Improved Zeroth-Order Variance Reduced Algorithms and Analysis for Nonconvex Optimization

Kaiyi Ji [1] **Zhe Wang** [1] **Yi Zhou** [2] **Yingbin Liang** [1]

Abstract

Two types of zeroth-order stochastic algorithms have recently been designed for nonconvex optimization respectively based on the first-order techniques SVRG and SARAH/SPIDER. This paper addresses several important issues that are still open in these methods. First, all existing SVRG-type zeroth-order algorithms suffer from worse function query complexities than either zeroth-order gradient descent (ZO-GD) or stochastic gradient descent (ZO-SGD). In this paper, we propose a new algorithm ZO-SVRG-Coord-Rand and develop a new analysis for an existing ZO-SVRG-Coord algorithm proposed in Liu et al. 2018b, and show that both ZO-SVRG-Coord-Rand and ZO-SVRG-Coord (under our new analysis) outperform other exiting SVRG-type zeroth-order methods as well as ZO-GD and ZO-SGD. Second, the existing SPIDER-type algorithm SPIDER-SZO (Fang et al., 2018) has superior theoretical performance, but suffers from the generation of a large number of Gaussian random variables as well as a $\sqrt{\epsilon}$-level stepsize in practice. In this paper, we develop a new algorithm ZO-SPIDER-Coord, which is free from Gaussian variable generation and allows a large constant stepsize while maintaining the same convergence rate and query complexity, and we further show that ZO-SPIDER-Coord automatically achieves a linear convergence rate as the iterate enters into a local PL region without restart and algorithmic modification.

1. Introduction

Zeroth-order optimization has recently gained increasing attention due to its wide usage in many applications where the explicit expressions of gradients of the objective function are expensive or infeasible to obtain and only function evaluations are accessible. Such a class of applications include black-box adversarial attacks on deep neural networks (DNNs) (Papernot et al., 2017; Chen et al., 2017; Kurakin et al., 2016), structured prediction (Taskar et al., 2005) and reinforcement learning (Choromanski et al., 2018).

Various zeroth-order algorithms have been developed to solve the following general finite-sum optimization problem

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

where d denotes the input dimension and $\{f_i(\cdot)\}_{i=1}^n$ denote smooth and nonconvex individual loss functions. Nesterov & Spokoiny 2011 introduced a zeroth-order gradient descent (ZO-GD) algorithm using a two-point Gaussian random gradient estimator, which yields a convergence rate of $\mathcal{O}(d/K)$ (where K is the number of iterations) and a function query complexity (i.e., the number of queried function values) of $\mathcal{O}(dn/\epsilon)$, to attain a stationary point $\mathbf{x}^\zeta$ such that $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$. Ghadimi & Lan 2013 proposed a zeroth-order stochastic gradient descent (ZO-SGD) algorithm using the same gradient estimation technique as in Nesterov & Spokoiny 2011, which has a convergence rate of $\mathcal{O}(\sqrt{d/K})$ and a function query complexity of $\mathcal{O}(d/\epsilon^2)$.

Furthermore, two types of zeroth-order stochastic variance reduced algorithms have been developed to further improve the convergence rate of ZO-SGD. The first type refers to the SVRG-based algorithm, which replaces the gradient in SVRG (Johnson & Zhang, 2013) by zeroth-order gradient estimators. In particular, Liu et al. 2018b proposed three zeroth-order SVRG-based algorithms, namely, ZO-SVRG based on a two-point random gradient estimator, ZO-SVRG-Ave based on an average random gradient estimator, and ZO-SVRG-Coord based on a coordinate-wise gradient estimator. The performances of the aforementioned algorithms are summarized in Table 1. Though existing studies appear comprehensive, two important questions are still left open

[1] Department of Electrical and Computer Engineering, The Ohio State University [2] Department of Electrical and Computer Engineering, Duke University. Correspondence to: Kaiyi Ji <ji.367@osu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

and require conclusive answers.

Q1.1 Although the existing zeroth-order SVRG-based algorithms have improved iteration rate of convergence (i.e., the dependence on K), their function query complexities are all larger than either ZO-GD or ZO-SGD. Whether there exist zeroth-order SVRG-based algorithms that outperform ZO-GD and ZO-SGD in terms of both the function query complexity and the convergence rate is an intriguing open question.

Q1.2 As shown in Liu et al. 2018b (see Table 1), ZO-SVRG-Coord suffers from approximately $\mathcal{O}(d)$ time more function queries than ZO-SVRG and ZO-SVRG-Ave. However, such inferior performance may be due to bounding technicality rather than algorithm itself. Intuitively, coordinate-wise estimator used in ZO-SVRG-Coord can estimate the gradient more accurately, and hence should require fewer iterations to convergence, so that its overall complexity can be comparable or superior than ZO-SVRG and ZO-SVRG-Ave. Thus, a refined convergence analysis is needed.

The second type of zeroth-order variance-reduced algorithms was proposed in Fang et al. 2018, named SPIDER-SZO, which replaces gradients in the SPIDER algorithm with zeroth-order gradient estimators. Differently from SVRG, SPIDER (Fang et al., 2018) and an earlier version SARAH (Nguyen et al., 2017a;b) are first-order stochastic variance-reduced algorithms whose inner-loop iterations *recursively* incorporate the fresh gradients to update the gradient estimator (see (10)). Fang et al. 2018 showed that SPIDER-SZO achieves an improved query complexity over SVRG-based zeroth-order algorithms. However, SPIDER-SZO requires the generation of a large number $\mathcal{O}(n^{1/2}d^2)$ of i.i.d. Gaussian random variables at *each* inner-loop iteration, and requires a very small stepsize $\eta = \mathcal{O}(\sqrt{\epsilon}/(\|\mathbf{v}^k\|L))$ (where $\mathbf{v}^k$ is an estimate of gradient $\nabla f(\mathbf{x}^k)$) to guarantee the convergence. Such two requirements can substantially restrict the performance of SPIDER-SZO in practice. Thus, the following two important questions arise.

Q2.1 Whether using coordinate-wise estimator for both inner and outer loops and at the same time enlarging the stepsize to the constant level provide competitive query complexity? If so, such a new zeroth-order SPIDER-based algorithm eliminates the aforementioned two restrictive requirements in SPIDER-SZO.

Q2.2 The existing study of zeroth-order SPIDER-based algorithms is only for smooth nonconvex optimization, which is far from comprehensive. We further want to understand their performance under specific geometries such as the Polyak-Łojasiewicz (PL) condition, convexity and for nonconvex nonsmooth composite optimization. Can SPIDER-based algorithms still outperform other existing zeroth-order algorithms for these cases?

In this paper, we provide comprehensive answers to the above questions.

1.1. Summary of Contributions

For SVRG-based algorithms, we provide affirmative answers to the questions Q1.1 and Q1.2. First, we propose a new zeroth-order SVRG-based algorithm ZO-SVRG-Coord-Rand and show that it achieves the function query complexity of $\mathcal{O}\big(\min\big\{dn^{2/3}\epsilon^{-1},\, d\epsilon^{-5/3}\big\}\big)$ for nonconvex optimization, which order-wisely improves the performance of not only all existing zeroth-order SVRG-based algorithms (see Table 1) but also ZO-GD and ZO-SGD. This for the first time establishes the order-wise complexity advantage of zeroth-order SVRG-based algorithms over the zeroth-order GD and SGD-based algorithms, and thus answers Q1.1. Furthermore, we provide a new convergence and complexity analysis for ZO-SVRG-Coord (Liu et al., 2018b) with order-wise tighter bound, and show that it achieves the same fantastic function query complexity as ZO-SVRG-Coord-Rand, which answers Q1.2. Furthermore, our new analysis allows a much larger stepsize for performance guarantee.

For SPIDER-based algorithms, we provide affirmative answers to the questions Q2.1 and Q2.2. To answer Q2.1, we first propose a novel zeroth-order algorithm ZO-SPIDER-Coord fully using coordinate-wise gradient estimators, and show that it achieves the same superior function query complexity as SPIDER-SZO (Fang et al., 2018). ZO-SPIDER-Coord is advantageous over SPIDER-SZO (Fang et al., 2018) by fully eliminating the cost of Gaussian random variable generation and allowing a much larger stepsize $\eta = \mathcal{O}(1)$ to enable a faster convergence in practice. Such two advantages are both due to a new convergence analysis we develop for ZO-SPIDER-Coord. To answer Q2.2, under the PL condition, we show that ZO-SPIDER-Coord achieves a linear convergence rate *without restart and algorithmic modification*. As a result, ZO-SPIDER-Coord automatically achieves a much faster convergence rate when the iterate enters a local region where the PL condition is satisfied.

Due to the space limitations, we relegate our results on zeroth-order *nonconvex nonsmooth* composite optimization and zeroth-order *convex* optimization to the supplementary materials, both of which outperform the corresponding existing algorithms with order-level improvement.

Our analysis reveals that for zero-order variance-reduced algorithms, although the coordinate-wise gradient estimator requires more queries than the two-point gradient estimator, it guarantees much higher estimation accuracy, which leads to a larger stepsize and a faster convergence rate.

1.2. Related Work

Zeroth-order convex optimization. Nemirovsky & Yudin 1983 first proposed a one-point random sampling scheme to

Table 1. Comparison of zeroth-order SVRG-based algorithms in terms of convergence rate and function query complexity for nonconvex optimization. $^{\clubsuit}$: ZO-SVRG, ZO-SVRG-Ave and ZO-SVRG-Coord in Liu et al. 2018b have no single-sample versions. $^{\spadesuit}$: p denotes the number of i.i.d. smoothing vectors for constructing the average random gradient estimator. $^{\bigstar}$: batch size $|\mathcal{S}_1| = \min\left\{n, \lceil (K/d)^{3/5} \rceil\right\}$.

Algorithms	Stepsize η	Convergence rate	Function query complexity
ZO-GD (Nesterov & Spokoiny, 2011)	$\mathcal{O}\left(\frac{1}{d}\right)$	$\mathcal{O}\left(\frac{d}{K}\right)$	$\mathcal{O}\left(\frac{dn}{\epsilon}\right)$
ZO-SGD (Ghadimi & Lan, 2013)	$\mathcal{O}\left(\frac{1}{d}\right)$	$\mathcal{O}\left(\sqrt{\frac{d}{K}}\right)$	$\mathcal{O}\left(\frac{d}{\epsilon^2}\right)$
ZO-SVRG (mini-batch) (Liu et al., 2018b)$^{\clubsuit}$	$\mathcal{O}\left(\frac{1}{d}\right)$	$\mathcal{O}\left(\frac{d}{K} + \frac{1}{\lvert\mathcal{S}_2\rvert}\right)$	$\mathcal{O}\left(\frac{n}{\epsilon} + \frac{d}{\epsilon^2}\right)$
ZO-SVRG-Ave (mini-batch) (Liu et al., 2018b)	$\mathcal{O}\left(\frac{1}{d}\right)$	$\mathcal{O}\left(\frac{d}{K} + \frac{1}{\lvert\mathcal{S}_2\rvert \min\{d,p\}}\right)$	$\mathcal{O}\left(\frac{pn}{\epsilon} + \max\left\{1, \frac{p}{d}\right\}\frac{d}{\epsilon^2}\right)^{\spadesuit}$
ZO-SVRG-Coord (mini-batch) (Liu et al., 2018b)	$\mathcal{O}\left(\frac{1}{d}\right)$	$\mathcal{O}\left(\frac{d}{K}\right)$	$\mathcal{O}\left(dn + \frac{d^2}{\epsilon} + \frac{dn}{\epsilon}\right)$
ZO-SVRG-Coord (mini-batch) (our new analysis)	$\mathcal{O}(1)$	$\mathcal{O}\left(\frac{1}{K}\right)$	$\mathcal{O}\left(\min\left\{\frac{dn^{2/3}}{\epsilon}, \frac{d}{\epsilon^{5/3}}\right\}\right)$
ZO-SVRG-Coord-Rand (mini-batch)	$\mathcal{O}(1)$	$\mathcal{O}\left(\frac{1}{K}\right)$	$\mathcal{O}\left(\min\left\{\frac{dn^{2/3}}{\epsilon}, \frac{d}{\epsilon^{5/3}}\right\}\right)$
ZO-SVRG-Coord-Rand (single-sample)	$\mathcal{O}\left(\frac{1}{d\lvert\mathcal{S}_1\rvert^{2/3}}\right)^{\bigstar}$	$\mathcal{O}\left(\frac{d\lvert\mathcal{S}_1\rvert^{2/3}}{K}\right)$	$\mathcal{O}\left(\min\left\{\frac{dn^{2/3}}{\epsilon}, \frac{d}{\epsilon^{5/3}}\right\}\right)$

Table 2. Comparison of zeroth-order SPIDER-based algorithms in terms of function query complexity and Gaussian sample complexity for nonconvex optimization. $^{\clubsuit}$: SPIDER-SZO in Fang et al. 2018 has no single-sample version. $^{\spadesuit}$: Gaussian sample complexity refers to the total number of generated Gaussian random samples for constructing gradient estimators. $^{\bigstar}$: The epoch length $q = \min\left\{n, \lceil K^{2/3} \rceil\right\}$.

Algorithms	Stepsize η	Function query complexity	Gaussian sample complexity$^{\spadesuit}$
SPIDER-SZO (mini-batch) (Fang et al., 2018)$^{\clubsuit}$	$\mathcal{O}(\sqrt{\epsilon})$	$\mathcal{O}\left(\min\left\{\frac{dn^{1/2}}{\epsilon}, \frac{d}{\epsilon^{3/2}}\right\}\right)$	$\mathcal{O}\left(\frac{d^2 n^{1/2}}{\epsilon}\right)$
ZO-SPIDER-Coord (mini-batch)	$\mathcal{O}(1)$	$\mathcal{O}\left(\min\left\{\frac{dn^{1/2}}{\epsilon}, \frac{d}{\epsilon^{3/2}}\right\}\right)$	None
ZO-SPIDER-Coord (single-sample)	$\mathcal{O}\left(\frac{1}{\sqrt{q}}\right)^{\bigstar}$	$\mathcal{O}\left(\min\left\{\frac{dn^{1/2}}{\epsilon}, \frac{d}{\epsilon^{3/2}}\right\}\right)$	None

estimate the gradient $\nabla f(\mathbf{x})$ by querying $f(\cdot)$ at a random location close to $\mathbf{x}$. Such a technique was then used in many other areas, e.g., bandit optimization (Flaxman et al., 2005; Shamir, 2013) . Multi-point gradient estimation approach was then proposed by Agarwal et al. 2010; Nesterov & Spokoiny 2011, and further explored in Wainwright et al. 2008; Duchi et al. 2015; Ghadimi & Lan 2013; Wang et al. 2017. For example, based on a two-point Gaussian gradient estimator, Ghadimi & Lan 2013 developed a ZO-SGD type of method and Balasubramanian & Ghadimi 2018 proposed a zeroth-order conditional gradient type of algorithm.

Zeroth-order nonconvex optimization. Ghadimi & Lan 2013 and Nesterov & Spokoiny 2011 proposed ZO-GD and its stochastic counterpart ZO-SGD, respectively. In Lian et al. 2016, an asynchronous zeroth-order stochastic gradient (ASZO) algorithm was proposed for parallel optimization. Gu et al. 2018 further improved the convergence rate of ASZO by combining SVRG technique with coordinate-wise gradient estimators. Liu et al. 2018a proposed a stochastic zeroth-order method with variance reduction under Gaussian smoothing. More recently, Liu et al. 2018b provided a comprehensive analysis on SVRG-based zeroth-order algorithms under three different gradient estimators. Fang et al. 2018 further proposed a SPIDER-based zeroth-order method named SPIDER-SZO. Our study falls into this category, where we propose new algorithms that improve the performance of existing algorithms and develop new complexity bounds that improve existing analysis.

Stochastic first-order algorithms. Since stochastic zeroth-order algorithms have been developed based on various first-order algorithms, we briefly summarizes some of them, which include but not limited to SGD (Robbins & Monro, 1951), SAG (Roux et al., 2012), SAGA (Defazio et al., 2014), SVRG (Johnson & Zhang, 2013; Allen-Zhu & Hazan, 2016), SARAH (Nguyen et al., 2017a;b), SNVRG (Zhou et al., 2018), SPIDER (Fang et al., 2018) and Spider-Boost (Wang et al., 2018). If the nonconvex objective function further satisfies the PL condition, Reddi et al. 2016a;b proved the linear convergence for SVRG and its proximal version named ProxSVRG by incorporating a restart step. Li & Li 2018 proposed ProxSVRG+ as an improved version of ProxSVRG and proved its linear convergence without restart. This paper studies a zeroth-order SPIDER-based algorithm under the PL condition without restart.

Notations. We use $\mathcal{O}(\cdot)$ to hide absolute constants that are independent of problem parameters, and $\|\cdot\|$ to denote the Euclidean norm of a vector or the spectral norm of a matrix. We use $[n]$ to denote the set $\{1, 2, ..., n\}$, $|\mathcal{S}|$ to denote the cardinality of a given set $\mathcal{S}$, and $\mathbf{e}_i$ to denote the vector that has only one non-zero entry 1 at its i^{th} coordinate. Given a set $\mathcal{S}$ whose elements are drawn from $[n]$, define $f_{\mathcal{S}}(\cdot) := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} f_i(\cdot)$ and $\nabla f_{\mathcal{S}}(\cdot) := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \nabla f_i(\cdot)$.

2. SVRG-based Zeroth-order Algorithms for Nonconvex Optimization

In this section, we first propose a novel zeroth-order stochastic algorithm named ZO-SVRG-Coord-Rand, and analyze its convergence and complexity performance. We then provide an improved analysis for the existing ZO-SVRG-Coord algorithm proposed by Liu et al. 2018b.

2.1. ZO-SVRG-Coord-Rand Algorithm

We propose a new SVRG-based zeroth-order algorithm ZO-SVRG-Coord-Rand in Algorithm 1, which is conducted in a multi-epoch way. At the beginning of each epoch (i.e., each outer-loop iteration), we estimate the gradient $\nabla f_{\mathcal{S}_1}(\mathbf{x}^k)$ over a batch set $\mathcal{S}_1$ of data samples based on a deterministic coordinate-wise gradient estimator $\hat{\nabla}_{\text{coord}} f_{\mathcal{S}_1}(\mathbf{x}^k) = \sum_{i=1}^{d} \frac{(f_{\mathcal{S}_1}(\mathbf{x}^k+\delta\mathbf{e}_i)-f_{\mathcal{S}_1}(\mathbf{x}^k-\delta\mathbf{e}_i))\mathbf{e}_i}{2\delta}$. In the following inner-loop iterations, we construct the stochastic gradient estimator $\mathbf{v}^k$ based on a mini-batch $\mathcal{S}_2$ of data samples as

$$\begin{aligned}\mathbf{v}^k =& \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\left(\hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x}^k; \mathbf{u}_j^k) - \hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x}^{qk_0}; \mathbf{u}_j^k)\right)\\ &+ \hat{\nabla}_{\text{coord}} f_{\mathcal{S}_1}(\mathbf{x}^{qk_0}),\end{aligned} \tag{2}$$

where $\hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x}; \mathbf{u}_j^k) = \frac{d(f_{a_j}(\mathbf{x}+\beta\mathbf{u}_j^k)-f_{a_j}(\mathbf{x}))}{\beta}\mathbf{u}_j^k$ is a two-point random gradient estimate of $\nabla f_{a_j}(\mathbf{x})$ using a smoothing vector $\mathbf{u}_j^k$ and $k_0 = \lfloor k/q \rfloor$. The above construction of $\mathbf{v}^k$ is the core of our Algorithm 1, which isdifferent from the following estimator in ZO-SVRG (Liu et al., 2018b)

$$\begin{aligned}\mathbf{v}^k =& \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\left(\hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x}^k; \mathbf{u}^k) - \hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x}^{qk_0}; \mathbf{u}^{qk_0})\right)\\ &+ \hat{\nabla}_{\text{rand}} f(\mathbf{x}^{qk_0}; \mathbf{u}^{qk_0}),\end{aligned}$$

where $\mathbf{u}^k$ is generated from the uniform distribution over the unit sphere at the k^{th} iteration.

There are two key differences between our construction of $\mathbf{v}^k$ and the one in Liu et al. 2018b. First, our construction of $\mathbf{v}^k$ introduces $|\mathcal{S}_2|$ i.i.d. smoothing vectors $\{\mathbf{u}_j^k\}_{j=1}^{|\mathcal{S}_2|}$ in each inner-loop iteration to estimate both $\nabla f_{\mathcal{S}_2}(\mathbf{x}^k)$ and $\nabla f_{\mathcal{S}_2}(\mathbf{x}^{qk_0})$, whereas Liu et al. 2018b uses a single smoothing vector $\mathbf{u}^k$ to estimate $\nabla f_{\mathcal{S}_2}(\mathbf{x}^k)$ and a single vector $\mathbf{u}^{qk_0}$ to estimate $\nabla f_{\mathcal{S}_2}(\mathbf{x}^{qk_0})$. Second, we adopt a coordinate-wise gradient estimator in each outer-loop iteration, whereas Liu et al. 2018b use a two-point random gradient estimator. As shown in the next subsection, our treatment does not introduce extra function query cost but achieves a much tighter estimation of $\nabla f(\mathbf{x}^k)$ by $\mathbf{v}^k$.

Algorithm 1 ZO-SVRG-Coord-Rand

```
Input: $q$, $K = qh, h \in \mathbb{N}$, $|\mathcal{S}_1|, |\mathcal{S}_2|, \mathbf{x}^0, \delta, \beta > 0, \eta$
for $k = 0$ to $K$ do
    if $k \mod q = 0$ then
        Sample $\mathcal{S}_1$ from $[n]$ without replacement
        Compute $\mathbf{v}^k = \hat{\nabla}_{\text{coord}} f_{\mathcal{S}_1}(\mathbf{x}^k)$
    else
        Sample $\mathcal{S}_2 = \{a_1, a_2, ..., a_{|\mathcal{S}_2|}\}$ from $[n]$ with replacement
        Draw i.i.d. $\mathbf{u}_1^k, ..., \mathbf{u}_{|\mathcal{S}_2|}^k$ from uniform distribution over unit sphere
        Compute $\mathbf{v}^k$ according to (2)
    end if
    $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta\mathbf{v}^k$
end for
Output: $\mathbf{x}_\zeta$ from $\{\mathbf{x}_0, ..., \mathbf{x}_K\}$ uniformly at random
```

2.2. Complexity and Convergence Analysis

Throughout this paper, we adopt the following standard assumption for the objective function (Nesterov & Spokoiny, 2011; Lian et al., 2016; Gu et al., 2018; Liu et al., 2018b).

Assumption 1. *We assume that $f(\cdot)$ in (1) satisfies:*

(1) $0 < f(\mathbf{x}^0) - f(\mathbf{x}^*) < \infty$*, where* $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$.

(2) *Each $f_i(\cdot), i = 1, ..., n$ has a L-Lipschitz gradient, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|$.*

(3) *Assume that stochastic gradient $\nabla f_i(\cdot)$ has bounded variance, i.e., there exists a constant $\sigma > 0$ such that $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \sigma^2$.*

The item (3) of the variance boundedness assumption is only needed for the online case with $|S_1| < n$. For the finite-sum case (i.e., $|S_1| = n$), the the variance boundedness assumption is not needed.

The following lemma provides a tighter upper bound on the estimation variance $\mathbb{E}\|\mathbf{v}^k - \nabla f_\beta(\mathbf{x}^k)\|^2$.

Lemma 1. *Under Assumption 1, we have, for any $qk_0 \le k \le \min\{q(k_0+1)-1, qh\}$, $k_0 = 0, ..., h$,*

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^k - \nabla f_\beta(\mathbf{x}^k)\|^2 \le& \frac{6dL^2\|\mathbf{x}^k - \mathbf{x}^{qk_0}\|^2}{|\mathcal{S}_2|} + \frac{3L^2\beta^2d^2}{|\mathcal{S}_2|}\\ &+ \frac{18I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2d\delta^2 + \sigma^2\right) + 6L^2d\delta^2 + \frac{3\beta^2L^2d^2}{2}\end{aligned}$$

where $f_\beta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}}\left(f(\mathbf{x}+\beta\mathbf{u})\right)$ *with* $\mathbf{u}$ *drawn from the uniform distribution over the* d*-dimensional unit Euclidean ball, and* $I(A) = 1$ *if the event* A *occurs and* 0 *otherwise.*

The bound in Lemma 1 improves that in Proposition 1 of ZO-SVRG (Liu et al., 2018b) by eliminating its two additional error terms $\mathcal{O}(d\,\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2)$ and $\mathcal{O}(d/|\mathcal{S}_2|)$. Such an improvement is due to our development of a novel and tight inequality $\mathbb{E}_k\left\|\hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x}^k;\mathbf{u}_j^k) - \hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x}^{qk_0};\mathbf{u}_j^k)\right\|^2 \le 3d\|\nabla f_{a_j}(\mathbf{x}^k) - \nabla f_{a_j}(\mathbf{x}^{qk_0})\|^2 + \frac{3L^2d^2\beta^2}{2} \le 3dL^2\|\mathbf{x}^k - \mathbf{x}^{qk_0}\|^2 + \frac{3L^2d^2\beta^2}{2}$ (See Lemma 5 in the supplementary materials), which can be of independent interest for analyzing other zeroth-order methods. Based on Lemma 1, we show that ZO-SVRG-Coord-Rand algorithm achieves significant improvements both in the convergence rate and the function query complexity, as shown in the subsequent analysis.

Theorem 1. *Let Assumptions 1 hold, and define*

$$\lambda = \frac{\eta}{4} - \frac{4c\eta}{g} - \frac{3L\eta^2}{2}, \rho = \left(6\eta^2 L + \frac{c\eta}{g}\right) L^2 d^2 \beta^2,$$

$$\chi = \beta^2 L^2 d^2 + \frac{9I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + 3L^2 d\delta^2,$$

$$\tau = \left(\frac{\eta}{2} + \frac{2c\eta}{g} + 4c\eta^2 + 3L\eta^2\right)\chi + \rho, \tag{3}$$

where g *is a positive parameter and* c *is a constant such that*

$$0 < c = \frac{9dL^3\eta^2}{|\mathcal{S}_2|}\frac{\left(1+\eta g + 12\eta^2 dL^2/|\mathcal{S}_2|\right)^q - 1}{\eta g + 12\eta^2 dL^2/|\mathcal{S}_2|}. \tag{4}$$

Then, the output $\mathbf{x}^\zeta$ *of Algorithm 1 satisfies*

$$\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \le \frac{f_\beta(\mathbf{x}^0) - f_\beta(\mathbf{x}^*_\beta)}{\lambda}\frac{1}{K+1} + \frac{\tau}{\lambda} \tag{5}$$

where $\mathbf{x}^*_\beta = \arg\min_{\mathbf{x}} f_\beta(\mathbf{x})$.

Compared with the standard SVRG analysis (Theorem 2 in Reddi et al. 2016a), Theorem 1 involves an additional term τ/λ in the upper bound on $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2$. By choosing sufficiently small smoothing parameters as well as a large mini-batch size $|\mathcal{S}_1|$, we guarantee that such an error term is dominated by the first term in (5), as shown below.

Corollary 1 (mini-batch, $|\mathcal{S}_2| > 1$)**.** *Under the setting of Theorem 1, let* $g = 4000d\eta^2L^3q/|\mathcal{S}_2|$ *and choose*

$$\eta = \frac{1}{20L}, |\mathcal{S}_1| = \min\{n, K\}, q = \left\lceil |\mathcal{S}_1|^{1/3}\right\rceil$$
$$|\mathcal{S}_2| = dq^2, \beta = \frac{1}{Ld\sqrt{K}}, \delta = \frac{1}{L\sqrt{dK}}, \tag{6}$$

where e *is the Euler's number. Then, Algorithm 1 satisfies* $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \le \mathcal{O}(1/K)$

To achieve an ϵ*-stationary point, i.e.,* $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \le \epsilon$*, the number of function queries required by Algorithm 1 is at most* $\mathcal{O}\left(\min\left\{n^{2/3}d\epsilon^{-1}, d\epsilon^{-5/3}\right\}\right)$.

Corollary 1 implies that mini-batch ZO-SVRG-Coord-Rand achieves a convergence rate of $\mathcal{O}(1/K)$, which improves the best rate of existing zeroth-order algorithms for nonconvex optimization by a factor of $\mathcal{O}(d)$. In particular, the function query complexity of our ZO-SVRG-Coord-Rand algorithm improves upon that of ZO-SGD by a factor of $\mathcal{O}(\epsilon^{-1/3})$, and that of ZO-GD by a factor of $\mathcal{O}(n^{1/3})$. As far as we know, this is the first SVRG-based zeroth-order algorithm that outperforms both ZO-GD and ZO-SGD in terms of the function query complexity.

The mini-batching strategy in Corollary 1 may require a parallel computation of $\mathbf{v}^k$. For nonparallel scenarios, we provide the following single-sample ZO-SVRG-Coord-Rand, which achieves the same function query complexity as mini-batch ZO-SVRG-Coord-Rand.

Corollary 2 (Single-sample, $|\mathcal{S}_2| = 1$)**.** *Under the setting of Theorem 1, let* $g = 4000dq\eta^2L^3$ *and*

$$|\mathcal{S}_1| = \min\left\{n, \left\lceil (K/d)^{3/5}\right\rceil\right\}, q = |\mathcal{S}_1|d,\ \beta = \frac{|\mathcal{S}_1|^{1/3}}{L\sqrt{dK}}$$
$$\eta = \frac{1}{20d^{1/3}q^{2/3}L}, \delta = \frac{|\mathcal{S}_1|^{1/3}}{L\sqrt{K}}, \tag{7}$$

where e *is the Euler's number. Then, Algorithm 1 satisfies* $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \le \mathcal{O}\left(d|\mathcal{S}_1|^{2/3}/K\right)$.

To achieve an ϵ*-stationary point, i.e.,* $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \le \epsilon$*, the number of function queries required by Algorithm 1 is at most* $\mathcal{O}\left(\min\left\{n^{2/3}d\epsilon^{-1}, d\epsilon^{-5/3}\right\}\right)$.

2.3. New Analysis for ZO-SVRG-Coord

In this subsection, we provide an improved analysis for the ZO-SVRG-Coord algorithm proposed by Liu et al. 2018b, which adopts the same outer-loop iteration as mini-batch ZO-SVRG-Coord-Rand, i.e., Algorithm 1, but updates the inner-loop estimator $\mathbf{v}^k$ coordinate-wisely by

$$\mathbf{v}^k = \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}^k) - \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}^{qk_0}) + \hat{\nabla}_{\text{coord}} f_{\mathcal{S}_1}(\mathbf{x}^{qk_0}). \tag{8}$$

We first show that although the coordinate-wise gradient estimator in (8) requires d times more function queries than the two-point random gradient estimator at each inner-loop iteration, it achieves more accurate gradient estimation, as stated in the following lemma.

Lemma 2. *Under Assumption 1, we have, for any* $qk_0 \le k \le \min\{q(k_0+1)-1, qh\}$, $k_0 = 0, ..., h$,

$$\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 \le \frac{12L^2d\delta^2}{|\mathcal{S}_2|} + \frac{6L^2}{|\mathcal{S}_2|}\|\mathbf{x}^k - \mathbf{x}^{qk_0}\|^2 + \frac{6I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2d\delta^2 + \sigma^2\right).$$

It can be seen that the above bound in Lemma 2 contains a tighter error term $\frac{6L^2}{|\mathcal{S}_2|}\|\mathbf{x}^k - \mathbf{x}^{qk_0}\|^2$ than that in Lemma 1 by a factor of $\mathcal{O}(d)$. More importantly, this bound is tighter than that in Theorem 3 in Liu et al. 2018b for ZO-SVRG-Coord by a factor of $\mathcal{O}(d)$. Based on Lemma 2, we have the following theorem.

Theorem 2. *Let Assumption 1 hold, and select*

$$\eta = \frac{1}{15L}, |\mathcal{S}_1| = \min\{n, K\}, q = \left\lceil |\mathcal{S}_1|^{1/3} \right\rceil$$
$$|\mathcal{S}_2| = q^2, \delta = \frac{1}{L\sqrt{dK}}, \tag{9}$$

where e is the Euler's number. Then, ZO-SVRG-Coord satisfies $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \mathcal{O}(1/K)$.

To achieve an ϵ-stationary point, i.e., $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, the number of function queries required by ZO-SVRG-Coord is at most $\mathcal{O}\left(\min\left\{n^{2/3}d\epsilon^{-1}, d\epsilon^{-5/3}\right\}\right)$.

Theorem 2 order-wisely improves the complexity bound in Liu et al. 2018b by a factor of $\mathcal{O}(\max\{n^{1/3}, dn^{-2/3}\})$ due to our new analysis. Furthermore, Theorem 2 shows that ZO-SVRG-Coord achieves the same performance as ZO-SVRG-Coord-Rand, both of which order-wisely improves ZO-GD and ZO-SGD in the convergence rate as well as the function query complexity for nonconvex optimization. Moreover, both ZO-SVRG-Coord-Rand and ZO-SVRG-Coord (under our new analysis) allows a much larger stepsize $\mathcal{O}(1)$ than $\eta = \mathcal{O}(1/d)$ used in ZO-SGD and all zeroth-order SVRG-based algorithms in Liu et al. 2018b, and hence converges much faster in practice, as demonstrated in our experiments.

3. ZO-SPIDER-Coord Algorithm for Nonconvex Optimization

Recently, Nguyen et al. 2017a;b and Fang et al. 2018 proposed a new first-order variance-reduced stochastic gradient estimator named SARAH and SPIDER respectively, which estimates stochastic gradients in a *recursive* way as

$$\mathbf{v}^k = \nabla f_{\mathcal{S}_2}(\mathbf{x}^k) - \nabla f_{\mathcal{S}_2}(\mathbf{x}^{k-1}) + \mathbf{v}^{k-1}. \tag{10}$$

In this section, we explore the performance of this estimator in zeroth-order nonconvex optimization. Motivated by our new analysis for ZO-SVRG-Coord, we propose a zeroth-order SPIDER-based algorithm ZO-SPIDER-Coord, as shown in Algorithm 2. Our ZO-SPIDER-Coord extends the estimator (10) for zeroth-order optimization by

$$\mathbf{v}^k = \hat{\nabla}_{\text{coord}} f_{\mathcal{S}_2}(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f_{\mathcal{S}_2}(\mathbf{x}^{k-1}) + \mathbf{v}^{k-1}. \tag{11}$$

where $\hat{\nabla}_{\text{coord}} f_{\mathcal{S}_2}(\mathbf{x}) = \sum_{i=1}^{d} \frac{(f_{\mathcal{S}_2}(\mathbf{x}+\delta\mathbf{e}_i) - f_{\mathcal{S}_2}(\mathbf{x}-\delta\mathbf{e}_i))\mathbf{e}_i}{2\delta}$. Differently from the existing SPIDER-based zeroth-order algorithm SPIDER-SZO proposed in Fang et al. 2018, which requires to generate totally $\mathcal{O}(d^2\sqrt{n}\epsilon^{-1})$ Gaussian random variables (see Theorem 8 in Fang et al. 2018), our ZO-SPIDER-Coord eliminates Gaussian variable generation due to the utilization of coordinate-wise gradient estimator, and still achieves the same complexity performance as SPIDER-SZO, as shown in the next subsection. In addition, mini-batch ZO-SPIDER-Coord allows a large constant stepsize (see Corollary 3), as apposed to the small stepsize $\mathcal{O}(\sqrt{\epsilon}/\|\mathbf{v}^k\|)$ used in SPIDER-SZO for guaranteeing the convergence. Similar idea has also been used in SpiderBoost (Wang et al., 2018) to enhance the stepsize of SPIDER (Fang et al., 2018).

Algorithm 2 ZO-SPIDER-Coord

```
Input: $K$, $q$, $|\mathcal{S}_1| \leq n$, $|\mathcal{S}_2|$, $\mathbf{x}^0$, $\delta > 0$, $\eta > 0$.
for $k = 0$ to $K$ do
    if $k \mod q = 0$ then
        Sample $\mathcal{S}_1$ from $[n]$ without replacement
        Compute $\mathbf{v}^k = \hat{\nabla}_{\text{coord}} f_{\mathcal{S}_1}(\mathbf{x}^k)$
    else
        Sample $\mathcal{S}_2 = \{a_1, a_2...., a_{|\mathcal{S}_2|}\}$ from $[n]$ with replacement
        Compute $\mathbf{v}^k$ according to (11)
    end if
    $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta\mathbf{v}^k$
end for
Output: $\mathbf{x}_\zeta$ from $\{\mathbf{x}_0, ..., \mathbf{x}_K\}$ uniformly at random.
```

3.1. Convergence and Complexity Analysis

The following theorem provides the convergence guarantee for ZO-SPIDER-Coord.

Theorem 3. *Let Assumption 1 hold, and define*

$$\phi = \frac{\eta}{2} - \frac{\eta^2 L}{2} - \frac{3L^2\eta^3 q}{|\mathcal{S}_2|},$$
$$\theta = \frac{3qL^2 d\delta^2}{|\mathcal{S}_2|} + \frac{3I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right). \tag{12}$$

Then, the output $\mathbf{x}^\zeta$ of Algorithm 2 satisfies

$$\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq 3L^2 d\delta^2 + 3\theta$$
$$+ \left(\frac{9q\eta^2 L^2}{\phi|\mathcal{S}_2|} + \frac{3}{\phi}\right)\left(\frac{\Delta}{K} + \eta\left(\theta + L^2 d\delta^2\right)\right), \tag{13}$$

where $\Delta := f(\mathbf{x}^0) - f(\mathbf{x}^*)$ with $\mathbf{x}^* := \arg\min_{\mathbf{x}} f(\mathbf{x})$.

Based on Theorem 3, we provide an analysis on mini-batch ZO-SPIDER-Coord.

Corollary 3 (Mini-batch, $|\mathcal{S}_2| > 1$)**.** *Under the setting of Theorem 3, we choose stepsize $\eta = \frac{1}{4L}$ and*

$$|\mathcal{S}_1| = \min\{n, K\}, |\mathcal{S}_2| = q = \lceil |\mathcal{S}_1|^{1/2} \rceil, \delta = \frac{1}{\sqrt{KdL}}. \tag{14}$$

Then, Algorithm 2 satisfies $\mathbb{E}\|\nabla f(\mathbf{x}^{\zeta})\|^2 \leq \mathcal{O}(1/K)$.

To achieve an ϵ*-stationary point, i.e.,* $\mathbb{E}\|\nabla f(\mathbf{x}^{\zeta})\|^2 \leq \epsilon$, *the number of function queries required by Algorithm 2 is* $\mathcal{O}\left(\min\left\{n^{1/2}d\epsilon^{-1}, d\epsilon^{-3/2}\right\}\right)$.

As shown in Corollary 3, mini-batch ZO-SPIDER-Coord achieves the convergence rate of $\mathcal{O}(1/K)$, and improves the function query complexity of ZO-SVRG-Coord-Rand by a factor of $\min\{\epsilon^{-1/6}, n^{1/6}\}$. The following corollary analyzes single-sample ZO-SPIDER-Coord, which achieves the same query complexity as mini-batch ZO-SPIDER-Coord.

Corollary 4 (Single-sample, $|\mathcal{S}_2| = 1$)**.** *Under the setting of Theorem 3, we choose* $\eta = \frac{1}{4L\sqrt{q}}, \delta = \frac{1}{\sqrt{qKd}L}, q = |\mathcal{S}_1| = \min\left\{n, \lceil K^{2/3}\rceil\right\}$. *Then, Algorithm 2 satisfies* $\mathbb{E}\|\nabla f(\mathbf{x}^{\zeta})\|^2 \leq \mathcal{O}(\sqrt{|\mathcal{S}_1|}/K)$.

To achieve an ϵ*-stationary point, i.e.,* $\mathbb{E}\|\nabla f(\mathbf{x}^{\zeta})\|^2 \leq \epsilon < 1$, *the number of function queries required by Algorithm 2 is* $\mathcal{O}\left(\min\left\{n^{1/2}d\epsilon^{-1}, d\epsilon^{-3/2}\right\}\right)$.

3.2. ZO-SPIDER-Coord under PL without Restart

Many nonconvex machine learning and deep learning problems satisfy the following Polyak-Łojasiewicz (PL) (i.e., gradient dominance) condition in local regions around global minimizers (Zhou et al., 2016; Zhong et al., 2017).

Definition 1 ((Polyak, 1963))**.** *Let* $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. *Then, the function* f *is said to be* γ*-gradient dominated if for any* $\mathbf{x} \in \mathbb{R}^d$, $f(\mathbf{x}) - f(\mathbf{x}^*) \leq \gamma\|\nabla f(\mathbf{x})\|^2$.

In this subsection, we explore whether ZO-SPIDER-Coord algorithm achieves a faster convergence rate when it enters the local areas where the loss function satisfies the PL condition. The following theorem provides an affirmative answer. For the simplicity of presentation, we choose $|\mathcal{S}_1| = n$.

Theorem 4. *Under the parameters selected in Corollary 3, we take* $|\mathcal{S}_2| = \lceil\gamma L B_\gamma\rceil$ *and* $\delta = \mathcal{O}(\sqrt{\epsilon}/(L\sqrt{\gamma d}))$. *We further assume that* $\gamma > q/(b_\gamma L)$, *where* b_γ *and* B_γ *are two positive constants satisfying* $\frac{1}{8}\left(1 - \frac{b_\gamma}{16q}\right)^q - \frac{3}{B_\gamma} > 0$. *Then, ZO-SPIDER-Coord satisfies*

$$\mathbb{E}(f(\mathbf{x}^K) - f(\mathbf{x}^*)) \leq \left(1 - \frac{1}{16L\gamma}\right)^K (f(\mathbf{x}^0) - f(\mathbf{x}^*)) + \epsilon,$$

and requires $\mathcal{O}\left(d(\gamma n^{1/2} + \gamma^2)\log\left(\frac{1}{\epsilon}\right)\right)$ *function queries.*

The assumption that $\gamma > q/(b_\gamma L) = \mathcal{O}(n^{1/2}/L)$ has been widely adopted in optimization under the PL condition, e.g., in Reddi et al. 2016a; Li & Li 2018. In contrast to the restart technique commonly used in the first-order algorithms, e.g., GD-SVRG (Reddi et al., 2016a), for proving the convergence under the PL condition, our proof of the linear convergence rate for ZO-SPIDER-Coord does not require restart and algorithmic modification. This implies that ZO-SPIDER-Coord can be initialized in a general nonconvex landscape and then automatically achieves a faster convergence rate as it enters a PL landscape. In addition, unlike SPIDER (Fang et al., 2018) and SpiderBoost (Wang et al., 2018), our proof of Theorem 4 does not need to upper-bound $\sum_{k=1}^{K}\mathbb{E}\|\mathbf{v}^k\|^2$, which is much simpler and can also be applied to both SPIDER and SpiderBoost for first-order nonconvex optimization under the PL condition.

4. Experiments

In this section, we compare the empirical performance of our proposed ZO-SVRG-Coord-Rand, ZO-SPIDER-Coord and ZO-SVRG-Coord (for which we provide improved analysis that allows a larger stepsize) with ZO-SGD (Ghadimi & Lan, 2013), ZO-SVRG-Ave (p=10)[1] (Liu et al., 2018b) and SPIDER-SZO (Fang et al., 2018). We conduct two experiments, i.e., generation of black-box adversarial examples and nonconvex logistic regression. The parameter settings for these algorithms are further specified in the supplementary materials due to the space limitations.

4.1. Generation of Black-Box Adversarial Examples

In image classification, adversary attack crafts input images with imperceptive perturbation to mislead a trained classifier. The resulting perturbed images are called adversarial examples, which are commonly used to understand the robustness of learning models. In the black-box setting, the attacker can access only the model evaluations, and hence the problem falls into the framework of zeroth-order optimization.

We use a well-trained DNN[2] $F(\cdot) = [F_1(\cdot), ..., F_K(\cdot)]$ for the MNIST handwritten digit classification as the target black-box model, where $F_k(\cdot)$ returns the prediction score of the k^{th} class. We attack a batch of n correctly-classified images $\{\mathbf{a}_i\}_{i=1}^n$ from the same class, and adopt the same black-box attacking loss as in Chen et al. 2017; Liu et al. 2018b. The i^{th} individual loss function $f_i(\mathbf{x})$ is given by

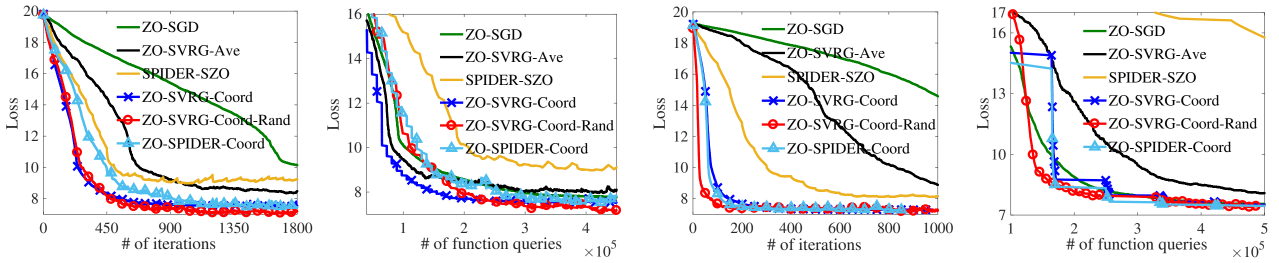
$$\begin{aligned} f_i(\mathbf{x}) = &\max\left\{\log F_{y_i}\left(\mathbf{a}_i^{adv}\right) - \max_{t\neq y_i}\log F_t\left(\mathbf{a}_i^{adv}\right), 0\right\} \\ &+ \lambda\|\mathbf{a}_i^{adv} - \mathbf{a}_i\|^2, \end{aligned}$$

where $\mathbf{a}_i^{adv} = 0.5\tanh\left(\tanh^{-1}(2\mathbf{a}_i) + \mathbf{x}\right)$ is the adversarial example of the i^{th} natural image $\mathbf{a}_i$, and y_i is the true label of image $\mathbf{a}_i$. In our experiment, we set the regularization parameter $\lambda = 1$ for digit "1" image class, and set $\lambda = 0.1$ for digit "4" class.

Fig. 1 and Fig. 3 (in the supplementary materials) provide comparison of the performance for the algorithms of interest. Two major observations can be made. First, our proposed two algorithms ZO-SVRG-Coord-Rand and ZO-SPIDER-

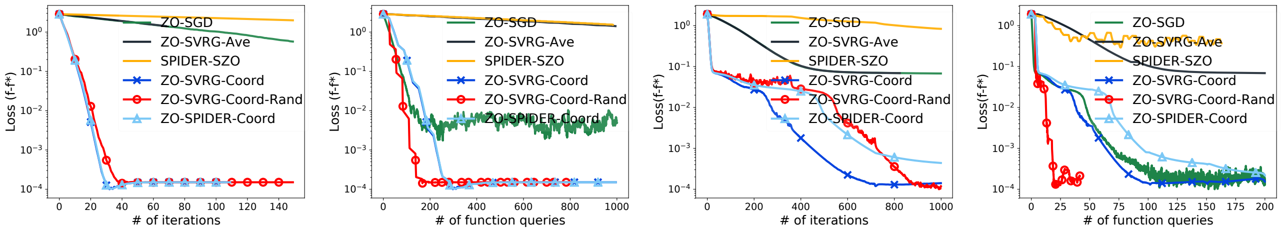
[1] ZO-SVRG-Ave (p=10) has the best performance among the three methods proposed by Liu et al. 2018b.

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



(a) Loss vs iterations: $n = 10$ (b) Loss vs queries: $n = 10$ (c) Loss vs iterations: $n = 100$ (d) Loss vs queries: $n = 100$

Figure 1. Comparison of different zeroth-order algorithms for generating black-box adversarial examples for digit "1" class



(a) Loss vs iterations: german (b) Loss vs queries: german (c) Loss vs iterations: ijcnn1 (d) Loss vs queries: ijcnn1

Figure 2. Comparison of different zeroth-order algorithms for logistic regression problem with a nonconvex regularizer

Coord as well as ZO-SVRG-Coord (with the large stepsize due to our improved analysis) have much better performance both in convergence rate (iteration complexity) and function query complexity than ZO-SGD, ZO-SVRG-Ave and SPIDER-SZO. Among them, ZO-SVRG-Coord-Rand achieves the best performance. Second, our ZO-SPIDER-Coord algorithm converges much faster than SPIDER-SZO in the initial optimization stage, and more importantly, has much lower function query complexity, which is largely due to the ϵ-level stepsize required by SPIDER-SZO. In addition, we present the generated adversarial examples for attacking digit "4" class in Table 3 in the supplementary materials, where our ZO-SVRG-Coord-Rand achieves the lowest image distortion.

Interestingly, though SPIDER-based algorithms have been shown to outperfom SVRG-based algorithms in theory, our experiments suggest that SVRG-based algorithms in fact achieve comparable and sometimes even better performance in practice. The same observations have also been made in Fang et al. 2018 and Nguyen et al. 2017a;b.

4.2. Nonconvex Logistic Regression

In this subsection, we consider the following zeroth-order nonconvex logistic regression problem with two classes $\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}^T\mathbf{x}_i, y_i) + \alpha\sum_{i=1}^d \frac{w_i^2}{1+w_i^2}$, where $\mathbf{x}_i \in \mathbb{R}^d$ denote the features, $y_i \in \{\pm 1\}$ are the classification labels, ℓ is the cross-entropy loss, and we set $\alpha = 0.1$. For this problem, we use two datasets from LIBSVM (Chang & Lin, 2011): the german dataset ($n = 1000, d = 24$) and the ijcnn1 dataset ($n = 49990, d = 22$).

As shown in Fig. 2, ZO-SVRG-Coord-Rand, ZO-SVRG-Coord and ZO-SPIDER-Coord converges much faster than ZO-SGD , ZO-SVRG-Ave and SPIDER-SZO in terms of number of iterations for both datasets. In terms of function query complexity, ZO-SVRG-Coord converges much faster than ZO-SVRG-Ave for both datasets and slightly faster than ZO-SGD for ijcnn1 dataset, which corroborates our new complexity analysis for ZO-SVRG-Coord. The convergence and complexity performance of ZO-SPIDER-Coord is similar to ZO-SVRG-Coord. Among these algorithms, ZO-SVRG-Coord-Rand has the best function query complexity for both datasets.

5. Conclusion

In this paper, we developed two novel zeroth-order variance-reduced algorithms named ZO-SVRG-Coord-Rand and ZO-SPIDER-Coord as well as an improved analysis on ZO-SVRG-Coord proposed by Liu et al. 2018b. We showed that ZO-SVRG-Coord-Rand and ZO-SVRG-Coord (under our new analysis) outperform ZO-GD, ZO-SGD and all other existing SVRG-based zeroth-order algorithms. Furthermore, compared with SPIDER-SZO (Fang et al., 2018), our ZO-SPIDER-Coord allows a much larger constant stepsize and is free from the generation of a large number of Gaussian random variables while maintaining the same function query complexity. Our experiments demonstrate the superior performance of our proposed algorithms.

Acknowledgements

The work was supported in part by U.S. National Science Foundation under the grants CCF-1761506 and CCF-1801855.

References

Agarwal, A., Dekel, O., and Xiao, L. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory (COLT)*, pp. 28–40, 2010.

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning (ICML)*, pp. 699–707, 2016.

Balasubramanian, K. and Ghadimi, S. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3459–3468, 2018.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. 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*, pp. 15–26, 2017.

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

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1646–1654. 2014.

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

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 385–394, 2005.

Gao, X., Jiang, B., and Zhang, S. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, pp. 1–37, 2014.

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

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

Gu, B., Huo, Z., Deng, C., and Huang, H. Faster derivative-free stochastic algorithm for shared memory machines. In *International Conference on Machine Learning (ICML)*, pp. 1807–1816, 2018.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 315–323, 2013.

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

Li, Z. and Li, J. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1802.04477*, 2018.

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

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

Liu, S., Kailkhura, B., Chen, P.-Y., Ting, P., Chang, S., and Amini, L. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3731–3741, 2018b.

Nemirovsky, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

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

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning (ICML)*, pp. 2613–2621, 2017a.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017b.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519. ACM, 2017.

Polyak, B. T. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, pp. 314–323, 2016a.

Reddi, S. J., Sra, S., Poczos, B., and Smola, A. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1145–1153. 2016b.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 09 1951.

Roux, N. L., Schmidt, M., and Bach, F. R. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2663–2671. 2012.

Shamir, O. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory (COLT)*, pp. 3–24, 2013.

Taskar, B., Chatalbashev, V., Koller, D., and Guestrin, C. Learning structured prediction models: A large margin approach. In *International Conference on Machine Learning (ICML)*, pp. 896–903, 2005.

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Wang, Y., Du, S., Balakrishnan, S., and Singh, A. Stochastic zeroth-order optimization in high dimensions. *arXiv preprint arXiv:1710.10551*, 2017.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. In *International Conference on Machine Learning (ICML)*, pp. 4140–4149, 2017.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3921–3932, 2018.

Zhou, Y., Zhang, H., and Liang, Y. Geometrical properties and accelerated gradient solvers of non-convex phase retrieval. In *54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 331–335, 2016.