Efficient Zeroth-Order Proximal Stochastic Method for Nonconvex Nonsmooth Black-Box Problems

Ehsan Kazemi[1*] and Liqiang Wang[1]

[1]Department of Computer Science, University of Central Florida.

*Corresponding author(s). E-mail(s): ehsan_kazemy@knights.ucf.edu;
Contributing authors: lwang@cs.ucf.edu;

Abstract

Proximal gradient method has a major role in solving nonsmooth composite optimization problems. However, in some machine learning problems related to black-box optimization models, the proximal gradient method could not be leveraged as the derivation of explicit gradients are difficult or entirely infeasible. Several variants of zeroth-order (ZO) stochastic variance reduced such as ZO-SVRG and ZO-SPIDER algorithms have recently been studied for nonconvex optimization problems. However, almost all the existing ZO-type algorithms suffer from a slowdown and increase in function query complexities up to a small-degree polynomial of the problem size. In order to fill this void, we propose a new analysis for the stochastic gradient algorithm for optimizing nonconvex, nonsmooth finite-sum problems, called ZO-PSVRG+ and ZO-PSPIDER+. The main goal of this work is to present an analysis that brings the convergence analysis for ZO-PSVRG+ and ZO-PSPIDER+ into uniformity, recovering several existing convergence results for arbitrary minibatch sizes while improving the complexity of their ZO oracle and proximal oracle calls. We prove that the studied ZO algorithms under Polyak-Łojasiewicz condition in contrast to the existent ZO-type methods obtain a global linear convergence for a wide range of minibatch sizes when the iterate enters into a local PL region without restart and algorithmic modification. The current analysis in the literature is mainly limited to large minibatch sizes, rendering the existing methods unpractical for real-world problems due to limited computational capacity. In the empirical experiments for black-box models, we show that the new analysis provides superior performance and faster convergence to a solution of nonconvex nonsmooth problems compared to the existing ZO-type methods as they suffer from small-level stepsizes. As a byproduct, the proposed analysis is generic and can be exploited to the other variants of gradient-free variance reduction methods aiming to make them more efficient.

Keywords: Nonconvex optimization, Zeroth-order methods, Polyak-Łojasiewicz condition, Nonsmooth optimization, Query efficient methods, Adversarial attacks

Introduction

In this paper, we consider nonsmooth nonconvex optimization problems of the generic form

$$\min_{x\in\mathbb{R}^d} F(x) = f(x) + h(x), \quad f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{1}$$

where each $f_i(x)$ is possibly nonconvex and smooth function, and $h(x)$ is a nonsmooth convex function such as L_1-norm regularizer. The optimization problem (1) governs numerous machine learning frameworks, ranging from neural networks to generalized linear models and from convex problems like SVM and Lasso to highly nonconvex optimization problems such as loss functions tailored to deep neural models. In this work, we shall explore a set of variance-reduced stochastic zeroth-order (SZO) optimization algorithms for (1) to improve their oracle calls. Stochastic variance-reduced gradient (SVRG) is a powerful approach to decrease the variance inherited from stochastic sampling Johnson and Zhang (2013); Reddi et al (2016a); Nitanda (2016); Allen-Zhu and Yuan (2016). These papers demonstrate that SVRG enhances the rate of convergence for stochastic gradient descent (SGD) by a factor of $O(1/\epsilon)$ due to the decrease in the variance of the gradient. The underlying idea to develop zeroth-order variance reduction methods is to leverage similar ideas from the first-order methods to reduce the variance of stochastic optimization methods in order to improve the convergence rate. The major adversity of first-order methods is their dependency on first-order information from the problem, while there are settings where the first-order gradients are computationally infeasible or costly whereas zeroth-order information (function information) is accessible. For instance, in online auctions and advertisement selections, only zeroth-order information in the form of responses to the queries is accessible Wibisono et al (2012). This illustrates the importance of the development of zeroth-order optimization methods for solving many machine-learning problems. Currently, there are only a few zeroth-order stochastic methods for solving problem (1), e.g., Ghadimi et al (2016) and Huang et al (2019). Ghadimi et al (2016) introduced a gradient-free proximal projection (RSPGF) algorithm using a two-point Gaussian random gradient estimator for smooth composite optimization problem (1), which achieves a convergence rate of $O(\sqrt{d}/\sqrt{T})$ with T is the number of iterations) and stochastic zeroth-order query complexity of $O(\frac{d^2}{\epsilon^2})$, to reach a stationary point x^* such that $\|F(x^*)\| \leq \epsilon$. There are other variants of zeroth-order for nonsmooth problems, e.g., Huang et al (2019) and also momentum accelerated zeroth-order methods Chen et al (2019) to achieve a higher rate of convergence. In addition, several zeroth-order projection-free methods were developed in Sahu et al (2019); Huang et al (2020) as well as ADAM-based methods in Gao et al (2018).

The other variation of variance reduction optimization schemes was introduced in Fang et al (2018) called SPIDER. Being different from SVRG-type algorithms, SPIDER Fang et al (2018) and the earlier version Nguyen et al (2017a,b) are first-order stochastic variance-reduced schemes where in each inner-loop iteration gradient is estimated recursively by applying the variance-reduced gradient estimation in the past iteration (see (19)). A zeroth-order variant of SPIDER was proposed in Fang et al (2018), named SPIDER-SZO. This set of ZO variance-reduced methods replaces first-order gradients in the SPIDER algorithm with zeroth-order gradient estimators. In Fang et al (2018) it shows that SPIDER-SZO gains an improved zeroth-order query complexity compared to SVRG-based zeroth-order algorithms. Nevertheless, the original SPIDER-SZO algorithm requires generating a large number of Gaussian random samples of order $O(n^{1/2}d^2)$ at each epoch and the convergence guarantee is obtained with a stepsize of $O(\epsilon)$. In practice, these requirements significantly limit the performance of SPIDER-SZO.

Recently, Ji et al (2019) refined the ZO estimations to derive an improved SZO complexity with improved convergence rates. However, the analysis in their work is only for smooth functions based on a complicated parameter selection analysis which is only valid for particular and mainly large minibatch sizes. Towards improving the convergence analysis in the existing works, in this paper we address this open question: *Is this possible to extend the existing convergence analysis for ZO-SVRG and ZO-SPIDER to arbitrary minibatch sizes with improved ZO query complexity and under more general nonsmooth, nonconvex setting?* Aiming to answer this question, we employ variance reduction methods to design an accelerated ZO proximal method for nonsmooth composite optimization problem (1). This demonstrates an improvement to ZO iteration complexity up to a factor of d compared with Huang et al (2019) and extends the existing analysis for ZO variance-reduced framework to versatile studies which are valid for more general parameter settings.

In Tables 1 and 2, we compared the results from our analysis and other state-of-the-art ZO

optimization methods for three different items. We also listed the minibatch sizes that our methods can achieve the optimal SZO calls known in the literature. Table 1 shows that the convergence of ZO-PSVRG+ provides a better dependency on problem dimension d than RSPGF and ZO-ProxSVRG/SAGA for nonconvex nonsmooth optimization. It also shows that RGF has the largest query complexity yet has the worst convergence rate. ZO-SVRG-Coord and ZO-ProxSVRG/SAGA provide an improved rate of convergence $O(d/\epsilon)$ owning to applying variance reduction techniques. Further observation of Table 1 reveals that the existing SVRG type zeroth-order algorithms are highly affected by function query complexities compared to RSPGF, while ZO-PSVRG+ (our algorithm) could achieve better trade-offs between the convergence rate and the querying complexity.

Main Contributions

In this work, we present a novel uniform analysis for ZO-PSVRG+ (Algorithm 1) and ZO-PSPIDER+ (Algorithm 2) which are different from the comparable convergence studies. Although recently several accelerated variance reduction techniques for nonconvex problems have been proposed Wang et al (2019), in this work we mainly focus to improve the analysis of the basic aforesaid methods and the extension to the novel accelerated frameworks shall be addressed in our future studies. The main contributions of this paper are summarized in the following:

1) Our analysis for ZO-PSVRG+ yields iteration complexity $O(\frac{1}{\epsilon})$ compared to $O(\frac{d}{\epsilon^2})$ of RSPGF Ghadimi and Lan (2016) and $O(\frac{d}{\epsilon})$ of ZO-ProxSVRG/SAGA Huang et al (2019) (the existing variance-reduce SZO proximal algorithm for solving nonconvex nonsmooth problems). It shows that our results have improved the complexity of the terms dependent on d in contrast to the existing proximal variance-reduced SZO methods. ZO-PSVRG+ also matches the best result achieved by ZO-SVRG-Coord-Rand with minibatch size $b = dn^{2/3}$ and epoch size $m = n^{1/3}$ in Ji et al (2019), while our analysis is valid for *any minibatch sizes* as detailed in the following sections. In addition, our analysis ZO-PSPIDER+ for is superior compared to SPIDER-SZO Fang et al (2018) by entirely removing the requirement for i.i.d. Gaussian samples and permitting a larger stepsize $O(1)$ to enable a speedier convergence. These convergence advantages are due to a new analysis that we present throughout this work. As seen from Figure 1, our analyses for both algorithms are valid for the entire range of minibatch sizes i.e., $b \in [1, n]$

2) The convergence analysis for ZO-PSVRG+ in contrast to ZO-SVRG-Coord Liu et al (2018b); Ji et al (2019) is straightforward and yields simpler proofs. Our analysis achieves new iteration complexity bounds and improves the effectiveness of all the existing ZO-SVRG-based algorithms along with RSPGF for nonconvex nonsmooth composite optimization. Note that the convergence studies for RSPGF and ZO-ProxSVRG/SAGA rely on bounded gradient assumption, which is not our working assumption in this paper.

3) For the nonconvex functions under the Polyak-Łojasiewicz (PL) condition Polyak (1963), we show that ZO-PSVRG+ and ZO-PSPIDER+ obtain a global linear convergence rate equivalent to the first-order ProxSVRG. Thus, these algorithms can certainly achieve linear convergence in some zones without restarting. To the best of our knowledge, this is the first paper that leverages the PL condition for improving the convergence of the zeroth-order proximal variance-reduced method for problem (1) with arbitrary minibatch sizes. This analysis generalizes the results Duchi et al (2015) while showing linear convergence in contrast to the sublinear convergence rate in their paper. In Ji et al (2019), the authors show that ZO-SPIDER-Coord achieves linear convergence under PL condition but the analysis is limited to the minibatch of size $b = O(n^{1/2})$. Note that due to both computational and statistical efficiency, convergence analysis for minibatch of moderate sizes is essential (also see the remarks after Theorem 7 for more details).

Finally, to demonstrate the efficiency and adaptability of our approach in achieving a balance between the convergence rate and the number of SZO queries, we perform some experimental evaluations for two distinct applications: black-box binary classification and universal adversarial attacks on black-box deep neural network models. The empirical results and theoretical investigations verify the effectiveness of our algorithms. To present the results coherently, we postpone our proofs for convergence analysis to the supplementary materials.

Table 1: Summary of convergence rate and function query complexity of various SZO algorithms. S: Smooth, NS: Nonsmooth, NC: Nonconvex, C: Convex, SC: Strong Convexity, and PL: Polyak-Łojasiewicz Condition. b denotes the minibatch size, m denotes the epoch size, T is the total number of iterations, λ is the constant in PL condition (15) and $s_n = \min\{n, \frac{1}{\epsilon}\}$. The dash-line "-" shows that the result is valid for a generic b or m.

Method	Problem	Stepsize	Convergence rate	SZO complexity	b	m
ZO-GD (Nesterov and Spokoiny (2011))	S(NC)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{T}\right)$	$O\left(\frac{dn}{\epsilon}\right)$	NA	NA
ZO-SGD (Ghadimi and Lan (2013))	S(NC)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{\sqrt{d}}{\sqrt{T}}\right)$	$O\left(\frac{d}{\epsilon^2}\right)$	NA	NA
RSPGF (Ghadimi et al (2016))	S(NC) + S(C)	$O(1)$	$O\left(\frac{\sqrt{d}}{\sqrt{T}}\right)$	$O\left(\frac{d^2}{\epsilon^2}\right)$	-	1
ZO-SVRG-Coord (Liu et al (2018b))	S(NC)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{T}\right)$	$O(\frac{nd^2}{\epsilon} + \frac{d^2 b}{\epsilon})$	-	$O(d)$
ZO-SVRG-Coord-Rand (Ji et al (2019))	S(NC)	$O\left(\frac{1}{dn^{2/3}}\right)$	$O\left(\frac{dn^{2/3}}{T}\right)$	$O(\min\{\frac{dn^{2/3}}{\epsilon}, \frac{d}{\epsilon^{5/3}}\})$	$\mathcal{B}^{2/3}$	$\mathcal{B}^{1/3}$
ZO-ProxSVRG-Coord (Huang et al (2019))	S(NC)+NS(C)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{T}\right)$	$O(\frac{nd^2}{\epsilon\sqrt{b}} + \frac{md^2\sqrt{b}}{\epsilon})$	$\mathcal{B}^{2/3}$	$\mathcal{B}^{1/3}$
ZO-ProxSAGA-Coord (Huang et al (2019))	S(NC)+NS(C)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{T}\right)$	$O(\frac{(n+b)d^2}{\epsilon})$	$\mathcal{B}^{2/3}$	$\mathcal{B}^{1/3}$
ZO-PSVRG+ (Ours)	S(NC)+NS(C)	$\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12mL}\}$	$O\left(\frac{1}{\eta T}\right)$	$O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{bd}{\epsilon\eta}\right)$	-	-
		$\min\{\frac{1}{8L}, \frac{\sqrt{1}}{12L}\}$	$O\left(\frac{1}{T}\right)$	$O\left(\min\{\frac{dn^{2/3}}{\epsilon}, \frac{d}{\epsilon^{5/3}}\}\right)$	$\mathcal{B}^{2/3}$	$\mathcal{B}^{1/3}$
ZO-PSVRG+ (RandSGE) (Ours)	S(NC)+NS(C)	$\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12mL\sqrt{d}}\}$	$O\left(\frac{1}{\eta T}\right)$	$O\left(s_n \frac{d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right)$	-	-
ZO-PSVRG+ (Ours)	S(PL)+NS(C)	$\eta = \min\{\frac{1}{8L}, \frac{\sqrt{\gamma b}}{12mL}\}$	$O(\epsilon)$	$O(\frac{s_n d}{\lambda\eta m}\log\frac{1}{\epsilon} + \frac{bd}{\lambda\eta}\log\frac{1}{\epsilon})$	-	-
		$O(1)$	$O(\epsilon)$	$O(\min\{\frac{dn^{2/3}}{\lambda}\log\frac{1}{\epsilon}, \frac{d}{\lambda\epsilon^{2/3}}\log\frac{1}{\epsilon}\})$	$\mathcal{B}^{2/3}$	$\mathcal{B}^{1/3}$
		$O(\frac{1}{m})$	$O(\epsilon)$	$O(\min\{\frac{dn}{\lambda}\log\frac{1}{\epsilon}, \frac{d}{\lambda\epsilon}\log\frac{1}{\epsilon}\})$	1	-
ZO-PSVRG+ (RandSGE) (Ours)	S(PL)+NS(C)	$\eta = \min\{\frac{1}{8L}, \frac{\sqrt{\gamma b}}{12mL\sqrt{d}}\}$	$O(\epsilon)$	$O(s_n\frac{d\sqrt{d}}{\lambda}\log\frac{1}{\epsilon} + \frac{b\sqrt{d}}{\lambda}\log\frac{1}{\epsilon})$	-	-

Table 2: Summary of convergence rate and function query complexity of various ZO-SPIDER algorithms. The acronyms and notations are similar to Table 1. *: Refer to the paper for the definition of variables used.

Method	Problem	Stepsize	Convergence rate	SZO complexity	b	m
SPIDER-SZO (Fang et al (2018))	S(NC)	$O(\sqrt{\epsilon})$	$O(\frac{\sqrt{d}}{T})$	$O(\min\{\frac{dn^{1/2}}{\epsilon}, \frac{d}{\epsilon^{3/2}}\})$		
ZO-SPIDER-Coord (Ji et al (2019))	S(NC)	$O(1)$	$O(\frac{1}{T})$	$O(\min\{\frac{dn^{1/2}}{\epsilon}, \frac{d}{\epsilon^{3/2}}\})$	$\mathcal{B}^{1/2}$	$\mathcal{B}^{1/2}$
		$O\left(\frac{1}{\sqrt{m}}\right)$	$O(\frac{\sqrt{n}}{T})$	$O(\min\{\frac{dn^{1/2}}{\epsilon}, \frac{d}{\epsilon^{3/2}}\})$	1	n
ZO-SPIDER-Coord* (Ji et al (2019))	S(PL)	$O\left(\frac{1}{\sqrt{m}}\right)$	$O(\epsilon)$	$O(d(\gamma n^{1/2} + \gamma^2)\log(\frac{1}{\epsilon}))$	$\gamma L B_\gamma$	$\frac{\gamma}{b_{\gamma L}}$
ZO-PSPIDER+ (Ours)	S(NC) + NS(C)	$\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12\sqrt{m}L}\}$	$O\left(\frac{1}{\eta T}\right)$	$O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{bd}{\epsilon\eta}\right)$	-	-
		$O(1)$	$O\left(\frac{1}{T}\right)$	$O(\min\{\frac{dn^{1/2}}{\epsilon}, \frac{d}{\epsilon^{3/2}}\})$	$\mathcal{B}^{1/2}$	$\mathcal{B}^{1/2}$
		$O\left(\frac{1}{\sqrt{m}}\right)$	$O\left(\frac{\sqrt{m}}{T}\right)$	$O(\min\{\frac{dn^{1/2}}{\epsilon}, \frac{d}{\epsilon^{3/2}}\})$	1	$\mathcal{B}$
ZO-PSPIDER+ (Ours)	S(PL) + NS(C)	$\eta = \min\{\frac{1}{8L}, \frac{\lambda b}{32L^2}\}$	$O(\epsilon)$	$O(\frac{s_n d}{\lambda\eta m}\log\frac{1}{\epsilon} + \frac{bd}{\lambda\eta}\log\frac{1}{\epsilon})$	-	-
		$O(1)$	$O(\epsilon)$	$O(\frac{bd}{\lambda}\log\frac{1}{\epsilon})$	-	$\mathcal{B}$

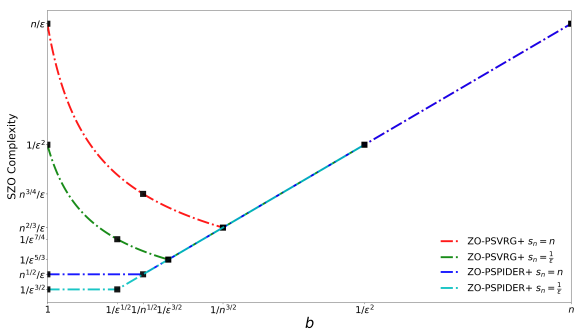


Fig. 1: SZO complexity versus minibatch size for $s_n = n$ and $s_n = \frac{1}{\epsilon}$ where $s_n = \min\{n, \frac{1}{\epsilon}\}$.

Implications of the Proposed Analysis

In application, the number of training data is at least in the order of $n \sim 10^7 - 10^9$ and so $n^{2/3} \sim 10^5 - 10^6$, which is quite large, given the capacity of modern computational infrastructures. On the other hand, if b is too small, the benefits from the parallelism of the algorithm would be forfeited. Thus, the question is if a convergence rate of almost optimal order could be achieved by applying moderate sizes b, which do not depend on n and d. Our results, in this case, match the best result achieved in Huang et al (2019) for $b = n^{2/3}$, despite the fact we did not apply the bounded gradient assumption similar to their work. Further, differently from Huang et al (2019), our analysis is more flexible as we avoided using all the samples to compute the full gradient and the gradients are computed on a subset of size $\mathcal{B}$. Note that avoiding calculating the full gradient besides

computational efficiency can improve the generalization performance of the model. As seen from Figure 1, we show lower SZO calls for $\mathcal{B} = \frac{1}{\epsilon} < n$ which advocates moderate minibatch sizes $\frac{1}{\epsilon^{2/3}}$ and $\frac{1}{\epsilon^{1/2}}$ versus $n^{2/3}$ and $n^{1/2}$, respectively for ZO-PSVRG+ and ZO-PSPIDER+. Our studies provide explicit and straightforward expressions for the algorithm parameters based on a notably new analysis, whereas the parameters for the analysis presented in Ji et al (2019) depend on implicit, unpractical, and complicated equations for only smooth functions. Furthermore, the analysis in Ji et al (2019) for the single batch case requires adopting a very large number of iterations in the inner loop, $m \sim O(nd)$ to improve the SZO complexity (see Corollary 2 in Ji et al (2019)). This will significantly increase the variance between the full gradient and the gradient over a minibatch, as the algorithm applies an outdated full gradient for many inner loop iterations.

Preliminaries

In the following, we illustrate preliminary details of zeroth-order gradient approximations. Considering the loss function f_i, a two-point random stochastic gradient estimator (RandSGE) $\hat{\nabla}_r f_i(x)$ is defined as Nesterov and Spokoiny (2017); Gao et al (2018)

$$\hat{\nabla}_r f_i(x, u_i) = \frac{d(f_i(x + \mu u_i) - f_i(x))}{\mu} u_i, \qquad i \in [n] \tag{2}$$

where d is the number of optimization variables, $\{u_i\}$ are i.i.d. random directions drawn from a uniform distribution over a unit sphere and $\mu > 0$ is the smoothing parameter Flaxman et al (2005); Shamir (2017); Gao et al (2018). RandSGE is a biased approximation to the true gradient $\nabla f_i(x)$, and its bias decreases as μ approach zero. Nevertheless, in practice, if μ is too small, the function variation could be magnified by the noise in the function evaluations when the rate of noise to signal is high Lian et al (2016). To obtain a more accurate approximation for ZO gradient, we could apply coordinate gradient estimation (CoordSGE) Gu et al (2018b,a); Liu et al (2018b) to approximate the gradients as:

$$\hat{\nabla} f_i(x) = \sum_{j=1}^{d} \frac{f_i(x + \mu e_j) - f_i(x - \mu e_j)}{2\mu} e_j, \quad i \in [n] \tag{3}$$

In this expression, e_j is a standard basis vector with 1 at its j-th coordinate and 0 otherwise, and μ is the smoothing parameter. In contrast to RandSGE, CoordSGE is deterministic and needs d times more ZO function calls. We will later show that ZO variance-reduced method using CoordSGE results in a larger stepsize and a speedier convergence, although the coordinate-wise gradient estimator requires more ZO calls compared to the two-point random gradient approximation.

Finally, we formulate the zeroth-order proximal gradient descent method using ZO gradient estimation (3):

$$x_t^s = \text{Prox}_{\eta h}(x_{t-1}^s - \eta \hat{\nabla} f(x_{t-1}^s)), \qquad t = 1, 2, \ldots \tag{4}$$

where s is epoch number, $\hat{\nabla} f = \frac{1}{n} \sum_{i=1}^{n} \hat{\nabla} f_i(x)$ and

$$\text{Prox}_{\eta h}(x) := \arg \min_{y \in \mathbb{R}^d} \left(h(y) + \frac{1}{2\eta} \|y - x\|^2 \right) \tag{5}$$

In the rest of the paper, we assume that the nonsmooth convex function $h(x)$ in (1) is well-defined, i.e., the proximal operator (5) can be computed effectively.

In general, for the convex problems the convergence is measured using the optimality gap $F(x) - F(x^*)$, where we let x^* denote the optimal solution of Problem (1). However, for general nonconvex problems, the gradient norm is commonly used as the convergence metric. For example, for smooth nonconvex optimization (i.e., $h(x) = 0$), in Ghadimi and Lan (2013); Reddi et al (2016a); Lei et al (2017); Liu et al (2018b) the norm of the function gradient $\|\nabla F(x)\|^2$ is applied as the convergence criterion. Aiming to investigate the convergence behavior for nonsmooth nonconvex problems, we define the gradient mapping metric as $g_\eta(x) = \frac{1}{\eta}(x - \text{Prox}_{\eta,h}(x - \eta \nabla f(x)))$. If $h(x)$ is a constant function, the gradient mapping reduces to the ordinary gradient: $g_\eta(x) = \nabla F(x) = \nabla f(x)$. In this work, we use the gradient mapping $g_\eta(x)$ as the convergence metric similar to Ghadimi and Lan

(2016); Reddi et al (2016b); Parikh et al (2014). For problems with nonconvex structure, the point x is called a stationary point if $g_\eta(x) = 0$, Parikh et al (2014). Therefore, we end up with the following definition for the convergence metric.

Definition 1 We call point $x \in \mathbb{R}^d$ as an ϵ-accurate point, if $\mathbb{E}\left\|g_\eta(x)\right\|^2 \leq \epsilon$, for some $\eta > 0$.

ZO Proximal Stochastic Method (ZO-PSVRG+)

Algorithm 1 Zeroth-Order Proximal Stochastic Method

1: **Input:** initial point x_0, batch size $\mathcal{B}$, minibatch size b, epoch length m, stepsize η
2: **Initialize:** $\tilde{x}^0 = x_0$
3: **for** $s = 1, 2, \ldots, S$ **do**
4: $\quad x_0^s = \tilde{x}^{s-1}$
5: $\quad \hat{g}^s = \frac{1}{\mathcal{B}} \sum_{j \in I_{\mathcal{B}}} \hat{\nabla} f_j(\tilde{x}^{s-1})$
6: $\quad$ **for** $t = 1, 2, \ldots, m$ **do**
7: $\quad\quad$ Compute $\hat{v}_{t-1}^s$ according to (7) or (8)
8: $\quad\quad x_t^s = \mathrm{Prox}_{\eta h}(x_{t-1}^s - \eta \hat{v}_{t-1}^s)$
9: $\quad$ **end for**
10: $\quad \tilde{x}^s = x_m^s$
11: **end for**
12: **Output:** $\hat{x}$ chosen uniformly from $\{x_t^s\}_{t \in [m], s \in [S]}$

The core idea in variance-reduced methods is to generate an additional sequence $\tilde{x}^{s-1}$ at which the full gradient is computed to obtain a more accurate stochastic gradient estimate

$$v_{t-1}^s = \frac{1}{b} \sum_{i \in I_b} \left(\nabla f_i(x_{t-1}^s) - \nabla f_i(\tilde{x}^{s-1})\right) + g^s \quad (6)$$

where v_{t-1}^s denotes the gradient estimate at x_{t-1}^s and $g^s = \frac{1}{\mathcal{B}} \sum_{i \in I_{\mathcal{B}}} \nabla f_i(\tilde{x}^{s-1})$. We study a proximal stochastic gradient algorithm based on the variance-reduced approach of ProxSVRG in Xiao and Zhang (2014); Reddi et al (2016b); Li and Li (2018). The description of ZO-PSVRG+ is presented in Algorithm 1. Our method has two types of random sampling. In the outer iteration, we calculate the gradient consisting of $\mathcal{B}$ samples. In the inner iteration, we randomly choose a minibatch of samples of size b to approximate the gradient over the minibatch. We call $\mathcal{B}$ and b, batch, and minibatch size, respectively. In our ZO framework, the mix gradient (6) is estimated by applying only function evaluations, given by

$$\hat{v}_{t-1}^s = \frac{1}{b} \sum_{i \in I_b} \left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) + \hat{g}^s \quad (7)$$

or

$$\hat{v}_{t-1}^s = \frac{1}{b} \sum_{i \in I_b} \left(\hat{\nabla}_r f_i(x_{t-1}^s, u_i) - \hat{\nabla}_r f_i(\tilde{x}^{s-1}, u_i)\right) + \hat{g}^s \quad (8)$$

where $\hat{g}^s = \frac{1}{\mathcal{B}} \sum_{i \in I_{\mathcal{B}}} \hat{\nabla} f_i(\tilde{x}^{s-1})$, $\hat{\nabla} f_i$ is a ZO gradient approximation using CoordSGE and $\hat{\nabla}_r f_i$ is a ZO gradient estimate using RandSGE. We let ZO-PSVRG+ and ZO-PSVRG+ (RandSGE) denote Algorithm 1 with gradient estimation (7) and (8), respectively. Note that, $\mathbb{E}_{I_b}[\hat{v}_{t-1}^s] = \hat{\nabla} f(x_{t-1}^s) \neq \nabla f(x_{t-1}^s)$, i.e., this stochastic gradient is a biased approximation of the true gradient. In other words, the unbiased assumption on gradient approximates utilized in ProxSVRG Reddi et al (2016b); Li and Li (2018) is no longer valid. Note that the biased ZO gradient estimation yields a fundamental challenge in analyzing ZO-PSVRG+. It means that adjusting the similar concepts from ProxSVRG to zeroth-order algorithm 1 is not effortless and requires an elaborated analysis of ZO-PSVRG+. To tackle this issue, we derive an upper bound for the variance of the gradient approximation $\hat{v}_t^s$ by selecting an appropriate stepsize η and smoothing parameter μ to control the variance of gradient estimation which will be discussed later.

The other major difference between ZO-PSVRG+ and ZO-ProxSVRG is that we avoid the evaluation of the total gradient for each epoch, i.e., the number of samples $\mathcal{B}$ is not necessarily equal to n (see Line 5 of Algorithm 1). If $\mathcal{B} = n$, ZO-PSVRG+ is equivalent to ZO-ProxSVRG, which indicates that our convergence studies yield a novel analysis for ZO-ProxSVRG-Coord (i.e, $\mathcal{B} = n$). In the next section, we will carefully study the convergence of ZO-PSVRG+ under different settings. The proofs of the main theorems are deferred to the extended paper.

Convergence Analysis for ZO-PSVRG+

Here we provide some minimal assumptions for problem (1):

Assumption 1 For $\forall i \in [n]$, gradient of the function f_i is Lipschitz continuous with a Lipschitz constant $L > 0$, such that

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\|, \ \forall x, y \in \mathbb{R}^d$$

Assumption 2 For $\forall x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\left\|\hat{\nabla} f_i(x) - \hat{\nabla} f(x)\right\|^2\right] \leq \sigma^2$$

where $\sigma > 0$ is a constant and $\hat{\nabla} f_i(x)$ is a CoordSGE gradient approximation of $\nabla f_i(x)$.

Assumptions 1 and 2 are standard assumptions applied in SZO optimization. The first assumption is for the convergence studies of the zeroth-order algorithms Ghadimi and Lan (2016); Nesterov and Spokoiny (2017); Liu et al (2018b). The second assumption specifies bounded variance for zeroth-order gradient approximations Lian et al (2016); Liu et al (2018a,b). Assumption 2 is essential in order to obtain a convergence result independent of n. Due to the error estimation for CoordSGE, this assumption is equivalent to the bounded variance of true gradients.

Assumption 2 is weaker than the assumption of bounded gradients in Liu et al (2017); Hajinezhad et al (2019), while, we are able to analyze the more complicated problem (1) involving a nonsmooth part and obtain faster convergence rates. Note that according to the error estimation for CoordSGE, this assumption is equivalent to the bounded variance of true gradients.

Theorem 1 *Suppose Assumptions 1 and 2 hold, and the ZO gradient estimator (7) for mix gradient $\hat{v}_k$ is used. The output $\hat{x}$ of Algorithm 1 satisfies*

$$\mathbb{E}[\|g_\eta(\hat{x})\|^2] \leq \frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} + \frac{I(\mathcal{B} < n) 12\sigma^2}{\mathcal{B}} + 21 L^2 d^2 \mu^2$$

where $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12mL}\}$ denotes the stepsize.

In contrast to the convergence rate of SVRG in Reddi et al (2016b), Theorem 1 presents two extra error terms $\frac{I(\mathcal{B}<n)\sigma^2}{\mathcal{B}}$ and $O(L^2 d^2 \mu^2)$, related to the batch gradient estimation $\mathcal{B} < n$ and the use of SZO gradient approximations, respectively. The error related to $\mathcal{B} < n$ is removed only when $\mathcal{B} = n$. Note that the stepsize η depends on the epoch length m, and the minibatch size b. The proof for Theorem 1 is significantly different from the proofs in the existing literature. For instance, the convergence analysis for ZO-SVRG-Coord and ZO-ProxSVRG/SAGA uses the notion of the Lyapunov function to show that the accumulated gradient mapping decreases with epoch s. However, in our analysis, we explicitly prove that $F(x^s)$ is descending. On the other hand, our convergence result is valid for a wide range of minibatch sizes and any epoch size m, whereas the analysis for ZO-SVRG-Coord is valid only for specific values of m with a complicated setting for parameter selection.

The next corollary demonstrates the convergence rate of ZO-PSVRG+ in terms of error of the solution $\hat{x}$, providing explicit descriptions for the parameters in Theorem 1.

Corollary 2 Let the batch size $\mathcal{B} = \min\{12\sigma^2/\epsilon, n\}$ and $\mu \leq \frac{\sqrt{\epsilon}}{5dL}$ denote the smoothing parameter. Suppose $\hat{x}$ in Algorithm 1 is an ϵ-accurate solution for problem (1). Recalling that CoordSGE requires $O(d)$ function queries, the number of SZO calls is at most

$$\begin{aligned} d(S\mathcal{B} + Smb) &= 6d\left(F(x_0) - F(x^*)\right)\left(\frac{\mathcal{B}}{\epsilon\eta m} + \frac{b}{\epsilon\eta}\right) \\ &= O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{bd}{\epsilon\eta}\right) \end{aligned} \tag{9}$$

and the number of PO calls is equal to $T = Sm = \frac{6(F(x_0)-F(x^*))}{\epsilon\eta} = O\left(\frac{1}{\epsilon\eta}\right)$. In particular, by setting $m = \sqrt{b}$ and $\eta = \frac{1}{12L}$, the number of SZO calls is at most

$$\begin{aligned} &72dL(F(x_0) - F(x^*))\left(\frac{\mathcal{B}}{\epsilon\sqrt{b}} + \frac{b}{\epsilon}\right) \\ &= O\left(s_n \frac{d}{\epsilon\sqrt{b}} + \frac{bd}{\epsilon}\right) \end{aligned} \tag{10}$$

where $s_n = \min\{n, \frac{1}{\epsilon}\}$ and the number of PO calls equals to $T = Sm = S\sqrt{b} = \frac{72L(F(x_0)-F(x^*))}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$.

Corollary 2 indicates that if the smoothing parameter μ is sufficiently small and the batch size $\mathcal{B}$ is large enough, then the errors induced from zeroth-order estimation and batch gradient approximation will decrease, leading to a non-dominant

term in the convergence rate of ZO-PSVRG+. Indeed, the error term induced by batch size is eliminated only when $\mathcal{B} = n$ (i.e., $I(\mathcal{B} < n) = 0$). In this case, Step 5 of Algorithm 1 converts to $\hat{g}^s = \hat{\nabla} f(\tilde{x}^{s-1})$ and consequently ZO-PSVRG+ changes to ZO-ProxSVRG. In fact equation (10) indicates that a large batch $\mathcal{B}$ for $\mathcal{B} \neq n$ reduces the error inherited by the variance of batch gradient and improves the convergence of ZO-PSVRG+. In summary, with the smoothing parameter and batch size chosen properly, we derive the error term $O(1/\epsilon)$, which is better than the convergence rate of the state-of-the-art SZO algorithms by the factor $\frac{1}{d}$. Moreover, ZO-PSVRG+ uses much fewer SZO oracle calls compared to the methods listed in Table 1. Further, it is worth mentioning that the stepsize η in Theorem 1 has a lower dimension-dependency than the existing SZO algorithms in Table 1.

Analysis for ZO-PSVRG+ (RandSGE)

In this section, we will study the convergence of ZO-PSVRG+ (RandSGE) under different settings. In particular, in the following theorem, we prove that ZO-PSVRG+ (RandSGE) improves the convergence rate and the function query complexity of the existing SZO methods.

Theorem 3 *Suppose Assumptions 1 and 2 hold, and the coordinate gradient estimator (8) is used to compute the mix gradient $\hat{v}_k$. The output $\hat{x}$ of Algorithm 1 satisfies*

$$\mathbb{E}[\|g_\eta(\hat{x})\|^2] \leq \frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} + \frac{I(\mathcal{B} < n) 12\sigma^2}{\mathcal{B}} + 21L^2 d^2 \mu^2 \tag{11}$$

where $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12mL\sqrt{d}}\}$ denotes the stepsize.

Corollary 4 We Let the batch size $\mathcal{B} = \min\{12\sigma^2/\epsilon, n\}$ and $\mu \leq \frac{\sqrt{\epsilon}}{5dL}$ denote the smoothing parameter. Suppose $\hat{x}$ returned by Algorithm 1 is an ϵ-accurate solution for problem (1). Recalling that CoordSGE and RandSGE require $O(d)$ and $O(1)$ function queries respectively, the number of SZO calls is at most

$$(dS\mathcal{B} + Smb) = 6\left(F(x_0) - F(x^*)\right)\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta}\right) = O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta}\right) \tag{12}$$

and the number of PO calls is equal to $T = Sm = \frac{6(F(x_0)-F(x^*))}{\epsilon\eta} = O\left(\frac{1}{\epsilon\eta}\right)$. In particular, by setting $m = \sqrt{b}$ and $\eta = \frac{1}{12L\sqrt{d}}$, the number of SZO calls is at most

$$72L(F(x_0) - F(x^*))\left(\frac{\mathcal{B}d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right) = O\left(s_n \frac{d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right) \tag{13}$$

where $s_n = \min\{n, \frac{1}{\epsilon}\}$ and the number of PO calls equals to $T = Sm = S\sqrt{b} = \frac{72L\sqrt{d}(F(x_0)-F(x^*))}{\epsilon} = O\left(\frac{\sqrt{d}}{\epsilon}\right)$.

Convergence Under PL Condition

In this section, we show the linear convergence of ZO-PSVRG+ under Polyak-Łojasiewicz (PL) assumption Polyak (1963). The classic structure of PL condition is, for all $x \in \mathbb{R}^d$,

$$\|\nabla f(x)\|^2 \geq 2\lambda(f(x) - f^*) \tag{14}$$

where $\lambda > 0$ and f^* denotes the optimal function value. This condition specifies the rate of increase of the loss function in the vicinity of optimal solutions. It is important to note that if f is λ-strongly convex then f fulfills the PL condition. We will show that under our analysis the complexity of ZO-PSVRG+ (Algorithm 1) under PL condition is improved. Due to the presence of the nonsmooth term $h(x)$ in problem (1), we utilize the gradient projection to characterize a more generic form of PL condition as follows,

$$\|g_\eta(x)\|^2 \geq 2\lambda(F(x) - F^*) \tag{15}$$

for some $\lambda > 0$ and for all $x \in \mathbb{R}^d$. In particular, if $h(x)$ is a constant function, the gradient projection changes to $g_\eta(x) = \nabla f(x)$. The PL condition has been thoroughly investigated in Karimi et al (2016) where the authors proved that PL condition is a milder condition than a large family of conditions for functions such as convexity. The revised PL condition (15) is feasibly natural and has been studied in several papers for problems with non-convex nonsmooth settings, e.g., Li and Li (2018). Similarly, a zeroth-order algorithm under PL condition for smooth functions has been analyzed in Ji et al (2019).

ZO-PSVRG+ Under PL Condition

In this section, we demonstrate the convergence analysis of ZO-PSVRG+ (Algorithm 1) under PL-condition. More specifically, we provide a generic analysis for enhancing the convergence rate of the existing SZO algorithms for functions satisfying the PL condition using variance-reduced techniques. It is worth noting that for functions satisfying PL condition (i.e. (15) holds), ZO-PSVRG+ can immediately use the final iteration $\tilde{x}^S$ as the output point rather than using a randomly chosen $\hat{x}$. The following theorem provides the convergence guarantee for ZO-PSVRG+ under the PL condition.

Theorem 5 *Given the Assumptions 1 and 2, suppose that in Algorithm 1 the ZO gradient estimator (7) is applied for mix gradient $\hat{v}_k$ with stepsize $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{\gamma b}}{12mL}\}$ where $\gamma = 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} > 0$. Then*

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I(\mathcal{B} < n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2d^2\mu^2}{2\lambda} \quad (16)$$

Theorem 5 shows that if the batch size and smoothing parameter are appropriately chosen, ZO-PSVRG+ has a dominant linear convergence rate without restart. Further, compared to Theorem 1, it is evident from (16) that the error term $\frac{6I(\mathcal{B}<n)\sigma^2}{\mathcal{B}} + \frac{7L^2d^2\mu^2}{2}$ is amplified by the factor $1/\lambda$. Thus, the error induced by these terms will be improved if $\lambda >> 1$.

We next explore the number of ZO queries in ZO-PSVRG+ under PL condition to obtain an ϵ-accurate solution, as formalized in Corollary 6.

Corollary 6 Suppose the final iteration point $\tilde{x}^S$ in Algorithm 1 satisfies $\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \epsilon$ under PL condition. Under Assumptions 1 and 2, we let batch size $\mathcal{B} = \min\{\frac{6\sigma^2}{\lambda\epsilon}, n\}$ and $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4Ld}$ denote the smoothing parameter. Then, the number of SZO calls is bounded by

$$d(S\mathcal{B} + Smb) = O\left(\frac{s_n d}{\lambda\eta m}\log\frac{1}{\epsilon} + \frac{bd}{\lambda\eta}\log\frac{1}{\epsilon}\right)$$

where $s_n = \min\{n, \frac{1}{\lambda\epsilon}\}$. The number of PO calls equals the total number of iterations T which is bounded by $T = Sm = O(\frac{1}{\lambda\eta}\log\frac{1}{\epsilon})$. In particular, under the setting $m = \sqrt{b}$ and $\eta = \frac{\sqrt{\gamma}}{12L}$, the number of SZO calls simplifies to $d(S\mathcal{B} + Smb) = O(\frac{\mathcal{B}d}{\lambda\sqrt{\gamma m}}\log\frac{1}{\epsilon} + \frac{bd}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$.

Here we provide an intuition regarding Corollary 6: this result actually indicates that leveraging the PL condition improves the dominant convergence rate when the error of order $O(1/\epsilon)$ in Corollary 2 is improved to $O(\log(1/\epsilon))$, resulting in a significant speedup. Compared to the sublinear convergence rate for ZO algorithms in Duchi et al (2015); Nesterov and Spokoiny (2017); Liu et al (2018b), the convergence performance of ZO-PSVRG+ under PL condition has a global linear convergence rate and therefore requires a lower number of ZO oracle calls. This also indicates that if ZO-PSVRG+ is initialized in a generic nonconvex domain, the rate of convergence can be automatically accelerated due to entering the PL area. It is an improved result compared with Reddi et al (2016a) where therein PL-SVRG/SAGA is used to restart ProxSVRG/SAGA to obtain a linear convergence rate under PL condition. On the other hand, note that the convergence analysis under PL condition in Ji et al (2019) has complex coupling structures which makes it difficult for practitioners to apply, while our proof is simple and the parameters are properly specified to be suitable for hyperparameter selections.

Remark 1 Compared to Theorem 1, the convergence rate of ZO-PSVRG+ in Theorem 5 admits additional parameter γ for parameter selection due to the PL condition. By assuming the condition number $\lambda/L \leq \frac{1}{n^{1/3}}$ and through choosing $m = n^{1/3}$ and $\eta = \frac{\rho}{L}$ with $\rho \leq \frac{1}{2}$, the definition of γ yields

$$\gamma = 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} \geq 1 - \rho \geq \frac{1}{2} \quad (17)$$

According to Theorem 5, equation (17) implies $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12\sqrt{2}mL}\}$. Hence, choosing $b = m^2$ implies the constant stepsize $\eta \leq \frac{1}{12\sqrt{2}L}$. Note that the assumption $\lambda/L \leq \frac{1}{n^{1/3}}$ on condition number is milder than the assumption $\lambda/L < \frac{1}{\sqrt{n}}$ in Reddi et al (2016b).

ZO-PSVRG+ (RandSGE) Under PL Condition

In the following theorem, we explore if ZO-PSVRG+ (RandSGE) achieves a linear convergence rate when it enters a local landscape where the loss function satisfies the PL condition.

Theorem 7 *Let Assumptions 1 and 2 hold, and ZO gradient estimator (8) for mix gradient $\hat{v}_k$ is used in Algorithm 1 with stepsize $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{\gamma b}}{12mL\sqrt{d}}\}$ where $\gamma = 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} > 0$. Then*

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I(\mathcal{B} < n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2d^2\mu^2}{2\lambda} \quad (18)$$

Corollary 8 Suppose the final iteration point $\tilde{x}^S$ in Algorithm 1 satisfies $\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \epsilon$ under PL condition. Under Assumptions 1 and 2, we let batch size $\mathcal{B} = \min\{\frac{6\sigma^2}{\lambda\epsilon}, n\}$ and the smoothing parameter $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4Ld}$. The number of SZO calls is bounded by

$$(dS\mathcal{B} + Smb) = O(\frac{s_n d}{\lambda\eta m}\log\frac{1}{\epsilon} + \frac{b}{\lambda\eta}\log\frac{1}{\epsilon})$$

where $s_n = \min\{n, \frac{1}{\lambda\epsilon}\}$. The number of PO calls equals the total number of iterations T which is bounded by

$$T = Sm = O(\frac{1}{\lambda\eta}\log\frac{1}{\epsilon})$$

In particular, given the setting $m = \sqrt{b}$ and $\eta = \frac{\sqrt{\gamma}}{12L\sqrt{d}}$, the number of SZO calls simplifies to $(dS\mathcal{B} + Smb) = O(\frac{\mathcal{B}d\sqrt{d}}{\lambda\sqrt{\gamma}m}\log\frac{1}{\epsilon} + \frac{b\sqrt{d}}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$.

Remark 2 It should be noted that by selecting $b = O(d)$ in Theorem 7, the stepsize η reduces to $O(1)$ with $O(s_n d\log\frac{1}{\epsilon})$ SZO queries. Note that the analysis for ZO-SPIDER-Coord in Ji et al (2019) has no single-sample version for functions satisfying PL condition and the authors only provided a rate of convergence for large minibatch sizes with involved parameter information.

ZO Proximal Stochastic Method (ZO-PSPIDER+)

Recently, a new variant of the variance reduction approaches is introduced by Nguyen et al (2017a,b) and Fang et al (2018). In this approach, the first-order stochastic gradient is estimated as

$$v_t^s = \frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i(x_t^s) - \nabla f_i(\tilde{x}^{s-1})\right) + v_{t-1}^s \quad (19)$$

In this section, we investigate the potency of the variance-reduced estimator (19) in zeroth-order optimization of (1). Inspired by our novel convergence analysis for ZO-SVRG+, we propose a new analysis for zeroth-order recursive-based algorithm ZO-PSPIDER+, presented in Algorithm 2. In our ZO-PSPIDER+ framework, the mix first-order stochastic gradient (19) is estimated recursively by applying only function evaluations, given by

$$\hat{v}_t^s = \frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(x_{t-1}^s)\right) + \hat{v}_{t-1}^s \quad (20)$$

or

$$\hat{v}_t^s = \frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_t^s, u_i) - \hat{\nabla}_r f_i(x_{t-1}^s, u_i)\right) + \hat{v}_{t-1}^s \quad (21)$$

where $\hat{v}_0^s = \frac{1}{\mathcal{B}}\sum_{i\in I_\mathcal{B}}\hat{\nabla} f_i(x_0^s)$, $\hat{\nabla} f_i$ is a ZO gradient approximation using CoordSGE and $\hat{\nabla}_r f_i$ is a ZO gradient estimate using RandSGE. We let ZO-PSPIDER+ and ZO-PSPIDER+ (RandSGE) denote Algorithm 2 with gradient estimation (20) and (21), respectively. Differently from the zeroth-order algorithm (SPIDER-SZO) introduced in Fang et al (2018) that requires a stepsize of $O(\epsilon)$ and generation of $O(\frac{d^2\sqrt{n}}{\epsilon})$ Gaussian samples at each inner-loop iteration, our ZO-PSPIDER+ similar to ZO-SPIDER-Coord Ji et al (2019) removes the requirement for generating random Gaussian samples as we utilize highly accurate coordinate-wise estimation of zeroth-order gradient without any compensation in the number of the SZO queries. Our analysis allows a constant stepsize of $O(1)$ for minibatches and $O(\frac{1}{m})$ for a single-sample minibatch, i.e., $b = 1$ in Algorithm 2. We further extend our new analysis under PL assumption to show the linear convergence of ZO-PSPIDER+ without restart. In this way, we bring the convergence analysis for ZO-PSVRG+ and ZO-PSPIDER+ into uniformity. We relegate the analysis of ZO-PSPIDER+ (RandSGE) in (21) to the appendix.

Algorithm 2 Zeroth-Order Proximal Stochastic Method (ZO-PSPIDER+)

Input: initial point x_0, batch size $\mathcal{B}$, minibatch size b, epoch length m, stepsize η
Initialize: $\tilde{x}^0 = x_0$
for $s = 1, 2, \ldots, S$ **do**
  $x_0^s = \tilde{x}^{s-1}$
  $\hat{v}_0^s = \frac{1}{\mathcal{B}} \sum_{j \in I_{\mathcal{B}}} \hat{\nabla} f_j(x_0^s)$
  for $t = 0, 1, 2, \ldots, m$ **do**
    Compute $\hat{v}_t^s$ according to (20) or (21)
    $x_{t+1}^s = \mathrm{Prox}_{\eta h}(x_t^s - \eta \hat{v}_t^s)$
  end for
  $\tilde{x}^s = x_{m+1}^s$
end for
Output: $\hat{x}$ chosen uniformly from $\{x_t^s\}_{t \in [m], s \in [S]}$

Convergence Analysis for ZO-PSPIDER+

Theorem 9 *Suppose Assumptions 1 and 2 hold, and the ZO gradient estimator (20) for mix gradient $\hat{v}_k$ is used. The output $\hat{x}$ of Algorithm 2 satisfies*

$$\mathbb{E}[\|g_\eta(\hat{x})\|^2] \leq \frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} + \frac{I(\mathcal{B} < n) 12\sigma^2}{\mathcal{B}} + 21 m L^2 d^2 \mu^2$$

where $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12\sqrt{m}L}\}$ denotes the stepsize.

Corollary 10 Let the batch size $\mathcal{B} = \min\{12\sigma^2/\epsilon, n\}$ and $\mu \leq \frac{\sqrt{\epsilon}}{5\sqrt{m}dL}$ denote the smoothing parameter. Suppose $\hat{x}$ in Algorithm 2 is an ϵ-accurate solution for problem (1). Recalling that CoordSGE requires $O(d)$ function queries, the number of SZO calls is at most

$$d(S\mathcal{B} + Smb) = 6d\left(F(x_0) - F(x^*)\right)\left(\frac{\mathcal{B}}{\epsilon\eta m} + \frac{b}{\epsilon\eta}\right) = O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{bd}{\epsilon\eta}\right) \quad (22)$$

and the number of PO calls is equal to $T = Sm = \frac{6(F(x_0) - F(x^*))}{\epsilon\eta} = O\left(\frac{1}{\epsilon\eta}\right)$. In particular, by setting $m = b$ and $\eta = \frac{1}{12L}$, the number of SZO calls is at most

$$72dL(F(x_0) - F(x^*))\left(\frac{\mathcal{B}}{\epsilon b} + \frac{b}{\epsilon}\right) = O\left(s_n \frac{d}{\epsilon b} + \frac{bd}{\epsilon}\right) \quad (23)$$

where $s_n = \min\{n, \frac{1}{\epsilon}\}$ and the number of PO calls equals to $T = Sm = Sb = \frac{72L(F(x_0) - F(x^*))}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$.

ZO-PSPIDER+ Under PL Condition

Theorem 11 *Given the Assumptions 1 and 2, suppose that in Algorithm 2 the ZO gradient estimator (20) is applied for mix gradient $\hat{v}_k$ with stepsize $\eta \leq \min\{\frac{1}{8L}, \frac{\lambda b}{32L^2}\}$ with $1 - \frac{\lambda\eta}{3} > 0$. Then*

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I(\mathcal{B} < n)\sigma^2}{\lambda\mathcal{B}} + \frac{21 m L^2 d^2 \mu^2}{2\lambda} \quad (24)$$

Corollary 12 Suppose the final iteration point $\tilde{x}^S$ in Algorithm 2 satisfies $\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \epsilon$ under PL condition. Under Assumptions 1 and 2, we let batch size $\mathcal{B} = \min\{\frac{6\sigma^2}{\lambda\epsilon}, n\}$ and the smoothing parameter $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4\sqrt{m}Ld}$. The number of SZO calls is bounded by

$$d(S\mathcal{B} + Smb) = O\left(\frac{s_n d}{\lambda\eta m} \log\frac{1}{\epsilon} + \frac{bd}{\lambda\eta} \log\frac{1}{\epsilon}\right)$$

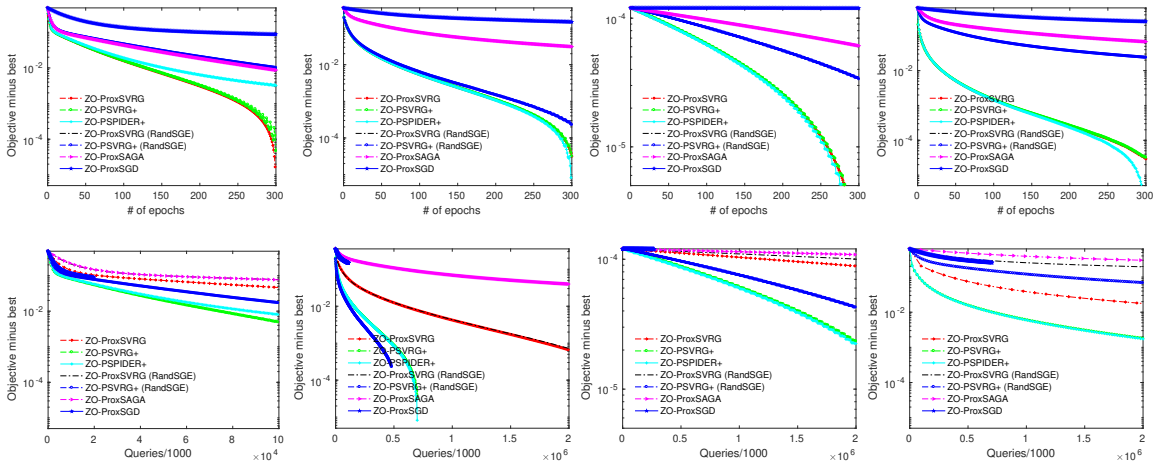
where $s_n = \min\{n, \frac{1}{\lambda\epsilon}\}$. The number of PO calls equals to the total number of iterations T which is bounded by

$$T = Sm = O\left(\frac{1}{\lambda\eta} \log\frac{1}{\epsilon}\right)$$

In particular, given the setting $\eta = \frac{\lambda}{32L^2}$, the number of SZO calls simplifies to $d(S\mathcal{B} + Smb) = O(\frac{\mathcal{B}d}{\lambda^2 m} \log\frac{1}{\epsilon} + \frac{bd}{\lambda^2} \log\frac{1}{\epsilon})$.

Experimental Results

In this section, we provide our experimental results. We compare the performance of our ZO-PSVRG+ and ZO-PSPIDER+ with 1) ZO-ProxSVRG (based on our improved analysis), 2) ZO-ProxSAGA-Coord Gu et al (2018a) and 3) ZO-ProxSGD Ghadimi and Lan (2016) over two empirical experiments: black-box binary classification and adversarial attacks for black-box deep neural networks (DNNs). We let ZO-ProxSGD denote RSPGF using CoordSGE (3) for gradient estimation. We also let ZO-ProxSVRG and ZO-ProxSVRG (RandSGE) denote ZO-PSVRG+ and ZO-PSVRG+ (RandSGE) with $\mathcal{B} = n$, respectively. The learning rates are tuned in the experiments for competitive algorithms according to their convergence guarantees in Tables 1 and 2, and the results shown in this section are based on the best learning rate we obtained for each algorithm. More specifically, Tables 1 and 2 show if the stepsize has to be dependent on the dimension d



(a) ijcnn (b) comparison on covtype (c) w8a (d) mnist

Fig. 2: Comparison of different zeroth-order algorithms for logistic regression loss residual $f(x) - f(x^*)$ versus the number of epochs (top) and ZO queries (bottom)

for the convergence guarantee. We set stepsize η and smoothing parameter μ for ZO-PSVRG+ and ZO-PSVRG+ (RandSGE) according to the convergence results that we obtained in the studied lemmas and theorems.

Binary Classification

In the first set of our experiments, we investigate the logistic regression loss function with L_1 and L_2 regularization for training the black-box binary classification problem. The problem can be described as the optimization problem (1) with $f_i(x) = \frac{1}{1+e^{y_i z_i^T x}}$, $h(x) = \lambda_1 \|x\|_1 + \frac{\lambda_2}{2}\|x\|^2$, where $z_i \in \mathbb{R}^d$ and y_i is the corresponding label for each i. The L_1 and L_2 regularization weights λ_1 and λ_2 in all the experiments are set respectively to 10^{-4} and 10^{-6}. We also set $\mathcal{B} = \lfloor \frac{n}{5} \rfloor$ for ZO-PSVRG+. We run our experiments on datasets from LIBSVM website[1], as listed in Table 3. The epoch size is chosen $m = 30$ over all the experiments and the minibatch size b is set to 50.

Table 3: Summary of training datasets.

Datasets	Data	Features
ijcnn	49990	22
a9a	32561	123
w8a	64,700	300
mnist	60000	784

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html

In Figure 2 (top), we show the training loss versus the number of epochs (i.e., iterations divided by the epoch length $m = 30$). Note that ZO-PSVRG+ is evaluated using mix gradient CoordSGE (3) and mix gradient RandSGE (2). Results in Figure 2 (bottom) compare the performance of ZO-PSVRG+ with the variants of ZO variance reduced stochastic gradient descent described earlier in Table 1 against the number of function queries. In these figures, ZO-PSVRG+ shows a faster convergence rate compared to ZO-PSVRG+ (RandSGE). Note that ZO-ProxSVRG based on our improved analysis compared to Huang et al (2019) presents a better convergence than both ZO-ProxSAGA and ZO-ProxSGD. On the other hand, the application of $\mathcal{B} = \lfloor \frac{n}{5} \rfloor$ in ZO-PSVRG+ significantly improves ZO-ProxSVRG with respect to the number of ZO-queries (see Table 1), leading to a non-dominant factor $O(I_{\{\mathcal{B}<n\}}/\mathcal{B})$ in the convergence rate of ZO-PSVRG+. In particular, ZO-PSVRG+ exhibits better performance in terms of the number of function queries than ZO-ProxSAGA using CoordSGE. The degradation in the convergence of ZO-ProxSAGA is due to the requirement for small stepsizes $O(\frac{1}{d})$. Furthermore, the large number of function queries to construct coordinate-wise gradient estimates significantly increases the number of SZO queries for ZO-ProxSVRG. On the other hand, ZO-ProxSGD consumes an extremely large number of iterations while exhibiting marginal convergence compared with variance-reduced algorithms. Thus, ZO-PSVRG+ and ZO-PSPIDER+

obtain the best trade-offs between the iteration and the function query complexity. The figure shows the convergence from ZO-PSPIDER+ marginally outperforms the one from ZO-PSVRG+.

Adversarial Attacks on Black-Box DNNs

Adversarial examples in image classification are related to designing unperceptive perturbations such that they lead to misclassifying the target model when added to benign images. In the framework of zeroth-order attacks Chen et al (2017); Liu et al (2018b), the model parameters are hidden, and gradient estimation is not feasible, while the model output could be obtained. We formulate the task of generating a universal adversarial perturbation for n natural images as a ZO optimization problem (1). More precisely, we apply the zeroth-order algorithms to obtain a global adversarial perturbation $x \in \mathbb{R}^d$ that could mislead the classifier on samples $\{a_i \in \mathbb{R}^d, y_i \in \mathbb{N}\}_{i=1}^n$. This problem can be specified as the following elastic-net attacks to black-box DNNs problem:

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \max\{F_{y_i}(a_i^{adv}) - \max_{j\neq y_i} F_j(a_i^{adv}), 0\} + c\left\|a_i^{adv} - a_i\right\|^2 + \lambda_1 \|x\|_1 + \lambda_2 \|x\|^2 \tag{25}$$

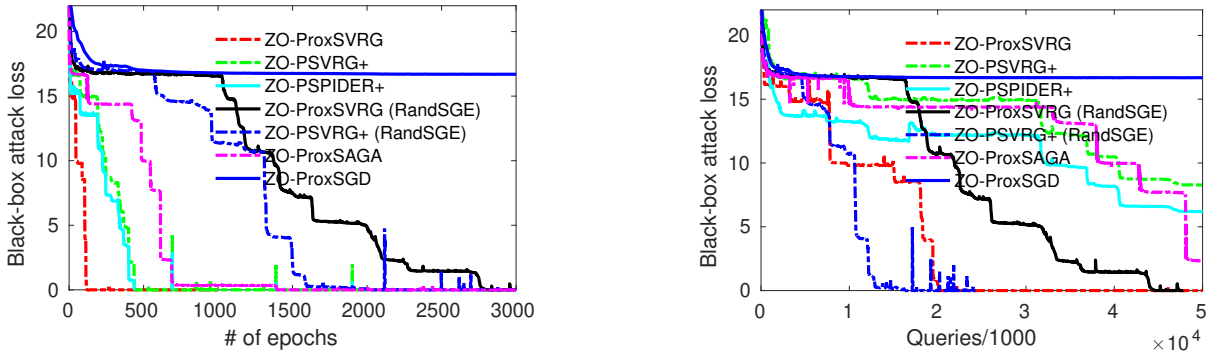
where $a_i^{adv} = 0.5\tanh(\tanh^{-1}(2a_i)+x)$ and λ_1 and λ_2 are nonnegative parameters to obtain consonance between attack success rate, distortion, and sparsity. Here $F(a) = [F_1(a), \ldots, F_K(a)] \in [0,1]^K$ describes a trained deep neural network (DNN) for K-class classification task, where $F_i(a)$ returns the prediction score of i-th class. The parameter c in (25) compensate the rate of adversarial success and the distortion of adversarial examples. In our experiment, we set the regularization parameter $c = 0.2$ and $\lambda_1 = \lambda_2 = 10^{-5}$ for MNIST dataset and $\lambda_1 = \lambda_2 = 0.1$ for CIFAR-10 dataset.

Our experiment setting is to generate universal black-box adversarial perturbation on $n = 10$ samples images from the same class over MNIST and CIFAR-10 datasets. We set the minibatch size to $b = 5$. We select the batch size $\mathcal{B} = \lfloor \frac{n}{2} \rfloor$ for ZO-PSVRG+ and ZO-PSPIDER+. Figures 3 and 4 show the performance of different ZO algorithms in this paper over MNIST and CIFAR-10 datasets, respectively. Figure 3 shows that our two algorithms ZO-PSVRG+ (RandSGE) and ZO-ProxSVRG (under our improved analysis) provide better performance both in convergence rate (iteration complexity) and function query complexity than ZO-ProxSGD and ZO-ProxSAGA. The performance of ZO-PSVRG+ (CoordSGE) algorithm degrades due to a large number of function queries for CoordSGE and the variance inherited by $\mathcal{B} \neq n$. ZO-PSVRG+ (RandSGE) shows faster convergence in the initial optimization stage, and more importantly, has much lower function query complexity, which is largely due to efficient ZO queries for computing mix gradient (8) and the $O(\frac{1}{\sqrt{d}})$-level stepsize required by ZO-PSVRG+ (RandSGE). ZO-ProxSAGA and ZO-PSVRG+ (CoordSGE) exhibit relatively similar convergence behaviors. Furthermore, the convergence performance of ZO-ProxSGD is poor compared to the other algorithms due to not using variance-reduced acceleration techniques. Similar observations are found in Figure 4 for CIFAR-10 dataset where the results for ZO-PSVRG+ and ZO-PSPIDER+ outperform the competitors in terms of convergence rate and function queries.

We visualize the pattern of generated universal perturbations and eventually the adversarial examples for class label "1" in Figure 5 and in Figure 1 in the supplementary materials. Figure 5 shows that the ZO algorithms with CoordSGE produce structured perturbations whereas the adversarial distortion produced with RandSGE scheme is unstructured and presents irregular patterns. We also observe that the patterns of the crafted universal perturbation form ZO-PSVRG+ and ZO-PSPIDER+ identify the most distinctive image segments corresponding to the ground-truth label "1". In addition, while each proximal ZO algorithm produces universal adversarial perturbation leading to mostly successful black-box attacks (misclassified samples), our attacks yield the best trade-offs between the attack success rate and the amount of the produced universal distortion in L_2-norm.

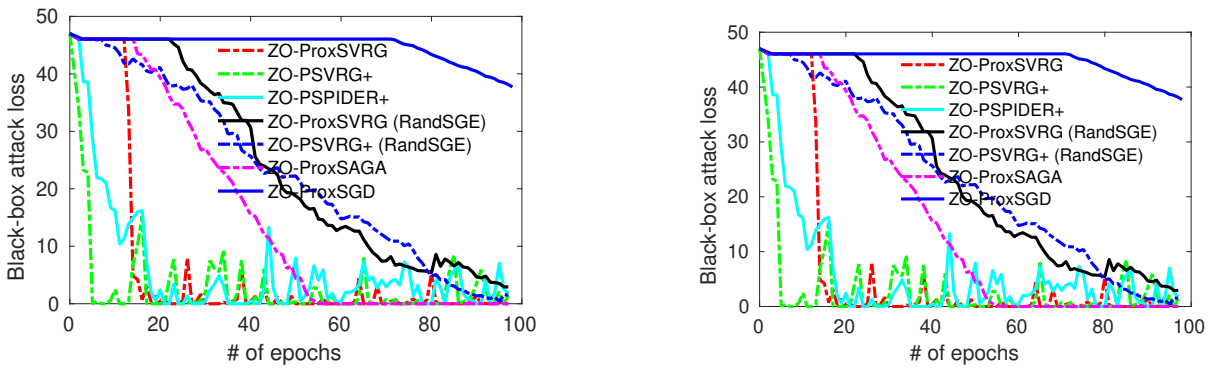
Conclusion

In this paper, we developed a novel analysis for two zeroth-order variance-reduced proximal algorithms, ZO-PSVRG+ and ZO-PSPIDER+ with CoordSGE and RangSGE zeroth-order gradient



(a) Loss vs iterations: $n = 10$ (b) Loss vs queries: $n = 10$

Fig. 3: Comparison of different zeroth-order algorithms for generating black-box adversarial examples from a black-box DNN over MNIST dataset



(a) Loss vs iterations: $n = 10$ (b) Loss vs queries: $n = 10$

Fig. 4: Comparison of different zeroth-order algorithms for generating black-box adversarial examples from a black-box DNN over CIFAR-10 dataset

estimation and bring the convergence analyses from all the studied algorithms into uniformity. Our convergence studies generalize and improve the analysis of several well-known convergence results, e.g., ZO-ProxSVRG and SPIDER-SZO. Compared with ZO-SVRG-Coord-Rand Ji et al (2019), our analyses provide convergence guarantee and SZO calls complexity for all the minibatch sizes with large stepsizes. Moreover, for nonconvex functions under Polyak-Łojasiewicz condition, we prove that ZO-PSVRG+ and ZO-PSPIDER+ obtain a global linear convergence rate for a wide range of minibatch sizes without restart. The empirical results demonstrate the effectiveness of our novel approaches. As a byproduct, our analysis provides the first step to lead the convergence studies of variance-reduced methods into compatible analyses which can be applied to other variance-reduced gradient-free approaches.

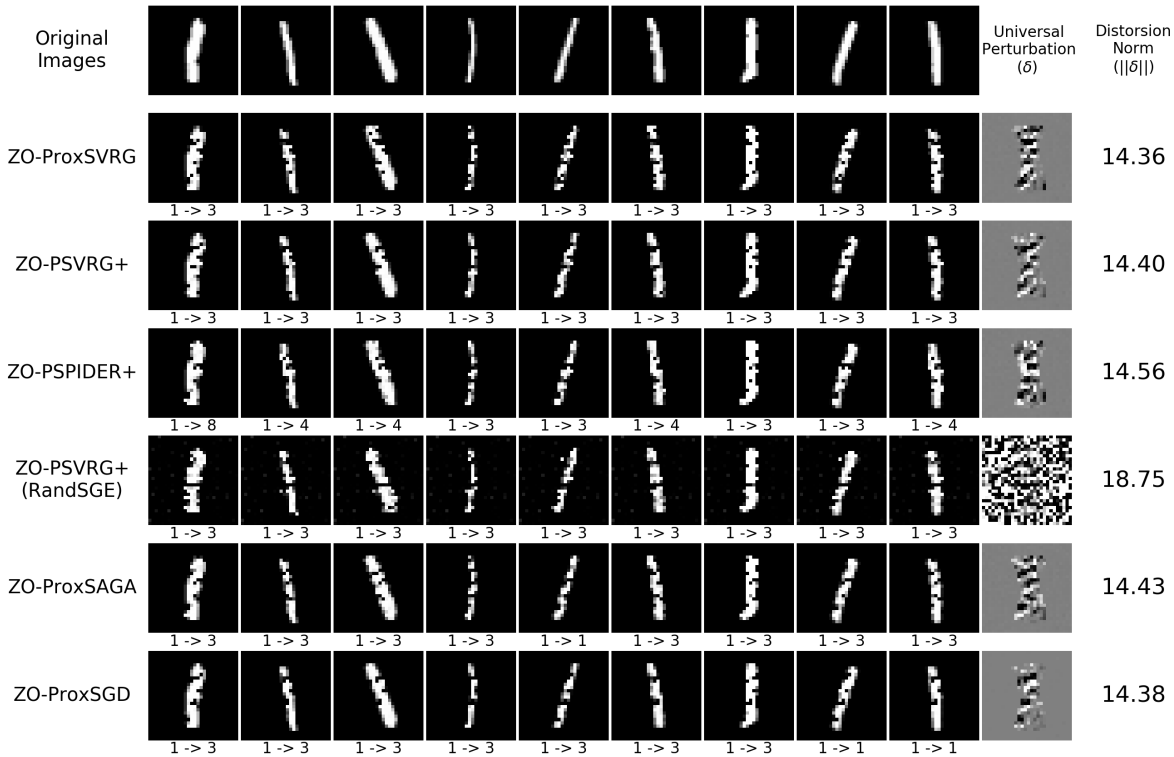


Fig. 5: Images of universal perturbation $\delta := a_i^{adv} - a_i$ from different proximal ZO algorithms and the subsequent generated adversarial images for the class label "1" of the MNIST dataset. The left nine columns show the generated adversarial examples crafted by different ZO algorithms. The second last columns present a visualization for the universal perturbation δ for each ZO algorithm and the last column shows the L_2-norm of the generated universal perturbation ($\|\delta\|$). The bottom of each subplot for adversarial attacks shows the label of the benign sample and the predicted label of the corresponding adversary (ground-truth label -> predicted label).

Funding Not applicable.
Availability of data and material Not applicable.
Declarations
Conflict of interest Not applicable.
Code availability Available.
Additional declarations Not applicable.
Ethics approval Not applicable.
Consent to participate Not applicable.
Consent for publication Not applicable

References

Allen-Zhu Z, Yuan Y (2016) Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In: International conference on machine learning, pp 1080–1089

Chen PY, Zhang H, Sharma Y, et al (2017) Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, ACM, pp 15–26

Chen X, Liu S, Xu K, et al (2019) Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. In: Advances in Neural Information Processing Systems, pp 7202–7213

Duchi JC, Jordan MI, Wainwright MJ, et al (2015) Optimal rates for zero-order convex optimization: The power of two function evaluations. IEEE Transactions on Information Theory 61(5):2788–2806

Fang C, Li CJ, Lin Z, et al (2018) Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In: Advances in Neural Information Processing Systems, pp 689–699

Flaxman AD, Kalai AT, McMahan HB (2005) Online convex optimization in the bandit setting: gradient descent without a gradient. In: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, pp 385–394

Gao X, Jiang B, Zhang S (2018) On the information-adaptive variants of the admm: an iteration complexity perspective. Journal of Scientific Computing 76(1):327–363

Ghadimi S, Lan G (2013) Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization 23(4):2341–2368

Ghadimi S, Lan G (2016) Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming 156(1-2):59–99

Ghadimi S, Lan G, Zhang H (2016) Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Mathematical Programming 155(1):267–305

Gu B, Huo Z, Deng C, et al (2018a) Faster derivative-free stochastic algorithm for shared memory machines. In: International Conference on Machine Learning, pp 1807–1816

Gu B, Wang D, Huo Z, et al (2018b) Inexact proximal gradient methods for non-convex and non-smooth optimization. In: Thirty-Second AAAI Conference on Artificial Intelligence

Hajinezhad D, Hong M, Garcia A (2019) Zone: Zeroth order nonconvex multi-agent optimization over networks. IEEE Transactions on Automatic Control

Huang F, Gu B, Huo Z, et al (2019) Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. In: AAAI

Huang F, Tao L, Chen S (2020) Accelerated stochastic gradient-free and projection-free methods. arXiv preprint arXiv:200712625

Ji K, Wang Z, Zhou Y, et al (2019) Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In: International Conference on Machine Learning, pp 3100–3109

Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in neural information processing systems, pp 315–323

Karimi H, Nutini J, Schmidt M (2016) Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp 795–811

Lei L, Ju C, Chen J, et al (2017) Non-convex finite-sum optimization via scsg methods. In: Advances in Neural Information Processing Systems, pp 2348–2358

Li Z, Li J (2018) A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In: Advances in Neural Information Processing Systems, pp 5564–5574

Lian X, Zhang H, Hsieh CJ, et al (2016) A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. In: Advances in Neural Information Processing Systems, pp 3054–3062

Liu L, Cheng M, Hsieh CJ, et al (2018a) Stochastic zeroth-order optimization via variance reduction method. arXiv preprint arXiv:180511811

Liu S, Chen J, Chen PY, et al (2017) Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. arXiv preprint arXiv:171007804

Liu S, Kailkhura B, Chen PY, et al (2018b) Zeroth-order stochastic variance reduction for nonconvex optimization. In: Advances in Neural Information Processing Systems, pp 3727–3737

Nesterov Y, Spokoiny V (2011) Random gradient-free minimization of convex functions. Tech. rep., Université catholique de Louvain, Center for Operations Research and Econometrics (CORE)

Nesterov Y, Spokoiny V (2017) Random gradient-free minimization of convex functions. Foundations of Computational Mathematics 17(2):527–566

Nguyen LM, Liu J, Scheinberg K, et al (2017a) Sarah: A novel method for machine learning problems using stochastic recursive gradient. In: International Conference on Machine Learning, PMLR, pp 2613–2621

Nguyen LM, Liu J, Scheinberg K, et al (2017b) Stochastic recursive gradient algorithm for nonconvex optimization. arXiv preprint arXiv:170507261

Nitanda A (2016) Accelerated stochastic gradient descent for minimizing finite sums. In: Artificial Intelligence and Statistics, pp 195–203

Parikh N, Boyd S, et al (2014) Proximal algorithms. Foundations and Trends® in Optimization 1(3):127–239

Polyak BT (1963) Gradient methods for minimizing functionals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki 3(4):643–653

Reddi SJ, Hefny A, Sra S, et al (2016a) Stochastic variance reduction for nonconvex optimization. In: International conference on machine learning, pp 314–323

Reddi SJ, Sra S, Póczos B, et al (2016b) Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In: Advances in Neural Information Processing Systems, pp 1145–1153

Sahu AK, Zaheer M, Kar S (2019) Towards gradient free and projection free stochastic optimization. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp 3468–3477

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

Wang Z, Ji K, Zhou Y, et al (2019) Spiderboost and momentum: Faster variance reduction algorithms. In: Advances in Neural Information Processing Systems, pp 2406–2416

Wibisono A, Wainwright MJ, Jordan MI, et al (2012) Finite sample convergence rates of zero-order stochastic optimization methods. In: Advances in Neural Information Processing Systems, pp 1439–1447

Xiao L, Zhang T (2014) A proximal stochastic gradient method with progressive variance reduction. SIAM Journal on Optimization 24(4):2057–2075

Supplemental Materials

In this section, we present the complete proofs of the above lemmas and theorems.

The following two lemmas are similar to (Reddi et al, 2016b, Lemma 1, Lemma 2), (Ghadimi and Lan, 2013, Lemma 1, Proposition 1) and (Li and Li, 2018, Lemma 1, Lemma 2). We added the proof for these lemmas for completeness.

Lemma 13 For a given $x \in \mathbb{R}^d$, let $\overline{x} = \text{Prox}_{\eta h}(x - \eta v)$, then we have for all $w \in \mathbb{R}^d$

$$\begin{aligned} F(\overline{x}) \leq F(w) + \langle \nabla f(x) - v, \overline{x} - w \rangle - \frac{1}{\eta} \langle \overline{x} - x, \overline{x} - w \rangle \\ + \frac{L}{2} \|\overline{x} - x\|^2 + \frac{L}{2} \|w - x\|^2 \end{aligned} \tag{S1}$$

Proof First, we recall the proximal operator

$$\text{Prox}_{\eta h}(x - \eta v) := \arg \min_{y \in \mathbb{R}^d} \left(h(y) + \frac{1}{2\eta} \|y - x\|^2 + \langle v, y \rangle \right) \tag{S2}$$

For the nonsmooth function $h(x)$, for all $w \in \mathbb{R}^d$ we have

$$\begin{aligned} h(\overline{x}) &\leq h(w) + \langle p, \overline{x} - w \rangle \\ &= h(w) - \left\langle v + \frac{1}{\eta}(\overline{x} - x), \overline{x} - w \right\rangle \end{aligned} \tag{S3}$$

where $p \in \partial h(\overline{x})$ such that $p + \frac{1}{\eta}(\overline{x} - x) + v = 0$ according to the optimality condition of (S2), and (S3) due to the convexity of h. In addition, since since $f(x)$ is L-Lipschitz continuous, we have

$$f(\overline{x}) \leq f(x) + \langle \nabla f(x), \overline{x} - x \rangle + \frac{L}{2} \|\overline{x} - x\|^2 \tag{S4}$$

and

$$f(x) \leq f(w) + \langle \nabla f(x), x - w \rangle + \frac{L}{2} \|w - x\|^2 \tag{S5}$$

This lemma is obtained by adding (S3), (S4), (S5), and using $F(x) = f(x) + h(x)$. □

Lemma 14 Let $x_t^s = \text{Prox}_{\eta h}(x_{t-1}^s - \eta \hat{v}_{t-1}^s)$ and $\overline{x}_t^s$ be the proximal projection using full true gradient, i.e., $\overline{x}_t^s = \text{Prox}_{\eta h}(x_{t-1}^s - \eta \nabla f(x_{t-1}^s))$. Then the following inequality holds

$$\left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \leq \eta \left\| \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s \right\|^2$$

Proof Based on inequality (S3) we obtain

$$h(x_t^s) \leq h(\overline{x}_t^s) - \left\langle \hat{v}_{t-1}^s + \frac{1}{\eta}(x_t^s - x_{t-1}^s), x_t^s - \overline{x}_t^s \right\rangle \tag{S6}$$

$$h(\overline{x}_t^s) \leq h(x_t^s) - \left\langle \nabla f(x_{t-1}^s) + \frac{1}{\eta}(\overline{x}_t^s - x_{t-1}^s), \overline{x}_t^s - x_t^s \right\rangle \tag{S7}$$

By summing (S6) and (S7), we have

$$\begin{aligned} \frac{1}{\eta} \left\langle x_t^s - \overline{x}_t^s, x_t^s - \overline{x}_t^s \right\rangle &\leq \left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\ \frac{1}{\eta} \left\| x_t^s - \overline{x}_t^s \right\|^2 &\leq \left\| \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s \right\| \left\| x_t^s - \overline{x}_t^s \right\| \end{aligned} \tag{S8}$$

where (S8) holds by Cauchy-Schwarz inequality. Thus, we obtain

$$\left\| x_t^s - \overline{x}_t^s \right\| \leq \eta \left\| \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s \right\| \tag{S9}$$

Now the proof is complete using Cauchy-Schwarz inequality and (S9). □

Below, we start by deriving an upper bound for the variance of estimated gradient $\hat{v}_{t-1}^s$ based on CoordSGE.

Lemma 15 Given the mix gradient estimation $\hat{v}_{t-1}^s = \frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) + \hat{g}^s$ with $\hat{g}^s = \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1})$, the following inequality holds.

$$\begin{aligned}\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2\right] \le \frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\ + 2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2}\end{aligned} \tag{S10}$$

where $I(A) = 1$ if the scenario A occurs and 0 otherwise.

Lemma 15 provides an upper bound for the variance of $\hat{v}_{t-1}^s$. We will show later that both points x_{t-1}^s and $\tilde{x}^{s-1}$ will converge to the same stationary point. This results in reducing the variance of stochastic gradient, however, the variance is not totally diminished due to the zeroth-order gradient estimation and the variance of the gradient over batch.

Proof We have

$$\begin{aligned}&\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2\right] \\ =&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \hat{g}^s\right)\right\|^2\right] \\ =&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1})\right)\right\|^2\right] \\ =&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right) + \left(\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \\ =&\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\nabla f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right) + \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \\ \le&2\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\hat{\nabla} f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right) + \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] \\ &+ 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad \text{(S11)} \\ =&2\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\hat{\nabla} f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right)\right\|^2\right] \\ &+ 2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1}) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right] + 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad \text{(S12)} \\ =&\frac{2\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\left(\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) - \left(\hat{\nabla} f(x_{t-1}^s) - \hat{\nabla} f(\tilde{x}^{s-1})\right)\right)\right\|^2\right]\end{aligned}$$

$$+ 2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla}f_j(\tilde{x}^{s-1})-\hat{\nabla}f(\tilde{x}^{s-1})\right)\right\|^2\right]+2\eta\mathbb{E}\left\|\hat{\nabla}f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \tag{S13}$$

$$\leq\frac{2\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\hat{\nabla}f_i(x_{t-1}^s)-\hat{\nabla}f_i(\tilde{x}^{s-1})\right\|^2\right]$$
$$+ 2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla}f_j(\tilde{x}^{s-1})-\hat{\nabla}f(\tilde{x}^{s-1})\right)\right\|^2\right]+2\eta\mathbb{E}\left\|\hat{\nabla}f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \tag{S14}$$

$$\leq\frac{3\eta L^2d^2\mu^2}{b}+\frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla}f_j(\tilde{x}^{s-1})-\hat{\nabla}f(\tilde{x}^{s-1})\right)\right\|^2\right]$$
$$+ 2\eta\mathbb{E}\left\|\hat{\nabla}f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \tag{S15}$$

$$\leq\frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+2\eta\mathbb{E}\left\|\hat{\nabla}f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2$$
$$+\frac{3\eta L^2d^2\mu^2}{b} \tag{S16}$$

$$\leq\frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}$$
$$+\eta\frac{L^2d^2\mu^2}{2}+\frac{3\eta L^2d^2\mu^2}{b} \tag{S17}$$

$$\leq\frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\eta\frac{7L^2d^2\mu^2}{2} \tag{S18}$$

where recalling that a deterministic gradient estimator is employed and the expectations are taken with respect to I_b and $I_{\mathcal{B}}$. The inequality (S11) holds by Jensen's inequality. (S12) and (S13) are due to $\mathbb{E}[\|x_1+x_2+\ldots+x_k\|^2]=\sum_{i=1}^k\mathbb{E}[\|x_i\|^2]$ if $x_1,x_2,\ldots,x_k$ are independent and of mean zero. Recall that I_b and $I_{\mathcal{B}}$ are also independent. (S14) applies the fact that for any random variable z, $\mathbb{E}[\|z-\mathbb{E}[z]\|^2]\leq\mathbb{E}[\|z\|^2]$. (S15) employs following inequality

$$\begin{aligned}\mathbb{E}\left\|\hat{\nabla}f_i(x_t^s)-\hat{\nabla}f_i(\tilde{x}^s)\right\|^2&=\mathbb{E}\left\|\hat{\nabla}f_i(x_t^s)-\nabla f_i(x_t^s)+\nabla f_i(x_t^s)-\hat{\nabla}f_i(\tilde{x}^s)+\nabla f_i(\tilde{x}^s)-\nabla f_i(\tilde{x}^s)\right\|^2\\&\leq 3\mathbb{E}\left\|\hat{\nabla}f_i(x_t^s)-\nabla f_i(x_t^s)\right\|^2+3\left\|\hat{\nabla}f_i(\tilde{x}^s)-\nabla f_i(\tilde{x}^s)\right\|^2\\&\quad+3\left\|\nabla f_i(x_t^s)-\nabla f_i(\tilde{x}^s)\right\|^2\\&\leq\frac{3L^2d^2\mu^2}{2}+3L^2\left\|x_t^s-\tilde{x}^s\right\|^2\end{aligned} \tag{S19}$$

where the last inequality used the fact that f_{i,μ_j} is L-smooth. (S16) is by Assumption 2 and (S17) uses Lemma 23. The proof is now complete. □

Below we present the counterpart of Lemma 15 for the mix gradient estimation in (8).

Lemma 16 Given the mix gradient estimation $\tilde{v}_{t-1}^s=\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right)+\hat{g}^s$ with $\hat{g}^s=\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla}f_j(\tilde{x}^{s-1})$, the following inequality holds.

$$\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s)-\tilde{v}_{t-1}^s\right\|^2\right]\leq\frac{6\eta L^2d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]$$
$$+2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\eta\frac{7L^2d^2\mu^2}{2} \tag{S20}$$

Proof We have

$$
\begin{aligned}
&\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s)-\tilde{v}_{t-1}^s\right\|^2\right]\\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\hat{g}^s\right)\right\|^2\right]\\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1})\right)\right\|^2\right]\\
=&\mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\hat{\nabla} f(\tilde{x}^{s-1})\right)+\left(\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]\\
=&\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right)-\left(\nabla f(x_{t-1}^s)-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right)+\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]\\
\le&2\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right)-\left(\hat{\nabla} f(x_{t-1}^s)-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right)+\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]\\
&+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad \text{(S21)}\\
=&2\eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right)-\left(\hat{\nabla} f(x_{t-1}^s)-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right)\right\|^2\right]\\
&+2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad \text{(S22)}\\
=&\frac{2\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\left(\left(\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right)-\left(\hat{\nabla} f(x_{t-1}^s)-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right)\right\|^2\right]\\
&+2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad \text{(S23)}\\
\le&\frac{2\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\hat{\nabla}_r f_i(x_{t-1}^s)-\hat{\nabla}_r f_i(\tilde{x}^{s-1})\right\|^2\right]+2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]\\
&+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad \text{(S24)}\\
\le&\frac{3\eta L^2d^2\mu^2}{b}+\frac{6\eta L^2d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]\\
&+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \qquad \text{(S25)}\\
\le&\frac{6\eta L^2d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2\\
&+\frac{3\eta L^2d^2\mu^2}{b} \qquad \text{(S26)}
\end{aligned}
$$

$$\leq \frac{6\eta L^2 d}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] + 2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{L^2d^2\mu^2}{2} + \frac{3\eta L^2 d^2\mu^2}{b} \tag{S27}$$

$$\leq \frac{6\eta L^2 d}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] + 2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \tag{S28}$$

where, the expectations are taking with respect to I_b and $I_{\mathcal{B}}$ and random directions $\{u_i\}$ in (2). The inequality (S21) holds by Jensen's inequality. (S22) and (S23) are based on $\mathbb{E}[\|x_1 + x_2 + \ldots + x_k\|^2] = \sum_{i=1}^k \mathbb{E}[\|x_i\|^2]$ if $x_1, x_2, \ldots, x_k$ are independent and of mean zero. Note that I_b and $I_{\mathcal{B}}$ are also independent. (S24) uses the fact that $\mathbb{E}[\|x - \mathbb{E}[x]\|^2] \leq \mathbb{E}[\|x\|^2]$, for any random variable x. (S25) holds due to Lemma 24. (S26) is by Assumption 2 and (S27) is by Lemma 23. (S28) uses $b \geq 1$. The proof is now complete. □

In the sequel, we frequently use the following inequality

$$\|x - z\|^2 \leq (1 + \frac{1}{\alpha})\|x - y\|^2 + (1 + \alpha)\|y - z\|^2, \forall \alpha > 0 \tag{S29}$$

Proof of Theorem 1

Proof Now, we apply Lemma 13 to prove Theorem 1. Let $x_t^s = \text{Prox}_{\eta h}(x_{t-1}^s - \eta\hat{v}_{t-1}^s)$ and $\overline{x}_t^s = \text{Prox}_{\eta h}(x_{t-1}^s - \eta\nabla f(x_{t-1}^s))$. By letting $x^+ = x_t^s$, $x = x_{t-1}^s$, $v = \hat{v}_{t-1}^s$ and $z = \overline{x}_t^s$ in (S1), we have

$$\begin{aligned} F(x_t^s) \leq & F(\overline{x}_t^s) + \left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle - \frac{1}{\eta}\left\langle x_t^s - x_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\ & + \frac{L}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 + \frac{L}{2}\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2. \end{aligned} \tag{S30}$$

Besides, by letting $x^+ = \overline{x}_t^s$, $x = x_{t-1}^s$, $v = \nabla f(x_{t-1}^s)$ and $z = x = x_{t-1}^s$ in (S1), we have

$$\begin{aligned} F(\overline{x}_t^s) \leq & F(x_{t-1}^s) - \frac{1}{\eta}\left\langle \overline{x}_t^s - x_{t-1}^s, \overline{x}_t^s - x_{t-1}^s \right\rangle + \frac{L}{2}\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 \\ = & F(x_{t-1}^s) - (\frac{1}{\eta} - \frac{L}{2})\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2. \end{aligned} \tag{S31}$$

Combining (S30) and (S31) we have

$$\begin{aligned} F(x_t^s) \leq & F(x_{t-1}^s) + \frac{L}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\ & - \frac{1}{\eta}\left\langle x_t^s - x_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\ = & F(x_{t-1}^s) + \frac{L}{2}\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\ & - \frac{1}{2\eta}\left(\left\|x_t^s - x_{t-1}^s\right\|^2 + \left\|x_t^s - \overline{x}_t^s\right\|^2 - \left\|\overline{x}_t^s - x_{t-1}^s\right\|^2\right) \\ = & F(x_{t-1}^s) - (\frac{1}{2\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{2\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\ & - \frac{1}{2\eta}\left\|x_t^s - \overline{x}_t^s\right\|^2 \\ \leq & F(x_{t-1}^s) - (\frac{1}{2\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{2\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\ & - \frac{1}{8\eta}\left\|x_t^s - x_{t-1}^s\right\|^2 + \frac{1}{6\eta}\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 \end{aligned} \tag{S32}$$

$$\begin{aligned} = & F(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\ \leq & F(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \eta\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2 \end{aligned} \tag{S33}$$

where the second inequality uses (S29) with $\alpha = 3$ and the last inequality holds due to the Lemma 14.

Note that $x_t^s = \text{Prox}_{\eta h}(x_{t-1}^s - \eta \hat{v}_{t-1}^s)$ is the iterated form in our algorithm. By taking the expectation with respect to all random variables in (S33) we obtain

$$\mathbb{E}[F(x_t^s)] \leq \mathbb{E}\left[F(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \eta\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2\right] \tag{S34}$$

In (S34), we further bound $\eta\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2$ using Lemma 15 to obtain

$$\begin{aligned}
&\mathbb{E}[F(x_t^s)] \\
\leq&\mathbb{E}\left[F(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2\right] \\
&+ \frac{6\eta L^2}{b}\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \\
=&\mathbb{E}\left[F(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ \frac{6\eta L^2}{b}\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \qquad \text{(S35)} \\
\leq&\mathbb{E}\left[F(x_{t-1}^s) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ (\frac{6\eta L^2}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \qquad \text{(S36)}
\end{aligned}$$

where recalling $\overline{x}_t^s := \text{Prox}_{\eta h}(x_{t-1}^s - \eta\nabla f(x_{t-1}^s))$, (S35) is based on the definition of gradient mapping $g_\eta(x_{t-1}^s)$. (S36) uses (S29) by choosing $\alpha = 2t - 1$.

Taking a telescopic sum for $t = 1, 2, \ldots, m$ in epoch s from (S36) and recalling that $x_m^s = \tilde{x}^s$ and $x_0^s = \tilde{x}^{s-1}$, we obtain

$$\begin{aligned}
&\mathbb{E}[F(\tilde{x}^s)] \\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \sum_{t=1}^{m}\frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ \sum_{t=1}^{m}(\frac{6\eta L^2}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 \\
&+ \sum_{t=1}^{m}\frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \sum_{t=1}^{m-1}\frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ \sum_{t=2}^{m}(\frac{6\eta L^2}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 \\
&+ \sum_{t=1}^{m}\frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \qquad \text{(S37)} \\
=&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \left(\frac{\eta}{3} - L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&- \sum_{t=1}^{m-1}\left((\frac{1}{2t} - \frac{1}{2t+1})(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b}\right)\mathbb{E}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 \\
&+ \sum_{t=1}^{m}\frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \left(\frac{\eta}{3} - L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{t=1}^{m-1} \left(\frac{1}{6t^2} (\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b}) \right) \mathbb{E} \left\| x_t^s - \tilde{x}^{s-1} \right\|^2 \\
&+ \sum_{t=1}^{m} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m} \eta \frac{7L^2 d^2 \mu^2}{2} \\
\leq &\mathbb{E} \left[F(\tilde{x}^{s-1}) - \frac{\eta}{6} \sum_{t=1}^{m} \left\| g_\eta(x_{t-1}^s) \right\|^2 \right] + \sum_{t=1}^{m} \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m} \eta \frac{7L^2 d^2 \mu^2}{2}
\end{aligned} \tag{S38}
$$

where (S37) holds since norm is always non-negative and $x_0^s = \tilde{x}^{s-1}$. In (S38) we have used the fact that $(\frac{1}{6t^2}(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b}) \geq 0$ for all $1 \leq t \leq m$ and $\frac{\eta}{6} \leq \frac{\eta}{3} - L\eta^2$ since $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12mL}\}$. Telescoping the sum for $s = 1, 2, \ldots, S$ in (S38), we obtain

$$
\begin{aligned}
0 \leq &\mathbb{E}[F(\tilde{x}^S) - F(x^*)] \\
\leq &\mathbb{E} \left[F(\tilde{x}^0) - F(x^*) - \sum_{s=1}^{S} \sum_{t=1}^{m} \frac{\eta}{6} \left\| g_\eta(x_{t-1}^s) \right\|^2 + \sum_{s=1}^{S} \sum_{t=1}^{m} (\frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta \frac{7L^2 d^2 \mu^2}{2}) \right]
\end{aligned}
$$

Thus, we have

$$
\mathbb{E}[\|g_\eta(\hat{x})\|^2] \leq \frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} + \frac{I(\mathcal{B} < n)12\sigma^2}{\mathcal{B}} + 21L^2 d^2 \mu^2 \tag{S39}
$$

where (S39) holds since we choose $\hat{x}$ uniformly randomly from $\{x_{t-1}^s\}_{t\in[m], s\in[S]}$. □

Proof of Corollary 2

Proof Using Theorem 1, we have $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12mL}\}$

$$
\begin{aligned}
\mathbb{E}[\|g_\eta(\hat{x})\|^2] \leq &\frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} \\
&+ \frac{I(\mathcal{B} < n)12\sigma^2}{\mathcal{B}} + 21L^2 d^2 \mu^2 = 3\epsilon
\end{aligned} \tag{S40}
$$

Now we derive the total number of iterations $T = Sm = \frac{6(F(x_0) - F(x^*))}{\epsilon\eta}$. Since $\mu \leq \frac{\sqrt{\epsilon}}{5dL}$, and for $\mathcal{B} = n$, the second term in the bound (S40) is 0, the proof is completed as the number of SZO call equals to $Sn + Smb = 6\left(F(x_0) - F(x^*)\right)(\frac{n}{\epsilon\eta m} + \frac{b}{\epsilon\eta})$. If $\mathcal{B} < n$ the number of SZO calls equal to $d(S\mathcal{B} + Smb) = 6d\left(F(x_0) - F(x^*)\right)(\frac{\mathcal{B}}{\epsilon\eta m} + \frac{b}{\epsilon\eta})$ by noting that $\frac{I(\mathcal{B}<n)12\sigma^2}{\mathcal{B}} \leq \epsilon$ due to $\mathcal{B} \geq 12\sigma^2/\epsilon$. The second part of the corollary is obtained by setting $m = \sqrt{b}$ in the first part. □

Proof of Theorem 5

Proof We start by recalling inequality (S35) from the proof of Theorem 1, i.e.,

$$
\begin{aligned}
&\mathbb{E}[F(x_t^s)] \\
\leq &\mathbb{E} \left[F(x_{t-1}^s) - \frac{1}{2t} (\frac{5}{8\eta} - \frac{L}{2}) \left\| x_t^s - \tilde{x}^{s-1} \right\|^2 - \left(\frac{\eta}{3} - L\eta^2 \right) \left\| g_\eta(x_{t-1}^s) \right\|^2 \right] \\
&+ (\frac{2\eta L^2 d}{b} + \frac{1}{2t-1} (\frac{5}{8\eta} - \frac{L}{2})) \mathbb{E} \left\| x_{t-1}^s - \tilde{x}^{s-1} \right\|^2 + \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta \frac{L^2 d^2 \mu^2}{2} \\
\leq &\mathbb{E} \left[F(x_{t-1}^s) - \frac{1}{2t} (\frac{5}{8\eta} - \frac{L}{2}) \left\| x_t^s - \tilde{x}^{s-1} \right\|^2 - \frac{\eta}{6} \left\| g_\eta(x_{t-1}^s) \right\|^2 \right] \\
&+ (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1} (\frac{5}{8\eta} - \frac{L}{2})) \mathbb{E} \left\| x_{t-1}^s - \tilde{x}^{s-1} \right\|^2 + \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta \frac{7L^2 d^2 \mu^2}{2}
\end{aligned} \tag{S41}
$$

where in (S41) inequality we applied $\eta L \leq \frac{1}{6}$. Moreover, substituting PL inequality, i.e.,

$$\|g_\eta(x)\|^2 \geq 2\lambda(F(x) - F^*) \tag{S42}$$

into (S41), we obtain

$$\begin{aligned}
&\mathbb{E}[F(x_t^s)] \\
\leq&\mathbb{E}\left[F(x_{t-1}^s) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \lambda\frac{\eta}{3}(F(x_{t-1}^s) - F^*)\right] \\
&+ (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2}
\end{aligned} \tag{S43}$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}[F(x_t^s)] \\
\leq&\mathbb{E}\left[(1 - \lambda\frac{\eta}{3})(F(x_{t-1}^s) - F^*) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
&+ (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2}
\end{aligned} \tag{S44}$$

Let $\beta := 1 - \lambda\frac{\eta}{3}$ and $\Psi_t^s := \frac{\mathbb{E}[F(x_t^s) - F^*]}{\beta^t}$. Combining these definitions with (S44), we have

$$\begin{aligned}
&\Psi_t^s \\
\leq&\Psi_{t-1}^s - \frac{1}{\beta^t}\mathbb{E}\left[\frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
&+ \frac{1}{\beta^t}\frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1}{\beta^t}\eta\frac{7L^2d^2\mu^2}{2}
\end{aligned} \tag{S45}$$

Similar to the proof of Theorem 1, summing (S45) for $t = 1, 2, \ldots, m$ in epoch s and remembering that $x_m^s = \tilde{x}^s$ and $x_0^s = \tilde{x}^{s-1}$, we have

$$\begin{aligned}
&\mathbb{E}[F(\tilde{x}^s) - F^*] \\
\leq&\beta^m\mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \beta^m\sum_{t=1}^m\frac{1}{\beta^t}\frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \beta^m\sum_{t=1}^m\frac{1}{\beta^t}\eta\frac{7L^2d^2\mu^2}{2} \\
&-\beta^m\mathbb{E}\left[\sum_{t=1}^m\frac{1}{2t\beta^t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^m\frac{1}{\beta^t}(\frac{6\eta L^2}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
\leq&\beta^m\mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^m}{1-\beta}\frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^m}{1-\beta}\frac{7\eta L^2d^2\mu^2}{2} \\
&-\beta^m\mathbb{E}\left[\sum_{t=1}^m\frac{1}{2t\beta^t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^m\frac{1}{\beta^t}(\frac{6\eta L^2}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
\leq&\beta^m\mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^m}{1-\beta}\frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^m}{1-\beta}\frac{7\eta L^2d^2\mu^2}{2} \\
&-\beta^m\mathbb{E}\left[\sum_{t=1}^{m-1}\frac{1}{2t\beta^t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
&+\beta^m\mathbb{E}\left[\sum_{t=2}^m\frac{1}{\beta^t}(\frac{6\eta L^2}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
\leq&\beta^m\mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^m}{1-\beta}\frac{2I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^m}{1-\beta}\frac{7\eta L^2d^2\mu^2}{2} \\
&-\beta^m\mathbb{E}\left[\sum_{t=1}^{m-1}\frac{1}{\beta^{t+1}}\left((\frac{\beta}{2t} - \frac{1}{2t+1})(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b}\right)\right)\left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right]
\end{aligned} \tag{S46}$$

$$\leq \beta^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^m}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^m}{1-\beta}\frac{7\eta L^2 d^2\mu^2}{2} \tag{S47}$$

where (S46) since $\|.\|^2$ always is non-negative and $x_0^s = \tilde{x}^{s-1}$. (S47) holds since it is sufficient to show $(\frac{\beta}{2t} - \frac{1}{2t+1})(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b} \geq 0$, for all $t = 1, 2, \ldots, m$. It is easy to see that this inequality is valid due to $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{\gamma b}}{12mL}\}$, where $\gamma = 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} > 0$. Similarly, let $\tilde{\beta} = \beta^m$ and $\tilde{\Psi}^s = \frac{\mathbb{E}[F(\tilde{x}^s) - F^*]}{\tilde{\beta}^s}$. Substituting these definitions into (S47), we have

$$\tilde{\Psi}^s \leq \tilde{\Psi}^{s-1} + \frac{1}{\tilde{\beta}^s}\frac{1-\tilde{\beta}}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1}{\tilde{\beta}^s}\frac{1-\tilde{\beta}}{1-\beta}\frac{7\eta L^2 d^2\mu^2}{2} \tag{S48}$$

Taking a telescopic sum from (S48) for all epochs $1 \leq s \leq S$, we obtain

$$\begin{aligned}
\mathbb{E}[F(\tilde{x}^S) - F^*] &\leq \tilde{\beta}^S \mathbb{E}[F(\tilde{x}^0) - F^*] + \tilde{\beta}^S \sum_{s=1}^{S} \frac{1}{\tilde{\beta}^s}\frac{1-\tilde{\beta}}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \tilde{\beta}^S \sum_{s=1}^{S} \frac{1}{\tilde{\beta}^s}\frac{1-\tilde{\beta}}{1-\beta}\frac{7\eta L^2 d^2\mu^2}{2} \\
&= \beta^{Sm}\mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{1-\tilde{\beta}^S}{1-\tilde{\beta}}\frac{1-\tilde{\beta}}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\tilde{\beta}^S}{1-\tilde{\beta}}\frac{1-\tilde{\beta}}{1-\beta}\frac{7\eta L^2 d^2\mu^2}{2} \\
&\leq \beta^{Sm}\mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{1}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1}{1-\beta}\frac{7\eta L^2 d^2\mu^2}{2} \\
&= \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2 d^2\mu^2}{2\lambda}
\end{aligned} \tag{S49}$$

where in (S49) we recall that $\beta = 1 - \frac{\lambda\eta}{3}$. □

Proof of Corollary 6

Proof From Theorem 5, we have

$$\begin{aligned}
\mathbb{E}[F(\tilde{x}^S) - F^*] &\leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] \\
&\quad + \frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2 d^2\mu^2}{2\lambda} = 3\epsilon
\end{aligned}$$

which gives the total number of iterations $T = Sm = O(\frac{1}{\lambda\eta}\log\frac{1}{\epsilon})$ and is equal to the number of PO calls. Since $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4Ld}$, we have $\frac{21L^2 d^2\mu^2}{2\lambda} \leq \epsilon$. The number of SZO calls is equal to $d(S\mathcal{B} + Smb) = O(\frac{\mathcal{B}d}{\lambda\eta m}\log\frac{1}{\epsilon} + \frac{bd}{\lambda\eta}\log\frac{1}{\epsilon})$. Note that if $\mathcal{B} < n$ then $\frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}} \leq \epsilon$ since $\mathcal{B} \geq 6\sigma^2/\lambda\epsilon$. With $m = \sqrt{b}$ and $\eta = \frac{\sqrt{\gamma}}{12L}$, the number of PO queries to $T = Sm = O(\frac{1}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$ and the number of SZO calls is equal to $d(S\mathcal{B} + Smb) = O(\frac{\mathcal{B}d}{\lambda\sqrt{\gamma}m}\log\frac{1}{\epsilon} + \frac{bd}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$. □

Proof of Corollary 8

Proof From Theorem 7, we have

$$\begin{aligned}
\mathbb{E}[F(\tilde{x}^S) - F^*] &\leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] \\
&\quad + \frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2 d^2\mu^2}{2\lambda} = 3\epsilon
\end{aligned} \tag{S50}$$

which gives the total number of iterations $T = Sm = O(\frac{1}{\lambda\eta}\log\frac{1}{\epsilon})$ and equals to the number of PO calls. Since $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4Ld}$, we have $\frac{21L^2 d^2\mu^2}{2\lambda} \leq \epsilon$. The number of SZO calls equals to $(S\mathcal{B}d + Smb) = O(\frac{\mathcal{B}d}{\lambda\eta m}\log\frac{1}{\epsilon} + \frac{b}{\lambda\eta}\log\frac{1}{\epsilon})$. Note that if $\mathcal{B} < n$ then $\frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}} \leq \epsilon$ since $\mathcal{B} \geq 6\sigma^2/\lambda\epsilon$. With $m = \sqrt{b}$ and $\eta = \frac{\sqrt{\gamma}}{12L\sqrt{d}}$, the number of PO calls equals to $T = Sm = O(\frac{\sqrt{d}}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$ and the number of SZO calls equals to $(S\mathcal{B}d + Smb) = O(\frac{\mathcal{B}d\sqrt{d}}{\lambda\sqrt{\gamma}m}\log\frac{1}{\epsilon} + \frac{b\sqrt{d}}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$. □

ZO Proximal Stochastic Method (ZO-PSPIDER+)

Lemma 17 Given the mix gradient estimation $\hat{v}_t^s = \frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(x_{t-1}^s)\right) + \hat{v}_{t-1}^s$ with $\hat{v}_0^s = \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(x_0^s)$, the following inequality holds.

$$\mathbb{E}\left[\eta\left\|\nabla f(x_t^s) - \hat{v}_t^s\right\|^2\right] \leq \frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left[\left\|x_l^s - x_{l-1}^s\right\|^2\right] + 2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2 t}{2} \tag{S51}$$

where $I(A) = 1$ if the scenario A occurs and 0 otherwise.

Proof Based on the property of square-integrable martingales Fang et al (2018), we have for $1 \leq t \leq m$

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{\nabla} f(x_t^s) - \hat{v}_t^s\right\|^2\right]\\ &= \mathbb{E}\left[\left\|\hat{\nabla} f(x_0^s) - \hat{v}_0^s\right\|^2\right] + \sum_{l=1}^{t}\mathbb{E}\left[\left\|\hat{v}_l^s - \hat{v}_{l-1}^s - (\hat{\nabla} f(x_l^s) - \hat{\nabla} f(x_{l-1}^s))\right\|^2\right]\end{aligned} \tag{S52}$$

We obtain from the above equality

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{\nabla} f(x_t^s) - \hat{v}_t^s\right\|^2\right]\\ &= \mathbb{E}\left[\left\|\hat{\nabla} f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2\right] + \mathbb{E}\left[\left\|\hat{v}_t^s - \hat{v}_{t-1}^s - (\hat{\nabla} f(x_t^s) - \hat{\nabla} f(x_{t-1}^s))\right\|^2\right]\end{aligned}$$

We have

$$\begin{aligned}&\mathbb{E}\left[\eta\left\|\hat{v}_t^s - \hat{v}_{t-1}^s - (\hat{\nabla} f(x_t^s) - \hat{\nabla} f(x_{t-1}^s))\right\|^2\right]\\ &= \mathbb{E}\left[\eta\left\|\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(x_{t-1}^s)\right) - \left(\hat{\nabla} f(x_t^s) - \hat{\nabla} f(x_{t-1}^s)\right)\right\|^2\right] &\text{(S53)}\\ &= \eta\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\left(\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(x_{t-1}^s)\right) - \left(\hat{\nabla} f(x_t^s) - \hat{\nabla} f(x_{t-1}^s)\right)\right)\right\|^2\right]\\ &= \frac{\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\left(\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(x_{t-1}^s)\right) - \left(\hat{\nabla} f(x_t^s) - \hat{\nabla} f(x_{t-1}^s)\right)\right\|^2\right] &\text{(S54)}\\ &\leq \frac{\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(x_{t-1}^s)\right\|^2\right] &\text{(S55)}\\ &\leq \frac{3\eta L^2d^2\mu^2}{2b} + \frac{3\eta L^2}{b}\mathbb{E}\left[\left\|x_t^s - x_{t-1}^s\right\|^2\right] &\text{(S56)}\end{aligned}$$

where recalling that a deterministic gradient estimator is employed and the expectations are taken with respect to I_b. The equality (S53) is implied by $\hat{v}_t^s - \hat{v}_{t-1}^s = \frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(x_{t-1}^s)\right)$. (S54) is due to $\mathbb{E}[\|x_1 + x_2 + \ldots + x_k\|^2] = \sum_{i=1}^{k}\mathbb{E}[\|x_i\|^2]$ if $x_1, x_2, \ldots, x_k$ are independent and of mean zero. (S55) applies the fact that for any random variable z, $\mathbb{E}[\|z - \mathbb{E}[z]\|^2] \leq \mathbb{E}[\|z\|^2]$. (S56) employs following inequality

$$\begin{aligned}\mathbb{E}\left\|\hat{\nabla} f_i(x_t^s) - \hat{\nabla} f_i(x_{t-1}^s)\right\|^2 &= \mathbb{E}\left\|\hat{\nabla} f_i(x_t^s) - \nabla f_i(x_{t-1}^s) + \nabla f_i(x_t^s) - \hat{\nabla} f_i(x_{t-1}^s) + \nabla f_i(x_{t-1}^s) - \nabla f_i(x_t^s)\right\|^2 &\text{(S57)}\\ &\leq 3\mathbb{E}\left\|\hat{\nabla} f_i(x_t^s) - \nabla f_i(x_t^s)\right\|^2 + 3\left\|\hat{\nabla} f_i(x_{t-1}^s) - \nabla f_i(x_{t-1}^s)\right\|^2\end{aligned}$$

$$
\begin{aligned}
&+3\left\|\nabla f_i(x_t^s)-\nabla f_i(x_{t-1}^s)\right\|^2\\
&\leq \frac{3L^2d^2\mu^2}{2}+3L^2\left\|x_t^s-x_{t-1}^s\right\|^2
\end{aligned}
\tag{S58}
$$

where the last inequality used the fact that f_{i,μ_j} is L-smooth and Lemma 23. Similarly, regarding the term $\mathbb{E}\left[\left\|\hat{\nabla} f(x_0^s)-\hat{v}_0^s\right\|^2\right]$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\eta\left\|\hat{\nabla} f(x_0^s)-\hat{v}_0^s\right\|^2\right]\\
=&\mathbb{E}\left[\eta\left\|\hat{\nabla} f(x_0^s)-\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(x_0^s)\right\|^2\right]\\
=&\mathbb{E}\left[\eta\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f(x_0^s)-\hat{\nabla} f_j(x_0^s)\right\|^2\right]\\
\leq&\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}
\end{aligned}
\tag{S59}
$$

From (S52), by summing on $l=0,\ldots,t$ we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\eta\left\|\hat{\nabla} f(x_t^s)-\hat{v}_t^s\right\|^2\right]\\
&=\mathbb{E}\left[\eta\left\|\hat{\nabla} f(x_0^s)-\hat{v}_0^s\right\|^2\right]+\sum_{l=1}^{t}\mathbb{E}\left[\eta\left\|\hat{v}_l^s-\hat{v}_{l-1}^s-(\hat{\nabla} f(x_l^s)-\hat{\nabla} f(x_{l-1}^s))\right\|^2\right]\\
&\leq\frac{3\eta L^2d^2\mu^2t}{2b}+\frac{3\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left[\left\|x_l^s-x_{l-1}^s\right\|^2\right]+\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}
\end{aligned}
\tag{S60}
$$

where (S60) uses (S58) and (S59). Finally, using the above inequality we obtain the following

$$
\begin{aligned}
&\mathbb{E}\left[\eta\left\|\nabla f(x_t^s)-\hat{v}_t^s\right\|^2\right]\\
&\leq 2\eta\mathbb{E}\left[\left\|\nabla f(x_t^s)-\hat{\nabla} f(x_t^s)\right\|\right]+2\eta\mathbb{E}\left[\left\|\hat{\nabla} f(x_t^s)-\hat{v}_t^s\right\|^2\right]\\
&\leq\frac{\eta L^2d^2\mu^2}{2}+\frac{3\eta L^2d^2\mu^2t}{b}+\frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left[\left\|x_l^s-x_{l-1}^s\right\|^2\right]+\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} \qquad \text{(S61)}\\
&=\frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left[\left\|x_l^s-x_{l-1}^s\right\|^2\right]+\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\frac{7\eta L^2d^2\mu^2t}{2} \qquad \text{(S62)}
\end{aligned}
$$

The proof is now complete. □

Below we present the counterpart of Lemma 17 for the mix gradient estimation in (21).

Lemma 18 Given the mix gradient estimation $\hat{v}_t^s=\frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_t^s)-\hat{\nabla} f_i(x_{t-1}^s)\right)+\hat{v}_{t-1}^s$ with $\hat{v}_0^s=\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(x_0^s)$, the following inequality holds.

$$
\begin{aligned}
\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s)-\tilde{v}_{t-1}^s\right\|^2\right]\leq&\frac{6\eta L^2d}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]\\
&+2\frac{I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}}+\eta\frac{7L^2d^2\mu^2}{2}
\end{aligned}
\tag{S63}
$$

Proof The proof is obtained by following the same steps as the proof of Lemma 17 except for the estimate in (S58) where we apply the approximation error from Lemma 24 instead of Lemma 23. □

Convergence Analysis for ZO-PSPIDER+

Proof of Theorem 9

Proof We start by recalling inequality (S34) from the proof of Theorem 1, i.e.,

$$\mathbb{E}[F(x_{t+1}^s)] \leq \mathbb{E}\left[F(x_t^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_{t+1}^s - x_t^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_{t+1}^s - x_t^s\right\|^2 + \eta\left\|\nabla f(x_t^s) - \hat{v}_t^s\right\|^2\right] \tag{S64}$$

In (S64), we further bound $\eta\left\|\nabla f(x_t^s) - \hat{v}_t^s\right\|^2$ using Lemma 17 to obtain

$$\begin{aligned}
&\mathbb{E}[F(x_{t+1}^s)] \\
\leq&\mathbb{E}\left[F(x_t^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_{t+1}^s - x_t^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_{t+1}^s - x_t^s\right\|^2\right] \\
&+ \frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2t}{2} \\
=&\mathbb{E}\left[F(x_t^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_{t+1}^s - x_t^s\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_t^s)\right\|^2\right] \\
&+ \frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2t}{2}
\end{aligned} \tag{S65}$$

where recalling $\overline{x}_{t+1}^s := \mathrm{Prox}_{\eta h}(x_t^s - \eta\nabla f(x_t^s))$, (S65) is based on the definition of gradient mapping $g_\eta(x_t^s)$. Taking a telescopic sum for $t = 0, 1, 2, \ldots, m$ in epoch s from (S65) and recalling that $x_{m+1}^s = \tilde{x}^s$ and $x_0^s = \tilde{x}^{s-1}$, we obtain

$$\begin{aligned}
&\mathbb{E}[F(\tilde{x}^s)] \\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1}) - (\frac{5}{8\eta} - \frac{L}{2})\sum_{t=0}^{m}\left\|x_{t+1}^s - x_t^s\right\|^2 - \sum_{t=0}^{m}\left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_t^s)\right\|^2\right] \\
&+ \frac{6\eta L^2}{b}\sum_{t=0}^{m}\sum_{l=1}^{t}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2 + \sum_{t=0}^{m}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=0}^{m}\eta\frac{7L^2d^2\mu^2t}{2} \qquad \text{(S66)} \\
=&\mathbb{E}\left[F(\tilde{x}^{s-1}) - (\frac{5}{8\eta} - \frac{L}{2})\sum_{t=0}^{m}\left\|x_{t+1}^s - x_t^s\right\|^2 - \sum_{t=0}^{m}\left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_t^s)\right\|^2\right] \\
&+ \frac{6\eta L^2 m}{b}\sum_{l=1}^{m}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2 + \sum_{t=0}^{m}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=0}^{m}\eta\frac{7L^2d^2\mu^2t}{2} \qquad \text{(S67)} \\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1}) - (\frac{5}{8\eta} - \frac{L}{2} - \frac{6\eta L^2 m}{b})\sum_{t=1}^{m}\left\|x_t^s - x_{t-1}^s\right\|^2 - \sum_{t=0}^{m}\left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_t^s)\right\|^2\right] \\
&+ \sum_{t=0}^{m}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=0}^{m}\eta\frac{7L^2d^2\mu^2t}{2} \qquad \text{(S68)} \\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \left(\frac{\eta}{3} - L\eta^2\right)\sum_{t=0}^{m}\left\|g_\eta(x_t^s)\right\|^2\right] \\
&- \left(\frac{5}{8\eta} - \frac{L}{2} - \frac{6\eta L^2 m}{b}\right)\sum_{t=1}^{m}\mathbb{E}\left\|x_t^s - x_{t-1}^s\right\|^2 \\
&+ \sum_{t=0}^{m}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=0}^{m}\eta\frac{7L^2d^2\mu^2t}{2} \\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \frac{\eta}{6}\sum_{t=0}^{m}\left\|g_\eta(x_t^s)\right\|^2\right] + \sum_{t=0}^{m}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \sum_{t=0}^{m}\eta\frac{7L^2d^2\mu^2t}{2} \qquad \text{(S69)}
\end{aligned}$$

where (S68) holds since $\frac{5}{8\eta} - \frac{L}{2} \geq 0$. In (S69) we have used the fact that $\frac{5}{8\eta} - \frac{L}{2} - \frac{6\eta L^2 m}{b}) \geq 0$ and $\frac{\eta}{6} \leq \frac{\eta}{3} - L\eta^2$ since $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12\sqrt{m}L}\}$. Telescoping the sum for $s = 1, 2, \ldots, S$ in (S69), we obtain

$$\begin{aligned} 0 &\leq \mathbb{E}[F(\tilde{x}^S) - F(x^*)] \\ &\leq \mathbb{E}\left[F(\tilde{x}^0) - F(x^*) - \sum_{s=1}^{S}\sum_{t=0}^{m}\frac{\eta}{6}\left\|g_\eta(x_t^s)\right\|^2 + \sum_{s=1}^{S}\sum_{t=0}^{m}(\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\sum_{s=1}^{S}\sum_{t=0}^{m}\frac{7L^2d^2\mu^2 t}{2})\right] \end{aligned}$$

Thus, we have

$$\mathbb{E}[\|g_\eta(\hat{x})\|^2] \leq \frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} + \frac{I(\mathcal{B}<n)12\sigma^2}{\mathcal{B}} + 21mL^2d^2\mu^2 \tag{S70}$$

where (S70) holds since we choose $\hat{x}$ uniformly randomly from $\{x_t^s\}_{t\in[m],s\in[S]}$. □

The next corollary illustrates the convergence rate of ZO-PSVRG+ in terms of error of the solution $\hat{x}$, providing explicit descriptions for the parameters in Theorem 9.

Proof of Corollary 10

Proof Using Theorem 9, we have $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12\sqrt{m}L}\}$

$$\begin{aligned} \mathbb{E}[\|g_\eta(\hat{x})\|^2] &\leq \frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} \\ &\quad + \frac{I(\mathcal{B}<n)12\sigma^2}{\mathcal{B}} + 21mL^2d^2\mu^2 = 3\epsilon \end{aligned} \tag{S71}$$

Now we derive the total number of iterations $T = Sm = \frac{6(F(x_0)-F(x^*))}{\epsilon\eta}$. Since $\mu \leq \frac{\sqrt{\epsilon}}{5\sqrt{m}dL}$, and for $\mathcal{B} = n$, the second term in the bound (S71) is 0, the proof is completed as the number of SZO call equals to $Sn + Smb = 6\left(F(x_0) - F(x^*)\right)(\frac{n}{\epsilon\eta m} + \frac{b}{\epsilon\eta})$. If $\mathcal{B} < n$ the number of SZO calls equal to $d(S\mathcal{B} + Smb) = 6d\left(F(x_0) - F(x^*)\right)(\frac{\mathcal{B}}{\epsilon\eta m} + \frac{b}{\epsilon\eta})$ by noting that $\frac{I(\mathcal{B}<n)12\sigma^2}{\mathcal{B}} \leq \epsilon$ due to $\mathcal{B} \geq 12\sigma^2/\epsilon$. The second part of the corollary is obtained by setting $m = b$ in the first part. □

Analysis for ZO-PSPIDER+ (RandSGE)

In this section, we study further the convergence of ZO-PSPIDER+ (RandSGE) under different settings. In particular, in the following theorem, we prove that ZO-PSPIDER+ (RandSGE) improves the convergence rate and the function query complexity of the existing SZO methods.

Theorem 19 *Suppose Assumptions 1 and 2 hold, and the coordinate gradient estimator (21) is used to compute the mix gradient $\hat{v}_k$. The output $\hat{x}$ of Algorithm 2 satisfies*

$$\begin{aligned} \mathbb{E}[\|g_\eta(\hat{x})\|^2] &\leq \frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} \\ &\quad + \frac{I(\mathcal{B}<n)12\sigma^2}{\mathcal{B}} + 21mL^2d^2\mu^2 \end{aligned}$$

where $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12\sqrt{m}L\sqrt{d}}\}$ denotes the stepsize.

Corollary 20 We Let the batch size $\mathcal{B} = \min\{12\sigma^2/\epsilon, n\}$ and $\mu \leq \frac{\sqrt{\epsilon}}{5\sqrt{m}dL}$ denote the smoothing parameter. Suppose $\hat{x}$ returned by Algorithm 2 is an ϵ-accurate solution for problem (1). Recalling that CoordSGE and RandSGE require $O(d)$ and $O(1)$ function queries respectively, the number of SZO calls is at most

$$(dS\mathcal{B} + Smb) = 6\left(F(x_0) - F(x^*)\right)(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta})$$

$$= O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta}\right) \tag{S72}$$

and the number of PO calls is equal to $T = Sm = \frac{6(F(x_0)-F(x^*))}{\epsilon\eta} = O\left(\frac{1}{\epsilon\eta}\right)$. In particular, by setting $m = b$ and $\eta = \frac{1}{12L\sqrt{d}}$, the number of SZO calls is at most

$$\begin{aligned} &72L(F(x_0) - F(x^*))\left(\frac{\mathcal{B}d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right) \\ &= O\left(s_n\frac{d\sqrt{d}}{\epsilon b} + \frac{b\sqrt{d}}{\epsilon}\right) \end{aligned} \tag{S73}$$

where $s_n = \min\{n, \frac{1}{\epsilon}\}$ and the number of PO calls equals to $T = Sm = Sb = \frac{72L\sqrt{d}(F(x_0)-F(x^*))}{\epsilon} = O\left(\frac{\sqrt{d}}{\epsilon}\right)$.

ZO-PSPIDER+ Under PL Condition

Proof of Theorem 11

Proof We start by recalling inequality (S65) from the proof of Theorem 9, i.e.,

$$\begin{aligned} &\mathbb{E}[F(x_{t+1}^s)] \\ \leq&\mathbb{E}\left[F(x_t^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_{t+1}^s - x_t^s\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_t^s)\right\|^2\right] \\ &+ \frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2t}{2} \end{aligned} \tag{S74}$$

$$\begin{aligned} \leq&\mathbb{E}\left[F(x_t^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_{t+1}^s - x_t^s\right\|^2 - \frac{\eta}{6}\left\|g_\eta(x_t^s)\right\|^2\right] \\ &+ \frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2t}{2} \end{aligned} \tag{S75}$$

where in (S75) inequality we applied $\eta L \leq \frac{1}{6}$. Moreover, substituting PL inequality, i.e.,

$$\left\|g_\eta(x)\right\|^2 \geq 2\lambda(F(x) - F^*) \tag{S76}$$

into (S75), we obtain

$$\begin{aligned} &\mathbb{E}[F(x_{t+1}^s)] \\ \leq&\mathbb{E}\left[F(x_t^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_{t+1}^s - x_t^s\right\|^2 - \lambda\frac{\eta}{3}(F(x_t^s) - F^*)\right] \\ &+ \frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2t}{2} \end{aligned} \tag{S77}$$

Thus, we have

$$\begin{aligned} &\mathbb{E}[F(x_{t+1}^s) - F^*] \\ \leq&\mathbb{E}\left[(1 - \lambda\frac{\eta}{3})(F(x_t^s) - -F^*) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_{t+1}^s - x_t^s\right\|^2\right] \\ &+ \frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2 + \frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2t}{2} \end{aligned} \tag{S78}$$

Let $\beta := 1 - \lambda\frac{\eta}{3}$ and $\Psi_t^s := \frac{\mathbb{E}[F(x_t^s)-F^*]}{\beta^t}$. Combining these definitions with (S44), we have

$$\Psi_{t+1}^s$$

$$\leq \Psi_t^s - \frac{1}{\beta^{t+1}}\mathbb{E}\left[(\frac{5}{8\eta}-\frac{L}{2})\left\|x_{t+1}^s - x_t^s\right\|^2 - \frac{6\eta L^2}{b}\sum_{l=1}^{t}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2\right] + \frac{1}{\beta^{t+1}}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1}{\beta^{t+1}}\eta\frac{7L^2d^2\mu^2t}{2} \tag{S79}$$

Similar to the proof of Theorem 9, summing (S79) for $t = 0, 1, 2, \ldots, m$ in epoch s and remembering that $x_{m+1}^s = \tilde{x}^s$ and $x_0^s = \tilde{x}^{s-1}$, we have

$$\begin{aligned}
&\mathbb{E}[F(\tilde{x}^s) - F^*] \\
&\leq \beta^{m+1}\mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \beta^{m+1}\sum_{t=0}^{m}\frac{1}{\beta^{t+1}}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \beta^{m+1}\sum_{t=0}^{m}\frac{1}{\beta^{t+1}}\eta\frac{7L^2d^2\mu^2t}{2} \\
&-\beta^{m+1}\mathbb{E}\left[(\frac{5}{8\eta}-\frac{L}{2})\sum_{t=0}^{m}\frac{1}{\beta^{t+1}}\left\|x_{t+1}^s - x_t^s\right\|^2 - \frac{6\eta L^2}{b}\sum_{t=0}^{m}\frac{1}{\beta^{t+1}}\sum_{l=1}^{t}\mathbb{E}\left\|x_l^s - x_{l-1}^s\right\|^2\right] \\
&\leq \beta^{m+1}\mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \beta^{m+1}\sum_{t=0}^{m}\frac{1}{\beta^{t+1}}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \beta^{m+1}\sum_{t=0}^{m}\frac{1}{\beta^{t+1}}\eta\frac{7L^2d^2\mu^2t}{2} \\
&-\beta^{m+1}\mathbb{E}\left[(\frac{5}{8\eta}-\frac{L}{2})\sum_{t=0}^{m}\left\|x_{t+1}^s - x_t^s\right\|^2 - \frac{6\eta L^2}{b}\frac{1}{1-\beta}\sum_{t=1}^{m}\mathbb{E}\left\|x_t^s - x_{t-1}^s\right\|^2\right] \\
&\leq \beta^{m+1}\mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^{m+1}}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^{m+1}}{1-\beta}\frac{7m\eta L^2d^2\mu^2}{2} \\
&\quad -\mathbb{E}\left[(\frac{5}{8\eta}-\frac{L}{2}-\frac{6\eta L^2}{b(1-\beta)})\sum_{t=1}^{m}\left\|x_t^s - x_{t-1}^s\right\|^2\right] \qquad \text{(S80)} \\
&\leq \beta^{m+1}\mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\beta^{m+1}}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\beta^{m+1}}{1-\beta}\frac{7m\eta L^2d^2\mu^2}{2} \qquad \text{(S81)}
\end{aligned}$$

where (S80) holds since $\frac{5}{8\eta} - \frac{L}{2} \geq 0$, $\|.\|^2$ always is non-negative. In (S80) also the equations $\beta^{m+1}/\beta^{t+1} \leq 1$ for $0 \leq t \leq m$ and $\beta^{m+1}\sum_{t=0}^{m}\frac{1}{\beta^{t+1}} \leq \frac{1}{1-\beta}$ are applied. (S81) holds since it is sufficient to show $\frac{5}{8\eta} - \frac{L}{2} - \frac{6\eta L^2}{b(1-\beta)} \geq 0$. It is easy to see that this inequality is valid due to $\eta \leq \min\{\frac{1}{8L}, \frac{\lambda b}{32L^2}\}$. Similarly, let $\tilde{\beta} = \beta^{m+1}$ and $\tilde{\Psi}^s = \frac{\mathbb{E}[F(\tilde{x}^s)-F^*]}{\tilde{\beta}^s}$. Substituting these definitions into (S81), we have

$$\tilde{\Psi}^s \leq \tilde{\Psi}^{s-1} + \frac{1}{\tilde{\beta}^s}\frac{1-\tilde{\beta}}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1}{\tilde{\beta}^s}\frac{1-\tilde{\beta}}{1-\beta}\frac{7\eta L^2d^2\mu^2}{2} \tag{S82}$$

Taking a telescopic sum from (S82) for all epochs $1 \leq s \leq S$, we obtain

$$\begin{aligned}
\mathbb{E}[F(\tilde{x}^S) - F^*] &\leq \tilde{\beta}^S\mathbb{E}[F(\tilde{x}^0) - F^*] + \tilde{\beta}^S\sum_{s=1}^{S}\frac{1}{\tilde{\beta}^s}\frac{1-\tilde{\beta}}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \tilde{\beta}^S\sum_{s=1}^{S}\frac{1}{\tilde{\beta}^s}\frac{1-\tilde{\beta}}{1-\beta}\frac{7m\eta L^2d^2\mu^2}{2} \\
&= \beta^{Sm+m}\mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{1-\tilde{\beta}^S}{1-\tilde{\beta}}\frac{1-\tilde{\beta}}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1-\tilde{\beta}^S}{1-\tilde{\beta}}\frac{1-\tilde{\beta}}{1-\beta}\frac{7m\eta L^2d^2\mu^2}{2} \\
&\leq \beta^{Sm+m}\mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{1}{1-\beta}\frac{2I(\mathcal{B}<n)\eta\sigma^2}{\mathcal{B}} + \frac{1}{1-\beta}\frac{7m\eta L^2d^2\mu^2}{2} \\
&= \left(1-\frac{\lambda\eta}{3}\right)^{Sm}\mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}} + \frac{21mL^2d^2\mu^2}{2\lambda} \qquad \text{(S83)}
\end{aligned}$$

where in (S83) we recall that $\beta = 1 - \frac{\lambda\eta}{3}$ and $\beta < 1$. □

Proof of Corollary 12

Proof From Theorem 11, we have

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1-\frac{\lambda\eta}{3}\right)^{Sm}\mathbb{E}[F(\tilde{x}^0) - F^*]$$

$$+\frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}}+\frac{21L^2d^2\mu^2m}{2\lambda}=3\epsilon$$

which gives the total number of iterations $T=Sm=O(\frac{1}{\lambda\eta}\log\frac{1}{\epsilon})$ and is equal to the number of PO calls. Since $\mu\leq\frac{\sqrt{\lambda\epsilon}}{4\sqrt{m}Ld}$, we have $\frac{21L^2d^2\mu^2m}{2\lambda}\leq\epsilon$. The number of SZO calls is equal to $d(S\mathcal{B}+Smb)=O(\frac{\mathcal{B}d}{\lambda\eta m}\log\frac{1}{\epsilon}+\frac{bd}{\lambda\eta}\log\frac{1}{\epsilon})$. Note that if $\mathcal{B}<n$ then $\frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}}\leq\epsilon$ since $\mathcal{B}\geq 6\sigma^2/\lambda\epsilon$. With $\eta=\frac{\lambda}{32L^2}$, the number of PO queries to $T=Sm=O(\frac{1}{\lambda^2}\log\frac{1}{\epsilon})$ and the number of SZO calls is equal to $d(S\mathcal{B}+Smb)=O(\frac{\mathcal{B}d}{\lambda^2m}\log\frac{1}{\epsilon}+\frac{bd}{\lambda^2}\log\frac{1}{\epsilon})$. □

ZO-PSPIDER+ (RandSGE) Under PL Condition

In the following theorem, we explore if ZO-PSPIDER+ (RandSGE) achieves a linear convergence rate when it enters a local landscape where the loss function satisfies the PL condition.

Theorem 21 *Let Assumptions 1 and 2 hold, and ZO gradient estimator (21) for mix gradient $\hat{v}_k$ is used in Algorithm 2 with $\eta\leq\min\{\frac{1}{8L},\frac{\lambda b}{32L^2\sqrt{d}}\}$ with $1-\frac{\lambda\eta}{3}>0$. Then*

$$\mathbb{E}[F(\tilde{x}^S)-F^*]\leq\left(1-\frac{\lambda\eta}{3}\right)^{Sm}\mathbb{E}[F(\tilde{x}^0)-F^*]+\frac{6I(\mathcal{B}<n)\sigma^2}{\lambda\mathcal{B}}+\frac{21mL^2d^2\mu^2}{2\lambda}\tag{S84}$$

Corollary 22 Suppose the final iteration point $\tilde{x}^S$ in Algorithm 2 satisfies $\mathbb{E}[F(\tilde{x}^S)-F^*]\leq\epsilon$ under PL condition. Under Assumptions 1 and 2, we let batch size $\mathcal{B}=\min\{\frac{6\sigma^2}{\lambda\epsilon},n\}$ and the smoothing parameter $\mu\leq\frac{\sqrt{\lambda\epsilon}}{4\sqrt{m}Ld}$. The number of SZO calls is bounded by

$$(dS\mathcal{B}+Smb)=O(\frac{s_nd}{\lambda\eta m}\log\frac{1}{\epsilon}+\frac{b}{\lambda\eta}\log\frac{1}{\epsilon})$$

where $s_n=\min\{n,\frac{1}{\lambda\epsilon}\}$. The number of PO calls equals the total number of iterations T which is bounded by

$$T=Sm=O(\frac{1}{\lambda\eta}\log\frac{1}{\epsilon})$$

In particular, given the setting $\eta=\frac{\lambda}{32L^2\sqrt{d}}$, the number of SZO calls simplifies to $(dS\mathcal{B}+Smb)=O(\frac{\mathcal{B}d\sqrt{d}}{\lambda m}\log\frac{1}{\epsilon}+\frac{b\sqrt{d}}{\lambda}\log\frac{1}{\epsilon})$.

Now we give some useful properties of CoordSGE and RandSGE that we previously applied to derive the estimation errors.

Lemma 23 (Liu et al (2018b)) Suppose that the function $f(x)$ is L-smooth. Let $\hat{\nabla}f(x)$ denote the estimated gradient defined by CoordSGE. Define $f_\mu=\mathbb{E}_{u\sim U[-\mu,\mu]}f(x+ue_j)$, where $U[-\mu,\mu]$ denotes the uniform distribution on the interval $[-\mu,\mu]$. Then for any $x\in\mathbb{R}^d$ we have

1. f_μ is L-smooth, and $\hat{\nabla}f(x)=\sum_{j=1}^d\frac{\partial f_\mu(x)}{\partial x_j}e_j$.
2. $|f_\mu(x)-f(x)|\leq\frac{L\mu^2}{2}$ and $|\frac{\partial f_\mu(x)}{\partial x_j}|\leq\frac{L\mu^2}{2}$.
3. $\left\|\hat{\nabla}f(x)-\nabla f(x)\right\|^2\leq\frac{L^2d^2\mu^2}{4}$.

Lemma 24 Assume that the function $f(x)$ is L-smooth. Let $\hat{\nabla}_rf(x)$ denote the estimated gradient defined by RandSGE. Define $f_\mu=\mathbb{E}_{u\sim U_S}[f(x+\mu u)]$, where U is uniform distribution over a d-dimensional unit ball S. Then, we have

1. For any $x \in \mathbb{R}^d$, $\nabla f_\mu(x) = \mathbb{E}_u[\hat{\nabla}_r f(x,u)]$.
2. $|f_\mu(x) - f(x)| \leq \frac{\mu^2 L}{2}$ and $\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\mu L d}{2}$ for any $x \in \mathbb{R}^d$.
3. $\mathbb{E}_u \left\| \hat{\nabla}_r f(x,u) - \hat{\nabla}_r f(y,u) \right\|^2 \leq 3dL^2 \|x-y\|^2 + \frac{3L^2 d^2 \mu^2}{2}$.

Proof The proof of items 1 and 2 can be found in Gao et al (2018). Item 3 is due to Lemma 5 in Ji et al (2019). □

Additional Empirical Results

Fig. 1: Images of universal perturbation $\delta := a_i^{adv} - a_i$ from different proximal ZO algorithms and the subsequent generated adversarial images for the class label "1" (automobile) of the CIFAR-10 dataset. The left nine columns show the generated adversarial examples crafted by different ZO algorithms. The second last columns present a visualization for the universal perturbation δ for each ZO algorithm and the last column shows the L_2-norm of the generated universal perturbation ($\|\delta\|$). The bottom of each subplot for adversarial attacks shows the label of the benign sample and the predicted label of the corresponding adversary (ground-truth label -> predicted label).