|F_2(x_t) - F_2(x_\infty)\|\}_t$). 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 $\mathbf{x}_0$ is random, the trajectory of GD is also random.



(a) (b) (c) (d)

Figure 2: 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 $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$ ($\mathbf{x}_\infty = 0$); Subfigures (b) and (d) show observed convergence rate results for $\{\|F_2(\mathbf{x}_t) - F_2(\mathbf{x}_\infty)\|\}_t$ ($F_2(\mathbf{x}_\infty) = 0$). 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 Figure 1, this figure plots errors (either $\|\mathbf{x}_t - \mathbf{x}_\infty\|$ or $f(\mathbf{x}_t) - f(\mathbf{x}_\infty)$) against number of function evaluations. For example, when $k = 10$, one iteration requires $10 \times 2 = 20$ functions evaluations.

7 Proof of Theorem 2

To start with, we need the following facts in Propositions 7 and 8.

Proposition 7 ([7]). *Let* $\mathbf{V} := [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k] \in \mathbb{R}^{n\times k}$ *be uniformly sampled from the Stiefel manifold* $St(n,k)$*. Then the marginal distribution for any* $\mathbf{v}_i$ *is uniform over the unit sphere* $\mathbb{S}^{n-1}$*.*

In words, the uniform measure over the Stiefel manifold $\mathrm{St}(n,k)$ 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 fact is the following proposition.

Proposition 8. *Let* $\mathbf{v}$ *be a vector uniformly randomly sampled from the unit sphere* $\mathbb{S}^{n-1}$*. Then it holds that* $\mathbb{E}\left[\mathbf{v}\mathbf{v}^\top\right] = \frac{1}{n}\mathbf{I}$*, where* $\mathbf{I} \in \mathbb{R}^{n\times n}$ *is the identity matrix.*

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

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

Taylor's expansion of f with integral form gives

$$f(\delta\mathbf{v}_i) = f(\mathbf{0}) + \delta\mathbf{v}_i^\top \nabla f(\mathbf{0}) + \int_0^\delta (\delta - t)\mathbf{v}_i^\top \nabla^2 f(t\mathbf{v}_i)\mathbf{v}_i \, dt$$

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

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

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

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

for all $i = 1,2,\cdots,k$. For any i,k,n, it holds that

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

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

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

For the variance of the gradient estimator, we have

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

where the last equation follows from the orthonormality of $\{\mathbf{v}_1, \mathbf{v}_2.\cdots, \mathbf{v}_k\}$. By Proposition 7, we know that $\mathbb{E}\left[\left(f(\delta\mathbf{v}_i) - f(-\delta\mathbf{v}_i)\right)v_i\right] = \mathbb{E}\left[\left(f(\delta\mathbf{v}_j) - f(-\delta\mathbf{v}_j)\right)\mathbf{v}_j\right]$ for all $i, j = 1, 2, \cdots, k$. Thus ③ gives

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

Combining ② and ④ gives

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

□

8 Conclusion

This paper studies the SZGD algorithm and its performance on Łojasiewicz functions. In particular, we establish 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 SZGD exhibits convergence behavior similar to its non-stochastic counterpart. Our results suggests $\{f(\mathbf{x}_t)\}_t$ tend to converge faster than $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$. 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 $\{\|\mathbf{x}_t - \mathbf{x}_\infty\|\}_t$.

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.