A Generic Approach for Accelerating Stochastic Zeroth-Order Convex Optimization

Xiaotian Yu[1,2], **Irwin King**[1,2], **Michael R. Lyu**[1,2], **Tianbao Yang**[3]
[1]Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2]Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications,
Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
{xtyu,king,lyu}@cse.cuhk.edu.hk
[3]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA
tianbao-yang@uiowa.edu

Abstract

In this paper, we propose a generic approach for accelerating the convergence of existing algorithms to solve the problem of stochastic zeroth-order convex optimization (SZCO). Standard techniques for accelerating the convergence of stochastic zeroth-order algorithms are by exploring multiple functional evaluations (e.g., two-point evaluations), or by exploiting global conditions of the problem (e.g., smoothness and strong convexity). Nevertheless, these classic acceleration techniques are necessarily restricting the applicability of newly developed algorithms. The key of our proposed generic approach is to explore a local growth condition (or called local error bound condition) of the objective function in SZCO. The benefits of the proposed acceleration technique are: (i) it is applicable to both settings with one-point evaluation and two-point evaluations; (ii) it does not necessarily require strong convexity or smoothness condition of the objective function; (iii) it yields an improvement on convergence for a broad family of problems. Empirical studies in various settings demonstrate the effectiveness of the proposed acceleration approach.

1 Introduction

We consider the following problem of stochastic convex optimization:

$$\min_{\mathbf{x}\in\Omega} f(\mathbf{x}) \triangleq \mathbb{E}_{\xi}[f(\mathbf{x};\xi)], \tag{1}$$

where $\Omega \subseteq \mathbb{R}^d$ is a closed compact convex set, $f(\mathbf{x};\xi)$ is a stochastic convex function depending on random noise ξ. This problem has broad applications in computer science and engineering. For example, many practical problems in machine learning can be cast into a stochastic convex optimization, where ξ denotes a random data point and $\mathbf{x}$ denotes the parameter of a prediction model. A standard approach for solving the problem of Eq. (1) is to adopt the stochastic (sub)gradient of $f(\mathbf{x};\xi)$ [Nemirovski *et al.*, 2009]. However, there exist situations where the first-order gradient information is computationally infeasible, expensive, or impossible, while the zeroth-order functional information can be easily obtained. For example, in online auctions and advertisement selections, only function values are revealed as feedbacks for algorithms [Wibisono *et al.*, 2012]. In stochastic structured predictions, explicit differentiations may be difficult to perform while the functional evaluations of predicted structures are easily obtained [Sokolov *et al.*, 2016]. The optimization problem of Eq. (1) in such situations is referred to SZCO.

A key concern in the development of iterative stochastic zeroth-oder algorithms for solving Eq. (1) is the order of the necessary number of functional evaluations in the form of $f(\mathbf{x};\xi)$, which is termed as sample complexity or iteration complexity. [Flaxman *et al.*, 2005] should be the first work related to SZCO. They studied a closely related setting namely online bandit convex optimization where only one-point evaluation (OPE) is available for the cost function at each iteration. Applied to the stochastic setting, their algorithm suffers from an iteration complexity of $O(d^2/\epsilon^4)$ for finding an ϵ-optimal solution $\widehat{\mathbf{x}}$ such that $\mathbb{E}[f(\widehat{\mathbf{x}}) - \min_{\mathbf{x}\in\Omega} f(\mathbf{x})] \leq \epsilon$. Since then, there have been a number of studies [Agarwal *et al.*, 2010; Shamir, 2013; Duchi *et al.*, 2015; Shamir, 2017; Nesterov and Spokoiny, 2017] trying to improve the iteration complexity of [Flaxman *et al.*, 2005] in online bandit setting or in stochastic optimization setting. A useful technique to accelerate the convergence of SZCO is by leveraging two-point evaluations (TPE) at each iteration. Another technique is to explore the strong convexity or the smoothness condition of the random function $f(\mathbf{x};\xi)$. Clearly, both techniques impose strong restrictions of their developed algorithms, and thus the applica-

bility of the resultant algorithms is limited.

Our Contributions. The goal of this paper is to design a generic approach for accelerating existing SZCO algorithms which is applicable to both settings with OPE and TPE, and to cases even without smoothness and strong convexity assumptions of the objective function. A novel contribution is to explore a generic local growth condition (or called local error bound condition) of the objective function, which specifies how fast the objective function grows in a local neighborhood of optimal solutions. In particular, we propose a generic algorithmic framework for accelerating existing SZCO algorithms in various settings by leveraging the local error bound condition. This is accomplished by a novel synthesis of existing SZCO algorithms and a multi-stage adaptive technique, which consists of three components: using a multi-stage scheme with each stage running existing algorithms, warm starting each stage using the solution from previous stage, and adaptively changing the algorithm's parameters after each stage (e.g., step size, the smoothing parameter). Depending on the local error bound (LEB) condition, the improvement over existing results is up to a factor of $1/\epsilon^2$. Empirical studies in various settings demonstrate the effectiveness of the proposed acceleration approach.

Related Work. A quick comparison between our obtained upper bounds of iteration complexities under different settings and previous upper bounds is shown in Table 1. Lower bounds for SZCO have been also studied in several works [Dani *et al.*, 2008; Shamir, 2013; Duchi *et al.*, 2015] in different settings. We will show that our proposed algorithm's performance in certain settings matches the existing lower bounds. For example, for stochastic zeroth-order *linear* optimization with OPE, our obtained upper bound of iteration complexity is $\widetilde{O}(d^2/\epsilon^2)$[1], which matches the lower bound in [Dani *et al.*, 2008]. In addition, the best upper bound in this paper for SZCO in the setting with OPE without smoothness assumption is $\widetilde{O}(d^2/\epsilon^2)$, which matches the lower bound in [Shamir, 2013] up to a logarithmic factor. It is also notable that the best upper bound achieved in this paper can be as good as $\min(O(d^2\log(1/\epsilon)), \widetilde{O}(d/\epsilon))$. However, we note that our result does not contradict to the lower bound in [Duchi *et al.*, 2015] because either their considered random functions do not necessarily have bounded gradients as assumed in this paper or their considered problem does not satisfy the LEB condition that yields our best result. Finally, we note that the LEB condition has been explored in (stochastic) convex optimization for improving the convergence of first-order methods [Yang and Lin, 2015; Xu *et al.*, 2017]. To the best of our knowledge, this is the first paper that leverages the LEB condition for improving the convergence of SZCO.

[1]We omit a poly-logarithmic factor for $\widetilde{O}(\cdot)$.

2 Notations and Preliminaries

In this section, we present some notations and preliminaries for SZCO. Let the ℓ-norm of a vector $\mathbf{x}$ be $\|\mathbf{x}\|_\ell$ (where $\ell \geq 1$). The inner product of two vectors $\mathbf{x}, \mathbf{y}$ is denoted by $\mathbf{x}^\top\mathbf{y} = \langle \mathbf{x}, \mathbf{y}\rangle$. The notation of $\mathbb{B}(\mathbf{x}, r)$ denotes a Euclidean ball centered at $\mathbf{x}$ with radius $r > 0$. The ceiling integer of a real number r is $\lceil r \rceil$.

Let $\partial f(\mathbf{x})$ and $\nabla f(\mathbf{x})$ denote, respectively, the subgradient of a non-smooth function and the gradient of a smooth function. $f(\mathbf{x})$ is G-Lipschitz continuous if $\exists G > 0$ such that $|f(\mathbf{x}) - f(\mathbf{y})| \leq G\|\mathbf{x}-\mathbf{y}\|_2, \forall \mathbf{x}, \mathbf{y} \in \Omega$, i.e., $\|\partial f(\mathbf{x})\|_2 \leq G, \forall \mathbf{x} \in \Omega$. $f(\mathbf{x})$ is L-smooth if it is differentiable and has L-Lipschitz-continuous gradient, i.e., $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x}-\mathbf{y}\|_2, \forall \mathbf{x}, \mathbf{y} \in \Omega$. $f(\mathbf{x})$ is convex if $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \partial f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle, \forall \mathbf{x}, \mathbf{y} \in \Omega$. $f(\mathbf{x})$ is σ-strongly convex if $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \partial f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \sigma\|\mathbf{x}-\mathbf{y}\|_2^2/2, \forall \mathbf{x}, \mathbf{y} \in \Omega$ and $\sigma \geq 0$.

Let $\mathbf{u} \sim \mathbb{B}(\mathbf{0}, 1)$ denote a noise vector uniformly sampled from a unit sphere, and $\mathbf{u} \sim \mathcal{N}(0, 1)$ denote a noise vector sampled from a standard Gaussian distribution. Given a noise vector $\mathbf{u}$, let $\hat{f}(\mathbf{x}; \xi) \triangleq \mathbb{E}_{\mathbf{u}}[f(\mathbf{x} + \delta\mathbf{u}; \xi)]$ denote a smoothed function with smoothing parameter $\delta > 0$ and $\hat{f}(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{u};\xi}[f(\mathbf{x}+\delta\mathbf{u}; \xi)]$. Let Ω_* denote the optimal solution set for Eq. (1), and $f_* \triangleq \min_{\mathbf{x}\in\Omega} f(\mathbf{x})$. In the sequel, we will make the following assumption.

Assumption 1. *Assume that there exist $\mathbf{x}_0 \in \Omega$ and $\epsilon_0 > 0$ such that $f(\mathbf{x}_0) - \min_{\mathbf{x}\in\Omega} f(\mathbf{x}) \leq \epsilon_0$. For any $\delta \in (0, +\infty)$, there exists $B > 0$ such that $|f(\mathbf{x} + \delta\mathbf{u}; \xi)| \leq B$ for any $\mathbf{x} \in \Omega$ and ξ, where $\mathbf{u} \sim \mathbb{B}(\mathbf{0}, 1)$.*

2.1 Noisy Gradient Estimators

The noisy gradient estimator in the setting with OPE proposed by [Flaxman *et al.*, 2005] is given as:

$$\mathbf{g}_t^{\mathrm{f}} = \frac{d}{\delta} f(\mathbf{x}_t + \delta\mathbf{u}_t; \xi_t)\mathbf{u}_t, \tag{2}$$

where $\mathbf{u}_t \sim \mathbb{B}(\mathbf{0}, 1)$ and $\delta > 0$. The properties of $\mathbf{g}_t^{\mathrm{f}}$ and $\hat{f}(\mathbf{x}; \xi)$ are stated below.

Lemma 1 ([Flaxman *et al.*, 2005])**.** *Given $\mathbf{u} \sim \mathbb{B}(\mathbf{0}, 1)$, we have $\mathbb{E}_{\mathbf{u}}[\mathbf{g}_t^f] = \nabla\hat{f}(\mathbf{x}_t; \xi_t)$, and $\|\mathbf{g}_t^f\|_2 \leq dB/\delta$. If $f(\mathbf{x}; \xi)$ is G-Lipschitz continuous, we have $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \leq G\delta$. If $f(\mathbf{x}; \xi)$ is L-smooth, we have $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \leq L\delta^2/2$.*

For the setting with TPE, there are different gradient estimators used in previous studies. For example, [Agarwal *et al.*, 2010; Shamir, 2017] used the following noisy gradient estimator with $\mathbf{u}_t \sim \mathbb{B}(\mathbf{0}, 1)$:

$$\mathbf{g}_t^{\mathrm{a}} = \frac{d}{2\delta}\big(f(\mathbf{x}_t + \delta\mathbf{u}_t; \xi_t) - f(\mathbf{x}_t - \delta\mathbf{u}_t; \xi_t)\big)\mathbf{u}_t. \tag{3}$$

[Nesterov and Spokoiny, 2017; Duchi *et al.*, 2015] considered the following noisy gradient estimator for TPE with $\mathbf{u}_t \sim \mathcal{N}(0, 1)$:

$$\mathbf{g}_t^{\mathrm{n}} = \frac{1}{\delta}(f(\mathbf{x}_t + \delta\mathbf{u}_t; \xi_t) - f(\mathbf{x}_t; \xi_t))\mathbf{u}_t. \tag{4}$$

Table 1: A comparison between our results and existing works for SZCO in the settings of OPE and TPE. LC: Lipschitz Continuous, SC: Strong Convexity, SM: SMoothness, and LEB: Local Error Bound.

setting	algorithm	assumption	iteration complexity	high probability or expectation
OPE	[Flaxman *et al.*, 2005]	LC	$O\left(\frac{d^2}{\epsilon^4}\right)$	expectation
	[Agarwal *et al.*, 2010]	LC + SC	$\widetilde{O}\left(\frac{d^2}{\epsilon^3}\right)$	expectation
		LC + SC + SM	$\widetilde{O}\left(\frac{d^2}{\epsilon^2}\right)$	expectation
	our work	LC + LEB	$\widetilde{O}\left(\frac{d^2}{\epsilon^{2(2-\theta)}}\right), \theta \in (0, \frac{1}{2}]$	expectation
			$\widetilde{O}\left(\frac{d^2}{\epsilon^{2(2-\theta)}}\right), \theta \in (0, 1]$	high probability
	our work	LC + LEB + SM	$\widetilde{O}\left(\frac{d^2}{\epsilon^{3-2\theta}}\right), \theta \in (0, \frac{1}{2}]$	expectation
			$\widetilde{O}\left(\frac{d^2}{\epsilon^{3-2\theta}}\right), \theta \in (0, 1]$	high probability
TPE	[Agarwal *et al.*, 2010]	LC	$O\left(\frac{d^2}{\epsilon^2}\right)$	high probability
		LC + SC	$\widetilde{O}\left(\frac{d^2}{\epsilon}\right)$	high probability
	[Nesterov and Spokoiny, 2017]	LC	$\widetilde{O}\left(\frac{d^2}{\epsilon^2}\right)$	expectation
		LC + SM	$O\left(\frac{d}{\epsilon^2}\right)$	expectation
	[Duchi *et al.*, 2015]	LC	$\widetilde{O}\left(\frac{d \log d}{\epsilon^2}\right)$	expectation
		LC + SM	$O\left(\frac{d}{\epsilon^2}\right)$	expectation
	[Shamir, 2017]	LC	$O\left(\frac{d}{\epsilon^2}\right)$	expectation
	our work	LC + LEB	$\widetilde{O}\left(\frac{d^2}{\epsilon^{2(1-\theta)}}\right), \theta \in (0, 1]$	high probability
	our work	LC + LEB	$\widetilde{O}\left(\frac{d}{\epsilon^{2(1-\theta)}}\right), \theta \in (0, \frac{1}{2}]$	expectation

The properties of estimators in Eqs. (3) and (4) are summarized as below.

Lemma 2 ([Agarwal *et al.*, 2010; Shamir, 2017])**.** *Given $\mathbf{u} \sim \mathbb{B}(\mathbf{0}, 1)$, we have $\mathbb{E}_{\mathbf{u}}[\mathbf{g}_t^a] = \nabla \hat{f}(\mathbf{x}_t; \xi_t)$. If $f(\mathbf{x}; \xi)$ is G-Lipschitz continuous, we have $\|\mathbf{g}_t^a\|_2 \leq Gd$, $\mathbb{E}_{\mathbf{u}}[\|\mathbf{g}_t^a\|_2^2] \leq db^2G^2C$, and $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \leq G\delta$, where C is a universal constant and b is a constant such that $(\mathbb{E}[\|\mathbf{u}\|_2^4])^{1/4} \leq b$. If $f(\mathbf{x}; \xi)$ is L-smooth, we have $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \leq L\delta^2/2$.*

Lemma 3 ([Nesterov and Spokoiny, 2017])**.** *Considering $\mathbf{u} \sim \mathcal{N}(0, 1)$, we have $\mathbb{E}_{\mathbf{u}}[\mathbf{g}_t^n] = \nabla \hat{f}(\mathbf{x}_t; \xi_t)$. If $f(\mathbf{x}; \xi)$ is G-Lipschitz continuous, we have $\mathbb{E}_{\mathbf{u}}[\|\mathbf{g}_t^n\|_2^2] \leq G^2(d+4)^2$, and $|f(\mathbf{x}; \xi_t) - \hat{f}(\mathbf{x}; \xi_t)| \leq \delta G d^{1/2}$. If $f(\mathbf{x}; \xi)$ is G-Lipschitz continuous and L-smooth, we have $\mathbb{E}_{\mathbf{u}}[\|\mathbf{g}_t^n\|_2^2] \leq \delta^2(d+6)^3L^2/2 + 2(d+4)G^2$, and $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \leq \delta^2 Ld/2$.*

Remark 1: The absolute upper bound of the noisy gradient estimators is needed for high probability analysis and the variance bound of the noisy gradient estimators is useful for expectational convergence analysis.

The iterative update in the previous studies takes the following form:

$$\mathbf{x}_{t+1} = \Pi_{\Omega}[\mathbf{x}_t - \eta \mathbf{g}_t], \tag{5}$$

where $\eta > 0$ is a step size, $\mathbf{g}_t$ is a gradient estimator and Π_{Ω} denotes the Euclidean projection onto the set Ω. We synthesize the convergence analysis of stochastic optimization in the following proposition, which, combined with properties of different gradient estimators, yields corresponding convergence results in previous studies.

Proposition 1. *Considering the update in Eq. (5) with an initial point of $\mathbf{x}_1 \in \Omega$, for any $\mathbf{x} \in \Omega$, we have*

$$\sum_{t=1}^{T} f(\mathbf{x}_t; \xi_t) - f(\mathbf{x}; \xi_t) \leq 2 \sum_{t=1}^{T} \sup_{\mathbf{x} \in \Omega} |f(\mathbf{x}; \xi_t) - \hat{f}(\mathbf{x}; \xi_t)|$$
$$+ \sum_{t=1}^{T} \mathbf{g}_t^{\top}(\mathbf{x}_t - \mathbf{x}) + (\nabla \hat{f}(\mathbf{x}_t; \xi_t) - \mathbf{g}_t)^{\top}(\mathbf{x}_t - \mathbf{x}),$$

and $\sum_{t=1}^{T} \mathbf{g}_t^{\top}(\mathbf{x}_t - \mathbf{x}) \leq \frac{\|\mathbf{x}_1 - \mathbf{x}\|_2^2}{2\eta} + \sum_{t=1}^{T} \frac{\eta \|\mathbf{g}_t\|_2^2}{2}$.

2.2 Local Error Bound (LEB) Condition

Definition 1. *A problem of Eq. (1) satisfies the LEB condition on a compact set Ω if there exist $\theta \in (0, 1]$ and $c > 0$ such that for any $\mathbf{x} \in \Omega$*

$$dist(\mathbf{x}, \Omega_*) \leq c(f(\mathbf{x}) - \min_{\mathbf{x} \in \Omega} f(\mathbf{x}))^{\theta}, \tag{6}$$

where dist$(\mathbf{x}, \Omega_*) \triangleq \min_{\mathbf{v}\in\Omega_*} \|\mathbf{v} - \mathbf{x}\|_2$.

Note that the LEB condition has been studied thoroughly in [Yang and Lin, 2015; Bolte *et al.*, 2015; Xu *et al.*, 2017]. It is satisfied for a broad family of problems. For example, when $f(\mathbf{x})$ is continuous and semi-algebraic (or sub-analytic), the LEB condition holds on any compact set [Bolte *et al.*, 2015]. Below, we consider several instances of problems that satisfy the LEB condition. More interesting examples in machine learning can be found in [Yang and Lin, 2015; Xu *et al.*, 2017].

Example 1: When $f(\mathbf{x};\xi) = \mathbf{x}^\top\xi$ is a linear function and Ω is a polyhedral set (e.g., hypercube), then the problem of Eq. (1) satisfies the LEB with $\theta = 1$ [Yang and Lin, 2015]. These functions are considered in online bandit linear optimization [Dani *et al.*, 2008]. More generally, if $f(\mathbf{x})$ is a polyhedral function and Ω is a polyhedral set, then LEB with $\theta = 1$ holds [Yang and Lin, 2015]. For instance, $f(\mathbf{x}) = \sum_{i=1}^n |\mathbf{a}_i^\top\mathbf{x} - b_i|/n$ and $\Omega = \{\|\mathbf{x}\|_1 \leq s\}$.

Example 2: When $f(\mathbf{x})$ is strongly convex, then the LEB condition holds with $\theta = 1/2$ [Xu *et al.*, 2017].

Example 3: Even when $f(\mathbf{x})$ is not strongly convex, the LEB condition with $\theta = 1/2$ may still hold, such as $f(\mathbf{x}) = \sum_{i=1}^n (\mathbf{a}_i^\top\mathbf{x} - b_i)^2/n$ and Ω is a polyhedral set.

3 Our Generic Approach

In this section, we propose a generic algorithm for accelerating the convergence of SZCO and its main results in various settings. In order to achieve improved high probability convergence results, we need to use the following update to control the last term in Proposition 1:

$$\mathbf{x}_{t+1} = \Pi_{\mathbb{D}}[\mathbf{x}_t - \eta\mathbf{g}_t], \tag{7}$$

where $\mathbb{D} = \Omega\cap\mathbb{B}(\mathbf{x}^1, D)$ with $\mathbf{x}^1$ being a reference point and D being the radius of the ball. The proposed acceleration framework is presented in Algorithm 1, which is a multi-stage adaptive scheme consisting of three key components: (i) a multi-stage scheme with each stage running existing updates, (ii) warm starting each stage using the solution from previous stage, and (iii) adaptively changing the parameters after each stage. Next, we present the iteration complexities of Algorithm 1 in various settings. Let $\epsilon_k = \epsilon_0/2^k$ be a sequence.

Theorem 1 (Results for OPE)**.** *Let Algorithm 1 run with Eq. (2) as the noisy gradient estimator and* $K = \lceil\log_2(\epsilon_0/\epsilon)\rceil$ *stages. We have the following results.*

- *R-I: if $f(\mathbf{x};\xi)$ is G-Lipschitz continuous, by employing Eq. (5) and setting $t = O(d^2/\epsilon^{2(2-\theta)})$, $\delta_k = \epsilon_k/(6G)$, $\eta_k = \epsilon_k^3/(54G^2d^2B^2)$, then Algorithm 1 enjoys an iteration complexity of $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$ in expectation for problems satisfy the LEB condition with $\theta \in (0, 1/2]$;*
- *R-II: if $f(\mathbf{x};\xi)$ is G-Lipschitz continuous and L-smooth, by employing Eq. (5) and setting $t = O(d^2/\epsilon^{3-2\theta})$, $\delta_k = \sqrt{\epsilon_k}/(\sqrt{3L})$, $\eta_k = 2\epsilon_k^2/(9Ld^2B^2)$, then Algorithm 1 enjoys an iteration complexity of $\widetilde{O}\big(d^2/\epsilon^{3-2\theta}\big)$ in expectation for problems satisfy the LEB condition with $\theta \in (0, 1/2]$;*
- *R-III: if $f(\mathbf{x};\xi)$ is G-Lipschitz continuous, by employing Eq. (7) and setting δ_k, η_k similarly as in R-I and $t = \widetilde{O}(d^2/\epsilon^{2(2-\theta)}), D_k = O(\epsilon_{k-1}^\theta)$, then Algorithm 1 enjoys an iteration complexity of $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$ with high probability $1-p$, where we set $p \in (0,1)$, for problems satisfy the LEB condition with $\theta \in (0, 1]$;*
- *R-IV: if $f(\mathbf{x};\xi)$ is G-Lipschitz continuous and L-smooth, by employing Eq. (7) and setting δ_k, η_k similarly as in R-II and $t = \widetilde{O}(d^2/\epsilon^{3-2\theta}), D_k = O(\epsilon_{k-1}^\theta)$, then Algorithm 1 enjoys an iteration complexity of $\widetilde{O}(d^2/\epsilon^{3-2\theta})$ with high probability $1-p$, where we set $p \in (0,1)$, for problems satisfy the LEB condition with $\theta \in (0, 1]$.*

Algorithm 1 A generic approach for accelerating SZCO

initialization $\mathbf{x}_0$, K, η_1, δ_1, D_1
for $k = 1, \cdots, K$ **do**
  $\mathbf{x}_k^1 = \mathbf{x}_{k-1}$, $\mathbb{D}_k = \Omega\cap\mathbb{B}(\mathbf{x}_k^1, D_k)$
  for $\tau = 1, \cdots, t$ **do**
    compute a gradient estimator in light of Eq. (2) or Eq. (3) or Eq. (4)
    compute $\mathbf{x}_k^\tau$ according to Eq. (5) or Eq. (7) with a step size η_k, a parameter δ_k, and a domain $\mathbb{D}_k$
  end for
  let $\mathbf{x}_k = \sum_{\tau=1}^t \mathbf{x}_k^\tau/t$
  update δ_{k+1}, D_{k+1} and η_{k+1}
end for
return $\mathbf{x}_K$

Remark 2: For the statement of high probability results, we omit a poly-logarithmic factor of $\log(K/p)$ in t. Our iteration complexities by leveraging the LEB condition are better than those in [Agarwal *et al.*, 2010; Flaxman *et al.*, 2005]. For LEB with $\theta = 1/2$ that is weaker than the strong convexity assumption, our iteration complexities match that in [Agarwal *et al.*, 2010] for strongly convex functions. For problems with $f(\mathbf{x};\xi)$ being a linear function and Ω being a polyhedral set, the LEB with $\theta = 1$ holds and we achieve an iteration complexity of $\widetilde{O}\big(d^2/\epsilon^2\big)$ with high probability, which matches the lower bound in [Dani *et al.*, 2008]. Besides, one may get expectational results for $\theta > 1/2$ from high probability results R-III and R-IV following the Corollary 3 in [Xu *et al.*, 2016].

Theorem 2 (Results for TPE)**.** *Let Algorithm 1 run with* $K = \lceil\log_2(\epsilon_0/\epsilon)\rceil$ *stages. We have the following results.*

- *R-I: if $f(\mathbf{x};\xi)$ is G-Lipschitz continuous, by employing the noisy gradient estimator of Eq. (3) and*

the update of Eq. (5) and setting $t = O(d/\epsilon^{2(1-\theta)})$, $\delta_k = \epsilon_k/(6G)$, $\eta_k = 2\epsilon_k/(3db^2G^2C)$ *where* b *and* C *are parameters discussed in Lemma 2, then Algorithm 1 enjoys an iteration complexity of* $\widetilde{O}(d/\epsilon^{2(1-\theta)})$ *in expectation for problems satisfy the LEB condition with* $\theta \in (0, 1/2]$*;*

- *R-II: if* $f(\mathbf{x};\xi)$ *is* G*-Lipschitz continuous and* L*-smooth, by employing the noisy gradient estimator of Eq. (4) and the update of Eq. (5) and setting* $t = O(d/\epsilon^{2(1-\theta)})$, $\delta_k = \sqrt{\epsilon_k}/(2\sqrt{dL})$, $\eta_k = \min\{\epsilon_k/(4(d+4)G^2), 2d/((d+6)^3L)\}$, *then Algorithm 1 enjoys an iteration complexity of* $\widetilde{O}(d/\epsilon^{2(1-\theta)})$ *in expectation for problems satisfy the LEB condition with* $\theta \in (0, 1/2]$*;*
- *R-III: if* $f(\mathbf{x};\xi)$ *is* G*-Lipschitz continuous, by employing the noisy gradient estimator of Eq. (3) and the update of Eq. (7) and setting* $\delta_k = \epsilon_k/(8G)$, $\eta_k = \epsilon_k/(2d^2G^2)$, $t = \widetilde{O}(d^2/\epsilon^{2(1-\theta)})$, *and* $D_k = O(\epsilon_{k-1}^{\theta})$, *then Algorithm 1 enjoys an iteration complexity of* $\widetilde{O}(d^2/\epsilon^{2(1-\theta)})$ *with high probability* $1-p$, *where we set* $p \in (0, 1)$, *for problems satisfy the LEB condition with* $\theta \in (0, 1]$.

Remark 3: It is notable that in the setting with TPE, the smoothness of the random function does not improve the convergence (by comparing R-I and R-II). The reason is that, for R-I, we utilize the refined analysis in [Shamir, 2017] to bound the variance of the noisy gradient estimator $\mathbb{E}[\|\mathbf{g}_t^{\mathrm{a}}\|_2^2] \leq O(d)$ (see Lemma 2), which is in the same order to that of the noisy gradient estimator $\mathbf{g}_t^{\mathrm{n}}$ with small enough δ as established in [Nesterov and Spokoiny, 2017] (see Lemma 3). The expectational results R-I and R-II have better dependence on d compared to the high probability result R-III. The reason is that, for high probability analysis, we have to use the absolute bound of $\mathbf{g}_t^{\mathrm{a}}$. However, the expectational results R-I and R-II cannot enjoy better dependence on ϵ for $\theta > 1/2$ as in the high probability result R-III. We notice that one can obtain similar expectational results for $\theta > 1/2$ in light of R-III with the same technique in Remark 2.

Finally, we would like to point out that although the above results require knowing the value of θ in the LEB condition, we can use another level of restarting on top of Algorithm 1 and an increasing sequence of t for the outer loop similar to that in [Yang and Lin, 2015; Xu *et al.*, 2017], which still enjoy improved iteration complexities compared with previous results. Due to limitation of space, this result and the related proofs are both omitted here.

Convergence Analyses. Due to limitation of space, we present proofs of results R-I and R-III in Theorem 1. However, we note that proofs of other results can be simply derived by using different variance bounds of the noisy gradient estimators and different bounds of $|f(\mathbf{x};\xi) - \hat{f}(\mathbf{x};\xi)|$ from Lemmas 1-3.

Proof of R-I. Based on Proposition 1 and Lemma 1, we have

$$\sum_{t=1}^{T} f(\mathbf{x}_t;\xi_t) - f(\mathbf{x};\xi_t) \leq 2TG\delta + \frac{\eta T d^2 B^2}{2\delta^2} + \frac{\|\mathbf{x}_1 - \mathbf{x}\|_2^2}{2\eta} + \sum_{t=1}^{T} (\nabla \hat{f}(\mathbf{x}_t;\xi_t) - \mathbf{g}_t^{\mathrm{f}})^\top (\mathbf{x}_t - \mathbf{x}).$$

By setting $\hat{\mathbf{x}}_T = \sum_{t=1}^{T} \mathbf{x}_t/T$ and taking the expectation over randomness in $\mathbf{u}$ and ξ, we have

$$\mathbb{E}[f(\hat{\mathbf{x}}_T) - f(\mathbf{x})] \leq \frac{\mathbb{E}[\|\mathbf{x}_1 - \mathbf{x}\|_2^2]}{2\eta T} + \frac{\eta d^2 B^2}{2\delta^2} + 2G\delta.$$

By adopting the generic framework in Algorithm 1, for the k-th stage, we have

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x})] \leq \frac{\mathbb{E}[\|\mathbf{x}_{k-1} - \mathbf{x}\|_2^2]}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k,$$

where we use t iterations in inner loops of Algorithm 1.

We will prove by induction that $\mathbb{E}[f(\mathbf{x}_k) - f_*] \leq \epsilon_k$. It is trivial for $k = 0$ due to Assumption 1. Conditioned on the inequality of $\mathbb{E}[f(\mathbf{x}_{k-1}) - f_*] \leq \epsilon_{k-1}$, we will show that $\mathbb{E}[f(\mathbf{x}_k) - f_*] \leq \epsilon_k$. Let $\mathbf{x}_{k-1,*} = \arg\min_{\mathbf{v}\in\Omega_*} \|\mathbf{v} - \mathbf{x}_{k-1}\|_2$. Then, we have

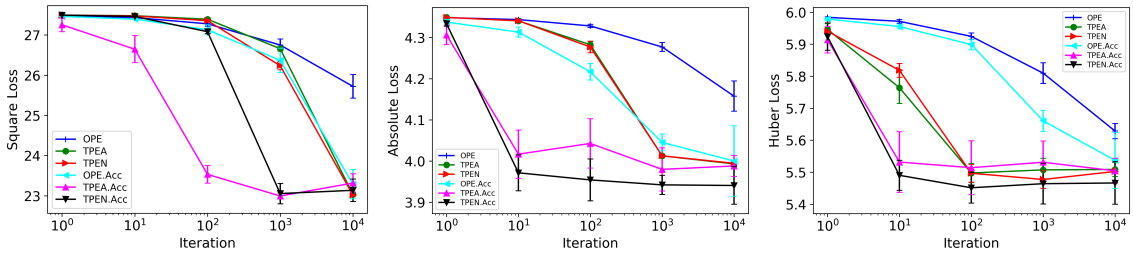
$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_{k-1,*})] \\
&\leq \frac{\mathbb{E}[\|\mathbf{x}_{k-1} - \mathbf{x}_{k-1,*}\|_2^2]}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k \\
&\leq \frac{c(\mathbb{E}[f(\mathbf{x}_{k-1}) - f(\mathbf{x}_{k-1,*})])^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k \\
&\leq \frac{c\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k,
\end{aligned}$$

where the second inequality uses the concavity of $(s)^{2\theta}$ for $\theta \leq 1/2$ and the Jensen's inequality. To establish $\mathbb{E}[f(\mathbf{x}_k) - f_*] \leq \epsilon_k$, we set

$$\begin{aligned}
&\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} \leq \frac{\epsilon_{k-1}}{6} \Rightarrow t \geq \frac{1296 d^2 B^2 G^2 c^2}{\epsilon_{k-1}^{2(2-\theta)}}, \\
&\frac{\eta_k d^2 B^2}{2\delta_k^2} \leq \frac{\epsilon_k}{3} \Rightarrow \eta_k \leq \frac{\epsilon_k^3}{54 G^2 d^2 B^2}, \\
&2G\delta_k \leq \frac{\epsilon_k}{3} \Rightarrow \delta_k \leq \frac{\epsilon_k}{6G}.
\end{aligned}$$

By setting $\epsilon_K = \epsilon_0/2^K = \epsilon$, we have $K = \lceil \log(\epsilon_0/\epsilon) \rceil$. Thus, we have $\mathbb{E}[f(\mathbf{x}_K) - f_*] \leq \epsilon_K \leq \epsilon$. As a result, the total iteration complexity is $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$.

Proof of R-III. First, $\sum_{t=1}^{T}(\hat{f}(\mathbf{x}_t) - \hat{f}(\mathbf{x})) \leq \sum_{t=1}^{T}\langle \mathbf{g}_t^{\mathrm{f}}, (\mathbf{x}_t - \mathbf{x})\rangle + \sum_{t=1}^{T}(\nabla\hat{f}(\mathbf{x}_t) - \mathbf{g}_t^{\mathrm{f}})^\top(\mathbf{x}_t - \mathbf{x})$. We can use the result in Proposition 1 and the absolute upper bound of the noisy gradient estimator to bound the first term in R.H.S. The second term can be bounded using martingale inequalities (please refer to Lemma 14 in [Hazan and Kale, 2014]). As a result, for any fixed $\mathbf{x} \in \Omega \cap \mathbb{B}(\mathbf{x}_1, D)$ and $\tilde{p} \in (0, 1)$, with a probability at



(a) $f(\mathbf{x}) = \frac{\sum_{i=1}^{N}(\mathbf{w}_i^\top \mathbf{x} - r_i)^2}{N}$, $\theta = 0.5$ (b) $f(\mathbf{x}) = \frac{\sum_{i=1}^{N}|\mathbf{w}_i^\top \mathbf{x} - r_i|}{N}$, $\theta = 1$ (c) $f(\mathbf{x})$ is averaged Huber loss, $\theta = 1$

Figure 1: Comparisons of convergence with three objective functions for music recommendation competition data and $T = 10^4$.

least $1 - \tilde{p}$, the following inequality holds

$$\hat{f}(\hat{\mathbf{x}}_T) - \hat{f}(\mathbf{x}) \leq \frac{\|\mathbf{x}_1 - \mathbf{x}\|_2^2}{2\eta T} + \frac{\eta d^2 B^2}{2\delta^2} + \frac{4dBD\sqrt{3\log(\frac{1}{\tilde{p}})}}{\sqrt{T}\delta},$$

where $\hat{\mathbf{x}}_T = \sum_{t=1}^{T} \mathbf{x}_t / T$. Then, we are ready to have

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k-1,*}) \leq \frac{c^2 \epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + \frac{4dBc\epsilon_{k-1}^{\theta}\sqrt{3\log(\frac{1}{\tilde{p}})}}{\sqrt{t}\delta_k} + 2G\delta_k,$$

where we use $D_k = c\epsilon_{k-1}^{\theta}$, and $\mathbf{x}_{k-1,*} \in \mathbb{B}(\mathbf{x}_{k-1}, D_k)$. We can easily establish $f(\mathbf{x}_k) - f_* \leq \epsilon_k$ with high probability by induction if we set $\delta_k = O(\epsilon_k)$, $\eta_k = O(\epsilon_k^3/d^2)$, $t = O(d^2 \log(1/\tilde{p})/\epsilon^{2(2-\theta)})$. Then, with $K = \lceil \log(\epsilon_0/\epsilon) \rceil$, we have the iteration complexity as $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$ with probability of $1-p$, where $\tilde{p} = p/K$.

4 Experiments

In this section, we conduct experiments on real-world datasets in various settings to demonstrate the superior performance of the proposed acceleration approach in Algorithm 1. We run experiments in a personal computer with Intel CPU@3.70GHz and 16GB memory.

To compare the efficiency of the acceleration framework with prior methods, we will show the evolution of function values with respect to the number of iterations. We adopt three baselines: the first is OPE from [Flaxman *et al.*, 2005]; the second is TPEA from [Agarwal *et al.*, 2010]; and the third is TPEN from [Nesterov and Spokoiny, 2017]. We add a term "Acc" to denote our acceleration version for each baseline in experiments. To show experimental results, we run experiments ten times with the same initialization point, and show the average of function values. For the first experiment on real-world datasets, we also show error bars of a standard variance.

4.1 Music Recommendation Competition Data

We consider the ensemble learning setting of recommendations as a black-box optimization problem discussed in [Lian *et al.*, 2016]. In particular, we blend the existing models in [Chen *et al.*, 2011] for music recommendation competition in KDD-Cup 2011, which turns out to be a linear regression problem. Since true ratings for the test set are unknown in competition, the feedback is the evaluation of the linear regression prediction of the blended model. Thus, this ensemble learning case is SZCO.

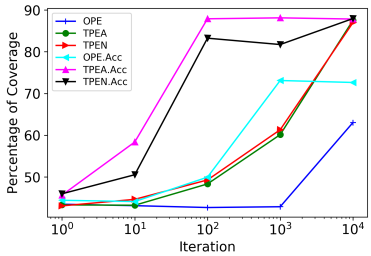


Figure 2: Growth of ceramic thin films with $T = 10^4$.

We get predicted ratings of individual models in [Chen *et al.*, 2011] for the test set in KDD-Cup 2011, with 237 models and 6,005,940 predictions for each model[2]. In addition to a square loss [Lian *et al.*, 2016], we also consider an absolute loss and a huber loss [Zadorozhnyi *et al.*, 2016] as objective functions. For better demonstrations of convergence rate, we sample 10 models from 237 models with predicted ratings denoted by $\mathbf{w} \in \mathbb{R}^{10}$ in ensemble learning, and set the number of training points as $N = 10^5$. The ground truth of sample i is denoted by r_i.

We show the superior convergence of our proposed acceleration approach with different objective functions in Figure 1. From the standard variance error bar, we clearly find that our approach stably accelerates the existing SZCO algorithms with order improvements.

4.2 Industrial Data on Ceramic Thin Films

We consider industrial data on crystallization of ceramic thin films in [Nakamura *et al.*, 2017]. The goal for the industrial application on crystallization of ceramic thin films is to determine an optimal setting for the volume of tetraethylene glycol (TEG), temperature (T), and the time of heat to a temperature in time (HTI), which is in fact a SZCO problem. The objective of the experiment is a quadratic function. For more details, please refer to [Nakamura *et al.*, 2017; Wang *et al.*, 2017].

By updating the values of TEG, T and HTI, we show the growth of ceramic thin films with the number of iterations in Figure 2. The superior performance of the acceleration via Algorithm 1 is clear. We also test different intensity of noises, and find that the acceleration

[2]We thank the authors of [Lian *et al.*, 2016] for providing the predicted ratings for us.

is robust. Note that, in ceramic thin films, we solve a concave function and thus the function value increases.

5 Conclusions

In this paper, we have developed a generic acceleration approach to solve the problem of SZCO. We tackled the SZCO problem with the core idea of exploring an LEB condition of objective functions, which is frequently encountered in real applications. The benefits of the proposed acceleration technique are three-fold: wide applicability, weak assumption and improvements on iteration complexity. With LEB condition, the best upper bound here can be $\min(O(d^2\log(1/\epsilon)), \widetilde{O}(d/\epsilon))$, and the improvement over existing results is up to a factor of $1/\epsilon^2$. Experimental results have shown superior and robust performance of the proposed acceleration approach.

Acknowledgments

The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14208815 and No. CUHK 14234416 of the General Research Fund), 2018 Microsoft Research Asia Collaborative Research Award, and National Science Foundation (NSF-1545995). Part of X. Yu's work was conducted at the University of Iowa when visiting T. Yang's research group.

References

[Agarwal *et al.*, 2010] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40, 2010.

[Bolte *et al.*, 2015] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce Suter. From error bounds to the complexity of first-order descent methods for convex functions. *CoRR*, abs/1510.08234, 2015.

[Chen *et al.*, 2011] Po-Lung Chen, Chen-Tse Tsai, Yao-Nan Chen, Ku-Chun Chou, Chun-Liang Li, Cheng-Hao Tsai, Kuan-Wei Wu, Yu-Cheng Chou, Chung-Yi Li, Wei-Shih Lin, et al. A linear ensemble of individual and blended models for music rating prediction. In *KDD-Cup*, pages 21–60, 2011.

[Dani *et al.*, 2008] Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *NIPS*, pages 345–352, 2008.

[Duchi *et al.*, 2015] 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.

[Flaxman *et al.*, 2005] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA*, pages 385–394, 2005.

[Hazan and Kale, 2014] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[Lian *et al.*, 2016] Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, and Ji Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. In *NIPS*, pages 3054–3062, 2016.

[Nakamura *et al.*, 2017] Nathan Nakamura, Jason Seepaul, Joseph B Kadane, and B Reeja-Jayan. Design for low-temperature microwave-assisted crystallization of ceramic thin films. *Applied Stochastic Models in Business and Industry*, 2017.

[Nemirovski *et al.*, 2009] 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.

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

[Shamir, 2013] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *COLT*, pages 3–24, 2013.

[Shamir, 2017] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

[Sokolov *et al.*, 2016] Artem Sokolov, Julia Kreutzer, Stefan Riezler, and Christopher Lo. Stochastic structured prediction under bandit feedback. In *NIPS*, pages 1489–1497, 2016.

[Wang *et al.*, 2017] Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. *arXiv preprint arXiv:1710.10551*, 2017.

[Wibisono *et al.*, 2012] Andre Wibisono, Martin J Wainwright, Michael I Jordan, and John C Duchi. Finite sample convergence rates of zero-order stochastic optimization methods. In *NIPS*, pages 1439–1447, 2012.

[Xu *et al.*, 2016] Yi Xu, Qihang Lin, and Tianbao Yang. Accelerated stochastic subgradient methods under local error bound condition. *arXiv preprint arXiv:1607.01027*, 2016.

[Xu *et al.*, 2017] Yi Xu, Qihang Lin, and Tianbao Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *ICML*, pages 3821–3830, 2017.

[Yang and Lin, 2015] Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *arXiv preprint arXiv:1512.03107*, 2015.

[Zadorozhnyi *et al.*, 2016] Oleksandr Zadorozhnyi, Gunthard Benecke, Stephan Mandt, Tobias Scheffer, and Marius Kloft. Huber-norm regularization for linear prediction models. In *ECML*, pages 714–730, 2016.

[Zinkevich, 2003] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

6 Appendix

In the appendix, we present the proofs of Proposition 1, Theorem 1 and Theorem 2.

6.1 Proof of Proposition 1

Proof. We adopt the update in Eq. (5) to find an ϵ-optimal solution for SZCO. After T iterations, we have

$$\sum_{t=1}^{T}\hat{f}(\mathbf{x}_t;\xi_t)-\hat{f}(\mathbf{x};\xi_t)\leq\sum_{t=1}^{T}\nabla\hat{f}(\mathbf{x}_t;\xi_t)^\top(\mathbf{x}_t-\mathbf{x})$$
$$\leq\sum_{t=1}^{T}(\nabla\hat{f}(\mathbf{x}_t;\xi_t)-\mathbf{g}_t)^\top(\mathbf{x}_t-\mathbf{x})+\sum_{t=1}^{T}\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x})$$
$$\leq\sum_{t=1}^{T}(\nabla\hat{f}(\mathbf{x}_t;\xi_t)-\mathbf{g}_t)^\top(\mathbf{x}_t-\mathbf{x})+$$
$$\sum_{t=1}^{T}\frac{\|\mathbf{x}-\mathbf{x}_t\|_2^2-\|\mathbf{x}-\mathbf{x}_{t+1}\|_2^2}{2\eta}+\sum_{t=1}^{T}\frac{\eta\|\mathbf{g}_t\|_2^2}{2},$$

where $\mathbf{x}\in\Omega$, η is the learning rate, and the last inequality is due to [Zinkevich, 2003]. By taking the upper bound between $f(\mathbf{x};\xi_t)$ and $\hat{f}(\mathbf{x};\xi_t)$, we have

$$\sum_{t=1}^{T}f(\mathbf{x}_t;\xi_t)-f(\mathbf{x};\xi_t)\leq\sum_{t=1}^{T}\hat{f}(\mathbf{x}_t;\xi_t)-\hat{f}(\mathbf{x};\xi_t)+$$
$$2\sum_{t=1}^{T}\sup_{\mathbf{x}\in\Omega}|f(\mathbf{x};\xi_t)-\hat{f}(\mathbf{x};\xi_t)|.$$

Thus, we are ready to have

$$\sum_{t=1}^{T}f(\mathbf{x}_t;\xi_t)-f(\mathbf{x};\xi_t)\leq 2\sum_{t=1}^{T}\sup_{\mathbf{x}\in\Omega}|f(\mathbf{x};\xi_t)-\hat{f}(\mathbf{x};\xi_t)|+$$
$$\frac{\|\mathbf{x}_1-\mathbf{x}\|_2^2}{2\eta}+\sum_{t=1}^{T}\frac{\eta\|\mathbf{g}_t\|_2^2}{2}+(\nabla\hat{f}(\mathbf{x}_t;\xi_t)-\mathbf{g}_t)^\top(\mathbf{x}_t-\mathbf{x}),$$

where $\mathbf{x}\in\Omega$. □

6.2 Proof of Theorem 1

Proof. We present the proof of the theorem based on different conditions as follows. For expectational results, we adopt the update in Eq. (5). For high probability results, we adopt the update in Eq. (7).

i) Proof of R-I: Expectational results when $f(\mathbf{x};\xi)$ is G-Lipschitz continuous.

If $f(\mathbf{x};\xi)$ is G-Lipschitz continuous, based on Proposition 1 and Lemma 1, we have

$$\sum_{t=1}^{T}f(\mathbf{x}_t;\xi_t)-f(\mathbf{x};\xi_t)\leq 2TG\delta+\frac{\eta Td^2B^2}{2\delta^2}$$
$$\frac{\|\mathbf{x}_1-\mathbf{x}\|_2^2}{2\eta}+\sum_{t=1}^{T}(\nabla\hat{f}(\mathbf{x}_t;\xi_t)-\mathbf{g}_t^{\mathrm{f}})^\top(\mathbf{x}_t-\mathbf{x}).$$

By setting $\hat{\mathbf{x}}_T=\sum_{t=1}^{T}\mathbf{x}_t/T$ and taking the expectation over randomness in $\mathbf{u}$ and ξ, we have

$$\mathbb{E}[f(\hat{\mathbf{x}}_T)-f(\mathbf{x})]\leq\frac{\mathbb{E}[\|\mathbf{x}_1-\mathbf{x}\|_2^2]}{2\eta T}+\frac{\eta d^2B^2}{2\delta^2}+2G\delta.$$

By adopting the generic framework in Algorithm 1, for the k-th stage, we have

$$\mathbb{E}[f(\mathbf{x}_k)-f(\mathbf{x})]\leq\frac{\mathbb{E}[\|\mathbf{x}_{k-1}-\mathbf{x}\|_2^2]}{2\eta_k t}+\frac{\eta_k d^2B^2}{2\delta_k^2}+2G\delta_k,$$

where we use t iterations in inner loops of Algorithm 1.

For $\theta\in(0,1/2]$, based on Jensen's inequality, we have $\mathbb{E}[\|\mathbf{x}-\mathbf{x}_*\|_2^2]\leq c^2\mathbb{E}[(f(\mathbf{x})-f_*)^{2\theta}]\leq c^2(\mathbb{E}[f(\mathbf{x})-f_*])^{2\theta}$, with $\mathbf{x}_*\in\Omega_*$ and $\mathbf{x}\in\Omega$. Note that here we adopt the LEB condition over Ω. Then, we show that $\mathbb{E}[f(\mathbf{x}_k)-f_*]\leq\epsilon_k$ holds by induction, where $\epsilon_k=\epsilon_0/2^k$.

If $k=0$, we clearly have $\mathbb{E}[f(\mathbf{x}_0)-f_*]\leq\epsilon_0$. Conditioned on the inequality of $\mathbb{E}[f(\mathbf{x}_{k-1})-f_*]\leq\epsilon_{k-1}$, we will show that $\mathbb{E}[f(\mathbf{x}_k)-f_*]\leq\epsilon_k$.

Let $\mathbf{x}_{k-1,*}=\arg\min_{\mathbf{v}\in\Omega_*}\|\mathbf{v}-\mathbf{x}_{k-1}\|_2$. We have

$$\mathbb{E}[f(\mathbf{x}_k)-f(\mathbf{x}_{k-1,*})]$$
$$\leq\frac{\mathbb{E}[\|\mathbf{x}_{k-1}-\mathbf{x}_{k-1,*}\|_2^2]}{2\eta_k t}+\frac{\eta_k d^2B^2}{2\delta_k^2}+2G\delta_k$$
$$\leq\frac{c(\mathbb{E}[f(\mathbf{x}_{k-1})-f(\mathbf{x}_{k-1,*})])^{2\theta}}{2\eta_k t}+\frac{\eta_k d^2B^2}{2\delta_k^2}+2G\delta_k$$
$$\leq\frac{c\epsilon_{k-1}^{2\theta}}{2\eta_k t}+\frac{\eta_k d^2B^2}{2\delta_k^2}+2G\delta_k.$$

To establish $\mathbb{E}[f(\mathbf{x}_k)-f_*]\leq\epsilon_k$, we set

$$\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t}\leq\frac{\epsilon_{k-1}}{6}\Rightarrow t\geq\frac{1296d^2B^2G^2c^2}{\epsilon_{k-1}^{2(2-\theta)}},$$
$$\frac{\eta_k d^2B^2}{2\delta_k^2}\leq\frac{\epsilon_k}{3}\Rightarrow\eta_k\leq\frac{\epsilon_k^3}{54G^2d^2B^2},$$
$$2G\delta_k\leq\frac{\epsilon_k}{3}\Rightarrow\delta_k\leq\frac{\epsilon_k}{6G}.$$

By setting $\epsilon_K=\epsilon_0/2^K=\epsilon$, we have $K=\lceil\log(\epsilon_0/\epsilon)\rceil$. Thus, we have $\mathbb{E}[f(\mathbf{x}_K)-f_*]\leq\epsilon_K=\epsilon$. As a result, the total iteration complexity is $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$.

ii) Proof of R-II: Expectational results when $f(\mathbf{x};\xi)$ is G-Lipschitz continuous and L-smooth.

If $f(\mathbf{x};\xi)$ is L-smooth, with Lemma 1, we have

$$\mathbb{E}[f(\mathbf{x}_k)-f(\mathbf{x}_{k-1,*})]\leq\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t}+\frac{\eta_k d^2B^2}{2\delta_k^2}+L\delta_k^2.$$

To establish $\mathbb{E}[f(\mathbf{x}_k)-f_*]\leq\epsilon_k$, we set

$$\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t}\leq\frac{\epsilon_{k-1}}{6}\Rightarrow t\geq\frac{54d^2B^2Lc^2}{\epsilon_{k-1}^{3-2\theta}},$$
$$\frac{\eta_k d^2B^2}{2\delta_k^2}\leq\frac{\epsilon_k}{3}\Rightarrow\eta_k\leq\frac{2\epsilon_k^2}{9Ld^2B^2},$$
$$L\delta_k^2\leq\frac{\epsilon_k}{3}\Rightarrow\delta_k\leq\frac{\sqrt{\epsilon_k}}{\sqrt{3L}}.$$

Thus, with $K=\lceil\log(\epsilon_0/\epsilon)\rceil$, the total iteration complex-

ity is $\widetilde{O}(d^2/\epsilon^{3-2\theta})$.

iii) Proof of R-III: High probability results when $f(\mathbf{x};\xi)$ is G-Lipschitz continuous.

We prove high probability convergence for $\theta \in (0,1]$. Similar to Proposition 1, we can derive

$$\sum_{t=1}^{T}\Big(\hat{f}(\mathbf{x}_t)-\hat{f}(\mathbf{x})\Big) \leq \sum_{t=1}^{T}\langle \mathbf{g}_t^{\mathrm{f}},(\mathbf{x}_t-\mathbf{x})\rangle+\sum_{t=1}^{T}(\nabla\hat{f}(\mathbf{x}_t)-\mathbf{g}_t^{\mathrm{f}})^{\top}(\mathbf{x}_t-\mathbf{x}).$$

Since $\mathbb{E}_{\mathbf{u},\xi}[\mathbf{g}_t]=\nabla\hat{f}(\mathbf{x}_t)$ and $\|\mathbf{g}_t^{\mathrm{f}}\|_2 \leq dB/\delta$, based on Lemma 14 in [Hazan and Kale, 2014], we have the following result. Given $\mathbf{x}_1 \in \Omega$, we apply T iterations of Eq. (2) and the update in Eq. (7). For any fixed $\mathbf{x} \in \Omega \cap \mathbb{B}(\mathbf{x}_1, D)$ and $\tilde{p} \in (0,1)$, with a probability at least $1-\tilde{p}$, the following inequality holds

$$\hat{f}(\hat{\mathbf{x}}_T)-\hat{f}(\mathbf{x}) \leq \frac{\|\mathbf{x}_1-\mathbf{x}\|_2^2}{2\eta T}+\frac{\eta d^2B^2}{2\delta^2}+\frac{4dBD\sqrt{3\log(\frac{1}{p})}}{\sqrt{T}\delta},$$

where $\hat{\mathbf{x}}_T=\sum_{t=1}^{T}\mathbf{x}_t/T$. Then, we are ready to have

$$f(\mathbf{x}_k)-f(\mathbf{x}_{k-1,*}) \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t}+\frac{\eta_k d^2B^2}{2\delta_k^2}+ \frac{4dBc\epsilon_{k-1}^{\theta}\sqrt{3\log(\frac{1}{p})}}{\sqrt{t}\delta_k}+2G\delta_k,$$

where we use $D_k = c\epsilon_{k-1}^{\theta}$. We can easily establish $f(\mathbf{x}_k)-f(\mathbf{x}_*) \leq \epsilon_k$ with high probability by induction. By setting $\delta_k = O(\epsilon_k)$, $\eta_k = O(\epsilon_k^3/d^2)$, $t = O(d^2\log(K/\tilde{p})/\epsilon^{2(2-\theta)})$ and $K = \lceil \log(\epsilon_0/\epsilon)\rceil$, we have the iteration complexity as $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$ with high probability of $1-p$, where $\tilde{p}=p/K$.

iv) Proof of R-IV: High probability results when $f(\mathbf{x};\xi)$ is G-Lipschitz continuous and L-smooth.

For smooth objective functions, we have

$$f(\mathbf{x}_k)-f(\mathbf{x}_{k-1,*}) \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t}+\frac{\eta_k d^2B^2}{2\delta_k^2}+ \frac{4dBc\epsilon_{k-1}^{\theta}\sqrt{3\log(\frac{1}{p})}}{\sqrt{t}\delta_k}+L\delta_k^2.$$

We establish $f(\mathbf{x}_k)-f(\mathbf{x}_*) \leq \epsilon_k$ with high probability by induction. By setting $D_k = c\epsilon_{k-1}^{\theta}$, $\delta_k = O(\sqrt{\epsilon_k})$, $\eta_k = O(\epsilon_k^2/d^2)$, $t = O(d^2\log(K/\tilde{p})/\epsilon^{3-2\theta})$ and $K = \lceil \log(\epsilon_0/\epsilon)\rceil$, we have the iteration complexity as $\widetilde{O}(d^2/\epsilon^{3-2\theta})$ with high probability of $1-p$, where $\tilde{p}=p/K$. □

6.3 Proof of Theorem 2

Proof. In the setting with TPE, we present the proofs as follows. Again, for expectational results, we adopt the update in Eq. (5). For high probability results, we adopt the update in Eq. (7).

i) Proof of R-I: Expectational results when $f(\mathbf{x};\xi)$ is G-Lipschitz continuous.

We consider the noisy gradient estimator of Eq. (3). Based on the LEB condition, for $\theta \in (0,1/2]$, we have $\mathbb{E}[\|\mathbf{x}-\mathbf{x}_*\|_2^2] \leq c^2\mathbb{E}[(f(\mathbf{x})-f_*)^{2\theta}] \leq c^2(\mathbb{E}[f(\mathbf{x})-f_*])^{2\theta}$. With results in Proposition 1 and Lemma 2, and the analysis of Theorem 1, we have

$$\mathbb{E}[f(\mathbf{x}_k)-f(\mathbf{x}_{k-1,*})] \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t}+\frac{\eta_k db^2G^2C}{2}+2G\delta_k,$$

where b and C are parameters discussed in Lemma 2.

Similarly, we can easily establish the relationship of $\mathbb{E}[f(\mathbf{x}_k)-f_*] \leq \epsilon_k$ by induction. We can set

$$\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} \leq \frac{\epsilon_{k-1}}{6} \Rightarrow t \geq \frac{9dG^2b^2Cc^2}{\epsilon_{k-1}^{2(1-\theta)}},$$
$$\frac{\eta_k db^2G^2C}{2} \leq \frac{\epsilon_k}{3} \Rightarrow \eta_k \leq \frac{2\epsilon_k}{3db^2G^2C},$$
$$2G\delta_k \leq \frac{\epsilon_k}{3} \Rightarrow \delta_k \leq \frac{\epsilon_k}{6G}.$$

Then, with $K = \lceil \log(\epsilon_0/\epsilon)\rceil$, the total iteration complexity is $\widetilde{O}(d/\epsilon^{2(1-\theta)})$.

ii) Proof of R-II: Expectational results when $f(\mathbf{x};\xi)$ is G-Lipschitz continuous and L-smooth.

If $f(\mathbf{x};\xi)$ is L-smooth, we adopt the noisy gradient estimator of Eq. (4) and the update of Eq. (5) to solve SZCO. For $\theta \in (0,1/2]$, with the results in Lemma 3, we are ready to have

$$\mathbb{E}[f(\mathbf{x}_k)-f(\mathbf{x}_{k-1,*})] \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k}{2}\left(\frac{\delta_k^2(d+6)^3L^2}{2}+2(d+4)G^2\right)+L\delta_k^2 d,$$

To establish the induction $\mathbb{E}[f(\mathbf{x}_k)-f_*] \leq \epsilon_k$, we set

$$\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} \leq \frac{\epsilon_{k-1}}{8} \Rightarrow t \geq \frac{32(d+4)G^2c^2}{\epsilon_{k-1}^{2(1-\theta)}},$$
$$\frac{\eta_k}{2}\left(\frac{\delta_k^2(d+6)^3L^2}{2}\right) \leq \frac{\epsilon_k}{4} \Rightarrow \eta_k \leq \frac{4d}{(d+6)^3L},$$
$$\frac{\eta_k}{2}(2(d+4)G^2) \leq \frac{\epsilon_k}{4} \Rightarrow \eta_k \leq \frac{\epsilon_k}{4(d+4)G^2},$$
$$L\delta_k^2 d \leq \frac{\epsilon_k}{4} \Rightarrow \delta_k \leq \frac{\sqrt{\epsilon_k}}{2\sqrt{dL}}.$$

Here we can set $\eta_k = \min\{\frac{\epsilon_k}{4(d+4)G^2}, \frac{2d}{(d+6)^3L}\}$. Since ϵ_k goes to ϵ, the term $\frac{\epsilon_k}{4(d+4)G^2}$ is dominant in iteration complexity and t is calculated via $\eta_k \leq \frac{\epsilon_k}{4(d+4)G^2}$. Thus, with $K = \lceil \log(\epsilon_0/\epsilon)\rceil$, the total iteration complexity is $\widetilde{O}(d/\epsilon^{2(1-\theta)})$.

iii) Proof of R-III: High probability results when $f(\mathbf{x};\xi)$ is G-Lipschitz continuous.

For high probability analysis, we adopt the noisy gradient estimator of Eq. (3) and the update of Eq. (7) to solve SZCO. Similar to the analysis of R-III in Theo-

rem 1, with high probability at least $1-\tilde{p}$,

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k-1,*}) \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 G^2}{2} + \frac{4dGc\epsilon_{k-1}^{\theta}\sqrt{3\log(\frac{1}{p})}}{\sqrt{t}} + 2G\delta_k,$$

where we set $D_k = c\epsilon_{k-1}^{\theta}$. To establish the induction $f(\mathbf{x}_k) - f_* \leq \epsilon_k$, we set $\delta_k = \epsilon_k/(8G)$, $\eta_k = \epsilon_k/(2d^2G^2)$, $t = O(d^2\log(K/\tilde{p})/\epsilon^{2(1-\theta)})$ and $K = \lceil \log(\epsilon_0/\epsilon) \rceil$, where $\tilde{p} = p/K$. As a result, the total iteration complexity is $\widetilde{O}(d^2/\epsilon^{2(1-\theta)})$ with probability of $1-p$. $\square$