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

Ehsan Kazemi, Deliang Fan, *Member, IEEE*, Liqiang Wang, *Member, IEEE*

***Abstract*—Proximal gradient method has an important role in solving nonsmooth composite optimization problems. However, in some machine learning problems related to black-box optimization models proximal gradient method could not be leveraged, where explicit gradient forms are difficult or infeasible to obtain. While first order methods are not suited for solving black-box optimization problems, zeroth-order (ZO) optimization methods can address these problems efficiently. Several varieties of zeroth-order variance reduced stochastic (ZO-SVRG) algorithms have recently been introduced for nonconvex optimization based on the first-order techniques of stochastic variance reduction. However, all existing ZO-SVRG type algorithms suffer from a slowdown and increase in function query complexities up to a small-degree polynomial of the problem size. To fill this gap, we propose a new stochastic gradient algorithm in the gradient-free regime for optimizing nonconvex, nonsmooth finite-sum problems, called ZO-PSVRG+. The analysis of ZO-PSVRG+ recovers several existing convergence results and improves their ZO oracle and proximal oracle calls. Furthermore, we prove that ZO-PSVRG+ under Polyak-Łojasiewicz condition in contrast to the existent ZO-SVRG type methods obtain a global linear convergence for a wide range of minibatch sizes. Our empirical experiments on black-box binary classification and black-box adversarial attack from black-box neural networks demonstrate that the studied algorithms under our new analysis exhibit superior performance and faster convergence to a solution of high accuracy with a lower query complexity compared to state-of-the-art ZO optimization methods for nonconvex nonsmooth problems.**

***Index Terms*—Block-box optimization, zeroth-order gradient estimator (ZO), stochastic variance reduced gradient descent (SVRG), adversarial examples.**

Introduction

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

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

where each $f_i(x)$ is possibly nonconvex and smooth function, and $h(x)$ is a nonsmooth convex function such as l_1-norm regularizer. The 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 [2], [3], [4], [5], [6]. It is known from SVRG that it enhances the rate of convergence for stochastic gradient descent (SGD) complexity by a factor of $O(1/\epsilon)$ due to the decrease in the variance of gradient. One may apply the comparable concepts and similar ideas in the first-order methods to reduce the variance in SZO optimization and enhance the convergence rate.

The major adversity of first-order methods for these accelerated methods is their design which is based on first-order information from the problem. However, there are circumstances where the first-order gradient evaluations are computationally unfeasible, 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 [7]. Similarly, in predictions with stochastic structure computing the derivatives is possibly complicated or forbidden, while the functional estimations of foreseen frameworks are accessible [8]. This highlights the necessity of derivative-free optimization method [9] to tackle these problems. The development of zeroth-order optimization methods has become significantly important to solve many machine learning problems in which calculation of the gradients explicitly is expensive or infeasible to derive. This procedure computes the full gradient using gradient approximation only based on function estimations, which will result to derivative-free optimization [10], [11]. Recently, zeroth-order optimization has attracted significant attention, e.g., black-box adversarial attacks on deep neural networks (DNNs)[12], [13], [14], reinforcement learning [15] and structured prediction [16]. Further applications cover time-varying constrained networks with restricted computation capacity [17], [18], and model inference with black-box setting [19], [20].

Currently, there are only a few number of zeroth-order stochastic methods for solving problem (1), e.g., [21] and [22]. In particular, [21] analyzed a zeroth-order gradient method called RSPGE. Nevertheless, due to high variance based on random vector sampling for two point gradient approximation, the iteration complexity of RSPGE (i.e., $O(\frac{d}{\epsilon^2})$) is notably worse than the best known rate $O(\frac{d}{\epsilon})$ for zeroth-order stochastic optimization. A major issue in the development of SZO algorithms for solving (1) is the order of the required number of function queries, namely SZO calls or iteration complexity. While the existing zeroth-order methods based

E. Kazemi and L. Wang are with the Department of Computer Science, University of Central Florida, Orlando, Florida, USA (e-mail: ehsan_kazemy@knights.ucf.edu; lwang@cs.ucf.edu). D. Fan is with the School of Electrical, Computer and Energy Engineering, Arizona State University (email: dfan@asu.edu). The work was supported in part by NSF-1741431.

on SVRG algorithms show higher convergence rate, they induce higher querying complexity than either of zeroth-order gradient descent (ZO-GD) and zeroth-order stochastic gradient descent (ZO-SGD) methods.

The term related to the dimension of the problem in the convergence studies (i.e., d) plays a key role on the efficiency of SZO optimization. For instance, [1] refined the ZO estimations to derive an improved ZO complexity with improved convergence rate. However, their 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 leveraging 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.

In Table I, we compared the results from our analysis and 8 other ZO optimization method from 4 perspectives: a) the type of gradient estimator, b) the setting of problem, c) stepsize, d) convergence rate, and e) function query complexity. Table 1 shows that for nonconvex nonsmooth optimization, the convergence of ZO-PSVRG+ achieves best dependency on d than RSPGF and and ZO-ProxSVRG/SAGA [22]. Table I shows that RGF [9] has the largest query complexity and yet has the worst convergence rate. ZO-SVRG-Coord [21] and ZO-ProxSVRG/SAGA provide an improved rate of convergence $O(d/\epsilon)$ owning to applying variance reduction techniques. On the other hand, existing SVRG type zeroth-order algorithms are highly affected by function query complexities compared to RSPGF, while our algorithm, ZO-PSVRG+, could achieve better trade-offs between the convergence rate and the querying complexity.

Despite the fact that proximal SVRG has indicated a huge promise for first-order algorithms, utilizing identical concepts to ZO optimization is not effortless. SZO algorithms have involved coupled stochastic structure which arises from both data sampling and the error induced by ZO gradient estimation. This makes the analysis of ZO optimization difficult in many settings. The other major challenge is related to the fact that ProxSVRG is based on the notion that stochastic gradient is an unbiased approximation of the actual full gradient, which is not valid in the ZO case. Considering the motivation of zeroth-order ProxSVRG and the success of proximal SVRG, one question arises that 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 fill the void between SZO optimization and ProxSVRG by improving the complexity of the exiting SZO variance reduced methods for problem (1).

Main contributions

In this paper, we present a novel analysis which is different from the existing convergence studies provided in [23], [1]. Through our new analysis we prove that ZO-PSVRG+ surpasses several state-of-the-art SVRG-type zeroth-order methods as well as RSPGF. In particular, we attempt to address several important open questions in ZO proximal variance reduced methods. More specifically, we address the open question if the dependence on the dimension d for the convergence analysis proposed in [23] is optimal. Our work provides an inclusive analysis on how ZO gradient approximations influence ProxSVRG on both convergence rate and function query complexity. This is conducted based on the novel structure of recently introduced SZO algorithms. Note that our analysis does not rely on bounded gradient assumption in [21], [22]. The convergence results are demonstrated with respect to the number of stochastic zeroth-order (SZO) queries and proximal oracle (PO) calls. We summarize the following results from this paper related to our new analysis:

1) Our analysis yields iteration complexity $O(\frac{1}{\epsilon})$ corresponding to $O(\frac{d}{\epsilon^2})$ of RSPGF [21] and $O(\frac{d}{\epsilon})$ of ZO-ProxSVRG/SAGA [22] (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. ZO-PSVRG+ also matches the best result achieved by ZO-SVRG-Coord-Rand with minibatch size $b = dn^{2/3}$ and epoch size $m = n^{1/3}$ in [1], while our results are valid for any minibatch sizes as detailed in the following sections. Indeed, it is necessary to analyze and study the convergence behavior of SZO optimization with minibatchs of single or moderate sizes, as practically many machine learning models are trained with intermediate minibatch sizes.

2) The convergence analysis for ZO-PSVRG+ in contrast to ZO-SVRG-Coord in [23], [1] is straightforward, and yields simpler proofs. Our analysis achieve new iteration complexity bounds and improve the effectiveness of all the existing ZO-SVRG-based algorithms along with RSPGF for nonconvex nonsmooth composite optimization, while it provides the best results to our best knowledge (see Table I). Note that the convergence studies for RSPGF and ZO-ProxSVRG/SAGA rely on bounded gradient assumption, which is not our working assumption in this paper.

3) For the nonconvex functions under Polyak-Łojasiewicz condition [24], 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 ZO-ProxSVRG for problem (1) with arbitrary minibatch sizes. This analysis generalizes the results of [25] while shows linear convergence versus the sublinear convergence rate in their paper. In [1], the authors show that ZO-SPIDER-Coord achieves linear convergence under PL condition but only for the minibatch of size $b = O(n^{1/2})$. Note that due to both computational and statistical efficiency, convergence analysis for minibatchs of moderate sizes is essential (Also see the remarks after Theorem 9 for more details).

Finally, to demonstrate the efficiency and adaptability of our approach for achieving a balance between the convergence rate and the number of SZO queries, we perform some experimental evaluations for two distinct applications: black-box binary classification and universal adversarial attacks on

black-box deep neural network models. The empirical results and theoretical investigations verify the effectiveness of our algorithms.

Related Works

Derivative-free (zeroth-order) methods have been efficiently applied 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 computes the gradient estimation $\hat{\nabla} f(x)$ by probing f at a single random point near to x, while the two-point estimator given by the difference of function values at random query points [26], [9]. In this paper, we focus on two-point gradient approximation since it has a lower variance and thus represents lower iteration complexity.

The recent studies show that ZO algorithms typically agree with the complexity of first-order algorithms up to a small-degree polynomial of the problem size d. The existing zeroth-order algorithms mostly target (strongly) convex problems, while the studies for nonconvex ZO methods are relatively limited. In particular, there are many nonconvex machine learning applications, where the explicit derivatives are not accessible, e.g., nonconvex black-box learning problems [14], [23]. Thus, developing zeroth-order stochastic methods for nonconvex optimization is indeed essential. [27] and [28] proposed ZO-GD and its corresponding stochastic algorithm ZO-SGD, respectively. [29] introduced a variance reduced stochastic zeroth-order method with Gaussian smoothing. More recently, [23] presented a thorough analysis based on SVRG algorithms. [1] elaborated the results in [23] and achieved improved bounds based on involved relations for parameter selection. However the downside of the improvement offered in their paper is its dependency on large minibatch sizes. More comprehensive discussion on SZO methods for nonconvex nonsmooth probelm (1) can be found in [22], where the authors proposed two algorithms called ZO-ProxSVRG and ZO-ProxSAGA based on the well-known variance reduction techniques ProxSVRG and ProxSAGA [30]. Before that, [21] also considered the stochastic case (here we denote it as RSPGF), which relies on increasing or large minibatch sizes, i.e., roughly of order $\Omega(1/\epsilon)$. However, due to the necessity for growing the size of minibatch sizes during iterations in their analysis, RSPGF may change to deterministic proximal gradient descent (ZO-ProxGD) after few iterations. Further, [29] also analyzed a zeroth-order algorithm for solving nonconvex nonsmooth problems, which are different from problem (1). Several asynchronous stochastic zeroth-order algorithms have also been studied for large-scale problems, e.g., [31], [20], [32]. In [20], an asynchronous ZO stochastic coordinate descent (ZO-SCD) was designed with a convergence rate of $O(d/\epsilon^2)$. In [32], the authors improved the onvergence rate of asynchronous SZO by integrating SVRG techniques with stochastic coordinate descent method. Moreover, ZO versions of Frank-Wolfe (FW) [33], [34] and alternating direction method of multipliers (ADMM) [18], [35] have been developed for constrained optimization.

In spite of the fact that the aforementioned zeroth-order stochastic algorithms can effectively solve problems with nonconvex structure, there are only a limited number of zeroth-order stochastic methods to solve nonconvex nonsmooth composite problems. We emphasize that, in contrast to existing ZO proximal methods, our analysis does not require bounded gradient assumption, which is not valid for many unconstrained optimization problems.

Preliminaries

In the following we 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 [9], [36]

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

where d is the number of optimization variables, $\{u_i\}$ are i.i.d. random directions drawn from a uniform distribution over a unit sphere and $\mu > 0$ is the smoothing parameter [37], [38], [36]. Typically, RandSGE is a biased approximation to the true 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 function evaluations when the rate of noise to signal is high [20]. To obtain a higher quality approximation for ZO gradient, one can apply coordinate gradient estimation (CoordSGE) [31], [32], [23] to evaluate the gradients as:

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

where e_j is a standard basis vector with 1 at its j-th coordinate and 0 otherwise, and μ is the smoothing parameter. In contrast to RandSGE, CoordSGE is deterministic and needs d times more ZO function calls. However, our studies indicate that for ZO variance-reduced methods based on CoordSGE provide a more accurate ZO estimation. This will result in a larger stepsize and a speedier convergence, although the coordinate-wise gradient estimator requires more ZO calls than the two-point random gradient approximation.

Since proximal gradient method needs to compute the gradient in each iteration, it cannot be applied to tackle the optimization problems where the computation of explicit gradient of function $f(x)$ is infeasible. We present a zeroth-order proximal gradient descent method based on ZO gradient estimation (3), which performs iterations of the form

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

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

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

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.

Generally, for convex problems the optimality gap $F(x) - F(x^*)$ is applied as the convergence metric, where throughout

Method	Problem	Stepsize	Convergence rate	SZO complexity
RGF ([9])	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 ([21])	S(NC)+NS(C)	$O(1)$	$O\left(\frac{d}{\epsilon^2}\right)$	$O\left(\frac{nd}{\epsilon^2}\right)$
ZO-SVRG-Coord ([23])	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 ([1])	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 ([22])	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 ([22])	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+ (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+ (Ours)	S(PL)+NS(C)	$O(1)$	$O\left(\log(\frac{1}{\epsilon})\right)$	$O(s_n\frac{d}{\lambda}\log\frac{1}{\epsilon}+\frac{bd}{\lambda}\log\frac{1}{\epsilon})$
ZO-PSVRG+ (RandSGE) (Ours)	S(PL)+NS(C)	$O\left(\frac{1}{\sqrt{d}}\right)$	$O\left(\sqrt{d}\log(\frac{1}{\epsilon})\right)$	$O(s_n\frac{d\sqrt{d}}{\lambda}\log\frac{1}{\epsilon}+\frac{b\sqrt{d}}{\lambda}\log\frac{1}{\epsilon})$

TABLE I: Summary of convergence rate and function query complexity of various SZO algorithms. S: Smooth, NS: Nonsmooth, NC: Nonconvex, C: Convex, SC: Strong Convexity, and PL: Polyak-Łojasiewicz Condition. b denotes the minibatch size, m denotes the epoch size, λ is the constant in PL condition (18) and $s_n = \min\{n, \frac{1}{\epsilon}\}$. *: The single-minibatch version.

the paper, we let x^* denote the optimal solution of Problem (1). However, for general nonconvex problems, the gradient norm is commonly used as the convergence metric. For instance, for smooth nonconvex optimization (i.e., $h(x) = 0$), [27], [3], [6], [23] 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 necessary to define the gradient mapping as illustrated in [21], [30], [22]:

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

If $h(x)$ is a constant function, the 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 [21], [30], [39]. For problems with nonconvex structure, if $g_\eta(x) = 0$, the point x is a stationary point [39]. With the aid of this notion, we can exploit the following definition as the convergence metric.

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

ZO Proximal Stochastic Method (ZO-PSVRG+)

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 $g^s = \frac{1}{\mathcal{B}}\sum_{i\in I_\mathcal{B}}\nabla f_i(\tilde{x}^{s-1})$. We study a proximal stochastic gradient algorithm based on variance reduced approach of ProxSVRG in [40], [30], [41]. The description of ZO-PSVRG+ is presented in Algorithm 1. Our method has two types of random sampling. In the outer iteration, we calculate the gradient consisting of $\mathcal{B}$ samples. In the inner iteration, we randomly

Algorithm 1 Zeroth-Order Proximal Stochastic Method

1: **Input:** initial point x_0, batch size $\mathcal{B}$, minibatch size b, epoch length m, stepsize η
2: **Initialize:** $\tilde{x}^0 = x_0$
3: **for** $s = 1, 2, \ldots, S$ **do**
4: $\quad x_0^s = \tilde{x}^{s-1}$
5: $\quad \hat{g}^s = \frac{1}{\mathcal{B}}\sum_{j\in I_\mathcal{B}}\hat{\nabla} f_j(\tilde{x}^{s-1})$
6: $\quad$ **for** $t = 1, 2, \ldots, m$ **do**
7: $\quad\quad$ Compute $\hat{v}_{t-1}^s$ according to (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 $\{x_t^s\}_{t\in[m],s\in[S]}$

choose a minibatch of samples of size b to approximate gradient over the minibatch. We call $\mathcal{B}$ and b, batch and minibatch size, respectively. In our ZO framework, the mix gradient (7) is estimated by applying only function evaluations, given by

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

or

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

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 using CoordSGE and $\hat{\nabla}_r f_i$ is a ZO gradient estimate using RandSGE. We let ZO-PSVRG+ and ZO-PSVRG+ (RandSGE) denote Algorithm 1 with gradient estimation (8) and (9), respectively. Note that, $\mathbb{E}_{I_b}[\hat{v}_{t-1}^s] = \hat{\nabla} f(x_{t-1}^s) \neq \nabla f(x_{t-1}^s)$, i.e., this stochastic gradient is a biased approximation of the true gradient. In other words, the unbiased assumption on gradient approximates utilized in ProxSVRG [30], [41] is no longer valid. Note that the biased ZO gradient

estimation yields a fundamental challenge in analyzing ZO-PSVRG+. It means that adjusting the similar concepts from ProxSVRG to zeroth-order algorithm 1 is not effortless and requires an elaborated analysis of ZO-PSVRG+. To tackle this issue, we derive an upper bound for the variance of the gradient approximation $\hat{v}_t^s$ by selecting an appropriate stepsize η and smoothing parameter μ to control variance of gradient estimation which is discussed later.

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

Convergence Analysis

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

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

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

Assumption 2. *For $\forall x \in \mathbb{R}^d$, $\mathbb{E}\left[\left\|\hat{\nabla} f_i(x) - \hat{\nabla} f(x)\right\|^2\right] \leq \sigma^2$, where $\sigma > 0$ is a constant and $\hat{\nabla} f_i(x)$ is a CoordSGE gradient approximation of $\nabla f_i(x)$.*

Assumptions 1 and 2 are standard assumptions applied in SZO optimization. The first assumption is for the convergence studies of the zeroth-order algorithms [21], [9], [23]. The second assumption specifies bounded variance for zeroth-order gradient approximations [20], [29], [23]. Assumption 2 is essential in order to obtain a convergence result independent of n. This assumption is weaker than the assumption of bounded gradients in [18], [42], while, we are capable to analyze the more complicated problem (1) involving a nonsmooth part and obtain faster convergence rates. Note that according to the error estimation for CoordSGE, this assumption is equivalent to the bounded variance of true gradients.

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

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

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

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

The proofs of this lemma, the lemmas and the theorems below are all included in the Supplementary Material.

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

Blow we present the counterpart of Lemma 1 for the mix gradient estimation in (9).

Lemma 2. *Given the mix gradient estimation $\tilde{v}_{t-1}^s = \frac{1}{b}\sum_{i\in I_b}\left(\hat{\nabla}_r f_i(x_{t-1}^s) - \hat{\nabla}_r f_i(\tilde{x}^{s-1})\right) + \hat{g}^s$ with $\hat{g}^s = \frac{1}{\mathcal{B}}\sum_{j\in I_{\mathcal{B}}}\hat{\nabla} f_j(\tilde{x}^{s-1})$, the following inequality holds.*

$$\mathbb{E}\left[\eta\left\|\nabla f(x_{t-1}^s) - \tilde{v}_{t-1}^s\right\|^2\right] \leq \frac{6\eta L^2 d}{b}\mathbb{E}\left[\left\|x_{t-1}^s - \tilde{x}^{s-1}\right\|^2\right] + 2\frac{I(\mathcal{B} < n)\eta\sigma^2}{\mathcal{B}} + \eta\frac{7L^2d^2\mu^2}{2} \tag{11}$$

Analysis for ZO-PSVRG+

In Theorem 3, we focus on the convergence rate of ZO-PSVRG+ and provide some corollaries.

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

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

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

In contrast to the convergence rate of SVRG in [30], Theorem 3 presents two extra error terms $\frac{I(\mathcal{B}<n)\sigma^2}{\mathcal{B}}$ and $O(L^2d^2\mu^2)$, related to the batch gradient estimation $\mathcal{B} < n$ and the use of SZO gradient approximations, respectively. The error related to $\mathcal{B} < n$ is removed only when $\mathcal{B} = n$. Note that the stepsize η depends on the epoch length m, and the minibatch size b.

The proof for Theorem 3 is significantly different from the proofs in the existing literature. For instance, the convergence analysis for ZO-SVRG-Coord and ZO-ProxSVRG/SAGA uses the notion of Lyapunov function to show that the accumulated gradient mapping decreases with epoch s. However, in our analysis we explicitly prove that $F(x^s)$ decreases by employing the inequalities from Lemma 1. On the other hand, our convergence result is valid for a wide range of minibatch sizes and any epoch size m, while the analysis for ZO-SVRG-Coord is valid only for specific values of m with a complicated setting for parameter selection.

In order to obtain an explicit description for the parameters in Theorem 3, the next corollary demonstrates the convergence rate of ZO-PSVRG+ in terms of precision at the solution $\hat{x}$.

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

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

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

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

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

Corollary 4 indicates that if the smoothing parameter μ is sufficiently small and the batch size $\mathcal{B}$ is large enough, then the errors induced from zeroth-order estimation and batch gradient approximation will decrease, leading to a non-dominant term in the convergence rate of ZO-PSVRG+. Indeed, the error term induced by batch size is eliminated only when $\mathcal{B} = n$ (i.e., $I(\mathcal{B} < n) = 0$). In this case, Step 5 of Algorithm 1 becomes $\hat{g}^s = \hat{\nabla} f(\tilde{x}^{s-1})$ and consequently ZO-PSVRG+ changes to ZO-ProxSVRG. Note that equation Theorem 3 indicates 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 properly, we derive the error term $O(1/\epsilon)$, which is better than the convergence rate of the state-of-the-art SZO algorithms by the factor $\frac{1}{d}$. Moreover, ZO-PSVRG+ uses much less SZO oracle calls compared to the methods listed in Table I. It is worth mentioning that the stepsize η in Theorem 3 has a milder dimension-dependency than the existing SZO algorithms in Table I.

Analysis for ZO-PSVRG+ (RandSGE)

In this section, we will study the convergence of ZO-PSVRG+ (RandSGE) under different settings using Lemma 2. In particular, in the following theorem, we prove that ZO-PSVRG+ (RandSGE) improves the convergence rate and the function query complexity of the existing SZO methods.

Theorem 5. *Suppose Assumptions 1 and 2 hold, and the coordinate gradient estimator* (9) *is used to compute the mix gradient* $\hat{v}_k$. *The output* $\hat{x}$ *of Algorithm 1 satisfies*

$$\mathbb{E}[\left\|g_\eta(\hat{x})\right\|^2] \leq \frac{6\left(F(x_0) - F(x^*)\right)}{\eta Sm} + \frac{I(\mathcal{B} < n)12\sigma^2}{\mathcal{B}} + 21L^2d^2\mu^2 \tag{14}$$

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

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

$$(dS\mathcal{B} + Smb) = 6\left(F(x_0) - F(x^*)\right)\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta}\right) = O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta}\right) \tag{15}$$

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

$$72L(F(x_0) - F(x^*))\left(\frac{\mathcal{B}d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right) = O\left(s_n\frac{d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon}\right) \tag{16}$$

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 1. *The results from Theorem 5 improves the convergence rate* $O(\frac{dn^{2/3}}{T})$ *for ZO-SVRG-Coord-Rand [1] in single-minibatch 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-minibatch 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 that is only achieved by selecting* $m = s_n^{1/3}$.

Convergence Under PL Condition

In this section, we show the linear convergence of Prox-SVRG+ under Polyak-Łojasiewicz (PL) assumption [24]. The classic structure of PL condition is, for all $x \in \mathbb{R}^d$

$$\|\nabla f(x)\|^2 \geq 2\lambda(f(x) - f^*) \tag{17}$$

where $\lambda > 0$ and f^* denotes 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 that the complexity of ZO-PSVRG+ (Algorithm 1) under PL condition will be 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,

$$\left\|g_\eta(x)\right\|^2 \geq 2\lambda(F(x) - F^*) \tag{18}$$

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 throughly investigated in [43] where the authors proved that PL condition is milder than a large class of functions. Moreover, the authors show that a function can be non-convex and still satisfy the PL condition. The revised PL condition (18) is controversially natural and studied in several papers for problems with nonconvex nonsmooth setting, e.g., [41]. A zeroth-order algorithm under PL condition for smooth functions has been analyzed in [1].

ZO-PSVRG+ Under PL Condition

In this section similarly as Theorem 3, we show the convergence result of ZO-PSVRG+ (Algorithm 1) under PL-condition. In particular, we provide a generic analysis for enhancing the convergence rate for existing SZO algorithms for

functions satisfying PL condition by applying variance reduced techniques. It is worth noting that for functions satisfying PL condition (i.e. (18) holds), ZO-PSVRG+ can immediately use the final iteration $\tilde{x}^S$ as the output point rather than using a randomly chosen $\hat{x}$. The following theorem provides the convergence guarantee for ZO-PSVRG+ under PL condition.

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

$$\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\tilde{x}^0) - F^*] + \frac{6I(\mathcal{B} < n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2d^2\mu^2}{2\lambda} \tag{19}$$

Theorem 7 shows that if the batch size and smoothing parameter are appropriately chosen, ZO-PSVRG+ has a dominant linear convergence rate without restart. Further, compared to Theorem 3, it is evident from (19) 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.

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

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

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

Here we provide some insights on Corollary 8. It actually indicates that leveraging the PL condition improves the dominant convergence rate, when the error of order $O(1/\epsilon)$ in Corollary 4 is improved to $O(\log(1/\epsilon))$, resulting to a significant speed up. Compared to the sub-linear convergence rate for ZO algorithms in [25], [9], [23], the convergence performance of PSVRG+ under PL condition has a global linear convergence rate and therefore requires lower number of ZO oracle calls. This also indicates that if ZO-PSVRG+ is initialized in a generic non-convex domain, the rate of convergence can be automatically accelerated because of entering in PL area. It is an improved result compared with [3] where the authors applied PL-SVRG/SAGA to restart ProxSVRG/SAGA in order to obtain a linear convergence rate under PL condition. On the other hand, note that the convergence analysis under PL condition in [1] has complex coupling structures which makes its application from practical perspective very difficult, while our proof is simple and the parameters are explicitly specified.

Remark 2. *Compared to Theorem 3, the convergence rate of ZO-PSVRG+ in Theorem 7 exhibits additional parameter γ for parameter selection due to the use of PL condition. By assuming the condition number $\lambda/L \leq \frac{1}{n^{1/3}}$ and choose $m = n^{1/3}$ and $\eta = \frac{\rho}{L}$ with $\rho \leq \frac{1}{2}$, the definition of γ yields*

$$\gamma = 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} \geq 1 - \rho \geq \frac{1}{2} \tag{20}$$

According to Theorem 7, equation (20) *implies $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12\sqrt{2}mL}\}$. Hence, choosing $b = m^2$ implying the constant stepsize $\eta \leq \frac{1}{12\sqrt{2}L}$. Note that the assumption $\lambda/L \leq \frac{1}{n^{1/3}}$ on condition number is milder than the assumption $\lambda/L < \frac{1}{\sqrt{n}}$ in [30].*

ZO-PSVRG+ (RandSGE) Under PL Condition

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

Theorem 9. *Let Assumptions 1 and 2 hold, and ZO gradient estimator* (9) *for mix gradient $\hat{v}_k$ is used in Algorithm 1 with stepsize $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{\gamma b}}{12mL\sqrt{d}}\}$ where $\gamma = 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} > 0$. Then*

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

Corollary 10. *Suppose the final iteration point $\tilde{x}^S$ in Algorithm 1 satisfies $\mathbb{E}[F(\tilde{x}^S) - F^*] \leq \epsilon$ under PL condition. Under Assumptions 1 and 2, we let batch size $\mathcal{B} = \min\{\frac{6\sigma^2}{\lambda\epsilon}, n\}$ and the smoothing parameter $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4Ld}$. The number of SZO calls is bounded by*

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

where $s_n = \min\{n, \frac{1}{\lambda\epsilon}\}$. The number of PO calls equals 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 $\eta = \frac{\sqrt{\gamma}}{12L\sqrt{d}}$, the number of SZO calls simplifies to $(S\mathcal{B}d + Smb) = O(\frac{\mathcal{B}d\sqrt{d}}{\lambda\sqrt{\gamma}m}\log\frac{1}{\epsilon} + \frac{b\sqrt{d}}{\lambda\sqrt{\gamma}}\log\frac{1}{\epsilon})$.

Remark 3. *Analysis for ZO-SPIDER-Coord in [1] 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 $b = O(d)$ in Theorem 9, the stepsize η reduces to $O(1)$ with $O(s_n d\log\frac{1}{\epsilon})$ SZO queries.*

Experimental Results

We provide our experimental results in this section. We compare the performance of our ZO-PSVRG+ with 1) ZO-ProxSVRG (based on our improved analysis), 2) ZO-ProxSAGA-Coord [32] and 3) ZO-ProxSGD [21] 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 respectively, ZO-PSVRG+ and ZO-PSVRG+ (RandSGE) with $\mathcal{B} = n$. The learning rates are tuned in the experiments for competitive algorithms according to their convergence guarantees in Table I, and the results shown in this section are based on the best learning rate we obtained for each algorithm. We set stepsize η and smoothing parameter μ for ZO-PSVRG+ and ZO-PSVRG+ (RandSGE) according to the convergence guarantee we derived in lemmas and theorems.

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 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 II. The epoch size is chosen as $m = 30$ in all of our experiments and the minibatch size b is fixed to 50.

TABLE II: Summary of training datasets.

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

In Figure 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 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+ than the counterpart ZO-PSVRG+ (RandSGE). 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 $\mathcal{B} = \lfloor \frac{n}{5} \rfloor$ in ZO-PSVRG+ significantly improves ZO-ProxSVRG with respect to the number of ZO-queries (see Table I), 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 $O(\frac{1}{d})$. Similarly, the large number of function queries to construct coordinate-wise gradient estimates increases significantly the number of SZO queries for ZO-ProxSVRG. On the other hand, ZO-ProxSGD consumes an extremely large number of iterations while exhibiting marginal convergence compared with variance reduced algorithms. Thus, ZO-PSVRG+ obtains the best tradeoffs between the iteration and the function query complexity.

Adversarial Attacks on Black-Box DNNs

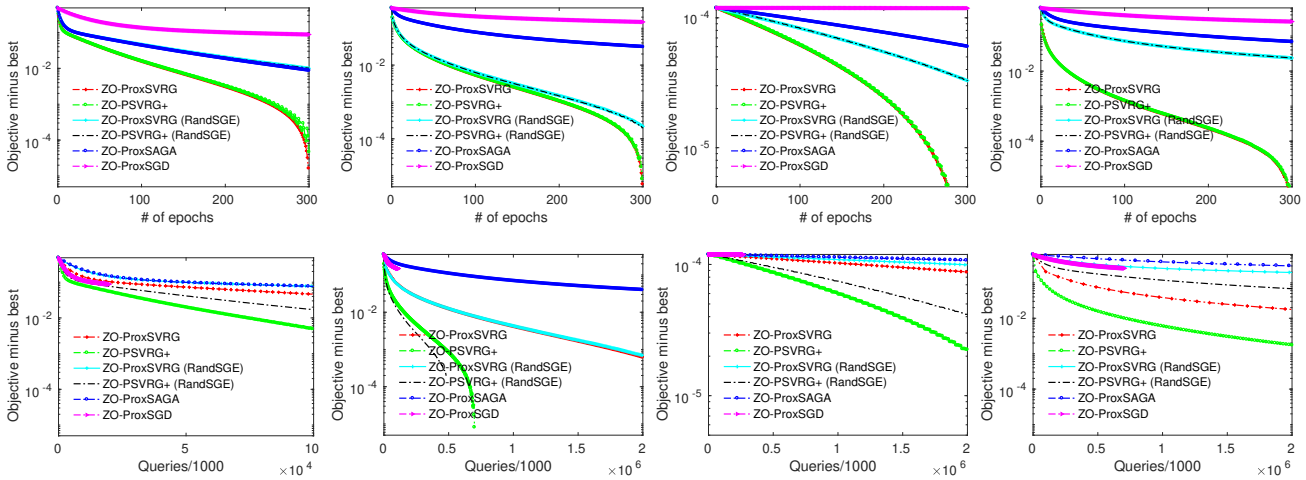
Adversarial examples in image classification are related to designing unperceptive perturbations such that they lead to misclassifying the target model by adding to the natural images. In the framework of zeroth-order attacks [14], [23], the model parameters are hidden and obtaining its gradient is not feasible, while 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 to black-box DNNs problem:

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \max\{F_{y_i}(a_i^{adv}) - \max_{j\neq y_i} F_j(a_i^{adv}), 0\} + c\left\|a_i^{adv} - a_i\right\|^2 + \lambda_1\|x\|_1 + \lambda_2\|x\|^2 \qquad (22)$$

where $a_i^{adv} = 0.5\tanh(\tanh^{-1}(2a_i) + x)$ and λ_1 and λ_2 are nonnegative parameters to obtain consistency between 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 (22) compensates the rate of adversarial success and the distortion of adversarial examples. In our experiment, we set the regularization parameter $c = 0.2$ and $\lambda_1 = \lambda_2 = 10^{-5}$. We perform two experiments by choosing $n = 10$ and $n = 100$ images from the same class, and set the minibatch sizes, respectively $b = 5$ and $b = 30$. We select the batch size $\mathcal{B} = \lfloor \frac{n}{2} \rfloor$ for ZO-PSVRG+. Figure 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+ (RandSGE) shows faster convergence in the initial optimization stage, and more importantly, has much lower function query complexity,

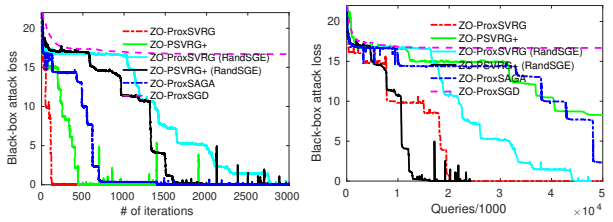
[1]https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html

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



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

Fig. 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)



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

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

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 relatively similar convergence behaviors. Furthermore, the convergence performance of ZO-ProxSGD is poor compared to other algorithms due to not using variance reduced techniques.

Conclusion

In this paper, we developed a novel analysis for two zeroth-order variance-reduced proximal algorithms named ZO-PSVRG+ and ZO-PSVRG+ (RangSGE). We prove that ZO-PSVRG+ improves and generalizes the analysis for several well-known convergence results, e.g., ZO-ProxSVRG. Compared with ZO-SVRG-Coord-Rand [1], our analysis allows single minibatch size and larger stepsizes while improving the function query complexity. Moreover, for nonconvex functions under Polyak-Łojasiewicz condition, we prove that ZO-PSVRG+ obtains global linear convergence rate for a wide range of minibatch sizes without restart. As a byproduct, our analysis provides the first step towards improving the query complexity of ZO methods for nonconvex optimization. Experimental results demonstrate the effectiveness of our novel approaches.

References

[1] K. Ji, Z. Wang, Y. Zhou, and Y. Liang, "Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization," in *International Conference on Machine Learning*, 2019, pp. 3100–3109.

[2] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in neural information processing systems*, 2013, pp. 315–323.

[3] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *International conference on machine learning*, 2016, pp. 314–323.

[4] A. Nitanda, "Accelerated stochastic gradient descent for minimizing finite sums," in *Artificial Intelligence and Statistics*, 2016, pp. 195–203.

[5] Z. Allen-Zhu and Y. Yuan, "Improved svrg for non-strongly-convex or sum-of-non-convex objectives," in *International conference on machine learning*, 2016, pp. 1080–1089.

[6] L. Lei, C. Ju, J. Chen, and M. I. Jordan, "Non-convex finite-sum optimization via scsg methods," in *Advances in Neural Information Processing Systems*, 2017, pp. 2348–2358.

[7] A. Wibisono, M. J. Wainwright, M. I. Jordan, and J. C. Duchi, "Finite sample convergence rates of zero-order stochastic optimization methods," in *Advances in Neural Information Processing Systems*, 2012, pp. 1439–1447.

[8] A. Sokolov, J. Kreutzer, S. Riezler, and C. Lo, "Stochastic structured prediction under bandit feedback," in *Advances in Neural Information Processing Systems*, 2016, pp. 1489–1497.

[9] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[10] R. P. Brent, *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

[11] J. C. Spall, *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2005, vol. 65.

[12] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[13] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. ACM, 2017, pp. 506–519.

[14] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. 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*. ACM, 2017, pp. 15–26.

[15] K. Choromanski, M. Rowland, V. Sindhwani, R. E. Turner, and A. Weller, "Structured evolution with compact architectures for scalable policy optimization," *arXiv preprint arXiv:1804.02395*, 2018.

[16] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin, "Learning structured prediction models: A large margin approach," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 896–903.

[17] T. Chen and G. B. Giannakis, "Bandit convex optimization for scalable and dynamic iot management," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1276–1286, 2019.

[18] S. Liu, J. Chen, P.-Y. Chen, and A. O. Hero, "Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications," *arXiv preprint arXiv:1710.07804*, 2017.

[19] M. C. Fu, "Optimization for simulation: Theory vs. practice," *INFORMS Journal on Computing*, vol. 14, no. 3, pp. 192–215, 2002.

[20] X. Lian, H. Zhang, C.-J. Hsieh, Y. Huang, and J. Liu, "A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order," in *Advances in Neural Information Processing Systems*, 2016, pp. 3054–3062.

[21] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Mathematical Programming*, vol. 156, no. 1-2, pp. 59–99, 2016.

[22] F. Huang, B. Gu, Z. Huo, S. Chen, and H. Huang, "Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization," in *AAAI*, 2019.

[23] S. Liu, J. Chen, P.-Y. Chen, and A. Hero, "Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 288–297.

[24] B. T. Polyak, "Gradient methods for minimizing functionals," *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 3, no. 4, pp. 643–653, 1963.

[25] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[26] A. Agarwal, O. Dekel, and L. Xiao, "Optimal algorithms for online convex optimization with multi-point bandit feedback." in *COLT*. Citeseer, 2010, pp. 28–40.

[27] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

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

[29] L. Liu, M. Cheng, C.-J. Hsieh, and D. Tao, "Stochastic zeroth-order optimization via variance reduction method," *arXiv preprint arXiv:1805.11811*, 2018.

[30] S. J. Reddi, S. Sra, B. Póczos, and A. J. Smola, "Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization," in *Advances in Neural Information Processing Systems*, 2016, pp. 1145–1153.

[31] B. Gu, D. Wang, Z. Huo, and H. Huang, "Inexact proximal gradient methods for non-convex and non-smooth optimization," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[32] B. Gu, Z. Huo, C. Deng, and H. Huang, "Faster derivative-free stochastic algorithm for shared memory machines," in *International Conference on Machine Learning*, 2018, pp. 1807–1816.

[33] K. Balasubramanian and S. Ghadimi, "Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates," in *Advances in Neural Information Processing Systems*, 2018, pp. 3455–3464.

[34] A. K. Sahu, M. Zaheer, and S. Kar, "Towards gradient free and projection free stochastic optimization," *arXiv preprint arXiv:1810.03233*, 2018.

[35] X. Chen, S. Liu, K. Xu, X. Li, X. Lin, M. Hong, and D. Cox, "Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization," in *Advances in Neural Information Processing Systems*, 2019, pp. 7202–7213.

[36] X. Gao, B. Jiang, and S. Zhang, "On the information-adaptive variants of the admm: an iteration complexity perspective," *Journal of Scientific Computing*, vol. 76, no. 1, pp. 327–363, 2018.

[37] A. D. Flaxman, A. T. Kalai, and H. B. 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*. Society for Industrial and Applied Mathematics, 2005, pp. 385–394.

[38] O. Shamir, "An optimal algorithm for bandit and zero-order convex optimization with two-point feedback." *Journal of Machine Learning Research*, vol. 18, no. 52, pp. 1–11, 2017.

[39] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[40] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 2057–2075, 2014.

[41] Z. Li and J. Li, "A simple proximal stochastic gradient method for nonsmooth nonconvex optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 5564–5574.

[42] D. Hajinezhad, M. Hong, and A. Garcia, "Zone: Zeroth order nonconvex multi-agent optimization over networks," *IEEE Transactions on Automatic Control*, 2019.

[43] H. Karimi, J. Nutini, and M. 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*. Springer, 2016, pp. 795–811.