An Efficient Derivative-Free Proximal Stochastic Gradient Method for Nonconvex Nonsmooth Optimization

March 2018

Abstract

Proximal gradient method has an important role in solving nonsmooth composite optimization problems. However, in some machine learning problems proximal gradient method could not be leveraged because the explicit gradients of these problems are not accessible. Associated with black-box models, these types of problems fall into zeroth-order (ZO) optimization. Several varieties of proximal zeroth-order variance reduced stochastic algorithms for nonconvex optimization have recently been introduced based on the first-order techniques of stochastic variance reduction. However, all existing zeroth-order SVRG-type algorithms suffer from function query complexities up to a small-degree polynomial of the problem size. To fill this gap, we analyze a new zeroth-order stochastic gradient algorithms for optimizing nonconvex, nonsmooth finite-sum problems, called ZO-PSVRG+. The analysis of ZO-PSVRG+ recovers several existing convergence results and improves theirs ZO oracle calls and proximal oracle calls. In particular, ZO-PSVRG+ yields simpler analysis for a wide range of minibatch sizes, while the improvement of ZO-SVRG in Ji et al. (2019) is only achieved for large minibatch sizes based on an involved parameter selection. We further prove ZO-PSVRG+ under Polyak-Łojasiewicz condition in contrast to existent ZO-SVRG type methods obtains a global linear convergence for a wide range of minibatch sizes. Our empirical experiments on black-box binary classification and black-box adversarial attack problem validate that the studied algorithms under our new analysis can achieve superior performance with a lower query complexity.

1 Introduction

1.1 what we want to do

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

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

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 general structure (1) covers numerous machine learning areas, ranged from neural networks to generalized linear models and from convex problems like SVM and Lasso to highly nonconvex optimization including minimizing loss function for deep learning. We will investigate and explore a set of accelerated variance reduced stochastic zeroth-order (SZO) optimization algorithms for (1). Stochastic variance reduced gradient (SVRG) is a generic and powerful methodology to decrease the variance induced by stochastic sampling Johnson and Zhang (2013); Reddi et al. (2016a); Nitanda (2016); Allen-Zhu and Yuan (2016); Lei et al. (2017). As a result of reduction in variance, it enhances the rate of convergence for stochastic gradient descent (SGD) complexity by a factor of $O(1/\epsilon)$. To reduce the variance in SZO optimization, one may apply the comparable concepts and similar ideas in the first-order methods.

1.2 Background in research

In recent years, there has been thorough studies for convex problems of the form (1) (see e.g., Nesterov (2013), Xiao and Zhang (2014); Defazio et al. (2014); Lan and Zhou (2017); Allen-Zhu (2017). In particular, in Beck and Teboulle (2009) a fast-converging class of proximal gradient (PG) schemes for problems with convex structure based on Nesterov's momentum acceleration are designed. Xiao and Zhang (2014) developed an algorithm called Prox-SVRG for large-scale problems, which obtains a linear rate of convergence when each f_i is strongly-convex. Several stochastic PG methods were developed in Bertsekas (2011); Xiao and Zhang (2014) to deal with the large-scale convex problems. Because of growing applications of deep neural networks, recently the studies for nonconvex case have been noticeably growing. Nevertheless, for the generic nonsmooth nonconvex problems, the analysis is still rather sparse. Li and Lin (2015) introduced a set of fast-converging PG algorithms for nonconvex structure problems. Similarly, Ghadimi and Lan (2016); Reddi et al. (2016b) studied stochastic PG methods for nonconvex optimization. Recently, Li and Li (2018) designed an algorithm by extending the results from Reddi et al. (2016b), leading to an improved iteration complexity for stochastic gradient method.

1.3 Reason to use zeroth–order techniques

The major adversity for these accelerated method is their designs on involving first-order information. Nevertheless , there are circumstances where the first-order gradient evaluations are computationally unrealizable, costly, or unachievable, while zeroth-order information (function information) are 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). Similarly, in predictions with stochastic structure, computing the derivatives is possibly complicated or prohibited to perform, while the functional estimations of foreseen frameworks are achievable Sokolov et al. (2016). For instance, in bandit Shamir (2017) and black-box intelligence Chen et al. (2017) settings, only the loss function evaluations are accessible as the derivatives

cannot be calculated directly. Thus, the derivative-free optimization algorithm Nesterov and Spokoiny (2017) is a viable option to tackle these issues. This procedure approximates the full gradient via gradient evaluator based on only the function estimations which end up in derivative-free optimization Brent (2013); Spall (2005). We describe the minimization problem (1) in this particular setting as stochastic proximal zeroth-order optimization. Recently, zeroth-order optimization has attracted significant attention due to its diverse applications, e.g., black-box adversarial attacks on deep neural networks (DNNs) Kurakin et al. (2016); Papernot et al. (2017); Chen et al. (2017), reinforcement learning Choromanski et al. (2018) and structured prediction Taskar et al. (2005). Further applications cover time-varying constrained networks with restricted computation capacity Chen and Giannakis (2019); Liu et al. (2017), and model inference with black-box setting Fu (2002); Lian et al. (2016).

1.4 Problem with existing methods

While a great number of SZO algorithms have recently been explored and studied Liu et al. (2017); Flaxman et al. (2005); Shamir (2013); Agarwal et al. (2010); Nesterov and Spokoiny (2017); Duchi et al. (2015); Shamir (2017); Dvurechensky et al. (2018); Wang et al. (2017), they are mostly employed for problems with convex structure, which confines their applications in the broad span of nonconvex problems.

Furthermore, these methods regularly degrade due to the large variances of SZO gradient evaluations, and consecutively, decaying rate of convergence. Hence, a practical method to obtain faster convergence for SZO is by exploiting variance reduced techniques. Recently, various SZO variance reduced methods have been investigated to additionally enhance the rate of convergence of ZO-SGD Liu et al. (2018a,b). These methods are primarily derived from SVRG algorithm, which substitutes the gradient in SVRG Johnson and Zhang (2013) by zeroth-order gradient approximations. Specifically, Liu et al. (2018b) introduced several zeroth-order methods, based on SVRG-type algorithms.

Currently, there are only a few number of zeroth-order stochastic methods for solving problem (1), e.g., Ghadimi and Lan (2016) and Huang et al. (2019). In particular, in Ghadimi and Lan (2016) a zeroth-order proximal stochastic gradient method has been analyzed. Nevertheless, as a result of the high variance of zeroth-order gradient approximation based on random sampling and vector sampling for two point derivative calculations, the iteration complexity of RSPGE ($O(\frac{d}{\epsilon^2})$) is notably worse than the best known rate $O(\frac{d}{\epsilon})$ for zeroth-order stochastic optimization. A major issue in the advancement of SZO algorithms for solving (1) is the order of the required number of function queries, namely SZO calls or iteration complexity. While the present zeroth-order based upon SVRG algorithms have higher convergence rate, the complexity of their function calls are all greater than each of ZO-GD and ZO-SGD. They also rely on a very small and some often diminishing stepsize of $O(\frac{1}{d})$. To perform an elaborated analysis, the term related to the dimension of the problem in the convergence studies (i.e., d) is a key factor with high impact on the efficiency of SZO optimization. Ji et al. (2019) refined the ZO estimations to achieve an improved ZO complexity and enhanced convergence rate. However, their improved

analysis is only for smooth functions based on a complicated parameter selection and it is only valid for large minibatch sizes. We design an accelerated ZO proximal variants by applying variance reduced gradient approximation for nonsmooth composite optimization. This provides a lower iteration complexity towards $O(1/\epsilon)$, which is to our knowledge the best iteration complexity bound obtained thus far for proximal ZO stochastic optimization with nonconvex structure. This demonstrates an improvement for ZO iteration complexity up to a factor of d.

1.5 Compare with other methods

We compared the results from our analysis and other comparable SZO algorithms in Table 4. It 4 indicates that RGF has the largest query complexity and yet has the worst convergence rate. ZO-SVRG-coord and ZO-ProxSVRG/SAGA provide an improved rate of convergence $O(d/\epsilon)$ due to using variance reduction techniques. On the other hand, existing SVRG type zeroth-order algorithms are affected by worse function query complexities compared with RSPGF, while ZO-PSVRG+ could achieve better trade-offs between the convergence rate and the query complexity.

2 Main Challenge

Despite the fact that proximal SVRG has indicated a huge promise for first-order algorithms, utilizing identical concepts to ZO optimization is not effortless. Due to the perturbation induced by ZO gradient estimation, SZO algorithms have complex joint structures, which make their analysis difficult in many settings. The other major difficulty is due to the fact that that ProxSVRG is based upon the notion that stochastic gradient is an unbiased approximation of the actual full gradient, which is not retained in the ZO case. Thus, it is a challenging question if the proximal ZO stochastic variance reduced gradient could accelerate the convergence of proximal ZO algorithms with arbitrary minibatch sizes. In this paper, we plan to address this question and in particular fill the void between SZO optimization and ProxSVRG by improving the complexity of exiting ZO variance reduced methods for problem (1).

3 Main contributions

We present a novel analysis for an existing ZO-SVRG-Coord algorithm introduced in Liu et al. (2018b); Ji et al. (2019), and prove that ZO-PSVRG+ based on our new analysis surpasses other state-of-the-art SVRG-type zeroth-order methods as well as RSPGF. We concentrate on several important debatable questions in these methods. To be specific, we somewhat address the open question if the dependence on the dimension d for the convergence analysis proposed in Liu et al. (2018b) is optimal. Our work provides an inclusive analysis on how ZO gradient approximation influence ProxSVRG on both convergence rate and function query complexity. This is performed based on the novel structure of recently introduced SZO algorithms. Note that problem (1) does not necessarily satisfy bounded gradient assumption in Ghadimi and Lan (2016); Huang

et al. (2019). We prove that compared to ProxSVRG, ZO-PSVRG+ obtains a sublinear convergence with SZO complexity of $O(1/\epsilon)$. The convergence results are declared with respect to the number of stochastic zeroth-order (SZO) queries and proximal oracle (PO) calls. Based on our new analysis, we summarize the following results from this parer:

1) Our analysis yields iteration complexity $O(\frac{1}{\epsilon})$ corresponding 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). Thus, our results have better or no dependence on d in contrast to the existing proximal variance-reduced SZO methods. Note that the number of PO calls equal to $O(1/\epsilon)$ and $O(1/\epsilon^2)$ for ZO-PSVRG+ and RSPGF, respectively. ZO-PSVRG+ also matches the best result achieved by ZO-SVRG-Coord-Rand with $b = dn^{2/3}$ for $m = n^{1/3}$ in Ji et al. (2019), while our results are valid for any minibatch sizes as detailed in the following sections. Indeed, since generally in training model parameters intermediate minibatch sizes are preferred, it is necessary to analyze and study the convergence behavior for minibatchs of moderate or single sizes.

2) The convergence analysis for ZO-PSVRG+ is not complicated in contrast to ZO-SVRG-Coord in Liu et al. (2018b); Ji et al. (2019), and yields simpler proofs. Our analysis achieves new iteration complexity bounds and improves the effectiveness of all the existing ZO-SVRG-based algorithms in addition to RSPGF for nonconvex nonsmooth composite optimization, which is the best results to our latest knowledge (see Table 4).

3) For the nonconvex functions under Polyak-Łojasiewicz condition Polyak (1963), we show that ZO-PSVRG+ obtains a global linear rate of convergence equivalent to first-order ProxSVRG. Thus, ZO-PSVRG+ 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 of ZO-ProxSVRG for problem (1) with arbitrary minibatch size. It is also noticeable that the convergence rate obtained in this exploration is comparable to the first-order ProxSVRG. This generalizes the results of Duchi et al. (2015) and achieves linear convergence compared 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 and with $b = O(n^{1/2})$. note that due to both computational and statistical efficiency, convergence analysis for practical minibatch sizes is demanding. Also see the remarks after Theorem 48 for more details.

Finally, to demonstrate the efficiency and adaptability of our approach to achieve a balance between the rate of convergence 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.

4 Related Works

Derivative-free (zeroth-order) methods have been efficiently utilized for solving numerous machine learning problems when the computation of the true gradient is infeasible. In ZO algorithms, a full gradient is generally estimated based on either a one-point or a two-point gradient approximation. The one-point estimator obtains a gradient estimate $\hat{\nabla} f(x)$ by probing f at a single random point near to x Flaxman et al. (2005); Shamir (2013), while the two-point estimator computes a difference of two random function probings Agarwal et al. (2010); Nesterov and Spokoiny (2017). In this paper, we concentrate on the two-point gradient approximation since it has a lower variance and thus amends the iteration complexity of ZO algorithms. Nesterov and Spokoiny (2017) proposed several stochastic derivative-free algorithms by employing Gaussian smoothing method. A zeroth-order mirror descent algorithm is analyzed in Duchi et al. (2015). More recently, Yu et al. (2018); Dvurechensky et al. (2018) introduced some accelerated zeroth-order algorithms for convex optimization. The recent studies confirmed that ZO algorithms typically agree with the complexity of first-order algorithms up to a small-degree polynomial of the problem size d.

These zeroth-order algorithms mostly target (strongly) convex problems. Despite the extensive studies for the convex structures, the studies for nonconvex ZO methods are relatively limited. Essentially, there are many nonconvex machine learning application, where the explicit derivatives are not accessible, e.g., nonconvex black-box learning problems Chen et al. (2017); Liu et al. (2018b). Thus, developing zeroth-order stochastic methods for the nonconvex optimization is indeed demanding. Ghadimi and Lan (2013) and Nesterov and Spokoiny (2011) proposed ZO-GD and its corresponding stochastic algorithm ZO-SGD, respectively. Liu et al. (2018a) introduced a variance reduced stochastic zeroth-order method with Gaussian smoothing. More recently, Liu et al. (2018b) presented a thorough analysis based on SVRG algorithms. Ji et al. (2019) elaborated the results in Liu et al. (2018b) and achieved improved bounds based on a complicated parameter selection, however their improvements relies on large minibatch sizes. Recently, for nonsmooth nonconvex problems Huang et al. (2019) provided two algorithms called ZO-ProxSVRG and ZO-ProxSAGA, which are based on the well-known variance reduction techniques ProxSVRG and ProxSAGA Reddi et al. (2016b). Before that, Ghadimi and Lan (2016) also considered the stochastic case (here we denote it as RSPGF). However, RSPGF requires increasing or large minibatch sizes i.e., $\Omega(1/\epsilon)$. Note that due to the growing minibatch sizes, RSPGF may change to deterministic proximal gradient descent (ZO-ProxGD) after few iterations. Further, Liu et al. (2018a) have also analyzed a zeroth-order algorithm for solving nonconvex nonsmooth problems, which are different from problem (1). In order to deal with the large-scale problems, several asynchronous stochastic zeroth-order algorithms have been studied, e.g., Gu et al. (2018b); Lian et al. (2016); Gu et al. (2018a). In Lian et al. (2016), an asynchronous ZO stochastic coordinate descent (ZO-SCD) was designed with a convergence rate of $O(d/\epsilon^2)$. The convergence rate of asynchronous SZO further improved in Gu et al. (2018a) by integrating SVRG techniques with stochastic coordinate descent method.

Even though the abovementioned zeroth-order stochastic algorithms can effectively solve the problems with nonconvex structure, there are limited number of zeroth-order

Method	Problem	Stepsize	Convergence rate	SZO complexity
RGF (Nesterov and Spokoiny (2017))	NS(C)	$O\left(\frac{1}{\sqrt{dT}}\right)$	$O\left(\frac{d^2}{\epsilon^2}\right)$	$O\left(\frac{nd^2}{\epsilon^2}b\right)$
RSPGF (Ghadimi and Lan (2016))	S(NC)+NS(C)	$O(1)$	$O\left(\frac{d}{\epsilon^2}\right)$	$O\left(\frac{nd}{\epsilon^2}\right)$
ZO-SVRG-Coord (Liu et al. (2018b))	S(NC)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{\epsilon}\right)$	$O(\frac{nd^2}{\epsilon}+\frac{d^2b}{\epsilon})$
ZO-SVRG-Coord-Rand (Ji et al. (2019))	S(NC)	$O\left(\frac{1}{dn^{2/3}}\right)$	$O\left(\frac{dn^{2/3}}{\epsilon}\right)$	$O(\min\{\frac{dn^{2/3}}{\epsilon},\frac{d}{\epsilon^{5/3}}\})$
ZO-ProxSVRG-Coord (Gu et al. (2018a))	S(NC)+NS(C)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{\epsilon}\right)$	$O(\frac{nd^2}{\epsilon\sqrt{b}}+\frac{md^2\sqrt{b}}{\epsilon})$
ZO-ProxSAGA-Coord (Gu et al. (2018a))	S(NC)+NS(C)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{\epsilon}\right)$	$O(\frac{nd^2}{\epsilon\sqrt{b}})$
ZO-PSVRG+ (CoordSGE) (Ours)	S(NC)+NS(C)	$O(1)$	$O\left(\frac{1}{\epsilon}\right)$	$O\left(s_n\frac{d}{\epsilon\sqrt{b}}+\frac{bd}{\epsilon}\right)$
ZO-PSVRG+ (RandSGE) (Ours)	S(NC)+NS(C)	$O\left(\frac{1}{\sqrt{d}}\right)$	$O\left(\frac{\sqrt{d}}{\epsilon}\right)$	$O\left(s_n\frac{d\sqrt{d}}{\epsilon\sqrt{b}}+\frac{b\sqrt{d}}{\epsilon}\right)$
ZO-PSVRG+ (CoordSGE) (Ours)	S(PL)+NS(C)	$O(1)$	$O\left(\log(\frac{1}{\epsilon})\right)$	$O(s_n\frac{d}{\lambda\sqrt{\gamma}m}\log\frac{1}{\epsilon}+\frac{bd}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$
ZO-PSVRG+ (RandSGE) (Ours)	S(PL)+NS(C)	$O\left(\frac{1}{\sqrt{d}}\right)$	$O\left(\sqrt{d}\log(\frac{1}{\epsilon})\right)$	$O(s_n\frac{d\sqrt{d}}{\lambda\sqrt{\gamma}m}\log\frac{1}{\epsilon}+\frac{b\sqrt{d}}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$

Table 1: Summary of convergence rate and function query complexity of SZO algorithms. S: Smooth, NS: Nonsmooth, NC: Nonconvex, C: Convex, SC: Strong Convexity, and PL: Polyak-Łojasiewicz Condition. $s_n = \min\{n, \frac{1}{\epsilon}\}$

.

stochastic methods for nonconvex nonsmooth composite problems. We emphasize that, in contrast to existing ZO proximal methods, our analysis do not require bounded gradient assumption, which is not valid for many unconstrained optimization problems. It should be highlighted that computing full-gradient may not be effective for large-scale machine learning problems. Thus, we focus on studying a more general framework of ZO-ProxSVRG with different gradient estimators.

5 Preliminary

In the following we specify and illustrate some details on ZO gradient approximations. Considering a single 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 a smoothing parameter Flaxman et al. (2005); Shamir (2017); Gao et al. (2018). Typically, RandSGE is a biased estimation to the actual gradient $\nabla f_i(x)$, and its bias decreases as μ approaches zero. Nevertheless, in practice, if μ is too small, the function variation could be signified by the noise in the system when the rate of noise to signal is high Lian et al. (2016). To obtain a higher quality approximation for ZO gradient, one can apply coordinate

gradient estimation (CoordSGE) Gu et al. (2018b,a); Liu et al. (2018b) to evaluate the gradients as shown:

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

where e_j is a standard basis vector with 1 at its j-th coordinate and 0 otherwise, and μ is a smoothing parameter. In contrast to RandSGE, CoordSGE is deterministic and needs d times more ZO function calls. However, our studies reveal that for ZO variance-reduced methods, although the coordinate-wise gradient estimator demands more ZO calls than the two-point random gradient approximation, it assures a more accurate ZO estimation, which results in a larger stepsize and a speedier convergence. More details on ZO gradient estimation can be found in Kazemi and Wang (2018).

Since proximal gradient method requires to compute the gradient in each iteration, it cannot be used to tackle the optimization problems where the computation of explicit gradient of function $f(x)$ is infeasible. Based on the ZO gradient estimation (3), we present a zeroth-order proximal gradient descent method, which undertakes iterations of the form:

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

where $\hat{\nabla} f = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_i(x)$ 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 following we assume that the nonsmooth convex function $h(x)$ in (1) is well-defined, i.e., the proximal operator (5) can be computed effectively.

5.1 Gradient Mapping

For convex problems, generally the optimality gap $F(x) - F(x^*)$ is applied as the convergence metric. But for general nonconvex problems, the gradient norm is generally employed as the convergence metric. For instance, for smooth nonconvex optimization (i.e., $h(x) = 0$), Ghadimi and Lan (2013); Reddi et al. (2016a); Lei et al. (2017); Liu et al. (2018b) applied $\|\nabla F(x)\|^2$ (i.e., $\|\nabla f(x)\|^2$) as the convergence criterion. Aiming to investigate the convergence behavior for nonsmooth nonconvex problems, it is needed to define the gradient mapping as illustrated in Ghadimi and Lan (2016); Reddi et al. (2016b); Huang et al. (2019):

$$g_\eta(x) = \frac{1}{\eta}(x - \text{Prox}_{\eta,h}(x - \eta \nabla f(x))) \tag{6}$$

If $h(x)$ is a constant function , it is noted that this gradient mapping reduces to the ordinary gradient: $g_\eta(x) = \nabla F(x) = \nabla f(x)$. In this paper, 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 the problems with nonconvex structure, if $g_\eta(x) = 0$, the point x is a stationary point (Parikh, Boyd, and others, 2014). Hence, we can exploit the following definition as the convergence metric.

Definition 5.1. The point x is referred to an ϵ-accurate, if $\mathbb{E}\left\|g_\eta(x)\right\|^2 \leq \epsilon$, for some $\eta > 0$.

6 ZO Proximal Stochastic Method (ZO-PSVRG+)

Algorithm 1 ZO-PSVRG+

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 (8) or (9)
8: $\quad\quad x_t^s = \text{Prox}_{\eta h}(x_{t-1}^s - \eta \hat{v}_{t-1}^s)$
9: $\quad \tilde{x}^s = x_m^s$
10: **Output:** $\hat{x}$ chosen uniformly from $\{x_{t-1}^s\}_{t \in [m], s \in [S]}$

In this section, we present a proximal stochastic gradient algorithm called ZO-PSVRG+ based on 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 6. 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. We call $\mathcal{B}$ and b the batch and minibatch sizes, respectively.

The major difference of our ZO-PSVRG+ and ZO-ProxSVRG is that we avoid the evaluation of the total gradient before each epoch, i.e., the number of samples $\mathcal{B}$ may not equal to n (see Line 5 of Algorithm 6). However, in ZO-SVRG-Coord and ZO-ProxSVRG-Coord it is assumed $\mathcal{B} = n$. If $\mathcal{B} = n$, ZO-PSVRG+ is equivalent to ZO-ProxSVRG since Step 7 of Algorithm 6. Our analysis implies the results for ZO-ProxSVRG-Coord (i.e, $\mathcal{B} = n$) are novel and improve the existing analysis. In addition, ZO-PSVRG+ generalizes the existing variance-reduced methods to a more general nonsmooth nonconvex setting using a direct analysis of zeroth-order methods and provides simpler analysis. The main idea in variance-reduced algorithms is to construct an additional sequence $\tilde{x}^{s-1}$ at which the full gradient is computed for obtaining a revised stochastic gradient estimate

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

where v_{t-1}^s represents the gradient estimate at x_{t-1}^s and $g^s = \frac{1}{\mathcal{B}} \sum_{i \in I_{\mathcal{B}}} \nabla f_i(\tilde{x}^{s-1})$. The main characteristic of (7) is that v_{t-1}^s is an unbiased gradient approximation of $\nabla f(x_{t-1}^s)$. In the ZO framework, the mix gradient (7) is estimated by applying only function

evaluations, specified by

$$\hat{v}_{t-1}^s = \frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})\right) + \hat{g}^s \qquad \text{ZO-PSVRG+} \tag{8}$$

or

$$\hat{v}_{t-1}^s = \frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_{t-1}^s, u_i) - \hat{\nabla}_r f_i(\tilde{x}^{s-1}, u_i)\right) + \hat{g}^s \qquad \text{ZO-PSVRG+ (RandSGE)} \tag{9}$$

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 given by CoordSGE and $\hat{\nabla}_r f_i$ is a ZO gradient estimate given by RandSGE.

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 actual gradient. In the other words, the unbiased assumption on gradient approximates utilized in ProxSVRG (Reddi et al. (2016b); Li and Li (2018)) is no longer valid. We emphasize that the biased ZO gradient estimation yields a fundamental challenge in the analyzing ZO-PSVRG+. Hence, adjusting the similar concepts from ProxSVRG to zeroth-order algorithm 6 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 variance of gradient estimation which is discussed later.

7 Convergence Analysis

Now, we provide some minimal assumptions for problem (1) as demonstrated in the sequel:

Assumption 7.1. For $\forall i \in 1, 2, \ldots, n$, 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 7.2. 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 $\nabla f_i(x)$.

Both of Assumptions 7.1 and 7.2 are standard assumptions applied in nonconvex 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 provides the bounded variance of zeroth-order gradient approximates Lian et al. (2016); Liu et al. (2018a,b); Hajinezhad et al. (2017). Note that due to the error estimation for CoordSGE, this assumption is equivalent to the bounded variance of true gradients. Assumption 7.2 is weaker than the assumption of bounded gradients Liu et al. (2017); Hajinezhad et al. (2017), while, we are capable to analyze more complicated problem (1) involving a non-smooth part and obtain faster convergence rates. Assumption 7.2 is essential in order to obtain a convergence result independent of n. Below, We start by deriving an upper bound for the variance of estimated gradient $\hat{v}_{t-1}^s$ based on CoordSGE.

Lemma 7.3. *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* $\hat{g}^s = \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1})$*, then the following inequality holds.*

$$\begin{aligned}\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2\right] \le & \frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\ & + 2\frac{I\{\mathcal{B}<n\}\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2}\end{aligned} \tag{10}$$

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 (11) \\
=&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 (12) \\
=&\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] \\
&+ 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 (13)
\end{aligned}$$

$$\leq \frac{2\eta}{b^2}\mathbb{E}\left[\sum_{i\in I_b}\left\|\hat{\nabla} f_i(x_{t-1}^s)-\hat{\nabla} f_i(\tilde{x}^{s-1})\right\|^2\right]$$

$$+2\eta\mathbb{E}\left[\left\|\frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\left(\hat{\nabla} f_j(\tilde{x}^{s-1})-\hat{\nabla} f(\tilde{x}^{s-1})\right)\right\|^2\right]+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \tag{14}$$

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

$$+2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s)-\nabla f(x_{t-1}^s)\right\|^2 \tag{15}$$

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

$$+\frac{3\eta L^2 d^2\mu^2}{b} \tag{16}$$

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

$$+\eta\frac{L^2 d^2\mu^2}{2}+\frac{3\eta L^2 d^2\mu^2}{b} \tag{17}$$

$$\leq \frac{6\eta L^2}{b}\mathbb{E}\left[\left\|x_{t-1}^s-\tilde{x}^{s-1}\right\|^2\right]+2\frac{I\{\mathcal{B}<n\}\eta\sigma^2}{\mathcal{B}}+\eta\frac{7L^2 d^2\mu^2}{2} \tag{18}$$

where, recalling that a deterministic gradient estimator is employed and the expectations are taking with respect to I_b and $I_{\mathcal{B}}$. The inequality (11) holds by the Jensen's inequality. (12) and (13) 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. (14) applies the fact that $\mathbb{E}[\|x-\mathbb{E}[x]\|^2]\leq\mathbb{E}[\|x\|^2]$, for any random variable x. (15) holds due to the following inequality

$$\begin{aligned}\mathbb{E}\left\|\hat{\nabla} f_i(x_t^s)-\hat{\nabla} f_i(\tilde{x}^s)\right\|^2 &= \mathbb{E}\left\|\hat{\nabla} f_i(x_t^s)-\nabla f_i(x_t^s)+\nabla f_i(x_t^s)-\hat{\nabla} f_i(\tilde{x}^s)+\nabla f_i(\tilde{x}^s)-\nabla f_i(\tilde{x}^s)\right\|^2\\ &\leq 3\mathbb{E}\left\|\hat{\nabla} f_i(x_t^s)-\nabla f_i(x_t^s)\right\|^2+3\left\|\hat{\nabla} f_i(\tilde{x}^s)-\nabla f_i(\tilde{x}^s)\right\|^2\\ &\quad+3\left\|\nabla f_i(x_t^s)-\nabla f_i(\tilde{x}^s)\right\|^2\\ &\leq \frac{3L^2 d^2\mu^2}{2}+3L^2\left\|x_t^s-\tilde{x}^s\right\|^2\end{aligned} \tag{19}$$

where the last inequality used the fact that f_{i,μ_j} is L-smooth. (16) is by Assumption 7.2 and (17) uses Lemma 10.1. The proof is now complete. □

Lemma 7.3 indicates that variance of $\hat{v}_{t-1}^s$ has an upper bound. By increasing the number of iterations, we will show both x_{t-1}^s and $\tilde{x}^{s-1}$ will approach the same stationary point x^*. This results in decreasing the variance of stochastic gradient, but due to the zeroth-order gradient estimation and the variance of the gradient on batch, it does not

diminish. In the sequel we frequently use the following inequality

$$\|x-z\|^2 \leq (1+\frac{1}{\beta})\|x-y\|^2 + (1+\beta)\|y-z\|^2, \forall \beta > 0 \tag{20}$$

Blow we present the counterpart of Lemma (7.3) for the mix gradient estimation in (9).

Lemma 7.4. *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.*

$$\begin{aligned}\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s) - \tilde{v}_{t-1}^s\right\|^2\right] \leq & \frac{6\eta L^2 d}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\ & + 2\frac{I\{\mathcal{B}<n\}\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2}\end{aligned} \tag{21}$$

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] \\
\leq&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 (22) \\
=&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 (23)
\end{aligned}$$

$$
\begin{aligned}
&= \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] \\
&\quad + 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 (24) \\
&\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] \\
&\quad + 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad (25) \\
&\le \frac{3\eta L^2 d^2\mu^2}{b} + \frac{6\eta L^2 d}{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] \\
&\quad + 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \qquad (26) \\
&\le \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}} + 2\eta\mathbb{E}\left\|\hat{\nabla} f(x_{t-1}^s) - \nabla f(x_{t-1}^s)\right\|^2 \\
&\quad + \frac{3\eta L^2 d^2\mu^2}{b} \qquad (27) \\
&\le \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}} \\
&\quad + \eta\frac{L^2 d^2\mu^2}{2} + \frac{3\eta L^2 d^2\mu^2}{b} \qquad (28) \\
&\le \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^2 d^2\mu^2}{2} \qquad (29)
\end{aligned}
$$

where, the expectations are taking with respect to I_b and $I_{\mathcal{B}}$ and random directions $\{u_i\}$ in (2). The inequality (22) holds by the Jensen's inequality. (23) and (24) 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). (25) uses the fact that $\mathbb{E}[\|x - \mathbb{E}[x]\|^2] \le \mathbb{E}[\|x\|^2]$, for any random variable x. (26) holds due to Lemma 10.2. (27) is by Assumption 7.2 and (28) is by Lemma 10.1. (29) uses $b \ge 1$. The proof is now complete. □

7.1 Analysis for ZO-PSVRG+

In Theorem 7.5, we concentrate on the convergence rate of ZO-PSVRG+ and provide some remarks.

Theorem 7.5. *Suppose Assumptions 7.1 and 7.2 hold, and the ZO gradient estimator (8) for mix gradient $\hat{v}_k$ is used. The output $\hat{x}$ of Algorithm 6 satisfies*

$$
\mathbb{E}[\left\|g_\eta(\hat{x})\right\|^2] \le \frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} + \frac{I\{\mathcal{B}<n\}12\sigma^2}{\mathcal{B}} + 21L^2 d^2\mu^2 \qquad (30)
$$

where $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12mL}\}$ *denotes the stepsize and* x^* *represents the optimal value of problem 1.*

Proof. Now, we apply Lemma 10.3 to prove Theorem 7.5. 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 (64), we have

$$\begin{aligned} F(x_t^s) \leq & F(\overline{x}_t^s) + \left\langle \nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle - \frac{1}{\eta} \left\langle x_t^s - x_{t-1}^s, x_t^s - \overline{x}_t^s \right\rangle \\ & + \frac{L}{2} \left\| x_t^s - x_{t-1}^s \right\|^2 + \frac{L}{2} \left\| \overline{x}_t^s - x_{t-1}^s \right\|^2. \end{aligned} \tag{31}$$

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 (64), we have

$$\begin{aligned} F(\overline{x}_t^s) \leq & F(x_{t-1}^s) - \frac{1}{\eta} \left\langle \overline{x}_t^s - x_{t-1}^s, \overline{x}_t^s - x_{t-1}^s \right\rangle + \frac{L}{2} \left\| \overline{x}_t^s - x_{t-1}^s \right\|^2 \\ = & F(x_{t-1}^s) - (\frac{1}{\eta} - \frac{L}{2}) \left\| \overline{x}_t^s - x_{t-1}^s \right\|^2. \end{aligned} \tag{32}$$

Combining (31) and (32) 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 + \| x_t^s - \overline{x}_t^s \|^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} \| x_t^s - \overline{x}_t^s \|^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 \qquad (33) \\ = & 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 \\ & \qquad (34) \end{aligned}$$

where the second inequality uses (20) with $\beta = 3$ and the last inequality holds due to the Lemma 10.4.

Note that $x_t^s = \mathrm{Prox}_{\eta h}(x_{t-1}^s - \eta \hat{v}_{t-1}^s)$ is the iterated form in our algorithm. By taking the expectation with respect to all random variables in (34) we obtain

$$\mathbb{E}[F(x_t^s)] \le \mathbb{E}\left[F(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2 + \eta\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2\right] \tag{35}$$

In (35), we further bound $\eta\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2$ using Lemma 7.3 to obtain

$$\begin{aligned}
&\mathbb{E}[F(x_t^s)] \\
\le&\mathbb{E}\left[F(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{1}{3\eta} - L\right)\left\|\overline{x}_t^s - x_{t-1}^s\right\|^2\right] \\
&+ \frac{6\eta L^2}{b}\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \\
=&\mathbb{E}\left[F(x_{t-1}^s) - (\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - x_{t-1}^s\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ \frac{6\eta L^2}{b}\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \qquad (36) \\
\le&\mathbb{E}\left[F(x_{t-1}^s) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ (\frac{6\eta L^2}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \qquad (37)
\end{aligned}$$

where recalling $\overline{x}_t^s := \mathrm{Prox}_{\eta h}(x_{t-1}^s - \eta\nabla f(x_{t-1}^s))$, (36) is based on the definition of gradient mapping $g_\eta(x_{t-1}^s)$. (37) uses (20) by choosing $\beta = 2t-1$.

Taking a telescopic sum for $t = 1, 2, \ldots, m$ in epoch s from (37) and recalling that $x_m^s = \tilde{x}^s$ and $x_0^s = \tilde{x}^{s-1}$, we obtain

$$\begin{aligned}
&\mathbb{E}[F(\tilde{x}^s)] \\
\le&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \sum_{t=1}^{m}\frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ \sum_{t=1}^{m}(\frac{6\eta L^2}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 \\
&+ \sum_{t=1}^{m}\frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \\
\le&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \sum_{t=1}^{m-1}\frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ \sum_{t=2}^{m}(\frac{6\eta L^2}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 \\
&+ \sum_{t=1}^{m}\frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \qquad (38)
\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \left(\frac{\eta}{3} - L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&- \sum_{t=1}^{m-1}\left((\frac{1}{2t} - \frac{1}{2t+1})(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b})\right)\mathbb{E}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 \\
&+ \sum_{t=1}^{m}\frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \left(\frac{\eta}{3} - L\eta^2\right)\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&- \sum_{t=1}^{m-1}\left(\frac{1}{6t^2}(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b})\right)\mathbb{E}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 \\
&+ \sum_{t=1}^{m}\frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \\
\leq&\mathbb{E}\left[F(\tilde{x}^{s-1}) - \frac{\eta}{6}\sum_{t=1}^{m}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] + \sum_{t=1}^{m}\frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \sum_{t=1}^{m}\eta\frac{7L^2d^2\mu^2}{2} \qquad (39)
\end{aligned}
$$

where (38) holds since norm is always non-negative and $x_0^s = \tilde{x}^{s-1}$. In (39) 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 (39), we obtain

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

Thus, we have

$$
\mathbb{E}[\left\|g_\eta(\hat{x})\right\|^2] \leq \frac{6\,(F(x_0) - F(x^*))}{\eta Sm} + \frac{I\{\mathcal{B} < n\}12\sigma^2}{\mathcal{B}} + 21L^2d^2\mu^2 \qquad (40)
$$

where (40) holds since we choose $\hat{x}$ uniformly randomly from $\{x_{t-1}^s\}_{t\in[m], s\in[S]}$. □

ZO-SVRG-Coord and ZO-ProxSVRG/SAGA used a Lyapunov function to show that the accumulated gradient mapping decreases with epoch s. In our analysis, we explicitly prove that $F(x^s)$ decreases and therefore, the proof for Theorem 7.5 is significantly different from proofs in the existing literature. This decent in function values is achieved through employing the inequalities using Lemma 7.3 which provides a more straightforward exploration for our ZO-PSVRG+ versus ZO-SVRG-Coord, ZO-ProxSVRG and ZO-ProxSAGA. In addition, our convergence result is valid for unfixed minibatch sizes and any epoch sizes m unlike ZO-SVRG-Coord which remains true only for specific values of m with a complicated parameter setting. We also do not compute the full gradient during the iterations, i.e., $\mathcal{B} \neq n$. (30) illustrates that a large

batch size $\mathcal{B}$ in fact decreases the variance of estimated full gradient and enhances the convergence of ZO-PSVRG+.

In contrast to the convergence rate of SVRG in Reddi et al. (2016b), Theorem 7.5 presents two extra error terms $\frac{I\{\mathcal{B}<n\}\sigma^2}{\mathcal{B}}$ and $O(L^2d^2\mu^2)$, attributed to 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 $\mathcal{B} = n$. It is also observed the stepsize η depends on the epoch length m, and the minibatch size b. In order to obtain an explicit description for the parameters in Theorem 7.5, Corollary 7.6 shows the convergence rate of ZO-PSVRG+ in terms of precision at the solution $\hat{x}$ and for specific parameter settings, as demonstrated below.

Corollary 7.6. *We set the batch size $\mathcal{B} = \min\{12\sigma^2/\epsilon, n\}$ and the smoothing parameter $\mu \leq \frac{\sqrt{\epsilon}}{5dL}$. Suppose $\hat{x}$ returned by Algorithm 6 is an ϵ-accurate solution for problem (1). Recalling that CoordSGE require $O(d)$ function queries, the number of SZO calls is at most*

$$d(S\mathcal{B} + Smb) = 6d\,(F(x_0) - F(x^*))\,(\frac{\mathcal{B}}{\epsilon\eta m} + \frac{b}{\epsilon\eta}) = O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{bd}{\epsilon\eta}\right). \tag{41}$$

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 ZO calls is at most

$$72dL(F(x_0) - F(x^*))\left(\frac{\mathcal{B}}{\epsilon\sqrt{b}} + \frac{b}{\epsilon}\right) = O\left(s_n\frac{d}{\epsilon\sqrt{b}} + \frac{bd}{\epsilon}\right). \tag{42}$$

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

Proof. Using Theorem 7.5, we have $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12mL}\}$

$$\mathbb{E}[\left\|g_\eta(\hat{x})\right\|^2] \leq \frac{6\,(F(x_0) - F(x^*))}{\eta Sm} + \frac{I\{\mathcal{B} < n\}12\sigma^2}{\mathcal{B}} + 21L^2d^2\mu^2 = 3\epsilon \tag{43}$$

Now we obtain 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 (43) is 0, the proof is finished as the number of SFO call equals to $Sn + Smb = 6\,(F(x_0) - F(x^*))\,(\frac{n}{\epsilon\eta m} + \frac{b}{\epsilon\eta})$. If $\mathcal{B} < n$ the number of SZO calls equal to $d(S\mathcal{B} + Smb) = 6d\,(F(x_0) - F(x^*))\,(\frac{\mathcal{B}}{\epsilon\eta m} + \frac{b}{\epsilon\eta})$ 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 corollary is obtained by setting $m = \sqrt{b}$ in the first part. □

Generally speaking, Corollary 7.6 indicates that if we select the smoothing parameter μ reasonably small and the batch size $\mathcal{B}$ sufficiently large, then the error originating from zeroth-order estimation and batch gradient approximation would reduce, leading to non-dominant effect on the convergence rate of ZO-PSVRG+. The error term induced by batch size is eliminated only when $\mathcal{B} = n$ (i.e., $I\{\mathcal{B} < n\} = 0$). In this

case, ZO-PSVRG+ changes to ZO-ProxSVRG since Step 5 of Algorithm 6 becomes $\hat{g}^s = \hat{\nabla} f(\tilde{x}^{s-1})$. Note that equation (42) shows that a large batch $\mathcal{B}$ for $\mathcal{B} \neq n$ indeed reduces the error inherited by the variance of batch gradient and improves the convergence of ZO-PSVRG+. In summation, if the smoothing parameter and batch size are chosen appropriately, we derive the error term $O(1/T)$, 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 less SZO oracle which is listed in Table 4. It is worth mentioning that the stepsize η in Theorem 7.5 is less restrictive than the existing SZO algorithms in Table 4, e.g., ZO-SVRG-Coord (Gu et al. (2018a)) requires $\eta = O(\frac{1}{d})$.

7.2 Analysis for ZO-PSVRG+ (RandSGE)

Based on Lemma 9, we indicate that ZO-PSVRG+ using the mix gradient (9) achieves significant improvements both in the convergence rate and the function query complexity, as demonstrated in the subsequent analysis.

Theorem 7.7. *Suppose Assumptions 7.1 and 7.2 hold, and the coordinate gradient estimator (9) for mix gradient $\hat{v}_k$ is used. The output $\hat{x}$ of Algorithm 6 satisfies*

$$\mathbb{E}[\left\|g_\eta(\hat{x})\right\|^2] \leq \frac{6\left(F(x_0) - F(x^*)\right)}{\eta S m} + \frac{I\{\mathcal{B} < n\}12\sigma^2}{\mathcal{B}} + 21L^2d^2\mu^2 \tag{44}$$

where $\eta = \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12mL\sqrt{d}}\}$ denotes the stepsize and x^* denotes the optimal value of problem 1.

Corollary 7.8. *We set the batch size $\mathcal{B} = \min\{12\sigma^2/\epsilon, n\}$ and the smoothing parameter $\mu \leq \frac{\sqrt{\epsilon}}{5dL}$. Suppose $\hat{x}$ returned by Algorithm 6 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*

$$(dS\mathcal{B} + Smb) = 6\left(F(x_0) - F(x^*)\right)\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta}\right) = O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta}\right). \tag{45}$$

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 ZO calls is at most

$$72L(F(x_0) - F(x^*))\left(\frac{\mathcal{B}d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right) = O\left(s_n\frac{d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right). \tag{46}$$

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

Remark 7.9. The results from Theorem 7.7 improves the convergence rate $O(\frac{dn^{2/3}}{T})$ for ZO-SVRG-Coord-Rand (Ji et al. (2019)) in single-batch setting and with the stepsize $O(\frac{1}{dn^{2/3}})$ to the convergence rate of $O(\frac{\sqrt{d}}{T})$ with the stepsize $O(\frac{1}{\sqrt{d}})$. Also note that ZO-SVRG-Coord-Rand in single-batch setting requires that the number of inner iterations is

equal to $m = d$. If we choose $b = dm^2$ for ProxSVRG+, then η reduces to $O(1)$ with the convergence rate $O(\frac{1}{\epsilon})$ which generalizes the best result for ZO-SVRG-Coord-Rand, which is only achieved by selecting $m = s_n^{1/3}$.

8 Convergence Under PL Condition

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

$$\|\nabla f(x)\|^2 \geq 2\lambda(f(x) - f^*) \tag{47}$$

where $\lambda > 0$ and f^* indicates the optimal function value. This condition specifies the rate of increasing of the loss function in a vicinity of optimal solutions. It is important to note that if f is λ-strongly convex then f fulfills the PL condition. We will prove the complexity of ZO-PSVRG+ (Algorithm 6) 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,

$$\left\|g_\eta(x)\right\|^2 \geq 2\lambda(F(x) - F^*) \tag{48}$$

for some $\lambda > 0$ and for all $x \in \mathbb{R}^d$. Note that if $h(x)$ is a constant function, the gradient projection changes to $g_\eta(x) = \nabla f(x)$. The PL condition has been investigated comprehensively in Karimi et al. (2016), where the authors proved that PL condition is milder than a large class of functions. The revised PL condition (48) is controversially natural and studied in several papers for problems with nonconvex nonsmooth setting, e.g., Li and Li (2018). A zeroth-order algorithm under PL condition for smooth functions has been analyzed in Ji et al. (2019).

8.1 ZO-PSVRG+ Under PL Condition

In the same way as Theorem 7.5, we show the convergence result of ZO-PSVRG+ (Algorithm 6) under PL-condition. Particularly, we present a generic convergence setting for enhancing the convergence rate for existing SZO algorithms for functions under PL condition using variance reduced methods. It is worth noting that for functions satisfying PL condition (i.e. (48) 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+ with mix gradient (8) under PL condition.

Theorem 8.1. *Let Assumptions 7.1 and 7.2 hold, and ZO gradient estimator (8) for mix gradient $\hat{v}_k$ is used in Algorithm 6 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*

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I\{\mathcal{B} < n\}\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2d^2\mu^2}{2\lambda} \tag{49}$$

Proof. We start by recalling inequality (36) from the proof of Theorem 7.5, i.e.,

$$\begin{aligned}
&\mathbb{E}[F(x_t^s)] \\
\leq&\mathbb{E}\left[F(x_{t-1}^s) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \left(\frac{\eta}{3} - L\eta^2\right)\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ (\frac{2\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \eta\frac{L^2d^2\mu^2}{2} \\
\leq&\mathbb{E}\left[F(x_{t-1}^s) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \frac{\eta}{6}\left\|g_\eta(x_{t-1}^s)\right\|^2\right] \\
&+ (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \qquad (50)
\end{aligned}$$

where in (50) inequality we applied $\eta L \leq \frac{1}{6}$. Moreover, substituting PL inequality, i.e.,

$$\left\|g_\eta(x)\right\|^2 \geq 2\lambda(F(x) - F^*) \qquad (51)$$

into (50), we obtain

$$\begin{aligned}
&\mathbb{E}[F(x_t^s)] \\
\leq&\mathbb{E}\left[F(x_{t-1}^s) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \lambda\frac{\eta}{3}(F(x_{t-1}^s) - F^*)\right] \\
&+ (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \qquad (52)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}[F(x_t^s)] \\
\leq&\mathbb{E}\left[(1 - \lambda\frac{\eta}{3})(F(x_{t-1}^s) - F^*) - \frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
&+ (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\mathbb{E}\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2 + \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \qquad (53)
\end{aligned}$$

Let $\alpha := 1 - \lambda\frac{\eta}{3}$ and $\Psi_t^s := \frac{\mathbb{E}[F(x_t^s) - F^*]}{\alpha^t}$. Combining these definitions with (53), we have

$$\begin{aligned}
&\Psi_t^s \\
\leq&\Psi_{t-1}^s - \frac{1}{\alpha^t}\mathbb{E}\left[\frac{1}{2t}(\frac{5}{8\eta} - \frac{L}{2})\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - (\frac{6\eta L^2 d}{b} + \frac{1}{2t-1}(\frac{5}{8\eta} - \frac{L}{2}))\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] \\
&+ \frac{1}{\alpha^t}\frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \frac{1}{\alpha^t}\eta\frac{7L^2d^2\mu^2}{2} \qquad (54)
\end{aligned}$$

Similar to the proof of Theorem 7.5, summing (54) for $t = 1, 2, \ldots, m$ in epoch s and recalling that $x_m^s = \tilde{x}^s$ and $x_0^s = \tilde{x}^{s-1}$, we have

$$\mathbb{E}[F(\tilde{x}^s) - F^*]$$

$$
\begin{aligned}
&\leq \alpha^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \alpha^m \sum_{t=1}^{m} \frac{1}{\alpha^t} \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \alpha^m \sum_{t=1}^{m} \frac{1}{\alpha^t} \eta \frac{7L^2 d^2 \mu^2}{2} \\
&-\alpha^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{1}{2t\alpha^t} (\frac{5}{8\eta} - \frac{L}{2}) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^{m} \frac{1}{\alpha^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 \alpha^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \frac{1-\alpha^m}{1-\alpha} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&-\alpha^m \mathbb{E}\left[\sum_{t=1}^{m} \frac{1}{2t\alpha^t} (\frac{5}{8\eta} - \frac{L}{2}) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2 - \sum_{t=1}^{m} \frac{1}{\alpha^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 \alpha^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \frac{1-\alpha^m}{1-\alpha} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&-\alpha^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{1}{2t\alpha^t} (\frac{5}{8\eta} - \frac{L}{2}) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
&+\alpha^m \mathbb{E}\left[\sum_{t=2}^{m} \frac{1}{\alpha^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] \qquad (55) \\
&\leq \alpha^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \frac{1-\alpha^m}{1-\alpha} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&-\alpha^m \mathbb{E}\left[\sum_{t=1}^{m-1} \frac{1}{\alpha^{t+1}} \left((\frac{\alpha}{2t} - \frac{1}{2t+1})(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b})\right) \left\|x_t^s - \tilde{x}^{s-1}\right\|^2\right] \\
&\leq \alpha^m \mathbb{E}\left[(F(\tilde{x}^{s-1}) - F^*)\right] + \frac{1-\alpha^m}{1-\alpha} \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \frac{1-\alpha^m}{1-\alpha} \frac{7\eta L^2 d^2 \mu^2}{2} \qquad (56)
\end{aligned}
$$

where (55) since $\|.\|^2$ always is non-negative and $x_0^s = \tilde{x}^{s-1}$. (56) holds since it is sufficient to show $(\frac{\alpha}{2t} - \frac{1}{2t+1})(\frac{5}{8\eta} - \frac{L}{2}) - \frac{6\eta L^2}{b} \geq 0$, for all $t = 1, 2, \ldots, m$. It is easy to see that this inequality holds since $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{\gamma b}}{12mL}\}$, where $\gamma = 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} > 0$. Similarly, let $\tilde{\alpha} = \alpha^m$ and $\tilde{\Psi}^s = \frac{\mathbb{E}[F(\tilde{x}^s) - F^*]}{\tilde{\alpha}^s}$. Substituting these definitions into (56), we have

$$
\tilde{\Psi}^s \leq \tilde{\Psi}^{s-1} + \frac{1}{\tilde{\alpha}^s} \frac{1-\tilde{\alpha}}{1-\alpha} \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \frac{1}{\tilde{\alpha}^s} \frac{1-\tilde{\alpha}}{1-\alpha} \frac{7\eta L^2 d^2 \mu^2}{2} \qquad (57)
$$

Taking a telescopic sum from (57) for all epochs $1 \leq s \leq S$, we obtain

$$
\begin{aligned}
\mathbb{E}[F(\tilde{x}^S) - F^*] &\leq \tilde{\alpha}^S \mathbb{E}[F(\tilde{x}^0) - F^*] + \tilde{\alpha}^S \sum_{s=1}^{S} \frac{1}{\tilde{\alpha}^s} \frac{1-\tilde{\alpha}}{1-\alpha} \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \tilde{\alpha}^S \sum_{s=1}^{S} \frac{1}{\tilde{\alpha}^s} \frac{1-\tilde{\alpha}}{1-\alpha} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&= \alpha^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{1-\tilde{\alpha}^S}{1-\tilde{\alpha}} \frac{1-\tilde{\alpha}}{1-\alpha} \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \frac{1-\tilde{\alpha}^S}{1-\tilde{\alpha}} \frac{1-\tilde{\alpha}}{1-\alpha} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&\leq \alpha^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{1}{1-\alpha} \frac{2I\{\mathcal{B} < n\}\eta\sigma^2}{\mathcal{B}} + \frac{1}{1-\alpha} \frac{7\eta L^2 d^2 \mu^2}{2} \\
&= \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I\{\mathcal{B} < n\}\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2 d^2 \mu^2}{2\lambda} \qquad (58)
\end{aligned}
$$

where in (58) we recall that $\alpha = 1 - \frac{\lambda\eta}{3}$. □

Theorem 8.1 shows that if the batch size and smoothing parameter are appropriately chosen, ZO-PSVRG+ has a dominant linear convergence rate. Further, comparing with Theorem 7.5, it is evident from (49) 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 8.2.

Corollary 8.2. *Suppose the final iteration point $\tilde{x}^S$ in Algorithm 6 satisfies $\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \epsilon$ under PL condition. Under Assumptions 7.1 and 7.2, 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*

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

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

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

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

Proof. From Theorem 8.1, we have

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm}\mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I\{\mathcal{B}<n\}\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2d^2\mu^2}{2\lambda} = 3\epsilon \quad (59)$$

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 $\frac{21L^2d^2\mu^2}{2\lambda} \leq \epsilon$. The number of SZO calls equals 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 calls equals to $T = Sm = O(\frac{1}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$ and the number of SZO calls equals to $d(S\mathcal{B} + Smb) = O(\frac{\mathcal{B}d}{\lambda\sqrt{\gamma}m}\log\frac{1}{\epsilon} + \frac{bd}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$. □

Corollary 8.2 indicates that leveraging the PL condition improves the dominant convergence rate, where the error of order $O(1/\epsilon)$ in Corollary 7.6 is amended to $O(\log(1/\epsilon))$, leading to a significant speed up. Compared to the sub-linear convergence rate for ZO algorithms in Duchi et al. (2015); Nesterov and Spokoiny (2017); Liu et al. (2018b), the convergence performance of PSVRG+ under PL condition has a global linear convergence rate and therefore has lower number of ZO oracle calls. This also indicates that if ZO-PSVRG+ is initialized in a generic non-convex domain, it can automatically achieve an accelerated convergence rate due to getting into a PL area.

It is an improved result compared with Reddi et al. (2016a) where they applied PL-SVRG/SAGA to restart ProxSVRG/SAGA in order to obtain a linear convergence rate

under PL condition. In addition, note that the convergence analysis under PL condition in Ji et al. (2019) has complex coupling structures which makes it hard to apply from practical perspective while our proof is simple and the parameters are directly specified.

Remark 8.3. Compared to Theorem 7.5, the convergence rate of ZO-PSVRG+ in Theorem 8.1 exhibits additional parameter γ for parameter selection due to the use of PL condition. If we assume the condition number $\lambda/L \leq \frac{1}{n^{1/3}}$ and choose $m = n^{1/3}$ and $\eta = \frac{\rho}{L}$ with $\rho \leq \frac{1}{2}$, then the definition of γ yields

$$\begin{aligned}\gamma &= 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} \\ &= 1 - \frac{2\lambda\rho}{3L}m - \frac{\lambda\rho}{3L} \\ &\geq 1 - \frac{2\rho}{3n^{1/3}}m - \frac{\rho}{3n^{1/3}} \\ &\geq 1 - \rho \geq \frac{1}{2}\end{aligned} \tag{60}$$

According to Theorem 8.1, equation (60) implies $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12\sqrt{2}mL}\}$. Hence, choosing $b = n^{2/3}$ leads to the constant stepsize $\eta = \frac{1}{24L}$. 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).

8.2 ZO-PSVRG+ (RandSGE) Under PL Condition

In the following theorem, we explore if ZO-PSVRG+ with mix gradient (9) achieves a linear convergence rate when it enters a local landscape where the loss function satisfying the PL condition.

Theorem 8.4. *Let Assumptions 7.1 and 7.2 hold, and ZO gradient estimator (9) for mix gradient $\hat{v}_k$ is used in Algorithm 6 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*

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I\{\mathcal{B} < n\}\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2d^2\mu^2}{2\lambda} \tag{61}$$

Corollary 8.5. *Suppose the final iteration point $\tilde{x}^S$ in Algorithm 6 satisfies $\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \epsilon$ under PL condition. Under Assumptions 7.1 and 7.2, 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*

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

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

Proof. From Theorem 8.1, we have

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I\{\mathcal{B} < n\}\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2d^2\mu^2}{2\lambda} = 3\epsilon \quad (62)$$

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 $\frac{21L^2d^2\mu^2}{2\lambda} \leq \epsilon$. 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 $(S\mathcal{B}d + Smb) = O(\frac{\mathcal{B}d\sqrt{d}}{\lambda\sqrt{\gamma}m}\log\frac{1}{\epsilon} + \frac{b\sqrt{d}}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$. □

Remark 8.6. 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 an involved parameter selection. In addition, it should be noted that by selecting $b = O(d)$, the stepsize η reduces to $O(1)$ with $O(s_n d\log\frac{1}{\epsilon})$ SZO queries.

9 Experimental Results

We provide our experimental results in this section. We compare the performance of our ZO-PSVRG+ with 1) ZO-ProxSVRG(based on improved analysis), 2) ZO-ProxSAGA-Coord Gu et al. (2018a) and 3) ZO-ProxSGD Ghadimi and Lan (2016) in experiments on two applications: black-box binary classification and adversarial attacks on black-box deep neural networks (DNNs). We let ZO-ProxSGD denote RSPGF based on CoordSGE (3) for gradient estimation. We also let ZO-ProxSVRG and ZO-ProxSVRG (RandSGE) denote ZO-PSVRG+ with $B = n$ based on the mix gradient approximation in (8) and (9), respectively.

9.1 Black-Box Binary Classification

In the first set of our experiments, we investigate 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 $f_i(x) = \log(1 + e^{-y_i z_i^T x})$, $h(x) = \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|^2$, where $z_i \in \mathbb{R}^d$ and y_i is the corresponding label for each i. The L_1 regularization and L_2 regularization weights λ_1 and λ_2 are set respectively to 10^{-4} and 10^{-6}, in all the experiments. 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 2. The epoch size is

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

chosen as $m = 30$ in all of our experiments and we fix the minibatch size b is fixed to 50. The learning rates are tuned in the experiments for competitive algorithms according to Table 4, and the results shown in this section are based on the best learning rate for each algorithm we achieved. We set stepsize η and μ according to our assumptions in lemmas and theorems for ZO-PSVRG+.

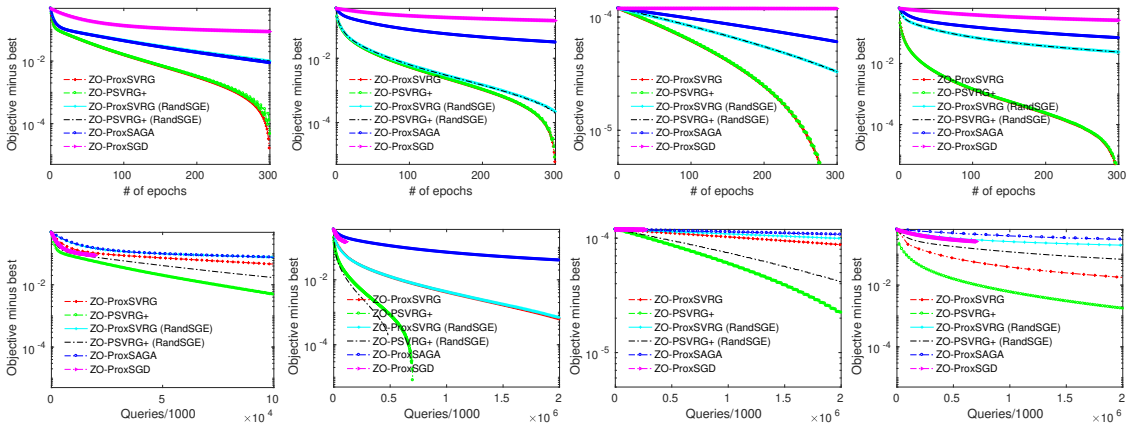
Table 2: Summary of training datasets.

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

In Figure 9.1 (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 9.1 (bottom) compare the performance of ZO-PSVRG+ with the variants of ZO variance reduced stochastic gradient descent described earlier in this section against the number of function queries. In these figures, we notice a relatively faster convergence rate for ZO-PSVRG+ using (8) than the counterpart employing (9). Note that ZO-ProxSVRG based on our improved analysis have faster convergence rate than ZO-ProxSAGA and also ZO-ProxSGD. On the other hand, the use of $\mathcal{B} = \frac{n}{5}$ in ZO-PSVRG+ significantly improves ZO-ProxSVRG with respect to the number of ZO-queries) (see Table 4), leading to a non-dominant factor $O(I_{\{\mathcal{B}<n\}}/\mathcal{B})$ in the convergence rate of ZO-PSVRG+. Particularly ZO-PSVRG+ exhibits better performance in terms of number of function queries than ZO-ProxSAGA using CoordSGE. The degradation in the convergence of ZO-ProxSAGA is due to the requisite for small stepsizes of order $\frac{1}{d}$. Similarly, the large number of function queries to construct coordinate-wise gradient estimates decreases the speed of convergence 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+ obtains the best tradeoffs between the iteration and the function query complexity.

9.2 Adversarial Attacks on Black-Box DNNs

Adversarial examples in image classification is related to designing unperceptive perturbations such that, by adding to the natural images, lead to misclassifying the target model. In the framework of zeroth-order attacks Chen et al. (2017); Liu et al. (2018b), the model parameters are unexposed and obtaining its gradient is not feasible and only the model evaluations are available. We can then consider the task of producing a universal adversarial example with respect to n natural images as an ZO optimization problem of the form (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



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

Figure 1: 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)

to black-box DNNs problem:

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \max\{F_{y_i}(a_i^{adv}) - \max_{j\neq y_i} F_j(a_i^{adv}), 0\} + c\left\|a_i^{adv} - a_i\right\|^2 + \lambda_1\|x\|_1 + \lambda_2\|x\|^2 \tag{63}$$

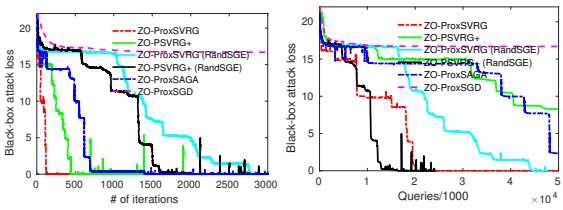
where $a_i^{adv} = 0.5\tanh(\tanh^{-1}(2a_i) + x)$ and λ_1 and λ_2 are nonnegative parameters to harmonize attack success rate, distortion and sparsity. Here $F(a) = [F_1(a), \ldots, F_K(a)] \in [0,1]^K$ describes a trained DNN[2] for the MNIST handwritten digit classification, where $F_i(a)$ returns the prediction score of i-th class. The parameter c in (63) compensate the rate of adversarial success and the distortion of adversarial examples. In our experiment, we set the regularization parameter $c = 0.2$. In addition, we set $\lambda_1 = 10^{-5}$ and $\lambda_2 = 10^{-5}$ in the experiments.

We perfrom two experiments by choosing $n = 10$ and $n = 100$ images from the same class, and set the minibatch sizes, respectively $b = 5$ and $b = 30$. The stepsizes are selected $30/d$ and $30/d$ respectively for ZO-PSVRG+ and ZO-PSVRG+ (RandSGE), where $d = 28\times 28$ is the image dimension. The stepsize η for other algorithms are selected according to Table 4. We select the batch size $\mathcal{B} = \lfloor \frac{n}{2} \rfloor$ for ZO-PSVRG+.

Figure 9.2 shows the performance of different ZO algorithms considered in this paper. Our two algorithms ZO-PSVRG+ (RandSGE) and ZO-ProxSVRG (under our improved analysis) show 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 large number of function queries for CoordSGE and the variance inherited by $\mathcal{B} \neq n$. ZO-PSVRG+

[2]https://github.com/carlini/nn_robust_attacks

(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 (9) and the $O(\frac{1}{\sqrt{d}})$-level stepsize required by ZO-PSVRG+ (RandSGE). ZO-ProxSAGA and ZO-PSVRG+ (CoordSGE) exhibit rel-



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

Figure 2: Comparison of different zeroth-order algorithms for generating black-box adversarial examples from a black-box DNN

atively similar convergence behavior. Furthermore, the convergence performance of ZO-ProxSGD is poor compared to other algorithms due to using variance reduced algorithms.

10 Appendix

In this section, we present the complete proofs of the above lemmas and theorems. In the beginning, we give some useful properties of CoordSGE and RandSGE, respectively.

Lemma 10.1 (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. *$\left|f_\mu(x) - f(x)\right| \le \frac{L\mu^2}{2}$ and $\left|\frac{\partial f_\mu(x)}{\partial x_j}\right| \le \frac{L\mu^2}{2}$.*
3. *$\left\|\hat{\nabla} f(x) - \nabla f(x)\right\|^2 \le \frac{L^2 d^2 \mu^2}{4}$.*

Theorem 10.2. *Assume that the function $f(x)$ is L-smooth. Let $\hat{\nabla}_r f(x)$ 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. *$\left|f_\mu(x) - f(x)\right| \le \frac{\mu^2 L}{2}$ and $\left\|f_\mu(x) - f(x)\right\| \le \frac{\mu L d}{2}$ 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). □

Lemma 10.3. *For a given $x \in \mathbb{R}^d$, let $\overline{x} = Prox_{\eta h}(x - \eta v)$, then we have for all $w \in \mathbb{R}^d$*

$$\begin{aligned} F(\overline{x}) \leq F(w) + \langle \nabla f(x) - v, \overline{x} - w \rangle - \frac{1}{\eta} \langle \overline{x} - x, \overline{x} - w \rangle \\ + \frac{L}{2} \|\overline{x} - x\|^2 + \frac{L}{2} \|w - x\|^2 \end{aligned} \tag{64}$$

Proof. First, we recall the proximal operator

$$\text{Prox}_{\eta h}(x - \eta v) := \arg \min_{y \in \mathbb{R}^d} \left(h(y) + \frac{1}{2\eta} \|y - x\|^2 + \langle v, y \rangle \right) \tag{65}$$

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

$$\begin{aligned} h(\overline{x}) &\leq h(w) + \langle p, \overline{x} - w \rangle \\ &= h(w) - \left\langle v + \frac{1}{\eta}(\overline{x} - x), \overline{x} - w \right\rangle \end{aligned} \tag{66}$$

where $p \in \partial h(\overline{x})$ such that $p + \frac{1}{\eta}(\overline{x} - x) + v = 0$ according to the optimality condition of (65), and (66) due to the convexity of h. In addition, since since $f(x)$ is L-Lipschitz continuous, we have

$$f(\overline{x}) \leq f(x) + \langle \nabla f(x), \overline{x} - x \rangle + \frac{L}{2} \|\overline{x} - x\|^2 \tag{67}$$

and

$$f(x) \leq f(w) + \langle \nabla f(x), x - w \rangle + \frac{L}{2} \|w - x\|^2 \tag{68}$$

This lemma is obtained by adding (66), (67), (68), and using $F(x) = f(x) + h(x)$. □

Lemma 10.4. *Let $x_t^s = Prox_{\eta h}(x_{t-1}^s - \eta \hat{v}_{t-1}^s)$ and $\overline{x}_t^s$ be the proximal projection using full true gradient, i.e., $\overline{x}_t^s = Prox_{\eta h}(x_{t-1}^s - \eta \nabla f(x_{t-1}^s))$. 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 (66) we obtain

$$h(x_t^s) \leq h(\overline{x}_t^s) - \left\langle \hat{v}_{t-1}^s + \frac{1}{\eta}(x_t^s - x_{t-1}^s), x_t^s - \overline{x}_t^s \right\rangle \tag{69}$$

$$h(\overline{x}_t^s) \leq h(x_t^s) - \left\langle \nabla f(x_{t-1}^s) + \frac{1}{\eta}(\overline{x}_t^s - x_{t-1}^s), \overline{x}_t^s - x_t^s \right\rangle \tag{70}$$

By summing (69) and (70), we have

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

$$\frac{1}{\eta}\|x_t^s - \overline{x}_t^s\|^2 \leq \left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\| \|x_t^s - \overline{x}_t^s\| \tag{71}$$

where (71) holds by Cauchy-Schwarz inequality. Thus, we obtain

$$\|x_t^s - \overline{x}_t^s\| \leq \eta \left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\| \tag{72}$$

Now the proof is complete using Cauchy-Schwarz inequality and (72). □

References

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

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Zeyuan Allen-Zhu and Yang Yuan. Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In *International conference on machine learning*, pages 1080–1089, 2016.

Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163, 2011.

Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 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*, pages 15–26. ACM, 2017.

Tianyi Chen and Georgios B Giannakis. Bandit convex optimization for scalable and dynamic iot management. *IEEE Internet of Things Journal*, 6(1):1276–1286, 2019.

Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard E Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. *arXiv preprint arXiv:1804.02395*, 2018.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

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

Pavel Dvurechensky, Alexander Gasnikov, and Eduard Gorbunov. An accelerated method for derivative-free smooth stochastic convex optimization. *arXiv preprint arXiv:1802.09022*, 2018.

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

Michael C Fu. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing*, 14(3):192–215, 2002.

Xiang Gao, Bo Jiang, and Shuzhong Zhang. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76 (1):327–363, 2018.

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

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

Bin Gu, Zhouyuan Huo, Cheng Deng, and Heng Huang. Faster derivative-free stochastic algorithm for shared memory machines. In *International Conference on Machine Learning*, pages 1807–1816, 2018a.

Bin Gu, De Wang, Zhouyuan Huo, and Heng Huang. Inexact proximal gradient methods for non-convex and non-smooth optimization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018b.

Davood Hajinezhad, Mingyi Hong, and Alfredo Garcia. Zeroth order nonconvex multi-agent optimization over networks. *arXiv preprint arXiv:1710.09997*, 2017.

Feihu Huang, Bin Gu, Zhouyuan Huo, Songcan Chen, and Heng Huang. Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. *arXiv preprint arXiv:1902.06158*, 2019.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, pages 3100–3109, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

Hamed Karimi, Julie Nutini, and Mark Schmidt. 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*, pages 795–811. Springer, 2016.

Ehsan Kazemi and Liqiang Wang. A proximal zeroth-order algorithm for nonconvex nonsmooth problems. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 64–71. IEEE, 2018.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, pages 1–49, 2017.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2348–2358, 2017.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in neural information processing systems*, pages 379–387, 2015.

Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 5564–5574, 2018.

Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, and Ji Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. In *Advances in Neural Information Processing Systems*, pages 3054–3062, 2016.

Liu Liu, Minhao Cheng, Cho-Jui Hsieh, and Dacheng Tao. Stochastic zeroth-order optimization via variance reduction method. *arXiv preprint arXiv:1805.11811*, 2018a.

Sijia Liu, Jie Chen, Pin-Yu Chen, and Alfred O Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. *arXiv preprint arXiv:1710.07804*, 2017.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3727–3737, 2018b.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011.

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

Atsushi Nitanda. Accelerated stochastic gradient descent for minimizing finite sums. In *Artificial Intelligence and Statistics*, pages 195–203, 2016.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519. ACM, 2017.

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

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

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016a.

Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pages 1145–1153, 2016b.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.

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

Artem Sokolov, Julia Kreutzer, Stefan Riezler, and Christopher Lo. Stochastic structured prediction under bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1489–1497, 2016.

James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pages 896–903. ACM, 2005.

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

Andre Wibisono, Martin J Wainwright, Michael I Jordan, and John C Duchi. Finite sample convergence rates of zero-order stochastic optimization methods. In *Advances in Neural Information Processing Systems*, pages 1439–1447, 2012.

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

Xiaotian Yu, Irwin King, Michael R Lyu, and Tianbao Yang. A generic approach for accelerating stochastic zeroth-order convex optimization. In *IJCAI*, pages 3040–3046, 2018.