

Efficient Derivative-Free Proximal Stochastic Methods and Analysis for Nonconvex Nonsmooth Optimization

Paper ID: 5869

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 ZO-SVRG type algorithms suffer from function query complexities up to a small-degree polynomial of the problem size. To fill this gap, we design a new zeroth-order stochastic gradient algorithm 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 calls and proximal oracle calls. In particular, ZO-PSVRG+ yields simpler analysis for a wide range of minibatch sizes, while the improved analysis of ZO-SVRG for smooth functions in (Ji et al. 2019) is only achieved for large minibatch sizes based on an involved parameter selection. We further prove that ZO-PSVRG+ under Polyak-Łojasiewicz condition in contrast to the 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.

Introduction

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

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

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.

The major adversity for these accelerated methods 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, while the functional estimations of foreseen frameworks are achievable (Sokolov et al. 2016). As an example, 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. We compared the results from our analysis and other comparable SZO algorithms in Table 1. It 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 our algorithm, ZO-PSVRG+, could

achieve better trade-offs between the convergence rate and the query complexity.

Main contributions

We present a novel analysis beyond the existing convergence studies introduced in (Liu et al. 2018; 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. Particularly, we somewhat address the open question if the dependence on the dimension d for the convergence analysis proposed in (Liu et al. 2018) 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 performed based on the novel structure of recently introduced SZO algorithms. Note that our analysis does not rely on bounded gradient assumption in (Ghadimi and Lan 2016; Huang et al. 2019). 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 paper:

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. ZO-PSVRG+ also matches the best result achieved by ZO-SVRG-Coord-Rand with minibatch size $b = dn^{2/3}$ and epoch size $m = n^{1/3}$ in (Ji et al. 2019), while our results are valid for unspecified minibatch sizes as detailed in the following sections. Indeed, it is necessary to analyze and study the convergence behavior of SZO optimization with minibatches of single or moderate sizes, as practically many machine learning models are trained with intermediate minibatch sizes.

2) The convergence analysis for ZO-PSVRG+ is straightforward in contrast to ZO-SVRG-Coord in (Liu et al. 2018; 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 and it provides the best results to our best knowledge (see Table 1). 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 (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 ZO-ProxSVRG for problem (1) with arbitrary minibatch sizes. This analysis generalizes the results of (Duchi et al. 2015) while shows linear convergence versus the sublinear conver-

gence rate in their paper. In (Ji et al. 2019), the authors show that ZO-SPIDER-Coord achieves linear convergence under PL condition but only for the minibatch size $b = O(n^{1/2})$. Note that due to both computational and statistical efficiency, convergence analysis for minibatches of moderate sizes is demanding. Also see the remarks after Theorem 9 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.

Preliminary

In the following we illustrate and specify 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, Jiang, and Zhang 2018)

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

where d is the number of optimization variables, $\{u_i\}$ are i.i.d. random directions drawn from a uniform distribution over a unit sphere and $\mu > 0$ is the smoothing parameter (Flaxman, Kalai, and McMahan 2005; Shamir 2017; Gao, Jiang, and Zhang 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 function evaluations 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; 2018a; Liu et al. 2018) to evaluate the gradients as:

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

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 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. This results in a larger step-size and a speedier convergence.

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 conducts iterations of the form:

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

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. 2018))	S(NC)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{\epsilon}\right)$	$O\left(\frac{nd^2}{\epsilon} + \frac{d^2b}{\epsilon}\right)$
ZO-SVRG-Coord-Rand ((Ji et al. 2019))	S(NC)	$O\left(\frac{1}{dn^{2/3}}\right)$	$O\left(\frac{dn^{2/3}}{\epsilon}\right)$	$O(\min\{\frac{dn^{2/3}}{\epsilon}, \frac{d}{\epsilon^{2/3}}\})^*$
ZO-ProxSVRG-Coord ((Gu et al. 2018a))	S(NC)+NS(C)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{\epsilon}\right)$	$O\left(\frac{nd^2}{\epsilon\sqrt{b}} + \frac{md^2\sqrt{b}}{\epsilon}\right)$
ZO-ProxSAGA-Coord ((Gu et al. 2018a))	S(NC)+NS(C)	$O\left(\frac{1}{d}\right)$	$O\left(\frac{d}{\epsilon}\right)$	$O\left(\frac{nd^2}{\epsilon\sqrt{b}}\right)$
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)	$O(1)$	$O\left(\log\left(\frac{1}{\epsilon}\right)\right)$	$O(s_n \frac{d}{\lambda} \log \frac{1}{\epsilon} + \frac{bd}{\lambda} \log \frac{1}{\epsilon})$
ZO-PSVRG+ (RandSGE) (Ours)	S(PL)+NS(C)	$O\left(\frac{1}{\sqrt{d}}\right)$	$O\left(\sqrt{d} \log\left(\frac{1}{\epsilon}\right)\right)$	$O(s_n \frac{d\sqrt{d}}{\lambda} \log \frac{1}{\epsilon} + \frac{b\sqrt{d}}{\lambda} \log \frac{1}{\epsilon})$

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. $s_n = \min\{n, \frac{1}{\epsilon}\}$. *: The single-minibatch version.

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

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

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.

ZO Proximal Stochastic Method (ZO-PSVRG+)

Algorithm 1 Zeroth-Order Proximal Stochastic Method

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

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

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

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 (Xiao and Zhang 2014; Reddi et al. 2016b; Li and Li 2018). The description of ZO-PSVRG+ is presented in Algorithm 1. Our method has two types of random sampling. In the outer iteration, we calculate the gradient consisting of \mathcal{B} samples. In the inner iteration, we randomly choose a minibatch of samples of size b to approximate gradient over the minibatch. We call \mathcal{B} and b , batch and minibatch size, respectively. In our ZO framework, the mix gradient (6) is estimated by applying only function evaluations, given by

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

or

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

where $\hat{g}^s = \frac{1}{\mathcal{B}} \sum_{i \in I_{\mathcal{B}}} \hat{\nabla}f_i(\tilde{x}^{s-1})$, $\hat{\nabla}f_i$ is a ZO gradient approximation using CoordSGE and $\hat{\nabla}_r f_i$ is a ZO gradient estimate using RandSGE. We let ZO-PSVRG+ and ZO-PSVRG+ (RandSGE) denote Algorithm 1 with gradient estimation (7) and (8), respectively. Note that, $\mathbb{E}_{I_b}[\hat{v}_{t-1}^s] = \hat{\nabla}f(x_{t-1}^s) \neq \nabla f(x_{t-1}^s)$, i.e., this stochastic gradient is a biased approximation of the actual gradient. In other words, the unbiased assumption on gradient approximates utilized in ProxSVRG (Reddi et al. 2016b; Li and Li 2018) is no longer valid.

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. Nevertheless, our convergence studies yield a novel analysis for ZO-ProxSVRG-Coord (i.e., $\mathcal{B} = n$).

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

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

Assumption 2. For $\forall x \in \mathbb{R}^d$, $\mathbb{E} \left[\|\hat{\nabla} f_i(x) - \hat{\nabla} f(x)\|^2 \right] \leq \sigma^2$, 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. Assumption 2 is weaker than the assumption of bounded gradients (Liu et al. 2017; Hajinezhad, Hong, and Garcia 2019), 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 $\hat{v}_{t-1}^s = \frac{1}{b} \sum_{i \in I_b} (\hat{\nabla} f_i(x_{t-1}^s) - \hat{\nabla} f_i(\tilde{x}^{s-1})) + \hat{g}^s$ with $\hat{g}^s = \frac{1}{B} \sum_{j \in I_B} \hat{\nabla} f_j(\tilde{x}^{s-1})$, then the following inequality holds.

$$\begin{aligned} \mathbb{E} \left[\eta \|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\|^2 \right] &\leq \frac{6\eta L^2}{b} \mathbb{E} \left[\|x_{t-1}^s - \tilde{x}^{s-1}\|^2 \right] \\ &+ 2 \frac{I(\mathcal{B} < n)\eta\sigma^2}{B} + \eta \frac{7L^2 d^2 \mu^2}{2} \end{aligned} \quad (9)$$

Lemma 1 provides an upper bound for the variance of \hat{v}_{t-1}^s . 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.

Below we present the counterpart of Lemma 1 for the mix gradient estimation in (8).

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

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

Analysis for ZO-PSVRG+

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

Theorem 3. Suppose Assumptions 1 and 2 hold, and the ZO gradient estimator (7) for mix gradient \hat{v}_k is used. The output \hat{x} of Algorithm 1 satisfies

$$\mathbb{E}[\|g_\eta(\hat{x})\|^2] \leq \frac{6(F(x_0) - F(x^*))}{\eta S m}$$

$$+ \frac{I(\mathcal{B} < n)12\sigma^2}{B} + 21L^2 d^2 \mu^2$$

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

The proof for Theorem 3 is significantly different from the proofs in the existing literature. 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 parameter setting. 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} for specific parameter settings.

Corollary 4. 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 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

$$\begin{aligned} d(\mathcal{S}\mathcal{B} + Smb) &= 6d(F(x_0) - F(x^*)) \left(\frac{\mathcal{B}}{\epsilon\eta m} + \frac{b}{\epsilon\eta} \right) \\ &= O \left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{bd}{\epsilon\eta} \right) \end{aligned} \quad (11)$$

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

$$\begin{aligned} 72dL(F(x_0) - F(x^*)) &\left(\frac{\mathcal{B}}{\epsilon\sqrt{b}} + \frac{b}{\epsilon} \right) \\ &= O \left(s_n \frac{d}{\epsilon\sqrt{b}} + \frac{bd}{\epsilon} \right) \end{aligned} \quad (12)$$

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

Generally speaking, Corollary 4 indicates that if we select the smoothing parameter μ reasonably small and the batch size \mathcal{B} sufficiently large, then the errors induced 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 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 1. It is worth mentioning that the stepsize η in Theorem 3 is less restrictive than the existing SZO algorithms in Table 1.

Analysis for ZO-PSVRG+ (RandSGE)

Based on Lemma 8, we indicate that ZO-PSVRG+ (RandSGE) achieves improvements both in the convergence rate and the function query complexity compared to the existing SZO methods based on RandSGE, as demonstrated in the subsequent analysis.

Theorem 5. Suppose Assumptions 1 and 2 hold, and the coordinate gradient estimator (8) for mix gradient \hat{v}_k is used. The output \hat{x} of Algorithm 1 satisfies

$$\mathbb{E}[\|g_\eta(\hat{x})\|^2] \leq \frac{6(F(x_0) - F(x^*))}{\eta Sm} + \frac{I(\mathcal{B} < n)12\sigma^2}{\mathcal{B}} + 21L^2d^2\mu^2 \quad (13)$$

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 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 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

$$\begin{aligned} (dS\mathcal{B} + Smb) &= 6(F(x_0) - F(x^*)) \left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta} \right) \\ &= O\left(\frac{\mathcal{B}d}{\epsilon\eta m} + \frac{b}{\epsilon\eta} \right) \end{aligned} \quad (14)$$

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

$$\begin{aligned} 72L(F(x_0) - F(x^*)) \left(\frac{\mathcal{B}d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon} \right) \\ = O\left(s_n \frac{d\sqrt{d}}{\epsilon\sqrt{b}} + \frac{b\sqrt{d}}{\epsilon} \right) \end{aligned} \quad (15)$$

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 (Ji et al. 2019) 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 ProxSVRG+ under Polyak-Łojasiewicz (PL) assumption (Polyak 1963). The classic structure of PL condition is, for all $x \in \mathbb{R}^d$

$$\|\nabla f(x)\|^2 \geq 2\lambda(f(x) - f^*) \quad (16)$$

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 is improved. Due to the presence of the nonsmooth term $h(x)$ in problem (1), we utilize the gradient projection to characterize a more generic form of PL condition as follows,

$$\|g_\eta(x)\|^2 \geq 2\lambda(F(x) - F^*) \quad (17)$$

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 authors in (Karimi, Nutini, and Schmidt 2016) proved that the set of functions satisfying PL condition includes a large class of functions. A zeroth-order algorithm under PL condition for smooth functions has been analyzed in (Ji et al. 2019).

ZO-PSVRG+ Under PL Condition

In the same way as Theorem 3, we show the convergence result of ZO-PSVRG+ (Algorithm 1) under PL-condition.

Theorem 7. Let Assumptions 1 and 2 hold, and ZO gradient estimator (7) 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

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

Theorem 7 shows that if the batch size and smoothing parameter are appropriately chosen, ZO-PSVRG+ has a dominant linear convergence rate without restart. 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 the smoothing parameter $\mu \leq \frac{\sqrt{\lambda\epsilon}}{4Ld}$. The number of SZO calls is bounded by

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

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

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

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

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. 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

$$\begin{aligned}\gamma &= 1 - \frac{2\lambda\eta}{3}m - \frac{\lambda\eta}{3} \\ &\geq 1 - \rho \geq \frac{1}{2}\end{aligned}\quad (19)$$

According to Theorem 7, equation (19) implies $\eta \leq \min\{\frac{1}{8L}, \frac{\sqrt{b}}{12\sqrt{2}mL}\}$. Hence, choosing $b = m^2$ leads to the constant stepsize $\eta \leq \frac{1}{12\sqrt{2}L}$. Note that the assumption $\lambda/L \leq \frac{1}{n^{1/3}}$ on condition number is milder than the assumption $\lambda/L < \frac{1}{\sqrt{n}}$ in (Reddi et al. 2016b).

ZO-PSVRG+ (RandSGE) Under PL Condition

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

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

$$\begin{aligned}\mathbb{E}[F(\bar{x}^S) - F^*] &\leq \left(1 - \frac{\lambda\eta}{3}\right)^{Sm} \mathbb{E}[F(\bar{x}^0) - F^*] \\ &\quad + \frac{6I(\mathcal{B} < n)\sigma^2}{\lambda\mathcal{B}} + \frac{21L^2d^2\mu^2}{2\lambda}\end{aligned}\quad (20)$$

Remark 3. *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 $b = O(d)$ 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 (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 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 1, and the results shown in this section are based on the best learning rate we achieved for each algorithm. We set stepsize η and smoothing parameter μ for ZO-PSVRG+ and ZO-PSVRG+ (RandSGE) according to the convergence guarantee we obtained 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 website¹, as listed in Table 2. The epoch size is chosen as $m = 30$ in all of our experiments and the minibatch size b is fixed to 50.

Table 2: 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 1), 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, by adding to the natural images, lead to misclassifying the target model. In the framework of zeroth-order attacks (Chen et al. 2017;

¹<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html>

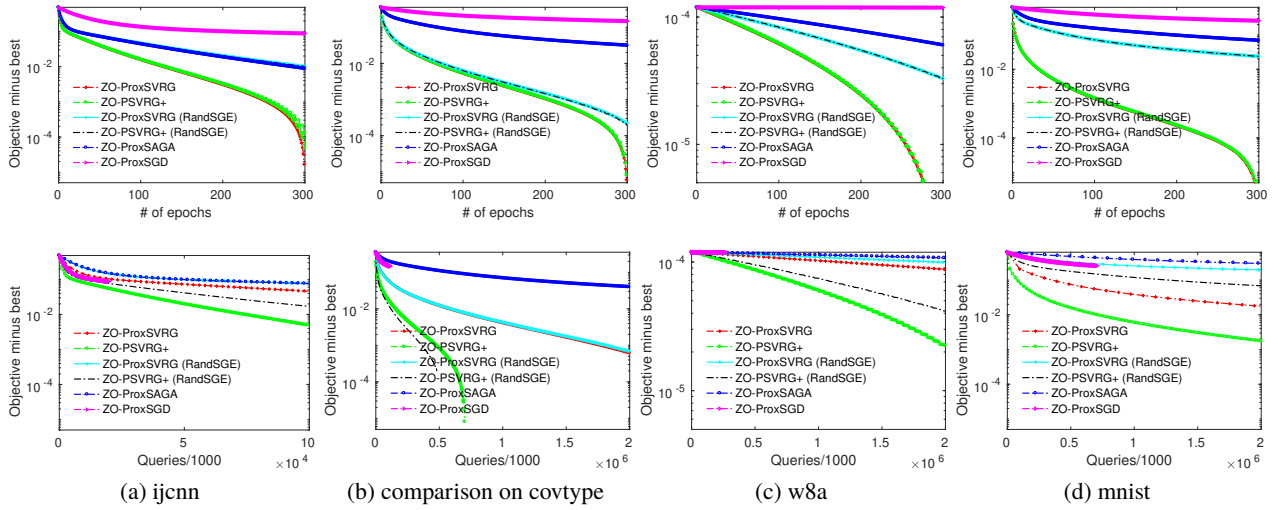
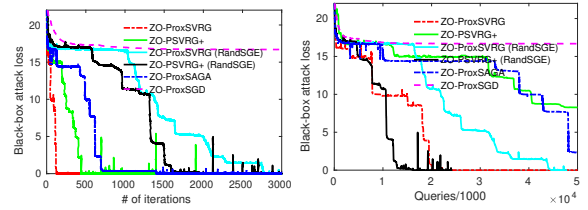


Figure 1: Comparison of different zeroth-order algorithms for logistic regression loss residual $f(x) - f(x^*)$ versus the number of epochs (top) and ZO queries (bottom)

Liu et al. 2018), the model parameters are unexposed 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:

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \max\{F_{y_i}(a_i^{adv}) - \max_{j \neq y_i} F_j(a_i^{adv}), 0\} + c \|a_i^{adv} - a_i\|^2 + \lambda_1 \|x\|_1 + \lambda_2 \|x\|^2 \quad (21)$$

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 $F(a) = [F_1(a), \dots, F_K(a)] \in [0, 1]^K$ describes a trained DNN² for the MNIST handwritten digit classification, where $F_i(a)$ returns the prediction score of i -th class. The parameter c in (21) 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)



(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

shows faster convergence in the initial optimization stage, and more importantly, has much lower function query complexity, which is largely due to efficient ZO queries for computing mix gradient (8) and the $O(\frac{1}{\sqrt{d}})$ -level stepsize required by ZO-PSVRG+ (RandSGE). ZO-ProxSAGA and ZO-PSVRG+ (CoordSGE) exhibit relatively similar convergence behaviors. Furthermore, the convergence performance of ZO-ProxSGD is poor compared to 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+ (RandSGE). 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 (Ji et al. 2019), our analysis allows single minibatch size and larger stepsizes while improving the function query complexity. Moreover, for non-convex functions under Polyak-Łojasiewicz condition, we prove that ZO-PSVRG+ obtains global linear convergence rate for a wide range of minibatch sizes without restart. Experimental results demonstrate the effectiveness of our novel approaches.

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

References

- Allen-Zhu, Z., and Yuan, Y. 2016. Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In *International conference on machine learning*, 1080–1089.
- Brent, R. P. 2013. *Algorithms for minimization without derivatives*. Courier Corporation.
- Chen, P.-Y.; Zhang, H.; Sharma, Y.; Yi, J.; and Hsieh, C.-J. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 15–26. ACM.
- Duchi, J. C.; Jordan, M. I.; Wainwright, M. J.; and Wibisono, A. 2015. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory* 61(5):2788–2806.
- Flaxman, A. D.; Kalai, A. T.; and McMahan, H. B. 2005. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 385–394. Society for Industrial and Applied Mathematics.
- Gao, X.; Jiang, B.; and Zhang, S. 2018. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing* 76(1):327–363.
- Ghadimi, S., and Lan, G. 2016. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming* 156(1-2):59–99.
- Gu, B.; Huo, Z.; Deng, C.; and Huang, H. 2018a. Faster derivative-free stochastic algorithm for shared memory machines. In *International Conference on Machine Learning*, 1807–1816.
- Gu, B.; Wang, D.; Huo, Z.; and Huang, H. 2018b. Inexact proximal gradient methods for non-convex and non-smooth optimization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Hajinezhad, D.; Hong, M.; and Garcia, A. 2019. Zone: Zeroth order nonconvex multi-agent optimization over networks. *IEEE Transactions on Automatic Control*.
- Huang, F.; Gu, B.; Huo, Z.; Chen, S.; and Huang, H. 2019. Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. In *AAAI*.
- Ji, K.; Wang, Z.; Zhou, Y.; and Liang, Y. 2019. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, 3100–3109.
- Johnson, R., and Zhang, T. 2013. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, 315–323.
- Karimi, H.; Nutini, J.; and Schmidt, M. 2016. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 795–811. Springer.
- Lei, L.; Ju, C.; Chen, J.; and Jordan, M. I. 2017. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, 2348–2358.
- Li, Z., and Li, J. 2018. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, 5564–5574.
- Lian, X.; Zhang, H.; Hsieh, C.-J.; Huang, Y.; and Liu, J. 2016. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. In *Advances in Neural Information Processing Systems*, 3054–3062.
- Liu, S.; Chen, J.; Chen, P.-Y.; and Hero, A. O. 2017. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. *arXiv preprint arXiv:1710.07804*.
- Liu, S.; Kailkhura, B.; Chen, P.-Y.; Ting, P.; Chang, S.; and Amini, L. 2018. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 3727–3737.
- Nesterov, Y., and Spokoiny, V. 2017. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics* 17(2):527–566.
- Nitanda, A. 2016. Accelerated stochastic gradient descent for minimizing finite sums. In *Artificial Intelligence and Statistics*, 195–203.
- Polyak, B. T. 1963. Gradient methods for minimizing functionals. *Zhurnal Vychislitel’noi Matematiki i Matematicheskoi Fiziki* 3(4):643–653.
- Reddi, S. J.; Hefny, A.; Sra, S.; Póczos, B.; and Smola, A. 2016a. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, 314–323.
- Reddi, S. J.; Sra, S.; Póczos, B.; and Smola, A. J. 2016b. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, 1145–1153.
- Shamir, O. 2017. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research* 18(52):1–11.
- Sokolov, A.; Kreutzer, J.; Riezler, S.; and Lo, C. 2016. Stochastic structured prediction under bandit feedback. In *Advances in Neural Information Processing Systems*, 1489–1497.
- Spall, J. C. 2005. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons.
- Wibisono, A.; Wainwright, M. J.; Jordan, M. I.; and Duchi, J. C. 2012. Finite sample convergence rates of zero-order stochastic optimization methods. In *Advances in Neural Information Processing Systems*, 1439–1447.
- Xiao, L., and Zhang, T. 2014. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization* 24(4):2057–2075.