Journal of Machine Learning Research 22 (2021) 1-47 Submitted 6/20; Revised 4/21; Published 6/21

Black-Box Reductions for Zeroth-Order Gradient Algorithms to Achieve Lower Query Complexity

Bin Gu BIN.GU@MBZUAI.AC.AE
Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, UAE
JD Finance America Corporation
Xiyuan Wei XYWEI0905@GMAIL.COM
School of Computer & Software, Nanjing University of Information Science & Technology, China
Shangqian Gao SHG84@PITT.EDU
Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA
Ziran Xiong XIONGZIRAN@NUIST.EDU.CN
School of Computer & Software, Nanjing University of Information Science & Technology, China
Cheng Deng CHDENG.XD@GMAIL.COM
School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, 710071, China
Heng Huang HENG.HUANG@PITT.EDU
Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA
JD Finance America Corporation

Editor: Julien Mairal

Abstract

Zeroth-order (ZO) optimization has been the key technique for various machine learning applications especially for black-box adversarial attack, where models need to be learned in a gradient-free manner. Although many ZO algorithms have been proposed, the high function query complexities hinder their applications seriously. To address this challenging problem, we propose two stagewise black-box reduction frameworks for ZO algorithms under convex and non-convex settings respectively, which lower down the function query complexities of ZO algorithms. Moreover, our frameworks can directly derive the convergence results of ZO algorithms under convex and non-convex settings without extra analyses, as long as convergence results under strongly convex setting are given. To illustrate the advantages, we further study ZO-SVRG, ZO-SAGA and ZO-Varag under strongly-convex setting and use our frameworks to directly derive the convergence results under convex and non-convex settings. The function query complexities of these algorithms derived by our frameworks are lower than that of their vanilla counterparts without frameworks, or even lower than that of state-of-the-art algorithms. Finally we conduct numerical experiments to illustrate the superiority of our frameworks.

Keywords: Zeroth order optimization, black-box reduction, stagewise training, convex optimization, non-convex optimization

1. Introduction

In many machine learning applications, such as black-box adversarial attacks on deep neural networks (DNNs) (Papernot et al., 2017; Madry et al., 2017; Kurakin et al., 2016), bandit optimization (Flaxman et al., 2005) and reinforcement learning (Choromanski et al., 2018),

calculating the explicit gradients of objective function is computationally expensive or infeasible. Thus, zeroth-order (ZO) optimization methods are extremely important to these optimization problems, because ZO method estimates the gradient only by two function evaluations. With the applications of black-box adversarial attacks, bandit optimization and reinforcement learning becoming more and more popular in machine learning, ZO optimization has gained increasing attention recently.

Since ZO algorithms use function values to estimate the gradient which can be used to update the solutions of the objective function, a critical assessment metric for algorithmic efficiency of ZO algorithms is the *function query complexity*, *i.e.*, the number of queried function values to converge to a specified accuracy. Take black-box adversarial attacks as an example, its ultimate goal is to use function estimation as little as possible to generate an adversarial example. Thus, in this paper, we will investigate different ZO algorithms in term of *function query complexity*.

Specifically, Nesterov and Spokoiny (2017) proposed the ZO gradient descent (ZO-GD) algorithm that used the Gaussian smoothing technique to construct a two-point gradient estimator. Its function query complexities are $\mathcal{O}\left(\frac{dn}{\epsilon}\right)$ for convex problems and $\mathcal{O}\left(\frac{dn}{\epsilon^2}\right)$ for non-convex problems respectively, where n and d are the sample and features sizes respectively. With the same gradient estimation technique, Ghadimi and Lan (2013) proposed a ZO stochastic gradient descent (ZO-SGD) algorithm with the function query complexity of $\mathcal{O}\left(\frac{d}{\epsilon^2}\right)$ and $\mathcal{O}\left(\frac{d}{\epsilon^4}\right)$ for convex and non-convex problems respectively. Since ZO gradient estimator have a high variance, variance reduction techniques have recently been used in developing new ZO algorithms. These include ZO-SVRG, ZO-SVRG-Ave and ZO-SVRG-Coord (Liu et al., 2018), SPIDER-SZO (Fang et al., 2018) and ZO-SPIDER-Coord-Rand (Ji et al., 2019). Their function query complexities for convex and non-convex problems are given in Table 1 and 2 respectively. Note that, variance reduction techniques were also extended to proximal algorithms (e.g. Huang et al. (2019)) which are beyond the scope of this paper.

Although many ZO algorithms have been proposed as mentioned above, the high function query complexities still hinder many ZO applications seriously especially for black-box adversarial attacks. As shown in (Tu et al., 2019), to achieve 97% attack success rate, ZO algorithm needs 10,000 more queries which is impractical in real-world applications of adversarial attacks because query count is normally limited in most machine learning systems. Thus, whether the function query complexities of ZO algorithms can be improved further is an important problem. To the best of our knowledge, it is still a vacancy to provide a framework to improve the function query complexities for different ZO algorithms under the convex and non-convex conditions.

To address this challenging problem, in this paper, we propose two ZO reduction frameworks, *AdaptRdct-C* and *AdaptRdct-NC* for handling convex and non-convex objectives, respectively. They work in a black-box manner, and only have minor requirement over the oracle optimizer. This means we can apply *AdaptRdct-C* and *AdaptRdct-NC* to various ZO algorithms to lower their function query complexity. To demonstrate the effectiveness of the proposed ZO reduction frameworks, we choose ZO-SVRG and ZO-SAGA, ZO extensions of two popular variance-reduced algorithms (Johnson and Zhang, 2013; Defazio et al., 2014), as the oracle optimizer for reduction. We theoretically study the function query complexity of applying our ZO reduction frameworks on ZO-SVRG and ZO-SAGA, and the

results are very promising: when ZO reduction frameworks are used, both ZO-SVRG and ZO-SAGA outperform their vanilla counterparts. We also empirically verify this conclusion with real-data experiments.

Besides lowering down the function query complexities of ZO algorithms, another advantage of our frameworks is that convergence results of ZO algorithms under convex and non-convex settings can be directly derived without extra analyses, as long as convergence results under strongly convex setting are given. At each stage of our frameworks, a ZO algorithm actually solves a strongly convex subproblem, instead of the original convex or non-convex problem. Thus with convergence results under strongly convex setting, our frameworks can directly derive convergence results under convex or non-convex setting. To illustrate this advantage, we study the convergence results of ZO-SVRG and ZO-SAGA under strongly convex setting, and use our frameworks to directly derive the convergence results under convex and non-convex settings. Specially, we also use the convergence results of ZO-Varag under strongly convex setting, which was derived by (Chen et al., 2020) (with only minor modifications), to directly derive the convergence results of ZO-Varag under convex and non-convex settings. Under convex setting, the function query complexity of *AdaptRdct-C* (ZO-Varag) matches that of vanilla ZO-Varag derived by (Chen et al., 2020). Under non-convex setting, the function query complexity of *AdaptRdct-NC* (ZO-Varag) solving σ-strongly non-convex problems outperforms that of state-of-the-arts algorithms when $n < \mathcal{O}\left(\frac{L}{\sigma}\right)$, where n is the number of individual functions and L is the Lipschitz constant.

Contributions. The main contributions of this paper are summarized as follows:

1. We propose two reduction frameworks for zeroth-order algorithms under convex and non-convex settings respectively, which lowers down the function query complexities of zeroth-order algorithms. Moreover, our frameworks can directly derive convergence results of zeroth-order algorithms under convex and non-convex settings without extra analyses, as long as convergence results under strongly convex setting are given.
2. We apply our frameworks to three zeroth-order algorithms to illustrate the advantages of our frameworks, which are ZO-SVRG, ZO-SAGA and ZO-Varag. To the best of our knowledge, we are the first to propose black-box reduction frameworks for zeroth-order algorithms and apply them to zeroth-order optimization.
3. As a by-product, we are the first to provide the convergence rates and function query complexities of ZO-SVRG and ZO-SAGA under both of the convex and strongly convex conditions.

2. Related Work

Black-Box Reduction Techniques: Allen-Zhu and Hazan (2016) proposed a black-box reduction method for convex problems. The reduction method is in a black-box manner, which means the analyses and results work for a wide range of convex problems and first-order algorithms. Chen et al. (2018) proposed a universal stagewise optimization framework for non-convex problems. The framework originates from the proximal point method in convex optimization (Rockafellar, 1976; Güler, 1992; Frostig et al., 2015; Lin et al., 2018). Our study also falls into developing black-box reduction methods for both convex and

Table 1: Comparison of ZO algorithms solving *convex* problems in terms of convergence rate and function query complexity. (n denotes the number of individual loss function, d denotes the dimension of problem space, S denotes the number of epochs, m denotes the update frequency, $\rho_{i\in[6]} \in (0,1)$ denotes six different constants and ϵ denotes the accuracy. C_1 denote the case that $n \geq \mathcal{O}\left(\frac{1}{\epsilon}\right)$, C_2 denotes the case that $n < \mathcal{O}\left(\frac{1}{\epsilon}\right)$. $\tilde{\mathcal{O}}$ hides a logarithmic factor.)

Algorithms	Smoothing parameter	Convergence rate	Function query complexity
ZO-GD (Nesterov and Spokoiny)	$\mathcal{O}\left(\frac{\sqrt{\epsilon}}{d}\right)$	$\mathcal{O}\left(\frac{d}{S}\right)$	$\mathcal{O}\left(\frac{dn}{\epsilon}\right)$
ZO-SGD (Ghadimi and Lan)	$\mathcal{O}\left(\frac{1}{\sqrt{d}}\right)$	$\mathcal{O}\left(\sqrt{\frac{d}{S}}\right)$	$\mathcal{O}\left(\frac{d}{\epsilon^2}\right)$
ZO-SVRG-Coord-Rand-C (Ji et al.)	$\mathcal{O}\left(\frac{\epsilon}{d}\right)$	$\mathcal{O}\left(\rho_1^S + \frac{1}{m}\right)$	$\mathcal{O}(\min\{dn, \frac{d}{\epsilon}\}\log\frac{1}{\epsilon})$
ZO-SPIDER-Coord-C (Ji et al.)	$\mathcal{O}\left(\frac{\sqrt{\epsilon}}{\sqrt{d}}\right)$	$\mathcal{O}\left(\rho_2^S + \frac{1}{m}\right)$	$\mathcal{O}(\min\{dn, \frac{d}{\epsilon}\}\log\frac{1}{\epsilon})$
ZO-Varag (Chen et al.)	$\begin{cases}\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right), & C_1\\ \mathcal{O}\left(\frac{\sqrt{n}\epsilon^{\frac{3}{2}}}{\sqrt{d}}\right), & C_2\end{cases}$	$\begin{cases}\mathcal{O}\left(\rho_3^S\right), & C_1\\ \mathcal{O}\left(\frac{1}{nS^2}\right), & C_2\end{cases}$	$\begin{cases}\mathcal{O}\left(dn\log\frac{1}{\epsilon}\right), & C_1\\ \mathcal{O}\left(dn\log n + d\sqrt{\frac{n}{\epsilon}}\right), & C_2\end{cases}$
ZO-SVRG/ZO-SAGA (**Ours**)	$\mathcal{O}\left(\frac{\sqrt{\epsilon}}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{d}{S}\right)$	$\mathcal{O}\left(\frac{n+d}{\epsilon}\right)$
AdaptRdct-C (**Ours**) (ZO-SVRG)	$\mathcal{O}\left(\frac{\sqrt{\epsilon}}{\sqrt{d}}\right)$	$\mathcal{O}\left(\rho_4^S\right)$	$\mathcal{O}(n\log\frac{1}{\epsilon} + \frac{d}{\epsilon})$
AdaptRdct-C (**Ours**) (ZO-SAGA)	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right)$	$\mathcal{O}\left(\rho_5^S\right)$	$\tilde{\mathcal{O}}(n\log\frac{1}{\epsilon} + \frac{d}{\epsilon})$
AdaptRdct-C (**Ours**) (ZO-Varag)	$\begin{cases}\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right), & C_1\\ \mathcal{O}\left(\frac{\sqrt{n}\epsilon^{\frac{3}{2}}}{\sqrt{d}}\right), & C_2\end{cases}$	$\mathcal{O}\left(\rho_6^S\right)$	$\begin{cases}\mathcal{O}\left(dn\log\frac{1}{\epsilon}\right), & C_1\\ \mathcal{O}\left(dn\log n + d\sqrt{\frac{n}{\epsilon}}\right), & C_2\end{cases}$

non-convex problems, but different from the methods above, our methods focus on ZO algorithms.

Zeroth-Order Optimization: ZO optimization is a classical problem in the optimization community. We first summarize the ZO algorithms for convex problems. Specifically, Nemirovski et al. (2009) first introduced a one-point random sampling scheme to approximate the true gradient $\nabla f(\mathbf{x})$ by querying $f(\mathbf{x})$ at a random location that is close to $\mathbf{x}$. After that, Agarwal et al. (2010); Nesterov and Spokoiny (2017) proposed multi-point gradient estimation approach. Many works were based on multi-point estimation. For example, Ghadimi and Lan (2013) presented ZO stochastic gradient descent (ZO-SGD) algorithm using a two-point Gaussian gradient estimator; Duchi et al. (2015) derived a ZO mirror descent algorithm; Ji et al. (2019) proposed ZO stochastic variance reduced gradient (ZO-SVRG-Coord-Rand-C) algorithm; Chen et al. (2020) proposed an accelerated ZO variance reduced gradient (ZO-Varag) algorithm.

Table 2: Comparison of ZO algorithms solving *σ-strongly non-convex* problems in terms of convergence rate and function query complexity. (n denotes the number of individual loss function, d denotes the dimension of problem space, L denotes the smoothness parameter, S denotes the number of epochs, m denotes the update frequency, and $K = mS$. $^{\spadesuit}$: $|\mathcal{S}_2|$ denotes the batch size for constructing the random gradient estimator. $^{\star\blacktriangleright}$: $p_{min} = \min\{d, q\}$ and $p_{max} = \max\{d, q\}$, where q denotes the number of i.i.d. smoothing vectors for constructing the average random gradient estimator. C_1 denote the case that $n \geq \mathcal{O}\left(\frac{L}{\sigma}\right)$, C_2 denotes the case that $n < \mathcal{O}\left(\frac{L}{\sigma}\right)$. $\tilde{\mathcal{O}}$ hides a logarithmic factor.)

Algorithms	Smoothing parameter	Convergence rate	Function query complexity
ZO-GD (Nesterov and Spokoiny)	$\mathcal{O}\left(\frac{\epsilon}{d}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{S}}\right)$	$\mathcal{O}\left(\frac{dn}{\epsilon^2}\right)$
ZO-SGD (Ghadimi and Lan)	$\mathcal{O}\left(\frac{\epsilon}{d^{\frac{3}{2}}}\right)$	$\mathcal{O}\left(\frac{d^{1/4}}{S^{1/4}}\right)$	$\mathcal{O}\left(\frac{d}{\epsilon^4}\right)$
ZO-SVRG-Rand (Liu et al.)	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{K}} + \frac{1}{\sqrt{\lvert\mathcal{S}_2\rvert}}\right)^{\spadesuit}$	$\mathcal{O}\left(\frac{nL}{\epsilon^2} + \frac{dL}{\epsilon^4}\right)$
ZO-SVRG-Ave (Liu et al.)	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{K}} + \frac{1}{\sqrt{\lvert\mathcal{S}_2\rvert p_{min}}}\right)^{\star}$	$\mathcal{O}\left(\frac{nqL}{\epsilon^2} + \frac{p_{max}L}{\epsilon^4}\right)^{\blacktriangleright}$
ZO-SVRG-Coord (Ji et al.)	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{K}}\right)$	$\mathcal{O}\left(\min\{\frac{dn^{2/3}L}{\epsilon^2}, \frac{dL}{\epsilon^{10/3}}\}\right)$
ZO-SPIDER-Coord (Ji et al.)	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}\lvert\mathcal{S}_1\rvert^{1/4}}{\sqrt{K}}\right)$	$\mathcal{O}\left(\min\{\frac{d\sqrt{n}L}{\epsilon^2}, \frac{dL}{\epsilon^3}\}\right)$
SPIDER-SZO (Fang et al.)	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{K}}\right)$	$\mathcal{O}\left(\min\{\frac{d\sqrt{n}L}{\epsilon^2}, \frac{dL}{\epsilon^3}\}\right)$
AdaptRdct-NC (**Ours**) (ZO-SVRG/ZO-SAGA)	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{S}}\right)$	$\tilde{\mathcal{O}}(\frac{n\sigma}{\epsilon^2} + \frac{dL}{\epsilon^2})$
AdaptRdct-NC (**Ours**) (ZO-Varag)	$\mathcal{O}\left(\frac{\epsilon^2}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{S}}\right)$	$\begin{cases}\tilde{\mathcal{O}}\left(\frac{d(n\sigma+L)}{\epsilon^2}\right), & C_1 \\ \tilde{\mathcal{O}}\left(\frac{d\sqrt{n\sigma L}}{\epsilon^2}\right), & C_2\end{cases}$

For non-convex problems, Nesterov and Spokoiny (2017) proposed ZO gradient descent (ZO-GD) algorithm. Then Ghadimi and Lan (2013) introduced its stochastic counterpart ZO-SGD. Lian et al. (2016) derived an asynchronous ZO stochastic gradient (ASZO) algorithm for parallel optimization. Gu et al. (2018a) further improved the convergence rate of ASZO by combining variance reduction technique with coordinate-wise gradient estimators. Liu et al. (2018) proposed ZO SVRG based algorithms using three different gradient estimators. Fang et al. (2018) presented a SPIDER based ZO method named SPIDER-SZO. Ji et al. (2019) further improved ZO SVRG based and SPIDER based algorithms.

3. Preliminaries

In this paper, we study the following finite-sum optimization problem

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

where $f_{i\in[n]}(\mathbf{x})$ are smooth individual loss functions. Problem (1) summarizes an extensive number of important (regularized) learning problems, such as, ℓ_2-regularized logistic regression (Conroy and Sajda, 2012), ridge regression (Shen et al., 2013), least-squares SVM (Suykens and Vandewalle, 1999).

For simplicity, we denote $||\cdot||$ to be the Euclidean norm $||\cdot||_2$, and denote $\mathbf{x}^*$ to be the optimal solution for our problem, *i.e.*, $f(\mathbf{x}^*) = \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. First we give some basic definitions as follows.

Definition 1 *For function $f : \mathbb{R}^d \to \mathbb{R}$, we have*

- *f is L-smooth with respect to the Euclidean norm if $\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it satisfies $|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle| \leq \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$.*
- *f is convex if $\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it satisfies $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle$.*
- *f is γ-strongly convex if $f(\mathbf{x}) - \frac{\gamma}{2}||\mathbf{x}||^2$ is convex,* i.e., *$\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it satisfies $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\gamma}{2}||\mathbf{y} - \mathbf{x}||^2$.*
- *f is σ-strongly non-convex if $f(\mathbf{x}) + \frac{\sigma}{2}||\mathbf{x}||^2$ is convex,* i.e., *$\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it satisfies $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle - \frac{\sigma}{2}||\mathbf{y} - \mathbf{x}||^2$.*

From Definition 1, we know that if $f(\mathbf{x})$ is L-smooth, then it is L-strongly non-convex. In consequence, for a function $f(\mathbf{x})$ which is both L-smooth and σ-strongly non-convex, we can ensure that $\sigma \leq L$. To evaluate the performance of an algorithm solving non-convex problems, we calculate its function query complexity to reach an ϵ-stationary point, which is defined as follows.

Definition 2 (ϵ-Stationary Point) *$\mathbf{x} \in \mathbb{R}^d$ is an ϵ-stationary point if $||\nabla f(\mathbf{x})|| \leq \epsilon$.*

3.1 Assumptions

From Definition 1, we list the following assumptions.

A 1 *$f_{i\in[n]}(\mathbf{x})$ is γ-strongly convex, and $f(\mathbf{x})$ is γ-strongly convex as well.*

A 2 *$f_{i\in[n]}(\mathbf{x})$ is convex, and $f(\mathbf{x})$ is convex as well.*

A 3 *$f_{i\in[n]}(\mathbf{x})$ is σ-strongly non-convex, and $f(\mathbf{x})$ is σ-strongly non-convex as well.*

A 4 *$f_{i\in[n]}(\mathbf{x})$ is L-smooth, and $f(\mathbf{x})$ is L-smooth as well.*

3.2 ZO Gradient Estimation

Given an individual problem $f_i(\mathbf{x})$, we use the following two-point gradient estimator with random direction (Gao et al., 2018), which is abbreviated as *random direction estimator*:

$$\hat{\nabla} f_i(\mathbf{x}) = \frac{d}{\mu}[f_i(\mathbf{x} + \mu\mathbf{u}) - f_i(\mathbf{x})]\mathbf{u}. \tag{2}$$

and the full gradient estimator with random direction is given by

$$\hat{\nabla} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_i(\mathbf{x}) \tag{3}$$

where μ is the smoothing parameter, $\mathbf{u} \in \mathbb{R}^d$ and $\{\mathbf{u}\}$ are i.i.d. random directions drawn from a uniform distribution over a unit sphere. In general, $\hat{\nabla} f_i(\mathbf{x})$ and $\hat{\nabla} f(\mathbf{x})$ are biased approximation of the true gradients, $\nabla f_i(\mathbf{x})$ and $\nabla f(\mathbf{x})$. It is clear that the bias is determined by smoothing parameter μ and will reduce as μ approaches zero. Ji et al. (2019) introduced another gradient estimator by setting the direction $\mathbf{u}$ in each coordinate, which is abbreviated as *coordinate-wise estimator*:

$$\hat{\nabla}_{coord} f_i(\mathbf{x}) = \sum_{j=1}^{d} \frac{f_i(\mathbf{x} + \mu\mathbf{e}_j) - f_i(\mathbf{x} - \mu\mathbf{e}_j)}{2\mu}\mathbf{e}_j. \tag{4}$$

where $\mathbf{e}_j$ denotes a standard basis vector with 1 at its j-th coordinate and 0 otherwise. The coordinate-wise estimator is more stable than random direction estimator, but it requires more function queries than the random direction estimator. Conversely, random direction estimator requires less function queries, but it introduces much higher error. In this paper, we mainly discuss random direction estimator. The coordinate-wise estimator only appears when we apply our frameworks to ZO-Varag in Section 5 since we use the convergence results derived by Chen et al. (2020), which are based on the coordinate-wise estimator.

4. Zeroth-Order Reduction Frameworks for Convex and Non-convex Problems

Recently, Allen-Zhu and Hazan (2016); Chen et al. (2018) proposed first-order reduction methods for solving convex and non-convex problems, respectively. Both methods add an quadratic regularizer to the original problem to ensure strong convexity. Then they call a first-order oracle to optimize the new strongly convex objective. Their theoretic analyses show that, in comparison to directly optimizing the original problem, algorithms optimizing the new problem leads to a better performance. Enlightened by their methods, we propose two ZO reduction frameworks (*i.e.*, *AdaptRdct-C* and *AdaptRdct-NC*) for solving convex and non-convex problems, respectively. Before presenting the two reduction frameworks, we first define a property on the ZO oracle algorithm.

Definition 3 (ZOOD Property) *We say an algorithm $\mathcal{A}(f, \mathbf{x})$ solving problem(1) satisfies the ZO Objective Decrease (ZOOD) property with complexity $\mathcal{C}(L, \gamma)$ if, for any starting*

point $\mathbf{x}$*, it produces an output* $\mathbf{x}' \leftarrow \mathcal{A}(f, \mathbf{x})$ *with function query complexity* $\mathcal{C}(L, \gamma)$*, such that*

$$\mathbb{E}\left[f(\mathbf{x}') - \min_{\mathbf{x}} f(\mathbf{x}) - \delta_\mu\right] \leq \mathcal{K}_0 \left[f(\mathbf{x}) - \min_{\mathbf{x}} f(\mathbf{x}) - \delta_\mu\right] \tag{5}$$

where δ_μ *is a fixed error introduced by smoothing parameter* μ *and* $\mathcal{K}_0 \in (0, 1)$*.*

4.1 *AdaptRdct-C*: Transforming Convex Problems into Strongly Convex Problems

In this subsection, we focus on solving problem (1) under convex setting, and propose *AdaptRdct-C*. We summarize *AdaptRdct-C* in Algorithm 1. Specifically, *AdaptRdct-C* consists of S epochs. At s-th epoch where $s = 1, ..., S$, we define a γ_s-strongly convex problem $f^{(s)}(\mathbf{x}) \stackrel{\text{def}}{=} f(\mathbf{x}) + \frac{\gamma_s}{2}||\mathbf{x} - \mathbf{x}_0||^2$, where $\mathbf{x}_0$ is a starting point. The parameter γ_s decreases at an exponential rate, *i.e.*, $\gamma_s = \sqrt{\mathcal{K}}\gamma_{s-1}$ for each $s \geq 1$. γ_1 is a specified parameter. In each epoch, we run a ZOOD algorithm $\mathcal{A}$ on $f^{(s)}(\mathbf{x})$ with starting point $\mathbf{x}_{s-1}$ which outputs a solution $\mathbf{x}_s$. After all epochs are finished, *AdaptRdct-C* outputs the latest result $\mathbf{x}_S$. Theorem 1 gives the convergence property of *AdaptRdct-C*. The proof is provided in the appendix.

Algorithm 1 *AdaptRdct-C* Framework

Input: Starting vector $\mathbf{x}_0$, epoch S, regularization parameter γ_1, ZOOD parameter $\mathcal{K}_0$

for $s = 1, ..., S$ **do**

$f^{(s)}(\mathbf{x}) \stackrel{\text{def}}{=} f(\mathbf{x}) + \frac{\gamma_s}{2}||\mathbf{x} - \mathbf{x}_0||^2$.

$\mathbf{x}_s \leftarrow \mathcal{A}(f^{(s)}, \mathbf{x}_{s-1})$ such that

$$\mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - \min_{\mathbf{x}} f^{(s)}(\mathbf{x}) - \delta_\mu\right] \leq \mathcal{K}_0 \left[f^{(s)}(\mathbf{x}_{s-1}) - \min_{\mathbf{x}} f^{(s)}(\mathbf{x}) - \delta_\mu\right]$$

$\gamma_s = \sqrt{\mathcal{K}_0}\gamma_{s-1}$.

end for

Output: $\mathbf{x}_S$.

Theorem 1 *Suppose Assumption 2 holds. Let* $\mathbf{x}_0$ *be a starting vector such that* $f(\mathbf{x}_0) - f(\mathbf{x}^*) \leq \Delta$*, and* $||\mathbf{x}_0 - \mathbf{x}^*||^2 \leq \Theta$*. For Algorithm1, we have*

$$\mathbb{E}[f(\mathbf{x}_S) - f(\mathbf{x}^*)] \leq \delta_\mu + \mathcal{K}_0^S[\Delta - \delta_\mu] + \left(\frac{1}{2} + \frac{2}{\sqrt{\mathcal{K}_0}}\right)\mathcal{K}_0^{\frac{S}{2}}\gamma_1\Theta \tag{6}$$

From the ZOOD property as defined in Definition 3, we can obtain the following corollary.

Corollary 1 *By applying* AdaptRdct-C*, the total function query complexity to solve the finite-sum problem is* $\sum_{s=1}^{S}\mathcal{C}(L, \gamma_s)$ *with* $S = \mathcal{O}\left(\log\frac{1}{\epsilon}\right)$*.*

In Section 5, we apply *AdaptRdct-C* on three ZO algorithms, which are ZO-SVRG, ZO-SAGA and ZO-Varag. Through theoretical analyses, we show that the *AdaptRdct-C* variants of ZO-SVRG and ZO-SAGA have lower function query complexities than their

original counterparts. Also, we show that the function query complexity of the *AdaptRdct-C* variant of ZO-Varag matches that of vanilla ZO-Varag.

4.2 *AdaptRdct-NC*: Transforming Non-convex Problems into Strongly Convex Problems

In this subsection, we focus on the reduction method for solving non-convex problems. Different from convex optimization, the convergence of non-convex optimization is usually identified by the ϵ-stationary points (*i.e.*, $||\nabla f(\mathbf{x})|| \leq \epsilon$). To propose our analysis, we introduce the following concepts.

Definition 4 (Moreau Envelope and Proximal Mapping) *For any function f and $\lambda > 0$, the following function is called a Moreau envelope of f*

$$f_\lambda(\mathbf{x}) = \min_{\mathbf{z}} f(\mathbf{z}) + \frac{1}{2\lambda}||\mathbf{z}-\mathbf{x}||^2 \tag{7}$$

Further, the optimal solution to the above problem is denoted as

$$Prox_{\lambda f}(\mathbf{x}) = \arg\min_{\mathbf{z}} f(\mathbf{z}) + \frac{1}{2\lambda}||\mathbf{z}-\mathbf{x}||^2 \tag{8}$$

It is known that $\nabla f_\lambda(\mathbf{x}) = \frac{\mathbf{x}-\mathrm{Prox}_{\lambda f}(\mathbf{x})}{\lambda}$ (see e.g. (Chen et al., 2019)). Also, for any $\mathbf{x} \in \mathbb{R}^d$, denote $\mathbf{x}^+ \overset{\text{def}}{=} Prox_{\lambda f}(\mathbf{x})$, we have

$$\begin{cases} f(\mathbf{x}^+) \leq f(\mathbf{x}) \\ \|\mathbf{x}-\mathbf{x}^+\| = \lambda\|\nabla f_\lambda(\mathbf{x})\| \\ \|\nabla f(\mathbf{x}^+)\| \leq \|\nabla f_\lambda(\mathbf{x})\| \end{cases} \tag{9}$$

Thus a point $\mathbf{x}$ satisfying $\|\nabla f_\lambda(\mathbf{x})\| \leq \epsilon$ is close to an ϵ-stationary point of f in distance of $O(\lambda\epsilon)$.

To solve a σ-strongly non-convex problem, *AdaptRdct-NC* works as follows (please see Algorithm 2). It consists of S epochs, at the beginning of each epoch $s = 1, 2, ..., S$, we define a σ-strongly convex problem $f^{(s)}(\mathbf{x}) \overset{\text{def}}{=} f(\mathbf{x}) + \sigma||\mathbf{x}-\mathbf{x}_{s-1}||^2$, where $\mathbf{x}_{s-1}$ is the output of the ZOOD algorithm of last epoch and σ is the non-convex parameter of the problem $f(\mathbf{x})$. In each epoch, we call a ZOOD algorithm $\mathcal{A}$ to solve $f^{(s)}(\mathbf{x})$ with starting point $\mathbf{x}_{s-1}$, and obtain the output $\mathbf{x}_s$ such that

$$\mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - \min_{\mathbf{x}} f^{(s)}(\mathbf{x}) - \delta_\mu\right] \leq \frac{\sigma^2\mathcal{K}_0}{L^2 s}\left[f^{(s)}(\mathbf{x}_{s-1}) - \min_{\mathbf{x}} f^{(s)}(\mathbf{x}) - \delta_\mu\right] \tag{10}$$

After all epochs are finished, *AdaptRdct-NC* outputs a randomly chosen result $\mathbf{x}_{\alpha+1}$ with α chosen from $0, ..., S$ with probability $p_\alpha = \alpha^\tau / \sum_{s=0}^{S-1} s^\tau$. The convergence property of *AdaptRdct-NC* is given in Theorem 2, where the proof is provided in the appendix.

Theorem 2 *Suppose Assumption 3 holds. Let $\mathbf{x}_0$ be a starting vector such that $f(\mathbf{x}_0) - f(\mathbf{x}^*) \leq \Delta$. For Algorithm 2, we have*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\alpha+1})\|^2\right] \leq \frac{16\sigma(2\mathcal{K}+1)}{(1-\mathcal{K})S}[f(\mathbf{x}_0) - f(\mathbf{x}^*)] + \left(\frac{64-16\mathcal{K}}{1-\mathcal{K}} + 4L^2\right)\sigma\delta_\mu$$

where $\mathcal{K} = \frac{\sigma^2\mathcal{K}_0}{L^2}$.

Algorithm 2 *AdaptRdct-NC* Framework

Input: Starting vector $\mathbf{x}_0$, epoch S, regularization parameter $\lambda = \frac{1}{2\sigma}$, ZOOD parameter $\mathcal{K}_0$, weight parameter $\tau > 0$.

for $s = 1, ..., S+1$ **do**

$f^{(s)}(\mathbf{x}) \stackrel{\text{def}}{=} f(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{x}_{s-1}\|^2$

$\mathbf{x}_s \leftarrow \mathcal{A}(f^{(s)}, \mathbf{x}_{s-1})$ such that

$$\mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - \min_{\mathbf{x}} f^{(s)}(\mathbf{x}) - \delta_\mu\right] \leq \frac{\sigma^2\mathcal{K}_0}{L^2 s}\left[f^{(s)}(\mathbf{x}_{s-1}) - \min_{\mathbf{x}} f^{(s)}(\mathbf{x}) - \delta_\mu\right]$$

end for

Output: $\mathbf{x}_{\alpha+1}$, where α is chosen from $0, ..., S$ with probability $p_\alpha = \alpha^\tau / \sum_{s=0}^{S-1} s^\tau$.

Note that the coefficient in (10) is $\frac{\sigma^2\mathcal{K}_0}{L^2 s}$, while the coefficient in *AdaptRdct-C* (Algorithm 3) is $\mathcal{K}_0$. In fact, if the coefficient in (10) is set to $\mathcal{K}_0$, then only the convergence of $\mathbb{E}\left[\|\nabla f_\lambda(\mathbf{x}_\alpha)\|^2\right]$ can be guaranteed (cf. the proof of Theorem 2). From (9) we can calculate $\mathbf{x}_\alpha^+ = Prox_\lambda f(\mathbf{x}_\alpha)$ to guarantee the convergence of $\mathbb{E}\left[\|\nabla f(\mathbf{x}_\alpha^+)\|^2\right]$, which is a measure of ϵ-stationary point. However, this calculation is impractical in zeroth-order optimization since we only have access to the function value. The choice of the coefficient in (10) helps us directly guarantee the convergence of $\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\alpha+1})\|^2\right]$, but this comes with an extra logarithmic cost in the function query complexity. To be specific, ZOOD property (3) claims that a zeroth-order algorithm produces an output $\mathbf{x}_s \leftarrow \mathcal{A}(f^{(s)}, \mathbf{x}_{s-1})$ with function query complexity $\mathcal{C}(L, \sigma)$, such that

$$\mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - \min_{\mathbf{x}} f^{(s)}(\mathbf{x}) - \delta_\mu\right] \leq \mathcal{K}_0\left[f^{(s)}(\mathbf{x}_{s-1}) - \min_{\mathbf{x}} f^{(s)}(\mathbf{x}) - \delta_\mu\right]$$

Then it is evident that the algorithm produces $\mathbf{x}_s$ with function query complexity $\mathcal{C}(L, \sigma)\log\frac{L^2 s}{\sigma^2}$ satisfying (10). Then we get the following corollary.

Corollary 2 *By applying* AdaptRdct-NC*, the total function query complexity to find an ϵ-stationary point is $\mathcal{O}\left(\mathcal{C}(L,\sigma)\sum_{s=1}^{S}\log\frac{L^2 s}{\sigma^2}\right)$ with $S = \mathcal{O}\left(\frac{\sigma}{\epsilon^2}\right)$. Specifically, we have*

$$\begin{aligned}\sum_{s=1}^{S}\log\frac{L^2 s}{\sigma^2} = \sum_{s=1}^{S}\left(\log\frac{L^2}{\sigma^2} + \log s\right) =& S\log\frac{L^2}{\sigma^2} + \log S \\ =& S\log\frac{L^2}{\sigma^2} + S\log S - S + \mathcal{O}\left(\log S\right)\end{aligned}$$

where the last equality comes from the Stirling's formula. Then the total function query complexity to find an ϵ-stationary point is $\tilde{\mathcal{O}}\left(\frac{\sigma\mathcal{C}(L,\sigma)}{\epsilon^2}\right)$, where $\tilde{\mathcal{O}}$ hides a logarithmic factor.

The quadratic terms in our two frameworks are different from each other. In *AdaptRdct-C*, the coefficient of the quadratic term (*i.e.*, γ_s) diminishes at an exponential rate. This feature contributes to the linear convergence of *AdaptRdct-C* for convex problems (cf. the proof of Theorem 1). The quadratic term is set to $\|\mathbf{x}-\mathbf{x}_0\|^2$ since it can be transformed into

another term that is much easier to deal with, *i.e.*, $\|\mathbf{x}_0-\mathbf{x}^*\|^2$, where $\mathbf{x}^*$ denotes the optimal solution to the original problem. Although it derives excellent results for convex problems, *AdaptRdct-C* is not applicable to non-convex problems. From Definition 1, it is clear that for a σ-strongly non-convex problem, a coefficient greater than $\sigma/2$ is necessary to ensure strong convexity. So *AdaptRdct-C* does not guarantee strong convexity for non-convex problems. For a σ-strongly non-convex problem, a natural idea is to fix the coefficient to a constant that is greater than $\sigma/2$. Thus in *AdaptRdct-NC*, the coefficient is set to σ. And the quadratic term is set to $\|\mathbf{x}-\mathbf{x}_{s-1}\|^2$ because it can be transformed into $\|\mathbf{x}_{s-1}-\mathbf{x}_s^*\|^2/\lambda$, which equals $\|\nabla f_\lambda(\mathbf{x}_{s-1})\|^2$, which is a key element in the proof since it is related to our convergence measure $\|\nabla f(\mathbf{x})\|^2$.

In Section 5, we apply *AdaptRdct-NC* on ZO-SVRG, ZO-SAGA and ZO-Varag, and present their convergence properties and function query complexities through theoretical analyses.

5. Applications of Reduction Frameworks on Zeroth-order Algorithms

The key idea of *AdaptRdct-C* and *AdaptRdct-NC* is to transform convex or non-convex problems into a strongly convex problem and call a ZO algorithm to solve the new problem. To keep our paper self-contained and evaluate our methods' performance, in this section, we give convergence analyses of three ZO variance reduced algorithms, and apply our reduction frameworks on them. Specifically, we consider the ZO versions of SVRG (Johnson and Zhang, 2013), SAGA (Defazio et al., 2014) and Varag (Lan et al., 2019), and provide their convergence rates along with function query complexities under both of the convex and strongly convex settings. To the best of our knowledge, we are the first to provide the convergence rates of ZO-SVRG and ZO-SAGA under both of the convex and strongly convex conditions. We apply *AdaptRdct-C* and *AdaptRdct-NC* on them, then compare our new methods with vanilla ZO-SVRG, ZO-SAGA and ZO-Varag. The results show that our *AdaptRdct-C* and *AdaptRdct-NC* greatly lower their query complexities for ZO-SVRG and ZO-SAGA.

5.1 Zeroth-Order SVRG Algorithm

Johnson and Zhang (2013) proposed first-order stochastic variance reduced gradient (SVRG) algorithm. The key step of SVRG is to generate a snapshot (denoted as $\tilde{\mathbf{x}}$) of $\mathbf{x}$ after a certain number of iterations, and the full gradient at $\tilde{\mathbf{x}}$ is used to build a modified stochastic gradient estimation, which is a gradient blending $\mathbf{v}=\nabla f_i(\mathbf{x})-\nabla f_i(\tilde{\mathbf{x}})+\nabla f(\tilde{\mathbf{x}})$, where $\mathbf{v}$ denotes the gradient estimate at $\mathbf{x}$, $i\in[n]$ is chosen uniformly randomly, and $\nabla f(\tilde{\mathbf{x}})=\frac{1}{n}\sum_{j=1}^n\nabla f_j(\tilde{\mathbf{x}})$. Under ZO setting, we use the following gradient estimator to approximate the gradient blending

$$\hat{\mathbf{v}}=\hat{\nabla} f_i(\mathbf{x})-\hat{\nabla} f_i(\tilde{\mathbf{x}})+\hat{\nabla} f(\tilde{\mathbf{x}}). \tag{11}$$

Recall from Section 3 that $\hat{\nabla} f_i(\mathbf{x})=\frac{d}{\mu}[f_i(\mathbf{x}+\mu\mathbf{u})-f_i(\mathbf{x})]\mathbf{u}$ and $\hat{\nabla} f(\mathbf{x})=\frac{1}{n}\sum_{j=1}^n\hat{\nabla} f_j(\mathbf{x})$. We present ZO-SVRG in Algorithm 3. Next we give the convergence property of ZO-SVRG for solving strongly convex problems, which shows that it can easily satisfy the ZOOD property.

Algorithm 3 ZO-SVRG Algorithm

Input: Stepsize η, update frequency m, epoch S, starting vector $\mathbf{x}_m^0 = \tilde{\mathbf{x}}_0 \in \mathbb{R}^d$.
  for $s = 1, 2, ..., S$ **do**
    Option I : $\mathbf{x}_0^s = \tilde{\mathbf{x}}_{s-1}$;
    Option II : $\mathbf{x}_0^s = \mathbf{x}_m^{s-1}$.
    Compute $\hat{\mathbf{g}}_s = \hat{\nabla} f(\tilde{\mathbf{x}}_{s-1})$.
    for $k = 0, 1, ..., m-1$ **do**
      Uniformly randomly choose $i_k \in [n]$ and $\mathbf{u}_k^s$ from a unit sphere.
      $\hat{\mathbf{v}}_k^s = \hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1}) + \hat{\mathbf{g}}_s$.
      $\mathbf{x}_{k+1}^s = \mathbf{x}_k^s - \eta \hat{\mathbf{v}}_k^s$.
    end for
    Set $\tilde{\mathbf{x}}_s = \mathbf{x}_k^s$ for randomly chosen $k \in \{0, 1, ..., m-1\}$.
  end for

Theorem 3 (Strongly Convex) *Suppose Assumptions 1 and 4 hold, denote* $f(\tilde{\mathbf{x}}_0) - f(\mathbf{x}^*) = \Delta$. *By using* ***Option I*** *in Algorithm 3, we have*

$$\mathbb{E}[f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*)] \leq \delta_\mu + \left(\frac{\beta_2}{\beta_1}\right)^s (\Delta - \delta_\mu) \tag{12}$$

where $\beta_1 = 2m\eta[1 - 24\eta dL]$, $\beta_2 = \frac{2}{\gamma} + 48m\eta^2 dL$, $\delta_\mu = \frac{2\eta m\mu^2 L(2\eta d^2 L+1)}{\beta_1 - \beta_2}$ *and* η, m *satisfy inequalities* $\eta < \frac{1}{48dL}$ *and* $m > \frac{1}{\gamma\eta(1-48\eta dL)}$.

Remark 1 *From Theorem 3, take* $S = 1$, *we have*

$$\mathbb{E}\left[f(\tilde{\mathbf{x}}_1) - f(\mathbf{x}^*) - \delta_\mu\right] \leq \frac{\beta_2}{\beta_1}\left[f(\tilde{\mathbf{x}}_0) - f(\mathbf{x}^*) - \delta_\mu\right] \tag{13}$$

then we know that our ZO-SVRG algorithm satisfies the ZOOD property after running one epoch, with function query complexity $\mathcal{O}\left(n + \frac{dL}{\gamma}\right)$, $\mathcal{K}_0 = \frac{\beta_2}{\beta_1}$ *and* $\delta_\mu = \frac{2\eta m\mu^2 L(2\eta d^2 L+1)}{\beta_1 - \beta_2}$.

Corollary 3 *Under strongly convex setting, if we take stepsize* $\eta = \frac{1}{112dL}$ *and* $m = \frac{896dL}{3\gamma}$, *then we have* $\frac{\beta_2}{\beta_1} = 0.75$, *and ZO-SVRG Algorithm has a convergence rate of* $\mathcal{O}\left(\frac{3}{4}\right)^S$ *and a function query complexity of* $\mathcal{O}\left((n+d)\log\frac{1}{\epsilon}\right)$.

For the purpose of comparison, we provide the convergence result of ZO-SVRG under the convex setting (Theorem 4). Based on Theorem 4, we provide the function query complexity of ZO-SVRG for solving convex problems (Corollary 4), where reduction is not used.

Theorem 4 (Convex) *Suppose Assumptions 2 and 4 hold,* $f(\tilde{\mathbf{x}}_0) - f(\mathbf{x}^*) \leq \Delta$ *and* $||\tilde{\mathbf{x}}_0 - \mathbf{x}^*||^2 \leq \Theta$. *Using* ***Option II*** *in Algorithm 3, we have*

$$\mathbb{E}[f(\mathbf{x}_\alpha) - f(\mathbf{x}^*)] \leq \frac{\Theta + 48m\eta^2 dL\Delta}{2mS\eta(1 - 48\eta dL)} + \frac{\mu^2 L(4\eta d^2 L + 1)}{1 - 48\eta dL} \tag{14}$$

where $\mathbf{x}_\alpha$ *is uniformly randomly chosen from* $\{\{\mathbf{x}_k^s\}_{k=0}^{m-1}\}_{s=1}^S$, *and* $\eta < \frac{1}{48dL}$.

Corollary 4 *Under convex setting, ZO-SVRG has a convergence rate of* $\mathcal{O}\left(\frac{d}{S}\right)$ *and a function query complexity of* $\mathcal{O}\left(\frac{n+d}{\epsilon}\right)$.

5.2 Zeroth-Order SAGA Algorithm

Defazio et al. (2014) proposed first-order SAGA algorithm. The key step of SAGA is to keep a table of gradients of previous results $\nabla f(\phi_{i\in[n]})$. The modified stochastic gradient estimate is $\mathbf{v} = \nabla f_i(\mathbf{x}) - \nabla f_i(\phi_i) + g(\phi)$, where $\mathbf{v}$ denotes the gradient estimate at $\mathbf{x}$, $i \in [n]$ is chosen uniformly randomly, and $g(\phi) = \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\phi_j)$. Under ZO setting, we use the following gradient estimator to approximate the gradient blending

$$\hat{\mathbf{v}} = \hat{\nabla} f_i(\mathbf{x}) - \hat{\nabla} f_i(\phi_i) + \hat{g}(\phi) \tag{15}$$

where $\hat{g}(\phi) = \frac{1}{n}\sum_{j=1}^{n} \hat{\nabla} f_j(\phi_j)$.

Algorithm 4 ZO-SAGA Algorithm

Input: Stepsize η, iteration K, starting vector $\mathbf{x}^0 \in R^d$, auxiliary vectors $\{\phi_i^0\}_{i=1}^n$ with $\phi_i^0 = x^0$ for each i.
Compute $\hat{\mathbf{g}}^0 = \hat{\nabla} f(\mathbf{x}^0) = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_i(\phi_i^0)$.
for $k = 0, 1, 2, ..., K-1$ **do**
  Uniformly randomly choose $i_k \in [n]$ and $\mathbf{u}^k$ from a unit sphere.
  $\hat{\mathbf{v}}^k = \hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k) + \hat{\mathbf{g}}^k$.
  $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \hat{\mathbf{v}}^k$.
  $\phi_{i_k}^{k+1} = \mathbf{x}^k$ and $\hat{\mathbf{g}}^{k+1} = \hat{\mathbf{g}}^k + \frac{1}{n}\left(\hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k)\right)$
end for

Next we give the convergence property of ZO-SAGA for solving strongly convex problems, which shows that it naturally satisfies the ZOOD property.

Theorem 5 (Strongly Convex) *Suppose assumptions 1 and 4 hold, denote* $f(\mathbf{x}^0) - f(\mathbf{x}^*) = \Delta$. *We have*

$$\mathbb{E}\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \leq \delta_\mu + \left(1 - \frac{\gamma}{112dL + n\gamma}\right)^K \left[\frac{L(2+c\gamma)}{2\gamma}\Delta - \delta_\mu\right] \tag{16}$$

where $c = \eta n(1 - 32\eta dL)$, $\eta = \frac{2}{112dL+n\gamma}$ *and* $\delta_\mu = \frac{2L^2\mu^2(2\eta L d^2+1)}{\gamma}$.

Remark 2 *From Theorem 3, we know that ZO-SAGA satisfies ZOOD property with function query complexity* $\mathcal{O}\left((n + \frac{dL}{\gamma})\log\frac{L}{\gamma}\right)$, $\mathcal{K}_0 = \frac{1}{e}$ *and* $\delta_\mu = \frac{2L^2\mu^2(2\eta L d^2+1)}{\gamma}$.

For the purpose of comparison, we provide the convergence result of ZO-SAGA under the convex setting (Theorem 6). Based on Theorem 6, we provide the function query complexity of ZO-SAGA for solving convex problems (Corollary 5), where reduction is not used.

Theorem 6 (Convex) *Suppose Assumptions 2 and 4 hold. For Algorithm 4, we have*

$$\mathbb{E}\left[f(\mathbf{x}^\tau) - f(\mathbf{x}^*)\right] \leq \frac{56dLT^0}{K} + 112\eta L\mu^2(2\eta Ld^2 + 1) \tag{17}$$

where $\eta = \frac{1}{56dL}, T^0 = ||\mathbf{x}^0 - \mathbf{x}^*||^2 + \frac{3}{7}\eta n\left[f(\mathbf{x}^0) - f(\mathbf{x}^*)\right]$ *and* $\mathbf{x}^\tau$ *is uniformly randomly chosen from* $\left\{\mathbf{x}^k\right\}_{k=1}^K$.

Corollary 5 *Under convex setting, ZO-SAGA has a convergence rate of* $\mathcal{O}\left(\frac{d}{K}\right)$ *and a function query complexity of* $\mathcal{O}\left(\frac{n+d}{\epsilon}\right)$.

5.3 Zeroth-order Varag Algorithm

Lan et al. (2019) proposed first-order Varag algorithm. Chen et al. (2020) introduced ZO-Varag, zeroth-order version of Varag. ZO-Varag is an accelerated algorithm which maintains three acceleration sequences (*i.e.*, $\{\underline{\mathbf{x}}\}$, $\{\mathbf{x}\}$ and $\{\bar{\mathbf{x}}\}$). It is also a variance reduced algorithm which leverages the same technique as SVRG. We present ZO-Varag in Algorithm 5.

Algorithm 5 ZO-Varag Algorithm

Input: Starting vector $\mathbf{x}^0 \in \mathbb{R}^d$, Epoch S, update frequency $\{m_s\}$, $\{\beta_s\}$, $\{\alpha_s\}$, $\{p_s\}$, $\{\theta_k\}$
  Set $\tilde{\mathbf{x}}^0 = \bar{\mathbf{x}}^0 = \mathbf{x}^0$.
  for $s = 1, 2, ..., S$ **do**
    Set $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_{s-1}$;
    Set $\mathbf{x}_0 = \mathbf{x}^{s-1}$, $\bar{\mathbf{x}}_0 = \tilde{\mathbf{x}}$.
    Compute $\hat{\mathbf{g}}_s = \hat{\nabla}_{coord} f(\tilde{\mathbf{x}})$.
    for $k = 1, 2, ..., m_s$ **do**
      $\underline{\mathbf{x}}_k = \left[(1+\gamma\beta_s)(1-\alpha_s-p_s)\bar{\mathbf{x}}_{k-1} + \alpha_s\mathbf{x}_{k-1} + (1+\gamma\beta_s)p_s\tilde{\mathbf{x}}\right] / \left[1+\gamma\beta_s(1-\alpha_s)\right]$
      Uniformly randomly choose $i_k \in [n]$
      $\hat{\mathbf{v}}_k^s = \hat{\nabla}_{coord} f_{i_k}(\underline{\mathbf{x}}_k) - \hat{\nabla}_{coord} f_{i_k}(\tilde{\mathbf{x}}_{s-1}) + \hat{\mathbf{g}}_s$.
      $\mathbf{x}_k = \left[\beta_s\gamma\underline{\mathbf{x}}_k + \mathbf{x}_{k-1} - \beta_s\hat{\mathbf{v}}_k^s\right] / \left[1+\gamma\beta_s\right]$.
      $\bar{\mathbf{x}}_k = (1-\alpha_s-p_s)\bar{\mathbf{x}}_{k-1} + \alpha_s\mathbf{x}_k + p_s\tilde{\mathbf{x}}$.
    end for
    Set $\mathbf{x}^s = \mathbf{x}_{m_s}$, $\bar{\mathbf{x}}^s = \bar{\mathbf{x}}_{m_s}$, and $\tilde{\mathbf{x}}^s = \sum_{k=1}^{m_s}(\theta_k\bar{\mathbf{x}}_k)/(\sum_{k=1}^{m_s}\theta_k)$.
  end for

Note that we present the algorithm with notations different from that in (Chen et al., 2020) to avoid ambiguity. Specifically, we use $\{\beta_s\}$ to replace $\{\gamma_s\}$ in their paper, $\{m_s\}$ to replace $\{T_s\}$, k to replace t and $\hat{\mathbf{v}}_k^s$ to replace G_t. The next assumption is the same as Assumption $A2_\nu$ in (Chen et al., 2020).

Assumption 5 *Let* $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. *For any epoch* s *of Algorithm 5, consider the inner-loop sequences* $\{\underline{\mathbf{x}}_k\}$ *and* $\{\bar{\mathbf{x}}_k\}$. *There exist a finite constant* $Z < \infty$, *potentially dependent on* L *and* d, *such that, for* μ *small enough,*

$$\sup_{s\geq 0}\max_{\mathbf{x}\in\{\underline{\mathbf{x}}_k\}\cup\{\bar{\mathbf{x}}_k\}}\mathbb{E}\left[\|\mathbf{x}-\mathbf{x}^*\|\right] \leq Z$$

The following Theorem corresponds to Theorem 8 in (Chen et al., 2020). We make minor modifications in the parameter setting so that it satisfies our ZOOD property. Note that Chen et al. (2020) set a threshold $s_0 \overset{\text{def}}{=} \lfloor \log n \rfloor + 1$. When $s > s_0$, their parameter setting is exactly the same as ours; when $s \leq s_0$, they chose a different parameter setting. We choose the following parameter setting because m_s is set to n, so that the algorithm satisfies our ZOOD property (which can be seen from the proof). Theorem 7 can be directly derived from Lemmas 8 and 9.

Theorem 7 (Strongly Convex) *Suppose Assumptions 1, 4 and 5 hold. Set*

$$m_s = n, \quad p_s = \frac{1}{2}, \quad \alpha_s = \min\left\{\sqrt{\frac{n\gamma}{12L}}, \frac{1}{2}\right\}, \quad \beta_s = \frac{1}{12L\alpha_s}$$

$$\Gamma_k = (1+\gamma\beta_s)^k, \qquad \theta_k = \begin{cases} \Gamma_{k-1} - (1-\alpha_s-p_s)\Gamma_k, & k \leq m_s - 1 \\ \Gamma_{k-1}, & k = m_s \end{cases}$$

We obtain

$$\begin{aligned} &\mathbb{E}\left[f(\tilde{\mathbf{x}}^S) - f(\mathbf{x}^*)\right] \\ \leq &\begin{cases} \left(\frac{1}{4}\right)^S \frac{4}{3}\mathbb{E}\left[f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*)\right] + \frac{1}{2}\mu^2 Ld + \frac{4}{3}L\sqrt{d}Z\mu, & n \geq \frac{18L}{\gamma} \\ \left(1+\sqrt{\frac{\gamma}{12nL}}\right)^{-nS} 3\mathbb{E}\left[f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*)\right] + \frac{S}{\Gamma_n}\left[\frac{3}{2}\mu^2 Ld + (4-2\alpha_s)L\sqrt{d}Z\mu\right], & n < \frac{18L}{\gamma} \end{cases} \end{aligned}$$

Remark 3 *From Theorem 7, we know that ZO-Varag satisfies ZOOD property with*

$$\begin{cases} FQC = \mathcal{O}(dn) \qquad \text{and } \mathcal{K}_0 = \frac{1}{4} \text{ and } \delta_\mu = \frac{2}{3}\mu^2 Ld + \frac{16}{9}L\sqrt{d}Z\mu, & n \geq \frac{18L}{\gamma} \\ FQC = \mathcal{O}\left(d\sqrt{\frac{nL}{\gamma}}\right) \text{and } \mathcal{K}_0 = \frac{1}{e} \text{ and } \delta_\mu = \sqrt{\frac{L}{n\gamma}}\left[\frac{3}{2}\mu^2 Ld + 4L\sqrt{d}Z\mu\right], & n < \frac{18L}{\gamma} \end{cases}$$

where FQC denotes function query complexity.

5.4 Applying Reduction Methods to ZO-SVRG, ZO-SAGA and ZO-Varag

With the above results, now we can present the function query complexity of applying *AdaptRdct-C* and *AdaptRdct-NC* on ZO-SVRG, ZO-SAGA and ZO-Varag:

Corollary 6 *Suppose Assumptions 2, 4 and 5 hold. Applying* AdaptRdct-C *on ZO-SVRG, from Corollary 1 and Remark 1 we know that the total function query complexity to solve the finite-sum problem is* $\mathcal{O}\left(n\log\frac{1}{\epsilon} + \frac{d}{\epsilon}\right)$. *Applying* AdaptRdct-C *on ZO-SAGA, from Corollary 1 and Remark 2 we know that the total function query complexity to solve the finite-sum problem is* $\mathcal{O}\left(\left(n\log\frac{1}{\epsilon} + \frac{d}{\epsilon}\right)\log\frac{1}{\epsilon^2}\right)$. *Applying* AdaptRdct-C *on ZO-Varag, from Corollary 1 and Remark 3 we know that the total function query complexity to solve the finite-sum problem is* $\begin{cases} \mathcal{O}\left(dn\log\frac{1}{\epsilon}\right), & n \geq \mathcal{O}\left(\frac{1}{\epsilon}\right) \\ \mathcal{O}\left(dn\log n + d\sqrt{\frac{n}{\epsilon}}\right), & n < \mathcal{O}\left(\frac{1}{\epsilon}\right) \end{cases}$

Corollary 6 entails the superiority of our reduction method *AdaptRdct-C* in efficiency: from Corollaries 4 and 5 above, we know that directly applying ZO-SVRG and ZO-SAGA

to solve convex problems yields function query complexity $\mathcal{O}\left(\frac{n+d}{\epsilon}\right)$ and $\mathcal{O}\left(\frac{nd}{\epsilon}\right)$ respectively. After applying *AdaptRdct-C*, we reduce the term $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ to $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ for ZO-SVRG and shave the factor n for ZO-SAGA. Furthermore, from Table 1 we know that vanilla ZO-Varag has a function query complexity $\begin{cases}\mathcal{O}\left(dn\log\frac{1}{\epsilon}\right), & n\geq\mathcal{O}\left(\frac{1}{\epsilon}\right)\\ \mathcal{O}\left(dn\log n+d\sqrt{\frac{n}{\epsilon}}\right), & n<\mathcal{O}\left(\frac{1}{\epsilon}\right)\end{cases}$. The result of ZO-Varag with *AdaptRdct-C* completely matches the result above, but our result is derived without extra analysing the convex setting, which saves much effort.

Corollary 7 *Suppose Assumptions 2, 4 and 5 hold. Applying* AdaptRdct-NC *on ZO-SVRG and ZO-SAGA, from Corollary 2, Remark 1 and 2, we know that to get an ϵ-stationary point, the total function query complexity is $\tilde{\mathcal{O}}\left(\frac{n\sigma+dL}{\epsilon^2}\right)$. Applying* AdaptRdct-NC *on ZO-Varag, from Corollary 2 and Remark 3 we know that to get an ϵ-stationary point, the total function query complexity is $\begin{cases}\tilde{\mathcal{O}}\left(\frac{d(n\sigma+L)}{\epsilon^2}\right), & n\geq\mathcal{O}\left(\frac{L}{\sigma}\right)\\ \tilde{\mathcal{O}}\left(\frac{d\sqrt{n\sigma L}}{\epsilon^2}\right), & n<\mathcal{O}\left(\frac{L}{\sigma}\right)\end{cases}$, where $\tilde{\mathcal{O}}$ hides a logarithmic factor.*

Corollary 7 also verifies that ZO-SVRG and ZO-SAGA with *AdaptRdct-NC* are better for solving non-convex problems. From Table 2, ZO-SVRG-Rand and ZO-SVRG-Ave with mini-batch technique have function query complexity of $\mathcal{O}\left(\frac{nL}{\epsilon^2}+\frac{dL}{\epsilon^4}\right)$ and $\mathcal{O}\left(\frac{nq}{\epsilon^2}+\frac{p_{max}L}{\epsilon^4}\right)$ respectively. It can be seen that our ZO-SVRG and ZO-SAGA with *AdaptRdct-NC* have much lower function query complexity. We attribute this acceleration to two factors and we use ZO-SVRG-Rand to illustrate these two factors. The first part in the function query complexity is reduced from $\frac{nL}{\epsilon^2}$ to $\frac{n\sigma}{\epsilon^2}$ (note that $\sigma\leq L$ always holds, cf. Definition 1), and this is due to our framework. The second part is reduced from $\frac{dL}{\epsilon^4}$ to $\frac{dL}{\epsilon^2}$, and this is because we control the term $\mathbb{E}\left[\|\hat{\mathbf{v}}_k^s\|^2\right]$ in a wiser way in the analysis. Note that the acceleration of the first part is huge when $\sigma\ll L$.

Also, we derive the convergence result of ZO-Varag for solving non-convex problems without extra effort, and this result is not studied by (Chen et al., 2020). Moreover, the known best function query complexity is $\mathcal{O}\left(\min\{\frac{d\sqrt{n}}{\epsilon^2},\frac{d}{\epsilon^3}\}\right)$, which is obtained by ZO-SPIDER-Coord (Ji et al., 2019) and SPIDER-SZO (Fang et al., 2018). When $\sigma<L$, ZO-Varag with *AdaptRdct-NC* outperforms the two SPIDER-based algorithms.

5.5 The Choice of the Smoothing Parameter

In Section 3.2 we introduce two different ZO gradient estimators. It can be seen that the smoothing parameter μ should be set as small as possible so that the difference between the gradient estimator and the true gradient is as small as possible. But in practice it is impossible to set μ to an arbitrarily small value as we want, since the accuracy of the computing system is limited. If μ is set to a small value, there is possibility that the result is effected by the hardware.

From the analyses in Sections 4 and 5 we know that μ introduces an error term δ_μ in the convergence results of our frameworks and ZO algorithms. In order to achieve ϵ-accuracy, δ_μ can be set to $\mathcal{O}(\epsilon)$, which means that μ is not required to be set to an arbitrarily small value. Different ZO algorithms have different δ_μ, thus the smoothing parameter μ also

differs. We list the value of the smoothing parameter of different ZO algorithms in Tables 1 and 2.

It can be seen that in all ZO algorithms listed in Tables 1 and 2, the value of the smoothing parameter μ has a dependence on the accuracy ϵ, which is difficult to implement in practice since we usually do not set the accuracy in advance. To solve this problem, Liu et al. (2018) set $\mu = \mathcal{O}\left(\frac{1}{\sqrt{dT}}\right)$, where T is the total number of iterations. Ji et al. (2019) set $\mu =$ 1e-2. Chen et al. (2020) set $\mu =$ 1e-3. Similar to their approaches, we set $\mu = \frac{1}{d}$ in our experiments. Compared with setting μ to a constant, our choice is more adaptive.

Finally we raise a point regarding *AdaptRdct-C*. When $n < \frac{18L}{\gamma}$, ZO-Varag satisfies ZOOD property with $\delta_\mu = \sqrt{\frac{L}{n\gamma}}\left[\frac{3}{2}\mu^2 Ld + 4L\sqrt{d}Z\mu\right]$, which is dependent on $\frac{1}{\sqrt{\gamma}}$. This can be a problem for *AdaptRdct-C*. Note that in Algorithm 1, γ_s diminishes at a linear rate. From Theorem 1 we know that at S-th stage, we achieve an accuracy of $\mathcal{O}\left(\sqrt{\mathcal{K}_0}^S\right) = \mathcal{O}(\epsilon)$, and $\gamma_S = \sqrt{\mathcal{K}_0}^{S-1}\gamma_1 = \mathcal{O}(\epsilon)$. Thus the smoothing parameter μ of *AdaptRdct-C* (ZO-Varag) need to be chosen $\mathcal{O}(\sqrt{\epsilon})$ smaller (cf. Table 1). But this result still matches that of vanilla ZO-Varag derived by Chen et al. (2020). ZO-SAGA has such dependence similarly. But this is not a problem for *AdaptRdct-NC* since σ does not diminish in Algorithm 2.

6. Experiments

In this section, we compare the performance of our reduction methods with other popular ZO algorithms. We conduct experiments on ZO-SVRG, ZO-SAGA and ZO-Varag with and without our reduction methods. We conduct two experiments with real-world datasets. The first experiment is generation of black-box adversarial examples for non-convex objectives, and the second experiment addresses logistic regression under convex and non-convex settings, respectively.

6.1 Generation of Black-Box Adversarial Examples

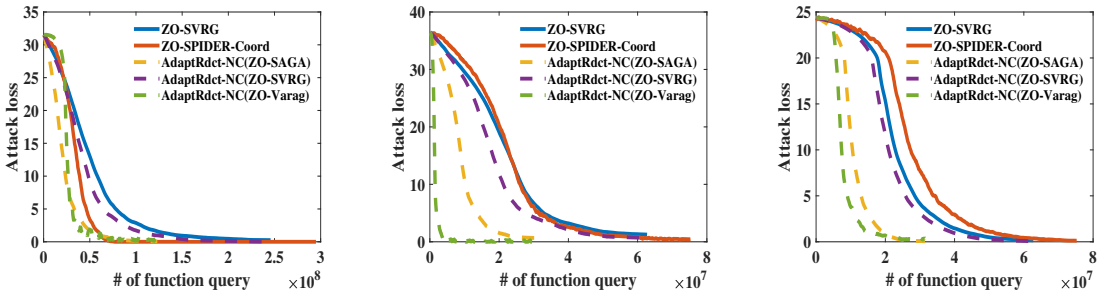
In image classification, adversary attack crafts input images with imperceptive perturbation to mislead a trained classifier. The resulting perturbed images are called *adversarial examples*, which are commonly used to understand the robustness of learning models. Under the black-box setting, the attackers only have access to the function value. It is obvious that this problem falls into the framework of ZO optimization.

For the target black-box model, we choose three well-trained DNNs $F(\cdot) = [F_1(\cdot), ..., F_K(\cdot)]$, where $F_k(\cdot)$ returns the prediction score of the k-th class. They are trained on three datasets, *i.e.*, Cifar-10, Fashion mnist (Fmnist) and Mnist. For each model, we attack $n = 50$ correctly-classified images $\{\mathbf{a}_i\}_{i=1}^n$ from the same class, and adopt the following black-box attacking loss. The i-th individual loss function $f_i(\mathbf{x})$ is given by

$$f_i(\mathbf{x}) = \max\left\{\log F_{y_i}\left(\mathbf{a}_i^{adv}\right) - \log F_{y_{tar}}\left(\mathbf{a}_i^{adv}\right), 0\right\} + \lambda\|\mathbf{a}_i^{adv} - \mathbf{a}_i\|^2 \tag{18}$$

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 $\mathbf{a}_i$, y_{tar} is the target attack class, λ is set to 1e-1. Note that the loss function is non-convex, so in this experiment we only examine the performance of *AdaptRdct-NC*.

In the experiment, we compare the performance of our frameworks applied to ZO-SVRG, ZO-SAGA and ZO-Varag. Fig. 1 shows that ZO-SVRG with *AdaptRdct-NC* has a better performance than its counterpart without reduction. Also, ZO-Varag and ZO-SAGA outperform other algorithms. Choice of parameters and more results of our frameworks applied to the three zeroth-order algorithms under different parameter settings can be found in Appendix F.



(a) Cifar-10 (b) Fmnist (c) Mnist

Figure 1: Comparison of black-box attack methods on three well-trained DNNs.

6.2 Convex and Non-convex Logistic Regression

In this subsection, we mainly consider logistic regression and its variant. To conduct experiments on *AdaptRdct-C*, we first choose the classical logistic regression problem

$$f_1(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} -\big(y_i \log s(-\mathbf{x}^T\mathbf{a}_i) + (1-y_i)\log s(\mathbf{x}^T\mathbf{a}_i)\big) \tag{19}$$

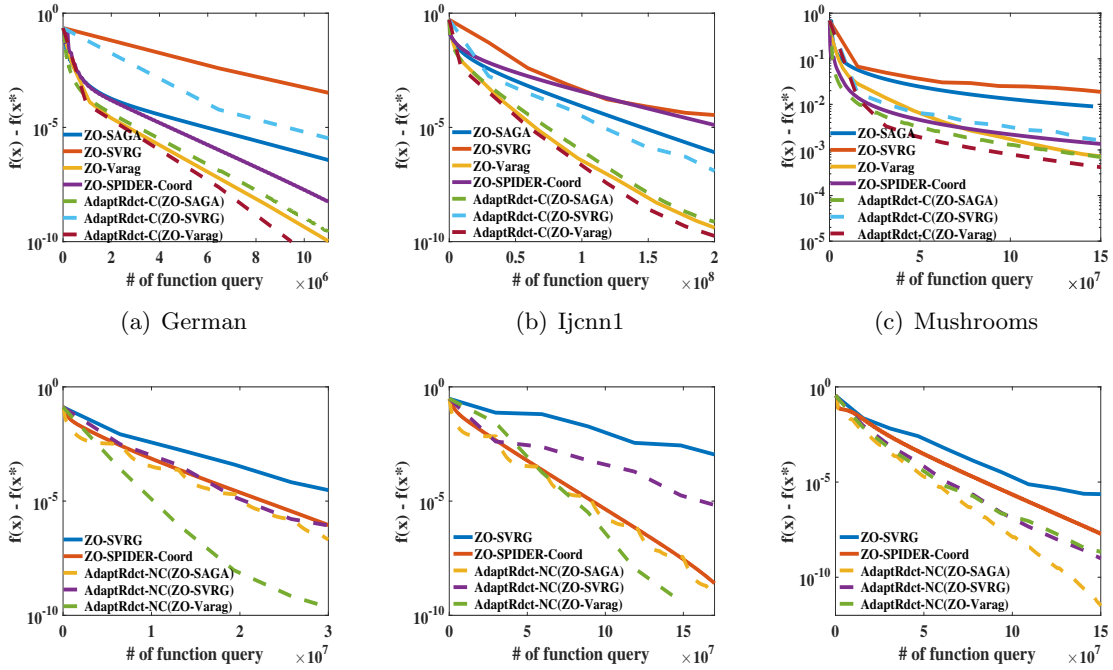
where $\mathbf{a}_i \in \mathbb{R}^d$ denote the features, $y_i \in \{0,1\}$ are the classification labels and $s(z) = 1/(1+\exp(-z))$ is the sigmoid function. It is obvious that the problem is convex. For non-convex problems, we add a non-convex regularizer $\sum_{i=1}^{d} \frac{\mathbf{x}_i^2}{1+\mathbf{x}_i^2}$ to the convex problem $f_1(\mathbf{x})$. Then we get

$$f_2(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} -\big(y_i \log s(-\mathbf{x}^T\mathbf{a}_i) + (1-y_i)\log s(\mathbf{x}^T\mathbf{a}_i)\big) + \lambda\sum_{i=1}^{d}\frac{\mathbf{x}_i^2}{1+\mathbf{x}_i^2} \tag{20}$$

where λ is set to 1e-1. For these problems, we conduct the experiments on three LIBSVM datasets (Chang and Lin, 2011), *i.e.*, the German ($n = 1,000$, $d = 24$), Ijcnn1 ($n = 49,990$, $d = 22$) and Mushrooms ($n = 8124$, $d = 112$) datasets.

Besides ZO-SVRG, ZO-SAGA and ZO-Varag, we also add SPIDER-SZO (Fang et al., 2018) for comparison. Figs. 2(a) - 2(c) show the convergence results in terms of suboptimality (the difference of objective function to the global optimal) of the algorithms solving the convex problem $f_1(\mathbf{x})$, and Figs. 2(d) - 2(f) show the convergence results in terms of suboptimality on the non-convex problem $f_2(\mathbf{x})$. It can be seen that under both convex

and non-convex settings, ZO-SVRG and ZO-SAGA equipped with reduction methods are much faster than the original algorithms without reduction. Choice of parameters and more results of our frameworks applied to the three zeroth-order algorithms under different parameter settings can be found in Appendix F.



(a) German (b) Ijcnn1 (c) Mushrooms

(d) German (e) Ijcnn1 (f) Mushrooms

Figure 2: Comparison of different ZO algorithms for logistic regression problems. (a) - (c) Convex. (d) - (f) Non-convex.

7. Conclusion

In this paper, we develop two reduction frameworks for ZO algorithms under convex and non-convex setting, respectively. Our frameworks work in a black-box manner, thus they can be applied to a wide range of ZO algorithms to further lower their function query complexities. Moreover, our frameworks can directly derive convergence results of ZO algorithms under convex and non-convex settings without extra analyses, as long as convergence results under strongly convex setting are given. To illustrate the advantages of our frameworks clearly, we have studied the performance of applying our frameworks to ZO versions of SVRG, SAGA and Varag. The results of ZO-SVRG and ZO-SAGA indicate that our approach has a lower query complexity than vanilla ZO algorithms under both convex and non-convex settings. Our theoretic study and experimental results highlight the advantages of combining ZO optimization with the reduction techniques. To the best of our knowledge, we are the first to propose black-box reduction frameworks for zeroth-order algorithms and apply them to zeroth-order optimization. In the future, we would like to extend this work

to the parallel computing (Gu et al., 2020b, 2019, 2018b) and federated learning (Gu et al., 2020a; Zhang et al., 2021) scenarios.

Appendix

Appendix A provides the proof of *AdaptRdct-C*.
Appendix B provides the proof of *AdaptRdct-NC*.
Appendix C provides the proof of ZO-SVRG.
Appendix D provides the proof of ZO-SAGA.
Appendix E provides the proof of ZO-Varag.
Appendix F provides choice of parameters in the experiment and additional experiment results of our frameworks applied to ZO-SVRG, ZO-SAGA and ZO-Varag under different parameter settings.

Appendix A. Proof of *AdaptRdct-C*

In this section we provide the proof of Theorem 1.

Theorem 1 *Suppose Assumption 2 holds. Let* $\mathbf{x}_0$ *be a starting vector such that* $f(\mathbf{x}_0) - f(\mathbf{x}^*) \leq \Delta$, *and* $||\mathbf{x}_0 - \mathbf{x}^*||^2 \leq \Theta$. *For Algorithm 1, we have*

$$\mathbb{E}[f(\mathbf{x}_S) - f(\mathbf{x}^*)] \leq \delta_\mu + \mathcal{K}_0^S[\Delta - \delta_\mu] + \left(\frac{1}{2} + \frac{2}{\sqrt{\mathcal{K}_0}}\right)\mathcal{K}_0^{\frac{S}{2}}\gamma_1\Theta$$

Proof Denote $\mathbf{x}_s^* = \arg\min_{\mathbf{x}} f^{(s)}(\mathbf{x})$. By the strong convexity of $f^{(s)}(\mathbf{x})$, we have

$$\mathbb{E}\left[f^{(s)}(\mathbf{x}_s^*) - f^{(s)}(\mathbf{x}^*)\right] \leq -\frac{\gamma_s}{2}\mathbb{E}\left[||\mathbf{x}_s^* - \mathbf{x}^*||^2\right] \tag{21}$$

Using the fact that $f^{(s)}(\mathbf{x}_s^*) \geq f(\mathbf{x}_s^*)$, as well as the definition $f^{(s)}(\mathbf{x}^*) = f(\mathbf{x}^*) + \frac{\gamma_s}{2}||\mathbf{x}^* - \mathbf{x}_0||^2$, we immediately have

$$\mathbb{E}\left[f(\mathbf{x}_s^*) - f(\mathbf{x}^*) - \frac{\gamma_s}{2}||\mathbf{x}^* - \mathbf{x}_0||^2\right] \leq -\frac{\gamma_s}{2}\mathbb{E}\left[||\mathbf{x}_s^* - \mathbf{x}^*||^2\right] \tag{22}$$

Rearranging the terms, we get

$$\frac{\gamma_s}{2}||\mathbf{x}_0 - \mathbf{x}^*||^2 - \frac{\gamma_s}{2}\mathbb{E}\left[||\mathbf{x}_s^* - \mathbf{x}^*||^2\right] \geq \mathbb{E}\left[f(\mathbf{x}_s^*) - f(\mathbf{x}^*)\right] \geq 0 \tag{23}$$

Thus we have

$$\mathbb{E}\left[\|\mathbf{x}_s^* - \mathbf{x}^*\|^2\right] \leq ||\mathbf{x}_0 - \mathbf{x}^*||^2 \tag{24}$$

Denote $\mathcal{K} = \mathcal{K}_0$, the ZOOD property of $\mathcal{A}$ ensures that

$$\mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - f^{(s)}(\mathbf{x}_s^*) - \delta_\mu\right] \leq \mathcal{K}\mathbb{E}\left[f^{(s)}(\mathbf{x}_{s-1}) - f^{(s)}(\mathbf{x}_s^*) - \delta_\mu\right] \tag{25}$$

Denote $D_s = \mathbb{E}\left[f^{(s)}(\mathbf{x}_{s-1}) - f^{(s)}(\mathbf{x}_s^*)\right]$. At the beginning, we have upper bound $D_1 = f^{(1)}(\mathbf{x}_0) - f^{(1)}(\mathbf{x}_1^*) \leq f(\mathbf{x}_0) - f(\mathbf{x}^*)$. For each epoch $s \geq 1$, we compute that

$$\begin{aligned}
D_{s+1} &= \mathbb{E}\left[f^{(s+1)}(\mathbf{x}_s) - f^{(s+1)}(\mathbf{x}_{s+1}^*)\right] \\
&\overset{①}{=} \mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - \frac{\gamma_s - \gamma_{s+1}}{2}||\mathbf{x}_s - \mathbf{x}_0||^2\right] - \mathbb{E}\left[f^{(s)}(\mathbf{x}_{s+1}^*) - \frac{\gamma_s - \gamma_{s+1}}{2}||\mathbf{x}_{s+1}^* - \mathbf{x}_0||^2\right] \\
&\overset{②}{\leq} \mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - \frac{\gamma_s - \gamma_{s+1}}{2}||\mathbf{x}_s - \mathbf{x}_0||^2\right] \\
&\quad - \mathbb{E}\left[f^{(s)}(\mathbf{x}_s^*) + \frac{\gamma_s}{2}||\mathbf{x}_{s+1}^* - \mathbf{x}_s^*||^2\right] + \frac{\gamma_s - \gamma_{s+1}}{2}\mathbb{E}\left[||\mathbf{x}_{s+1}^* - \mathbf{x}_0||^2\right] \\
&\leq \mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - f^{(s)}(\mathbf{x}_s^*)\right] + \frac{\gamma_s - \gamma_{s+1}}{2}\mathbb{E}\left[||\mathbf{x}_{s+1}^* - \mathbf{x}_0||^2\right] \\
&\overset{③}{\leq} \mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - f^{(s)}(\mathbf{x}_s^*)\right] + (\gamma_s - \gamma_{s+1})\mathbb{E}\left[||\mathbf{x}_{s+1}^* - \mathbf{x}^*||^2 + ||\mathbf{x}_0 - \mathbf{x}^*||^2\right] \\
&\overset{④}{\leq} \mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - f^{(s)}(\mathbf{x}_s^*)\right] + 2(1 - \sqrt{\mathcal{K}})\gamma_s||\mathbf{x}_0 - \mathbf{x}^*||^2
\end{aligned} \tag{26}$$

Above, ① uses the definition of $f^{(s+1)}(\mathbf{x})$; ② uses the convexity of $f^{(s)}(\mathbf{x})$ and the fact that $\mathbf{x}_s^*$ is the minimizer of $f^{(s)}(\mathbf{x})$; ③ uses the inequality $||\mathbf{a} - \mathbf{b}||^2 \leq 2||\mathbf{a}||^2 + 2||\mathbf{b}||^2$; ④ follows from (24) and $\gamma_{s+1} = \sqrt{\mathcal{K}}\gamma_s$. That implies

$$\begin{aligned}
D_{s+1} - \delta_\mu &\leq \mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - f^{(s)}(\mathbf{x}_s^*) - \delta_\mu\right] + 2(1 - \sqrt{\mathcal{K}})\gamma_s||\mathbf{x}_0 - \mathbf{x}^*||^2 \\
&\leq \mathcal{K}(D_s - \delta_\mu) + 2(1 - \sqrt{\mathcal{K}})\gamma_s||\mathbf{x}_0 - \mathbf{x}^*||^2
\end{aligned} \tag{27}$$

The second inequality follows from (25). Recursively applying the above inequality, we have

$$\begin{aligned}
D_{S+1} - \delta_\mu &\leq \mathcal{K}^S(D_1 - \delta_\mu) + ||\mathbf{x}_0 - \mathbf{x}^*||^2 \cdot 2 \cdot (1 - \sqrt{\mathcal{K}})[\gamma_S + \mathcal{K}\gamma_{S-1} + ... + \mathcal{K}^{S-1}\gamma_1] \\
&= \mathcal{K}^S(D_1 - \delta_\mu) + ||\mathbf{x}_0 - \mathbf{x}^*||^2 \cdot 2 \cdot (1 - \sqrt{\mathcal{K}})\gamma_S[1 + \sqrt{\mathcal{K}} + ... + \sqrt{\mathcal{K}}^{S-1}] \\
&\leq \mathcal{K}^S(D_1 - \delta_\mu) + ||\mathbf{x}_0 - \mathbf{x}^*||^2 \cdot 2 \cdot \gamma_S \\
&= \mathcal{K}^S(D_1 - \delta_\mu) + ||\mathbf{x}_0 - \mathbf{x}^*||^2 \cdot \frac{2}{\sqrt{\mathcal{K}}} \cdot \gamma_{S+1}
\end{aligned} \tag{28}$$

Finally, we obtain

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_S) - f(\mathbf{x}^*) - \delta_\mu] &\overset{①}{\leq} \mathbb{E}[f^{(S+1)}(\mathbf{x}_S) - f^{(S+1)}(\mathbf{x}^*) - \delta_\mu + \frac{\gamma_{S+1}}{2}||\mathbf{x}_0 - \mathbf{x}^*||^2] \\
&\overset{②}{\leq} \mathbb{E}[f^{(S+1)}(\mathbf{x}_S) - f^{(S+1)}(\mathbf{x}_{S+1}^*) - \delta_\mu + \frac{\gamma_{S+1}}{2}||\mathbf{x}_0 - \mathbf{x}^*||^2] \\
&\overset{③}{\leq} \mathcal{K}^S[f(\mathbf{x}_0) - f(\mathbf{x}^*) - \delta_\mu] + (\frac{1}{2} + \frac{2}{\sqrt{\mathcal{K}}})\gamma_{S+1}||\mathbf{x}_0 - \mathbf{x}^*||^2 \\
&\overset{④}{\leq} \mathcal{K}^S[f(\mathbf{x}_0) - f(\mathbf{x}^*) - \delta_\mu] + (\frac{1}{2} + \frac{2}{\sqrt{\mathcal{K}}})\mathcal{K}^{\frac{S}{2}}\gamma_1||\mathbf{x}_0 - \mathbf{x}^*||^2
\end{aligned} \tag{29}$$

The ① comes from the definition of $f^{(S+1)}(\mathbf{x})$ and the fact that $f(\mathbf{x}) \leq f^{(S+1)}(\mathbf{x})$; ② comes from the fact that $\mathbf{x}_{S+1}^*$ is the minimizer of $f^{(S+1)}(\mathbf{x}_S)$; ③ comes from (28), and ④ comes

from the update rule $\gamma_{s+1} = \sqrt{\mathcal{K}}\gamma_s$. Note that $f(\mathbf{x}_0) - f(\mathbf{x}^*) \leq \Delta$, $||\mathbf{x}_0 - \mathbf{x}^*||^2 \leq \Theta$ and $\mathcal{K} = \mathcal{K}_0$, then we have

$$\mathbb{E}[f(\mathbf{x}_S) - f(\mathbf{x}^*)] \leq \delta_\mu + \mathcal{K}_0^S[\Delta - \delta_\mu] + (\frac{1}{2} + \frac{2}{\sqrt{\mathcal{K}_0}})\mathcal{K}_0^{\frac{S}{2}}\gamma_1\Theta \tag{30}$$

Then we complete the proof. ■

Appendix B. Proof of *AdaptRdct-NC*

In this section we provide the proof of Theorem 2.

Theorem 2 *Suppose Assumption 3 holds. Let* $\mathbf{x}_0$ *be a starting vector such that* $f(\mathbf{x}_0) - f(\mathbf{x}^*) \leq \Delta$. *For Algorithm 2, we have*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\alpha+1})\|^2\right] \leq \frac{16\sigma(2\mathcal{K}+1)}{(1-\mathcal{K})S}[f(\mathbf{x}_0) - f(\mathbf{x}^*)] + \left(\frac{64-16\mathcal{K}}{1-\mathcal{K}} + 4L^2\right)\sigma\delta_\mu$$

where $\mathcal{K} = \frac{\sigma^2\mathcal{K}_0}{L^2}$

Proof With the definition of Moreau envelope, we have $f_\lambda(\mathbf{x}_{s-1}) = \min_{\mathbf{x}} f^{(s)}(\mathbf{x})$. Denote $\mathbf{x}_s^* = \arg\min_{\mathbf{x}} f^{(s)}(\mathbf{x}) = Prox_{\lambda f}(\mathbf{x}_{s-1})$, we get $\nabla f_\lambda(\mathbf{x}_{s-1}) = \frac{\mathbf{x}_{s-1} - \mathbf{x}_s^*}{\lambda}$. Then with the definition of $f^{(s)}(\mathbf{x})$, we have

$$f(\mathbf{x}_{s-1}) = f^{(s)}(\mathbf{x}_{s-1}) \geq f^{(s)}(\mathbf{x}_s^*) = f(\mathbf{x}_s^*) + \frac{1}{2\lambda}||\mathbf{x}_{s-1} - \mathbf{x}_s^*||^2 \tag{31}$$

Denote $\mathcal{K} = \frac{\sigma^2\mathcal{K}_0}{L^2}$, Algorithm 2 ensures that

$$\mathbb{E}\left[f^{(s)}(\mathbf{x}_s) - f^{(s)}(\mathbf{x}_s^*) - \delta_\mu\right] \leq \mathcal{K}\left[f^{(s)}(\mathbf{x}_{s-1}) - f^{(s)}(\mathbf{x}_s^*) - \delta_\mu\right] \tag{32}$$

which implies

$$\mathbb{E}\left[f^{(s)}(\mathbf{x}_s)\right] \leq f^{(s)}(\mathbf{x}_s^*) + \mathcal{K}\left[f^{(s)}(\mathbf{x}_{s-1}) - f^{(s)}(\mathbf{x}_s^*)\right] - (\mathcal{K}-1)\delta_\mu \tag{33}$$

Thus we have

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_s) + \frac{1}{2\lambda}||\mathbf{x}_s - \mathbf{x}_{s-1}||^2\right] &= \mathbb{E}\left[f^{(s)}(\mathbf{x}_s)\right] \\ &\leq f^{(s)}(\mathbf{x}_s^*) + \mathcal{K}\left[f^{(s)}(\mathbf{x}_{s-1}) - f^{(s)}(\mathbf{x}_s^*)\right] - (\mathcal{K}-1)\delta_\mu \\ &\leq f(\mathbf{x}_{s-1}) + \mathcal{K}\left[f^{(s)}(\mathbf{x}_{s-1}) - f^{(s)}(\mathbf{x}_s^*)\right] - (\mathcal{K}-1)\delta_\mu\end{aligned} \tag{34}$$

The last inequality follows from (31). On the other hand, we have

$$\begin{aligned}||\mathbf{x}_s - \mathbf{x}_{s-1}||^2 &= ||\mathbf{x}_s - \mathbf{x}_s^* + \mathbf{x}_s^* - \mathbf{x}_{s-1}||^2 \\ &= ||\mathbf{x}_s - \mathbf{x}_s^*||^2 + ||\mathbf{x}_s^* - \mathbf{x}_{s-1}||^2 + 2\left\langle \mathbf{x}_s - \mathbf{x}_s^*, \mathbf{x}_s^* - \mathbf{x}_{s-1}\right\rangle \\ &\geq -||\mathbf{x}_s - \mathbf{x}_s^*||^2 + \frac{1}{2}||\mathbf{x}_{s-1} - \mathbf{x}_s^*||^2\end{aligned} \tag{35}$$

The inequality follows from Young's inequality. Combining with (34), we then get

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{4\lambda}||\mathbf{x}_{s-1}-\mathbf{x}_s^*||^2\right] \leq \mathbb{E}\left[\frac{1}{2\lambda}||\mathbf{x}_s-\mathbf{x}_{s-1}||^2+\frac{1}{2\lambda}||\mathbf{x}_s-\mathbf{x}_s^*||^2\right] \\
&\leq \mathbb{E}\left[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)+\mathcal{K}[f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s^*)]-(\mathcal{K}-1)\delta_\mu+\frac{1}{2\lambda}||\mathbf{x}_s-\mathbf{x}_s^*||^2\right] \\
&\overset{①}{\leq} \mathbb{E}\left[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)+\mathcal{K}[f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s^*)]-(\mathcal{K}-1)\delta_\mu+\frac{1}{\lambda(\lambda^{-1}-\sigma)}[f^{(s)}(\mathbf{x}_s)-f^{(s)}(\mathbf{x}_s^*)]\right] \\
&\overset{②}{\leq} \mathbb{E}\left[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)+\frac{2-\lambda\sigma}{1-\lambda\sigma}\left(\mathcal{K}\left[f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s^*)\right]-(\mathcal{K}-1)\delta_\mu\right)\right] \\
&\overset{③}{=} \mathbb{E}\left[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)+3\left(\mathcal{K}\left[f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s^*)\right]-(\mathcal{K}-1)\delta_\mu\right)\right]
\end{aligned} \tag{36}$$

① holds because $f^{(s)}(\mathbf{x})$ is $(\lambda^{-1}-\sigma)$-strongly convex, ② holds due to (32), and ③ holds due to $\lambda=\frac{1}{2\sigma}$. Next, we bound $f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s^*)$ given that $\mathbf{x}_{s-1}$ is fixed. According to the definition of $f^{(s)}(\mathbf{x})$, we have

$$\begin{aligned}
f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s^*) &= f^{(s)}(\mathbf{x}_s)-f^{(s)}(\mathbf{x}_s^*)+f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s) \\
&= f^{(s)}(\mathbf{x}_s)-f^{(s)}(\mathbf{x}_s^*)+f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)-\frac{1}{2\lambda}||\mathbf{x}_s-\mathbf{x}_{s-1}||^2 \\
&\leq f^{(s)}(\mathbf{x}_s)-f^{(s)}(\mathbf{x}_s^*)+f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)
\end{aligned} \tag{37}$$

Taking expectation over randomness in the s-th stage on both sides, we have

$$\begin{aligned}
f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s^*) &\leq \mathbb{E}\left[f^{(s)}(\mathbf{x}_s)-f^{(s)}(\mathbf{x}_s^*)\right]+\mathbb{E}\left[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)\right] \\
&\leq \mathcal{K}\left[f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s^*)\right]-(\mathcal{K}-1)\delta_\mu+\mathbb{E}[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)]
\end{aligned} \tag{38}$$

The second inequality follows from (32). Rearranging the terms, we have

$$f^{(s)}(\mathbf{x}_{s-1})-f^{(s)}(\mathbf{x}_s^*) \leq \frac{1}{1-\mathcal{K}}\mathbb{E}[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)]+\delta_\mu \tag{39}$$

Plug this upper bound into (36), then we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{4\lambda}||\mathbf{x}_{s-1}-\mathbf{x}_s^*||^2\right] &\leq \mathbb{E}\left[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)+\frac{3\mathcal{K}}{1-\mathcal{K}}[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)]+3\delta_\mu\right] \\
&= \frac{2\mathcal{K}+1}{1-\mathcal{K}}\mathbb{E}\left[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)\right]+3\delta_\mu
\end{aligned} \tag{40}$$

Since $||\nabla f_\lambda(\mathbf{x}_{s-1})||=\frac{||\mathbf{x}_{s-1}-\mathbf{x}_s^*||}{\lambda}$, we have

$$\frac{\lambda}{4}\mathbb{E}\left[||\nabla f_\lambda(\mathbf{x}_{s-1})||^2\right] \leq \frac{2\mathcal{K}+1}{1-\mathcal{K}}\mathbb{E}\left[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)\right]+3\delta_\mu \tag{41}$$

Multiplying both sides with weight $w_{s-1}\overset{\text{def}}{=}(s-1)^\tau$, we get

$$\frac{\lambda}{4}w_{s-1}\mathbb{E}\left[||\nabla f_\lambda(\mathbf{x}_{s-1})||^2\right] \leq \frac{2\mathcal{K}+1}{1-\mathcal{K}}w_{s-1}\mathbb{E}\left[f(\mathbf{x}_{s-1})-f(\mathbf{x}_s)\right]+3w_{s-1}\delta_\mu \tag{42}$$

By summing over $s = 1, ..., S+1$, we get

$$\frac{\lambda}{4}\sum_{s=1}^{S+1} w_{s-1}\mathbb{E}\left[||\nabla f_\lambda(\mathbf{x}_{s-1})||^2\right] \leq \frac{2\mathcal{K}+1}{1-\mathcal{K}}\sum_{s=1}^{S+1} w_{s-1}\mathbb{E}\left[f(\mathbf{x}_{s-1}) - f(\mathbf{x}_s)\right] + 3\sum_{s=1}^{S+1} w_{s-1}\delta_\mu \quad (43)$$

With the choice of $\mathbf{x}_\alpha$, we have

$$\begin{aligned}
&\frac{\lambda}{4}\sum_{s=1}^{S+1} w_{s-1}\mathbb{E}[||\nabla f_\lambda(\mathbf{x}_\alpha)||^2] \\
\leq&\frac{2\mathcal{K}+1}{1-\mathcal{K}}\mathbb{E}\left[w_0 f(\mathbf{x}_0) - w_S f(\mathbf{x}_{S+1}) + \sum_{s=1}^{S+1}(w_{s-1}-w_s) f(\mathbf{x}_{s-1})\right] + 3\sum_{s=1}^{S+1} w_{s-1}\delta_\mu \\
\overset{①}{=}&\frac{2\mathcal{K}+1}{1-\mathcal{K}}\mathbb{E}\left[\sum_{s=1}^{S+1}(w_{s-1}-w_s)\left[f(\mathbf{x}_{s-1}) - f(\mathbf{x}_{S+1})\right]\right] + 3\sum_{s=1}^{S+1} w_{s-1}\delta_\mu \\
\overset{②}{\leq}&\frac{2\mathcal{K}+1}{1-\mathcal{K}}\mathbb{E}\left[\sum_{s=1}^{S+1}(w_{s-1}-w_s)\left[f(\mathbf{x}_0) + (s-1)\delta_\mu - f(\mathbf{x}_{S+1})\right]\right] + 3\sum_{s=1}^{S+1} w_{s-1}\delta_\mu \\
\leq&\frac{2\mathcal{K}+1}{1-\mathcal{K}} w_S\mathbb{E}[f(\mathbf{x}_0) - f(\mathbf{x}_{S+1})] + \frac{2\mathcal{K}+1}{1-\mathcal{K}}\sum_{s=1}^{S+1} w_{s-1}\delta_\mu + 3\sum_{s=1}^{S+1} w_{s-1}\delta_\mu
\end{aligned} \quad (44)$$

where ① holds because $w_0 = 0$, ② holds because from ZOOD property we have $f(\mathbf{x}_{s-1}) - f(\mathbf{x}_0) \leq \delta_\mu$. Since $\lambda = \frac{1}{2\sigma}$ and $\sum_{s=1}^{S+1} w_{s-1} \geq S$, we have

$$\mathbb{E}[||\nabla f_\lambda(\mathbf{x}_\alpha)||^2] \leq \frac{8\sigma(2\mathcal{K}+1)}{(1-\mathcal{K})S}[f(\mathbf{x}_0) - f(\mathbf{x}^*)] + \frac{32-8\mathcal{K}}{1-\mathcal{K}}\sigma\delta_\mu \quad (45)$$

Note that $f(\mathbf{x}_0) - f(\mathbf{x}_{S+1}) \leq f(\mathbf{x}_0) - f(\mathbf{x}^*) = \Delta$. Denote $\mathbf{x}_\alpha^+ \overset{\text{def}}{=} Prox_\lambda f(\mathbf{x}_\alpha) = \arg\min_{\mathbf{x}} f^{(\alpha+1)}(\mathbf{x})$, then

$$\begin{aligned}
\|\nabla f(\mathbf{x}_{\alpha+1})\|^2 \overset{①}{\leq}&2\|\nabla f(\mathbf{x}_{\alpha+1}) - \nabla f(\mathbf{x}_\alpha^+)\|^2 + 2\|\nabla f(\mathbf{x}_\alpha^+)\|^2 \\
\overset{②}{\leq}&2L^2\|\mathbf{x}_{\alpha+1} - \mathbf{x}_\alpha^+\|^2 + 2\|\nabla f_\lambda(\mathbf{x}_\alpha)\|^2
\end{aligned} \quad (46)$$

where ① comes from the fact that $\|\mathbf{a}+\mathbf{b}\|^2 \leq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$, ② comes from the smoothness of f and (9). For any $s = 0, ..., S$, we have

$$\begin{aligned}
L^2\|\mathbf{x}_{s+1} - \mathbf{x}_s^+\|^2 \overset{①}{\leq}&\frac{2L^2}{\sigma}\left[f^{(s+1)}(\mathbf{x_{s+1}}) - f^{(s+1)}(\mathbf{x_s}^+)\right] \\
\overset{②}{\leq}&\frac{2\sigma\mathcal{K}}{s+1}\left[f^{(s+1)}(\mathbf{x_s}) - f^{(s+1)}(\mathbf{x_s}^+)\right] + 2L^2\delta_\mu \\
\overset{③}{=}&\frac{2\sigma\mathcal{K}}{s+1}\left[f(\mathbf{x}_s) - f(\mathbf{x}_s^+) - \sigma\|\mathbf{x}_s - \mathbf{x}_s^+\|^2\right] + 2L^2\delta_\mu \\
\overset{④}{\leq}&\frac{2\sigma\mathcal{K}}{s+1}\left[f(\mathbf{x}_s) - f(\mathbf{x}^*)\right] + 2L^2\delta_\mu \\
\overset{⑤}{\leq}&\frac{2\sigma\mathcal{K}}{s+1}\Delta + \left(\mathcal{K} + 2L^2\right)\delta_\mu
\end{aligned} \quad (47)$$

where ① comes from the strong convexity of $f^{s+1}(\mathbf{x})$ and $\mathbf{x}_s^+$ is its minimizer, ② comes from Algorithm 2, ③ comes from the definition of $f^{s+1}(\mathbf{x})$, ④ holds since $\mathbf{x}^*$ minimizes f, ⑤ comes from the ZOOD property. Multiplying both sides with w_{s+1}, summing over $s = 0, ...S$, following similar analysis above, we have

$$L^2 \sum_{s=0}^{S} w_{s+1} \mathbb{E}\left[\|\mathbf{x}_{\alpha+1} - \mathbf{x}_\alpha^+\|^2\right] \leq \sum_{s=0}^{S} w_{s+1} \frac{2\sigma\mathcal{K}}{s+1}\Delta + \sum_{s=0}^{S} w_{s+1}\left(\mathcal{K} + 2L^2\right)\delta_\mu \tag{48}$$

with the choice of w_s we have

$$L^2 \mathbb{E}\left[\|\mathbf{x}_{\alpha+1} - \mathbf{x}_\alpha^+\|^2\right] \leq \frac{2\sigma\mathcal{K}}{S}\Delta + \left(\mathcal{K} + 2L^2\right)\delta_\mu \tag{49}$$

Combining the above inequality with (45) and (46), we get

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\alpha+1})\|^2\right] \leq \frac{16\sigma(2\mathcal{K}+1)}{(1-\mathcal{K})S}[f(\mathbf{x}_0) - f(\mathbf{x}^*)] + \left(\frac{64 - 16\mathcal{K}}{1-\mathcal{K}} + 4L^2\right)\sigma\delta_\mu \tag{50}$$

Then we complete the proof ■

Appendix C. Proof of ZO-SVRG

In this section we provide proof of Theorem 3 and 4. First we present four auxiliary lemmas, Lemma 1, 2, 3 and 4. We introduce a smoothing function. Define $f_\mu(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbf{U}_b}[f(\mathbf{x}+\mu\mathbf{u})]$ where $\mathbf{U}_b$ is a uniform distribution over the unit Euclidean ball. Then we have :

Lemma 1 *Suppose Assumptions 2 and 4 hold, we have*

1) $f_\mu(\mathbf{x})$ *is also L-smooth and convex, and*

$$\nabla f_\mu(\mathbf{x}) = \mathbb{E}_{\mathbf{u}}[\hat{\nabla} f(\mathbf{x})]$$

2) $\forall\, \mathbf{x} \in \mathbb{R}^d$,

$$|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq \frac{L\mu^2}{2}$$

$$||\nabla f_\mu(\mathbf{x}) - \nabla f(\mathbf{x})||^2 \leq \frac{\mu^2 L^2 d^2}{4}$$

3) $\forall\, \mathbf{x} \in \mathbb{R}^d$,

$$\mathbb{E}_{\mathbf{u}}\left[||\hat{\nabla} f(\mathbf{x}) - \nabla f_\mu(\mathbf{x})||^2\right] \leq \mathbb{E}_{\mathbf{u}}\left[||\hat{\nabla} f(\mathbf{x})||^2\right] \leq 2d||\nabla f(\mathbf{x})||^2 + \frac{\mu^2 L^2 d^2}{2}$$

Proof See (Liu et al., 2018, Lemma 1). ■

Lemma 2 *Suppose Assumption 4 holds, we have*

$$\mathbb{E}_i[||\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)||^2] \leq 2L[f(\mathbf{x}) - f(\mathbf{x}^*)]$$

Proof See (Johnson and Zhang, 2013, Theorem 1). ■

Lemma 3 *Suppose Assumption 4 holds, $\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $i \in [n]$ and $\beta > 0$, we have*

$$\begin{aligned}&\mathbb{E}\left[||\hat{\nabla} f_i(\mathbf{x}) - \hat{\nabla} f_i(\mathbf{y})||^2\right]\\ \leq& 3d\mathbb{E}\left[(1+\beta)||\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)||^2 + (1+\frac{1}{\beta})||\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x}^*)||^2\right] + \frac{3L^2d^2\mu^2}{2}\end{aligned}$$

Proof From (Ji et al., 2019, Lemma 5), we have

$$\mathbb{E}\left[||\hat{\nabla} f_i(\mathbf{x}) - \hat{\nabla} f_i(\mathbf{y})||^2\right] \leq 3d\mathbb{E}\left[||\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})||^2\right] + \frac{3L^2d^2\mu^2}{2} \tag{51}$$

Using Young's inequality, we have

$$\begin{aligned}&\mathbb{E}\left[||\hat{\nabla} f_i(\mathbf{x}) - \hat{\nabla} f_i(\mathbf{y})||^2\right]\\ \leq& 3d\mathbb{E}\left[(1+\beta)||\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)||^2 + (1+\frac{1}{\beta})||\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x}^*)||^2\right] + \frac{3L^2d^2\mu^2}{2}\end{aligned} \tag{52}$$

■

Lemma 4 *Suppose Assumptions 2 and 4 hold. Taking expectation with respect to all variables, we have :*

$$\mathbb{E}\left[||\hat{\mathbf{v}}_k^s||^2\right] \leq 48dL\left[f(\mathbf{x}_k^s) - f(\mathbf{x}^*)\right] + 48dL\left[f(\tilde{\mathbf{x}}_{s-1}) - f(\mathbf{x}^*)\right] + 8L^2d^2\mu^2$$

Proof From the definition of f_μ and the choice of i_k, we have

$$\mathbb{E}\left[\hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1})\right] = \nabla f_\mu(\mathbf{x}_k^s) - \nabla f_\mu(\tilde{\mathbf{x}}_{s-1}) \tag{53}$$

Then we can rewrite $\hat{\mathbf{v}}_k^s$ as

$$\begin{aligned}\hat{\mathbf{v}}_k^s =&\hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1}) + \hat{\nabla} f(\tilde{\mathbf{x}}_{s-1})\\ =&\hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1}) - \mathbb{E}\left[\hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1})\right] + \nabla f_\mu(\mathbf{x}_k^s) + \left(\hat{\nabla} f(\tilde{\mathbf{x}}_{s-1}) - \nabla f_\mu(\tilde{\mathbf{x}}_{s-1})\right)\end{aligned} \tag{54}$$

Then we have

$$\begin{aligned}&\mathbb{E}\left[||\hat{\mathbf{v}}_k^s||^2\right]\\ =&\mathbb{E}\left[||\hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1}) - \mathbb{E}\left[\hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1})\right] + \nabla f_\mu(\mathbf{x}_k^s) + \left(\hat{\nabla} f(\tilde{\mathbf{x}}_{s-1}) - \nabla f_\mu(\tilde{\mathbf{x}}_{s-1})\right)||^2\right]\\ \leq&3\mathbb{E}\left[||\hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1}) - \mathbb{E}_{i_k}[\hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1})]||^2\right] + 3\mathbb{E}\left[||\nabla f_\mu(x_k^s)||^2\right]\\ &+ 3\mathbb{E}\left[||\hat{\nabla} f(\tilde{\mathbf{x}}_{s-1}) - \nabla f_\mu(\tilde{\mathbf{x}}_{s-1})||^2\right]\\ \leq&3\mathbb{E}\left[||\hat{\nabla} f_{i_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1})||^2\right] + 6\mathbb{E}\left[||\nabla f(\mathbf{x}_k^s)||^2\right] + \frac{3\mu^2d^2L^2}{2}\\ &+ 6d\mathbb{E}\left[||\nabla f(\tilde{\mathbf{x}}_{s-1})||^2\right] + \frac{3\mu^2d^2L^2}{2}\end{aligned} \tag{55}$$

where the first inequality holds due to $||\mathbf{a}+\mathbf{b}+\mathbf{c}||^2 \leq 3||\mathbf{a}||^2+3||\mathbf{b}||^2+3||\mathbf{c}||^2$, the second inequality holds due to $\mathbb{E}\left[||\mathbf{a}-\mathbb{E}[\mathbf{a}]||^2\right] \leq \mathbb{E}\left[||\mathbf{a}||^2\right]$ and Lemma 1. From Lemma 3 with $\beta=1$, we bound the first term on the right hand side of (55) as

$$\begin{aligned}
&\mathbb{E}\left[||\hat{\nabla} f_{i_k}(\mathbf{x}_k^s)-\hat{\nabla} f_{i_k}(\tilde{\mathbf{x}}_{s-1})||^2\right] \\
&\leq 6d\mathbb{E}\left[||\nabla f_{i_k}(\mathbf{x}_k^s)-\nabla f_{i_k}(\mathbf{x}^*)||^2+||\nabla f_{i_k}(\tilde{\mathbf{x}}_{s-1})-\nabla f_{i_k}(\mathbf{x}^*)||^2\right]+\frac{3L^2d^2\mu^2}{2} \qquad (56)\\
&\leq 12dL\mathbb{E}\left[f(\mathbf{x}_k^s)-f(\mathbf{x}^*)+f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*)\right]+\frac{3L^2d^2\mu^2}{2}
\end{aligned}$$

From the smoothness of f and the fact that $\nabla f(\mathbf{x}^*)=\mathbf{0}$, we have for all $\mathbf{x}\in\mathbb{R}^d$

$$\mathbb{E}\left[||\nabla f(\mathbf{x})||^2\right] \leq 2L\mathbb{E}\left[f(\mathbf{x})-f(\mathbf{x}^*)\right] \qquad (57)$$

Substitute (56) and (57) into (55), we get

$$\mathbb{E}\left[||\hat{\mathbf{v}}_k^s||^2\right] \leq 48dL\left[f(\mathbf{x}_k^s)-f(\mathbf{x}^*)\right]+48dL\left[f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*)\right]+8L^2d^2\mu^2 \qquad (58)$$

Then we complete the proof. ■

Theorem 3 *Suppose Assumptions 1 and 4 hold, denote* $f(\tilde{\mathbf{x}}_0)-f(\mathbf{x}^*)=\Delta$. *By using* ***Option I*** *in Algorithm 3, we have*

$$\mathbb{E}\left[f(\tilde{\mathbf{x}}_s)-f(\mathbf{x}^*)\right] \leq \delta_\mu+\left(\frac{\beta_2}{\beta_1}\right)^s(\Delta-\delta_\mu)$$

where $\beta_1=2m\eta\left[1-24\eta dL\right]$, $\beta_2=\frac{2}{\gamma}+48m\eta^2dL$, $\delta_\mu=\frac{2\eta m\mu^2L(4\eta d^2L+1)}{\beta_1-\beta_2}$ *and* η, m *satisfy inequalities* $\eta<\frac{1}{48dL}$ *and* $m>\frac{1}{\gamma\eta(1-48\eta dL)}$.

Proof From Lemma 1, we have $\mathbb{E}[\hat{\mathbf{v}}_k^s]=\nabla f_\mu(\mathbf{x}_k^s)$. Conditioned on $\mathbf{x}_k^s$, we have

$$\begin{aligned}
\mathbb{E}[||\mathbf{x}_{k+1}^s-\mathbf{x}^*||^2] &= ||\mathbf{x}_k^s-\mathbf{x}^*||^2-2\eta(\mathbf{x}_k^s-\mathbf{x}^*)^T\mathbb{E}[\hat{\mathbf{v}}_k^s]+\eta^2\mathbb{E}[||\hat{\mathbf{v}}_k^s||^2]\\
&\leq ||\mathbf{x}_k^s-\mathbf{x}^*||^2-2\eta[f_\mu(\mathbf{x}_k^s)-f_\mu(\mathbf{x}^*)]+\eta^2\mathbb{E}[||\hat{\mathbf{v}}_k^s||^2] \qquad (59)\\
&\leq ||\mathbf{x}_k^s-\mathbf{x}^*||^2-2\eta[f(\mathbf{x}_k^s)-f(\mathbf{x}^*)]+\eta^2\mathbb{E}[||\hat{\mathbf{v}}_k^s||^2]+2\eta L\mu^2
\end{aligned}$$

where the first inequality comes from the convexity of $f_\mu(x)$, the last inequality holds due to Lemma 1. Substituting Lemma 4 into (59), we have

$$\begin{aligned}
\mathbb{E}[||\mathbf{x}_{k+1}^s-\mathbf{x}^*||^2] \leq &||\mathbf{x}_k^s-\mathbf{x}^*||^2-2\eta(1-24\eta dL)[f(\mathbf{x}_k^s)-f(\mathbf{x}^*)]\\
&+48\eta^2dL[f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*)]+8\eta^2L^2d^2\mu^2+2\eta L\mu^2
\end{aligned} \qquad (60)$$

Now we consider a fixed stage s, so that $\mathbf{x}_0^s=\tilde{\mathbf{x}}_{s-1}$ and $\tilde{\mathbf{x}}_s$ is selected after all of the updates have completed. With the choice of $\tilde{\mathbf{x}}_s$, we get $\sum_{i=1}^m[f(\mathbf{x}_k^s)-f(\mathbf{x}^*)]=m\mathbb{E}[f(\tilde{\mathbf{x}}_s)-f(\mathbf{x}^*)]$. By summing the previous inequality over $k=0,1,...,m-1$, we have

$$\begin{aligned}
&\mathbb{E}\left[||\mathbf{x}_m^s-\mathbf{x}^*||^2\right]+2m\eta(1-24\eta dL)\mathbb{E}\left[f(\tilde{\mathbf{x}}_s)-f(\mathbf{x}^*)\right]\\
&\leq \mathbb{E}\left[||\mathbf{x}_0^s-\mathbf{x}^*||^2\right]+48m\eta^2dL\left[f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*)\right]+2\eta m\mu^2L(4\eta d^2L+1)
\end{aligned} \qquad (61)$$

Note that using **option I**, we have $\mathbf{x}_0^s = \tilde{\mathbf{x}}_{s-1}$. With strong convexity of $f(\mathbf{x})$, we get

$$\begin{aligned}&\mathbb{E}[||\mathbf{x}_m^s - \mathbf{x}^*||^2] + 2m\eta(1-24\eta dL)\mathbb{E}[f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*)] \\ &\leq (\frac{2}{\gamma} + 48m\eta^2 dL)\mathbb{E}[f(\tilde{\mathbf{x}}_{s-1}) - f(\mathbf{x}^*)] + 2\eta m\mu^2 L(4d^2\eta^2 L + 1)\end{aligned} \tag{62}$$

Denote $\beta_1 = 2m\eta[1-24\eta dL]$, $\beta_2 = \frac{2}{\gamma} + 48m\eta^2 dL$, $\delta_\mu = \frac{2\eta m\mu^2 L(4\eta d^2 L+1)}{\beta_1 - \beta_2}$. This implies

$$\mathbb{E}[f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*)] - \delta_\mu \leq \frac{\beta_2}{\beta_1}\{\mathbb{E}[f(\tilde{\mathbf{x}}_{s-1}) - f(\mathbf{x}^*)] - \delta_\mu\} \tag{63}$$

Note that we need to choose $\eta < \frac{1}{48dL}$ and $m > \frac{1}{\gamma\eta(1-48\eta dL)}$ to ensure that $0 < \frac{\beta_2}{\beta_1} < 1$. Telescope the sum in $s = 1, 2, ..., S$, we get

$$\mathbb{E}\left[f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*)\right] \leq \delta_\mu + \left(\frac{\beta_2}{\beta_1}\right)^s (\Delta - \delta_\mu) \tag{64}$$

Then we complete the proof. ■

Theorem 4 *Suppose Assumptions 2 and 4 hold, $f(\tilde{\mathbf{x}}_0) - f(\mathbf{x}^*) \leq \Delta$ and $||\tilde{\mathbf{x}}_0 - \mathbf{x}^*||^2 \leq \Theta$. Using **Option II** in Algorithm 3, we have*

$$\mathbb{E}[f(\mathbf{x}_\alpha) - f(\mathbf{x}^*)] \leq \frac{\Theta + 48m\eta^2 dL\Delta}{2mS\eta(1-48\eta dL)} + \frac{\mu^2 L(4\eta d^2 L + 1)}{1 - 48\eta dL}$$

where $\mathbf{x}_\alpha$ is uniformly randomly chosen from $\{\{\mathbf{x}_k^s\}_{k=0}^{m-1}\}_{s=1}^S$, and $\eta < \frac{1}{48dL}$.

Proof From (61), we have

$$\begin{aligned}&\mathbb{E}[||\mathbf{x}_m^s - \mathbf{x}^*||^2] + 2m\eta(1-24\eta dL)\mathbb{E}[f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*)] \\ &\leq \mathbb{E}[||\mathbf{x}_0^s - \mathbf{x}^*||^2] + 48m\eta^2 dL[f(\tilde{\mathbf{x}}_{s-1}) - f(\mathbf{x}^*)] + 2\eta m\mu^2 L(4\eta d^2 L + 1) \\ &= \mathbb{E}[||\mathbf{x}_m^{s-1} - \mathbf{x}^*||^2] + 48m\eta^2 dL[f(\tilde{\mathbf{x}}_{s-1}) - f(\mathbf{x}^*)] + 2\eta m\mu^2 L(4\eta d^2 L + 1)\end{aligned} \tag{65}$$

The first equality holds because in **option II**, we set $\mathbf{x}_0^s = \mathbf{x}_m^{s-1}$. Now we define the Lyapunov function

$$P^s \stackrel{\text{def}}{=} \mathbb{E}\left[||\mathbf{x}_m^s - \mathbf{x}^*||^2\right] + 48m\eta^2 dL\mathbb{E}\left[f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*)\right]$$

Then we have

$$2m\eta(1-48\eta dL)\mathbb{E}\left[f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*)\right] \leq P^{s-1} - P^s + 2\eta m\mu^2 L(4\eta d^2 L + 1) \tag{66}$$

Telescope the sum in $s = 1, 2, ..., S$, we get

$$2m\eta(1-48\eta dL)\sum_{s=1}^S \mathbb{E}\left[f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*)\right] \leq P^0 - P^S + 2\eta mS\mu^2 L(4\eta d^2 L + 1) \tag{67}$$

With the definition of $\mathbf{x}_\alpha$ and the choice of $\tilde{\mathbf{x}}_s$, we have $\mathbb{E}[f(\mathbf{x}_\alpha)-f(\mathbf{x}^*)] = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}[f(\tilde{\mathbf{x}}_s)-f(\mathbf{x}^*)]$, thus we get

$$\begin{aligned}
&2mS\eta(1-48\eta dL)\mathbb{E}[f(\mathbf{x}_\alpha)-f(\mathbf{x}^*)] \\
&\leq P^0 + 4\eta mS\mu^2 L(2\eta d^2 L+1) \\
&= ||\tilde{\mathbf{x}}_0-\mathbf{x}^*||^2 + 48m\eta^2 dL[f(\tilde{\mathbf{x}}_0)-f(\mathbf{x}^*)] + 2\eta mS\mu^2 L(4\eta d^2 L+1)
\end{aligned} \tag{68}$$

which implies

$$\mathbb{E}[f(\mathbf{x}_\alpha)-f(\mathbf{x}^*)] \leq \frac{\Theta + 48m\eta^2 dL\Delta}{2mS\eta(1-48\eta dL)} + \frac{\mu^2 L(4\eta d^2 L+1)}{1-48\eta dL} \tag{69}$$

Thus we get Theorem 4. Note that we need to choose $\eta < \frac{1}{48dL}$. ■

Appendix D. Proof of ZO-SAGA

In this section we provide the proof of Theorem 5 and Theorem 6. We first present two auxiliary lemmas, Lemma 5 and 6.

Lemma 5 *Suppose Assumption 4 holds.* $\forall\, \mathbf{x}, \phi_i \in \mathbb{R}^d$, *we have :*

$$\frac{1}{n}\sum_{i=1}^{n}||\nabla f_i(\phi_i^k)-\nabla f_i(\mathbf{x}^*)||^2 \leq 2L\left[\frac{1}{n}\sum_{i=1}^{n}f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n}(\phi_i^k-\mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]$$

Proof See (Defazio et al., 2014, Lemma 6). ■

Lemma 6 *Suppose Assumption 4 holds. Take expection over all variables, we have*

$$\begin{aligned}
&\mathbb{E}\left[||\hat{\mathbf{v}}^k||^2\right] \\
\leq& 12dL\left((1+\beta)\left[f(\mathbf{x}^k)-f(\mathbf{x}^*)\right] + (1+\frac{1}{\beta})\left[\frac{1}{n}\sum_{i=1}^{n}f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n}(\phi_i^k-\mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]\right) \\
&+ 4L^2d^2\mu^2 + 8dL\left[f(\mathbf{x}^k)-f(\mathbf{x}^*)\right]
\end{aligned}$$

Proof Since i_k is uniformly randomly chosen, we can rewrite $\hat{\mathbf{v}}^k$ as

$$\hat{\mathbf{v}}^k = \hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k) - \mathbb{E}_{i_k}\left[\hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k)\right] + \hat{\nabla} f(\mathbf{x}^k) \tag{70}$$

Then we have

$$\begin{aligned}
\mathbb{E}\left[||\hat{\mathbf{v}}^k||^2\right] &= \mathbb{E}\left[||\hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k) - \mathbb{E}_{i_k}\left[\hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k)\right] + \hat{\nabla} f(\mathbf{x}^k)||^2\right] \\
&\leq 2\mathbb{E}\left[||\hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k) - \mathbb{E}_{i_k}\left[\hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k)\right]||^2\right] + 2\mathbb{E}\left[||\hat{\nabla} f(\mathbf{x}^k)||^2\right] \\
&\leq 2\mathbb{E}\left[||\hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k)||^2\right] + 4d\mathbb{E}\left[||\nabla f(\mathbf{x}^k)||^2\right] + \mu^2 d^2 L^2
\end{aligned} \tag{71}$$

where the first inequality holds due to $||\mathbf{a}+\mathbf{b}||^2 \leq 2||\mathbf{a}||^2 + 2||\mathbf{b}||^2$, the second inequality holds due to $\mathbb{E}[||\mathbf{a}-\mathbb{E}[\mathbf{a}]||^2] \leq \mathbb{E}[||\mathbf{a}||^2]$ and Lemma 1. From Lemma 3, we bound the first term on the right hand side of (71) as

$$\begin{aligned}&\mathbb{E}\left[||\hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k)||^2\right] \\ &\leq 3d\left((1+\beta)\mathbb{E}\left[||\nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\mathbf{x}^*)||^2 + (1+\frac{1}{\beta})\mathbb{E}\left[||\nabla f_{i_k}(\phi_{i_k}^k) - \nabla f_{i_k}(\mathbf{x}^*)||^2\right]\right]\right) + \frac{3L^2d^2\mu^2}{2}\end{aligned} \tag{72}$$

From Lemma 2, we have

$$\mathbb{E}\left[||\nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\mathbf{x}^*)||^2\right] \leq 2L\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \tag{73}$$

From Lemma 5 and the fact that $\mathbb{E}\left[||\nabla f_{i_k}(\phi_{i_k}^k) - \nabla f_{i_k}(\mathbf{x}^*)||^2\right] = \frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\phi_i^k) - \nabla f_i(\mathbf{x}^*)||^2$ we have

$$\mathbb{E}\left[||\nabla f_{i_k}(\phi_{i_k}^k) - \nabla f_{i_k}(\mathbf{x}^*)||^2\right] \leq 2L\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n (\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right] \tag{74}$$

Thus we have

$$\begin{aligned}&\mathbb{E}\left[||\hat{\nabla} f_{i_k}(\mathbf{x}^k) - \hat{\nabla} f_{i_k}(\phi_{i_k}^k)||^2\right] \\ &\leq 6dL\left((1+\beta)\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] + (1+\frac{1}{\beta})\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n (\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]\right) \\ &\quad + \frac{3L^2d^2\mu^2}{2}\end{aligned} \tag{75}$$

Since $f(\mathbf{x})$ is L-smooth and convex, we can bound the second term on the right hand side as

$$\mathbb{E}\left[||\nabla f(\mathbf{x}^k)||^2\right] \leq 2L\mathbb{E}\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \tag{76}$$

Substitute (75) and (76) into (71), we get

$$\begin{aligned}&\mathbb{E}\left[||\hat{\mathbf{v}}^k||^2\right] \\ &\leq 12dL\left((1+\beta)\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] + (1+\frac{1}{\beta})\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n (\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]\right) \\ &\quad + 4L^2d^2\mu^2 + 8dL\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right]\end{aligned} \tag{77}$$

Then we complete the proof. ■

Theorem 5 *Suppose Assumptions 1 and 4 hold, denote $f(\mathbf{x}^0) - f(\mathbf{x}^*) = \Delta$. We have*

$$\mathbb{E}\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \leq \delta_\mu + \left(1 - \frac{\gamma}{112dL + n\gamma}\right)^K\left[\frac{L(2+c\gamma)}{2\gamma}\Delta - \delta_\mu\right]$$

where $c = \eta n(1 - 32\eta dL)$, $\eta = \frac{2}{112dL+n\gamma}$ *and* $\delta_\mu = \frac{2L^2\mu^2(2\eta Ld^2+1)}{\gamma}$.

Proof From Lemma 1, we have $\mathbb{E}[\hat{\mathbf{v}}^k] = \nabla f_\mu(\mathbf{x}^k)$, Conditioned on $\mathbf{x}^k$, we have

$$\begin{aligned}
&\mathbb{E}\left[||\mathbf{x}^{k+1} - \mathbf{x}^*||^2\right] = ||\mathbf{x}^k - \mathbf{x}^*||^2 - 2\eta(\mathbf{x}^k - \mathbf{x}^*)^T\mathbb{E}\left[\hat{\mathbf{v}}^k\right] + \eta^2\mathbb{E}\left[||\hat{\mathbf{v}}_k^s||^2\right] \\
&\leq ||\mathbf{x}^k - \mathbf{x}^*||^2 - 2\eta\left[f_\mu(\mathbf{x}^k) - f_\mu(\mathbf{x}^*)\right] + \eta^2\mathbb{E}\left[||\hat{\mathbf{v}}^k||^2\right] \\
&\leq ||\mathbf{x}^k - \mathbf{x}^*||^2 - 2\eta\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] + 2\eta L\mu^2 \\
&\quad + \eta^2 12dL\left((1+\beta)\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] + (1+\frac{1}{\beta})\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n(\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]\right) \\
&\quad + 4\eta^2L^2d^2\mu^2 + 8\eta^2dL\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \\
&\leq (1 - \frac{\eta\gamma}{2})||\mathbf{x}^k - \mathbf{x}^*||^2 - \eta\left[1 - 4\eta dL(3\beta+5)\right]\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \\
&\quad + 12\eta^2dL(1+\frac{1}{\beta})\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n(\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right] + 2\eta L\mu^2(2\eta Ld^2+1)
\end{aligned} \tag{78}$$

where the second inequality comes from Lemma 1 and Lemma 6, and the last imequality holds because $-\eta\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \leq -\frac{\eta\gamma}{2}||\mathbf{x}^k - \mathbf{x}^*||^2$. Define

$$T^k \stackrel{\text{def}}{=} ||\mathbf{x}^k - \mathbf{x}^*||^2 + c\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n(\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]$$

Note that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^{k+1})\right] = \frac{1}{n}f(\mathbf{x}^k) + (1-\frac{1}{n})\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) \tag{79}$$

$$\begin{aligned}
\mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^n(\phi_i^{k+1} - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right] &= -\frac{1}{n}\left\langle\nabla f(\mathbf{x}^*), \mathbf{x}^k - \mathbf{x}^*\right\rangle - (1-\frac{1}{n})\frac{1}{n}\sum_{i=1}^n\left\langle\nabla f_i(\mathbf{x}^*), \phi_i^k - \mathbf{x}^*\right\rangle \\
&= -(1-\frac{1}{n})\frac{1}{n}\sum_{i=1}^n\left\langle\nabla f_i(\mathbf{x}^*), \phi_i^k - \mathbf{x}^*\right\rangle
\end{aligned} \tag{80}$$

Denote $\delta = \frac{4L\mu^2(2\eta Ld^2+1)}{\gamma}$, then from (78) by rearranging the terms we can immediately get

$$\begin{aligned}
&\mathbb{E}\left[T^{k+1} - \delta\right] \\
\leq&(1 - \frac{\eta\gamma}{2})\left[T^k - \delta\right] + \left(\frac{c}{n} - \eta\left[1 - 4\eta dL(3\beta+5)\right]\right)\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \\
&+ \left[12\eta^2dL(1+\frac{1}{\beta}) - (\frac{1}{n} - \frac{\eta\gamma}{2})c\right]\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n(\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]
\end{aligned} \tag{81}$$

Due to the convexity of $f_i(\mathbf{x})$, we have

$$\frac{1}{n}\sum_{i=1}^{n} f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n}(\phi_i^k - \mathbf{x}^*)^T \nabla f_i(\mathbf{x}^*) \geq 0 \tag{82}$$

By setting $\eta = \frac{2}{112dL+n\gamma}, \beta = 1, c = \eta n\left[1 - 4\eta dL(3\beta + 5)\right]$, we can ensure the second and third terms on the right hand side of (81) are non-positive, which implies

$$\mathbb{E}\left[T^{k+1} - \delta\right] \leq (1 - \frac{\eta\gamma}{2})\left[T^k - \delta\right] \tag{83}$$

Telescope the sum, we get

$$\begin{aligned}\mathbb{E}\left[T^K - \delta\right] &\leq (1 - \frac{1}{112dL + n\gamma})^K \left[T^0 - \delta\right] \\ &= (1 - \frac{1}{112dL + n\gamma})^K \left(||\mathbf{x}^k - \mathbf{x}^*||^2 + c\left[f(\mathbf{x}^0) - f(\mathbf{x}^*)\right] - \delta\right)\end{aligned} \tag{84}$$

From (82), we have $||\mathbf{x}^k - \mathbf{x}^*||^2 \leq T^k$. With the smoothness and strong convextiy of $f(\mathbf{x})$, we have

$$\begin{aligned}f(\mathbf{x}^k) - f(\mathbf{x}^*) &\leq \frac{L}{2}\left[||\mathbf{x}^k - \mathbf{x}^*||^2\right] \leq \frac{LT^k}{2} \\ &\leq \frac{L}{2}\delta + \frac{L}{2}(1 - \frac{1}{112dL + n\gamma})^K \left(\frac{2}{\gamma}\left[f(\mathbf{x}^0) - f(\mathbf{x}^*)\right] + c\left[f(\mathbf{x}^0) - f(\mathbf{x}^*)\right] - \delta\right)\end{aligned} \tag{85}$$

Denote $\delta_\mu = \frac{L}{2}\delta$, we get

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \delta_\mu + (1 - \frac{1}{112dL + n\gamma})^K \left(\frac{L(2 + c\gamma)}{2\gamma}\left[f(\mathbf{x}^0) - f(\mathbf{x}^*)\right] - \delta_\mu\right) \tag{86}$$

where $c = \eta n(1 - 32\eta dL)$ and $\eta = \frac{2}{112dL+n\gamma}$. Then we complete the proof. ■

Theorem 6 *Suppose Assumptions 2 and 4 hold. For Algorithm 4, we have*

$$\mathbb{E}\left[f(\mathbf{x}^\tau) - f(\mathbf{x}^*)\right] \leq \frac{56dLT^0}{K} + 112\eta L\mu^2(2\eta Ld^2 + 1)$$

where $\eta = \frac{1}{56dL}, T^0 = ||\mathbf{x}^0 - \mathbf{x}^*||^2 + \frac{3}{7}\eta n\left[f(\mathbf{x}^0) - f(\mathbf{x}^*)\right]$ *and* $\mathbf{x}^\tau$ *is uniformly randomly chosen from* $\left\{\mathbf{x}^k\right\}_{k=1}^K$.

Proof From Lemma 1, we have $\mathbb{E}[\hat{\mathbf{v}}^k] = \nabla f_\mu(\mathbf{x}^k)$, Conditioned on $\mathbf{x}^k$, we have

$$\begin{aligned}
&\mathbb{E}\left[||\mathbf{x}^{k+1} - \mathbf{x}^*||^2\right] = ||\mathbf{x}^k - \mathbf{x}^*||^2 - 2\eta(\mathbf{x}^k - \mathbf{x}^*)^T\mathbb{E}\left[\hat{\mathbf{v}}^k\right] + \eta^2\mathbb{E}\left[||\hat{\mathbf{v}}_k^s||^2\right] \\
&\leq ||\mathbf{x}^k - \mathbf{x}^*||^2 - 2\eta\left[f_\mu(\mathbf{x}^k) - f_\mu(\mathbf{x}^*)\right] + \eta^2\mathbb{E}\left[||\hat{\mathbf{v}}^k||^2\right] \\
&\leq ||\mathbf{x}^k - \mathbf{x}^*||^2 - 2\eta\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] + 2\eta L\mu^2 \\
&\quad + \eta^2 12dL\left((1+\beta)\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] + (1+\frac{1}{\beta})\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n(\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]\right) \\
&\quad + 4\eta^2L^2d^2\mu^2 + 8\eta^2dL\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \\
&= ||\mathbf{x}^k - \mathbf{x}^*||^2 - 2\eta\left[1 - 2\eta dL(3\beta+5)\right]\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] \\
&\quad + 12\eta^2dL(1+\frac{1}{\beta})\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n(\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right] + 2\eta L\mu^2(2\eta Ld^2 + 1)
\end{aligned} \tag{87}$$

The second inequality comes from Lemma 1 and Lemma 6.

Define $T^k \overset{\text{def}}{=} ||\mathbf{x}^k - \mathbf{x}^*||^2 + \alpha\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n(\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]$, which is the Lyapunov Function, and note that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^{k+1})\right] = \frac{1}{n}f(\mathbf{x}^k) + (1-\frac{1}{n})\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) \tag{88}$$

$$\begin{aligned}
\mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^n(\phi_i^{k+1} - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right] &= -\frac{1}{n}\left\langle\nabla f(\mathbf{x}^*), \mathbf{x}^k - \mathbf{x}^*\right\rangle - (1-\frac{1}{n})\frac{1}{n}\sum_{i=1}^n\left\langle\nabla f_i(\mathbf{x}^*), \phi_i^k - \mathbf{x}^*\right\rangle \\
&= -(1-\frac{1}{n})\frac{1}{n}\sum_{i=1}^n\left\langle\nabla f_i(\mathbf{x}^*), \phi_i^k - \mathbf{x}^*\right\rangle
\end{aligned} \tag{89}$$

Then we have

$$\begin{aligned}
\mathbb{E}\left[T^{k+1}\right] \leq& T^k + \left\{\frac{\alpha}{n} - 2\eta\left[1 - 2\eta dL(3\beta+5)\right]\right\}\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] + 2\eta L\mu^2(2\eta Ld^2+1) \\
&+ \left[12\eta^2dL(1+\frac{1}{\beta}) - \frac{\alpha}{n}\right]\left[\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n(\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*)\right]
\end{aligned} \tag{90}$$

Due to the convexity of $f_i(\mathbf{x})$, we have

$$\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^n(\phi_i^k - \mathbf{x}^*)^T\nabla f_i(\mathbf{x}^*) \geq 0 \tag{91}$$

By setting $\beta = 1, \eta = \frac{1}{4dL(3\beta+3\beta^{-1}+8)} = \frac{1}{56dL}, \alpha = \frac{3n}{392dL} = \frac{3}{7}\eta n$, we can ensure the last term on the right hand side of (90) is non-positive, which implies

$$\mathbb{E}\left[T^{k+1}\right] \leq T^k - \frac{1}{56dL}\left[f(\mathbf{x}^k) - f(\mathbf{x}^*)\right] + 2\eta L\mu^2(2\eta Ld^2+1) \tag{92}$$

Telescope the sum and rearrange the terms, we get

$$\frac{1}{56dL}\sum_{k=1}^{K}\left[f(\mathbf{x}^k)-f(\mathbf{x}^*)\right]\leq T^0-\mathbb{E}\left[T^{K+1}\right]+2K\eta L\mu^2(2\eta Ld^2+1)\leq T^0+2K\eta L\mu^2(2\eta Ld^2+1) \tag{93}$$

The second inequality comes from the fact that T^k is always positive. Since $\mathbb{E}\left[f(\mathbf{x}^\tau)\right]=\frac{1}{K}\sum_{k=1}^{K}\left[f(\mathbf{x}^k)-f(\mathbf{x}^*)\right]$, then we get

$$[f(\mathbf{x}^\tau)-f(\mathbf{x}^*)]\leq\frac{56dLT^0}{K}+112\eta L\mu^2(2\eta Ld^2+1) \tag{94}$$

From the definition of T^0, we know that $T^0=||\mathbf{x}^0-\mathbf{x}^*||^2+\frac{3}{7}\eta n\left[f(\mathbf{x}^0)-f(\mathbf{x}^*)\right]$. Then we complete the proof. ■

Appendix E. Proof of ZO-Varag

In this section, we provide the proof of Theorem 7. Specifically, Theorem 7 is a direct result of Lemma 8 and 9. The following Lemma corresponds to Lemma 25 in (Chen et al., 2020). They are the same except the notations.

Lemma 7 *Suppose Assumptions 1, 4 and 5 hold. Under the choice of parameters from Theorem 7, we have*

$$\begin{aligned}&\mathbb{E}\left[\frac{\beta_s}{\alpha_s}\left[f(\bar{\mathbf{x}}_k)-f(\mathbf{x}^*)\right]+\frac{1+\gamma\beta_s}{2}\|\mathbf{x}_k-\mathbf{x}^*\|^2\right]\\ \leq&\frac{\beta_s}{\alpha_s}(1-\alpha_s-p_s)\left[f(\bar{\mathbf{x}}_{k-1})-f(\mathbf{x}^*)\right]+\frac{\beta_s p_s}{\alpha_s}\left[f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right]+\frac{1}{2}\|\mathbf{x}_{k-1}-\mathbf{x}^*\|^2\\ &+\frac{\beta_s}{\alpha_s}\cdot\frac{3}{4}\mu^2Ld+\frac{\beta_s}{\alpha_s}(2-\alpha_s)L\sqrt{d}Z\mu\end{aligned}$$

The following Lemma corresponds to Lemma 27 in (Chen et al., 2020). We make minor modifications so that it satisfies our ZOOD property.

Lemma 8 *Suppose Assumptions 1, 4 and 5 hold. Under the choice of parameters from Theorem 7, if $n\geq\frac{18L}{\gamma}$, we have*

$$\mathbb{E}\left[f(\tilde{\mathbf{x}}^S)-f(\mathbf{x}^*)\right]\leq\left(\frac{1}{4}\right)^S\frac{4}{3}\mathbb{E}\left[f(\tilde{\mathbf{x}}^0)-f(\mathbf{x}^*)\right]+\frac{1}{2}\mu^2Ld+\frac{4}{3}L\sqrt{d}Z\mu$$

Proof For this case, $\alpha_s=\alpha=p_s=\frac{1}{2}$, $\beta_s=\beta=\frac{1}{6L}$, $m_s=n$. Based on Lemma 7, we have

$$\begin{aligned}&\mathbb{E}\left[\frac{\beta}{\alpha}\left[f(\bar{\mathbf{x}}_k)-f(\mathbf{x}^*)\right]+(1+\gamma\beta)\cdot\frac{1}{2}\|\mathbf{x}-\mathbf{x}^*\|^2\right]\\ \leq&\frac{\beta}{2\alpha}\left[f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right]+\frac{1}{2}\|\mathbf{x}_{k-1}-\mathbf{x}^*\|^2+\frac{\beta}{\alpha}\cdot\frac{3}{4}\mu^2Ld+\frac{\beta}{\alpha}2L\sqrt{d}Z\mu\end{aligned} \tag{95}$$

Multiplying both sides by $\Gamma_{k-1} = (1+\gamma\beta)^{k-1}$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\frac{\beta}{\alpha}\Gamma_{k-1}\left[f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^*)\right] + \frac{\Gamma_k}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2\right]\\ \leq&\frac{\beta}{2\alpha}\Gamma_{k-1}\left[f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right] + \frac{\Gamma_{k-1}}{2}\|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 + \frac{\beta}{\alpha}\Gamma_{k-1}\cdot\frac{3}{4}\mu^2 Ld + \frac{\beta}{\alpha}\Gamma_{k-1}\cdot 2L\sqrt{d}Z\mu\end{aligned} \tag{96}$$

Since $\theta_k = \Gamma_{k-1}$, the last inequality can be rewritten as

$$\begin{aligned}&\mathbb{E}\left[\frac{\beta}{\alpha}\theta_k\left[f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^*)\right] + \frac{\Gamma_k}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2\right]\\ \leq&\frac{\beta}{2\alpha}\theta_k\left[f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right] + \frac{\Gamma_{k-1}}{2}\|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 + \frac{\beta}{\alpha}\theta_k\cdot\frac{3}{4}\mu^2 Ld + \frac{\beta}{\alpha}\theta_k\cdot 2L\sqrt{d}Z\mu\end{aligned} \tag{97}$$

Summing up the inequality above from $k=1$ to m_s, we obtain

$$\begin{aligned}&\frac{\beta}{\alpha}\sum_{k=1}^{m_s}\theta_k\mathbb{E}\left[f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^*)\right] + \frac{\Gamma_{m_s}}{2}\mathbb{E}\left[\|\mathbf{x}_{m_s} - \mathbf{x}^*\|^2\right]\\ \leq&\frac{\beta}{2\alpha}\sum_{k=1}^{m_s}\theta_k\mathbb{E}\left[f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right] + \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{\beta}{\alpha}\cdot\frac{3}{4}\mu^2 Ld\sum_{k=1}^{m_s}\theta_k + \frac{\beta}{\alpha}\cdot 2L\sqrt{d}Z\mu\sum_{k=1}^{m_s}\theta_k\end{aligned} \tag{98}$$

and then

$$\begin{aligned}&4\left[\frac{\beta}{2\alpha}\sum_{k=1}^{m_s}\theta_k\mathbb{E}\left[f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^*)\right]\right] + \frac{\Gamma_{m_s}}{2}\mathbb{E}\left[\|\mathbf{x}_{m_s} - \mathbf{x}^*\|^2\right]\\ \leq&\frac{\beta}{2\alpha}\sum_{k=1}^{m_s}\theta_k\mathbb{E}\left[f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right] + \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{\beta}{\alpha}\cdot\frac{3}{4}\mu^2 Ld\sum_{k=1}^{m_s}\theta_k + \frac{\beta}{\alpha}\cdot 2L\sqrt{d}Z\mu\sum_{k=1}^{m_s}\theta_k\end{aligned} \tag{99}$$

which is based on the fact that

$$\Gamma_{m_s} = (1+\gamma\beta)^{m_s} \geq 1 + \gamma\beta m_s = 1 + \frac{\gamma n}{6L} \geq 4$$

where the last inequality is conditioned on $n \geq \frac{18L}{\gamma}$. Since $\tilde{\mathbf{x}}^s = \sum_{k=1}^{m_s}(\theta_k\bar{\mathbf{x}}_k)/(\sum_{k=1}^{m_s}\theta_k)$, $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}^{s-1}$, $\mathbf{x}_0 = \mathbf{x}^{s-1}$, $\mathbf{x}_{m_s} = \mathbf{x}^s$ in the epoch s and thanks to the convexity of f, (99) implies

$$\begin{aligned}&4\left[\frac{\beta}{2\alpha}\mathbb{E}\left[f(\tilde{\mathbf{x}}^s) - f(\mathbf{x}^*)\right]\right] + \frac{1}{2\sum_{k=1}^{m_s}\theta_k}\mathbb{E}\left[\|\mathbf{x}^s - \mathbf{x}^*\|^2\right]\\ \leq&\frac{\beta}{2\alpha}\mathbb{E}\left[f(\tilde{\mathbf{x}}^{s-1}) - f(\mathbf{x}^*)\right] + \frac{1}{2\sum_{k=1}^{m_s}\theta_k}\|\mathbf{x}^{s-1} - \mathbf{x}^*\|^2 + \frac{\beta}{\alpha}\cdot\frac{3}{4}\mu^2 Ld + \frac{\beta}{\alpha}\cdot 2L\sqrt{d}Z\mu\end{aligned} \tag{100}$$

Multiplying both sides with $\frac{2\alpha}{\beta}$ and applying this inequality recursively for $s = 1, ..., S$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{\mathbf{x}}^S) - f(\mathbf{x}^*)\right] + \frac{2\alpha}{\gamma\sum_{k=1}^{m_s}\theta_k} \cdot \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^S - \mathbf{x}^*\|^2\right] \\
\leq &\left(\frac{1}{4}\right)^S \left[\mathbb{E}\left[f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*)\right] + \frac{2\alpha}{\gamma\sum_{k=1}^{m_s}\theta_k} \cdot \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right]\right] + \sum_{s=1}^{S}\left(\frac{1}{4}\right)^s \left[\frac{3}{2}\mu^2 Ld + 4L\sqrt{d}Z\mu\right] \\
\leq &\left(\frac{1}{4}\right)^S \left[\mathbb{E}\left[f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*)\right] + \frac{2\alpha}{\gamma n} \cdot \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right]\right] + \frac{1}{2}\mu^2 Ld + \frac{4}{3}L\sqrt{d}Z\mu
\end{aligned} \tag{101}$$

where the last inequality holds since $\sum_{s=1}^{S}\left(\frac{1}{4}\right)^s \leq \frac{1}{4} \cdot \frac{1}{1-\frac{1}{4}} = \frac{1}{3}$ and $\sum_{k=1}^{m_s}\theta_k \geq m_s = n$. From the convexity of f and the fact that $n \geq \frac{18L}{\gamma}$, we have

$$\frac{2\alpha}{\gamma n} \cdot \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right] = \frac{6L}{n} \cdot \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right] \leq \frac{6L}{n\gamma}\mathbb{E}\left[f(\mathbf{x}^0) - f(\mathbf{x}^*)\right] \leq \frac{1}{3}\mathbb{E}\left[f(\mathbf{x}^0) - f(\mathbf{x}^*)\right]$$

Note that $\mathbf{x}^0 = \tilde{\mathbf{x}}^0$. Combining the above inequality with (101), we have

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{\mathbf{x}}^S) - f(\mathbf{x}^*)\right] \leq \mathbb{E}\left[f(\tilde{\mathbf{x}}^S) - f(\mathbf{x}^*)\right] + \frac{2\alpha}{\gamma\sum_{k=1}^{m_s}\theta_k} \cdot \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^S - \mathbf{x}^*\|^2\right] \\
\leq &\left(\frac{1}{4}\right)^S \left[\mathbb{E}\left[f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*)\right] + \frac{2\alpha}{\gamma n} \cdot \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right]\right] + \frac{1}{2}\mu^2 Ld + \frac{4}{3}L\sqrt{d}Z\mu \\
\leq &\left(\frac{1}{4}\right)^S \frac{4}{3}\mathbb{E}\left[f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*)\right] + \frac{1}{2}\mu^2 Ld + \frac{4}{3}L\sqrt{d}Z\mu
\end{aligned} \tag{102}$$

Then we complete the proof. ■

The following Lemma corresponds to Lemma 28 in (Chen et al., 2020). We make minor modifications so that it satisfies our ZOOD property.

Lemma 9 *Suppose Assumptions 1, 4 and 5 hold. Under the choice of parameters from Theorem 7, if $n < \frac{18L}{\gamma}$, we have*

$$\mathbb{E}\left[f(\tilde{\mathbf{x}}^S) - f(\mathbf{x}^*)\right] \leq \left(1 + \sqrt{\frac{\gamma}{12nL}}\right)^{-nS} 3\mathbb{E}\left[f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*)\right] + \frac{S}{\Gamma_n}\left[\frac{3}{2}\mu^2 Ld + (4 - 2\alpha_s)L\sqrt{d}Z\mu\right]$$

Proof For this case, $\alpha_s = \alpha = \sqrt{\frac{n\gamma}{12L}}$, $p_s = p = \frac{1}{2}$, $\beta_s = \beta = \frac{1}{\sqrt{12nL\gamma}}$, $m_s = n$. Based on Lemma 7, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\beta_s}{\alpha_s}\left[f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^*)\right] + \frac{1 + \gamma\beta_s}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] \\
\leq &\frac{\beta_s}{\alpha_s}(1 - \alpha_s - p_s)\left[f(\bar{\mathbf{x}}_{k-1}) - f(\mathbf{x}^*)\right] + \frac{\beta_s p_s}{\alpha_s}\left[f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right] + \frac{1}{2}\|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 \\
&+ \frac{\beta_s}{\alpha_s} \cdot \frac{3}{4}\mu^2 Ld + \frac{\beta_s}{\alpha_s}(2 - \alpha_s)L\sqrt{d}Z\mu
\end{aligned} \tag{103}$$

Multiplying both sides by $\Gamma_{k-1} = (1+\gamma\beta)^{k-1}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{\Gamma_{k-1}\beta_s}{\alpha_s}\left[f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^*)\right] + \frac{\Gamma_k}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] \\
\leq&\frac{\Gamma_{k-1}\beta_s}{\alpha_s}(1-\alpha_s-p_s)\left[f(\bar{\mathbf{x}}_{k-1}) - f(\mathbf{x}^*)\right] + \frac{\Gamma_{k-1}\beta_s p_s}{\alpha_s}\left[f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right] + \frac{\Gamma_{k-1}}{2}\|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 \\
&+ \frac{\Gamma_{k-1}\beta_s}{\alpha_s}\cdot\frac{3}{4}\mu^2 Ld + \frac{\Gamma_{k-1}\beta_s}{\alpha_s}(2-\alpha_s)L\sqrt{d}Z\mu
\end{aligned} \tag{104}$$

Summing up the inequality above from $k=1$ to m_s, we obtain

$$\begin{aligned}
&\frac{\beta}{\alpha}\sum_{k=1}^{m_s}\theta_k\mathbb{E}\left[f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^*)\right] + \frac{\Gamma_{m_s}}{2}\mathbb{E}\left[\|\mathbf{x}_{m_s} - \mathbf{x}^*\|^2\right] \\
\leq&\frac{\beta}{\alpha}\left(1-\alpha-p+p\sum_{k=1}^{m_s}\Gamma_{k-1}\right)\mathbb{E}\left[f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right] + \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \\
&+ \frac{\beta}{\alpha}\sum_{k=1}^{m_s}\Gamma_{k-1}\cdot\frac{3}{4}\mu^2 Ld + \frac{\beta}{\alpha}\sum_{k=1}^{m_s}\Gamma_{k-1}\cdot(2-\alpha_s)L\sqrt{d}Z\mu
\end{aligned} \tag{105}$$

Since $\tilde{\mathbf{x}}^s = \sum_{k=1}^{m_s}(\theta_k\bar{\mathbf{x}}_k)/(\sum_{k=1}^{m_s}\theta_k)$, $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}^{s-1}$, $\mathbf{x}_0 = \mathbf{x}^{s-1}$, $\mathbf{x}_{m_s} = \mathbf{x}^s$ in the epoch s and thanks to the convexity of f, it implies

$$\begin{aligned}
&\frac{\beta}{\alpha}\sum_{k=1}^{m_s}\theta_k\mathbb{E}\left[f(\tilde{\mathbf{x}}^s) - f(\mathbf{x}^*)\right] + \frac{\Gamma_{m_s}}{2}\mathbb{E}\left[\|\mathbf{x}^s - \mathbf{x}^*\|^2\right] \\
\leq&\frac{\beta}{\alpha}\left(1-\alpha-p+p\sum_{k=1}^{m_s}\Gamma_{k-1}\right)\mathbb{E}\left[f(\tilde{\mathbf{x}}^{s-1}) - f(\mathbf{x}^*)\right] + \frac{1}{2}\|\mathbf{x}^{s-1} - \mathbf{x}^*\|^2 \\
&+ \frac{\beta}{\alpha}\sum_{k=1}^{m_s}\Gamma_{k-1}\cdot\frac{3}{4}\mu^2 Ld + \frac{\beta}{\alpha}\sum_{k=1}^{m_s}\Gamma_{k-1}\cdot(2-\alpha_s)L\sqrt{d}Z\mu
\end{aligned} \tag{106}$$

Moreover, we have

$$\begin{aligned}
\sum_{k=1}^{m_s}\theta_k =&\Gamma_{m_s-1} + \sum_{k=1}^{m_s-1}\left(\Gamma_{k-1} - (1-\alpha-p)\Gamma_k\right) \\
=&\Gamma_{m_s}(1-\alpha-p) + \sum_{k=1}^{m_s}\left(\Gamma_{k-1} - (1-\alpha-p)\Gamma_k\right) \\
=&\Gamma_{m_s}(1-\alpha-p) + \left[1-(1-\alpha-p)(1+\gamma\beta)\right]\sum_{k=1}^{m_s}\Gamma_{k-1}
\end{aligned} \tag{107}$$

And $\alpha = \sqrt{\frac{n\gamma}{12L}} = \gamma\beta m_s$. Then we have

$$\begin{aligned}
1-(1-\alpha-p)(1+\gamma\beta) =&(1+\gamma\beta)(\alpha+p-\gamma\beta) + \gamma^2\beta^2 \\
\geq&(1+\gamma\beta)(\gamma\beta m_s + p - \gamma\beta) \\
=&p(1+\gamma\beta)\left(1+2(m_s-1)\gamma\beta\right) \\
\geq&p(1+\gamma\beta)^{m_s} = p\Gamma_{m_s}
\end{aligned} \tag{108}$$

Hence we obtain $\sum_{k=1}^{m_s}\theta_k \geq \Gamma_{m_s}\left(1-\alpha-p+p\sum_{k=1}^{m_s}\Gamma_{k-1}\right)$. Thus (106) implies

$$\begin{aligned}
&\Gamma_{m_s}\left[\frac{\beta}{\alpha}\left(1-\alpha-p+p\sum_{k=1}^{m_s}\Gamma_{k-1}\right)\mathbb{E}\left[f(\tilde{\mathbf{x}}^s)-f(\mathbf{x}^*)\right]+\frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^s-\mathbf{x}^*\|^2\right]\right]\\
\leq&\frac{\beta}{\alpha}\left(1-\alpha-p+p\sum_{k=1}^{m_s}\Gamma_{k-1}\right)\mathbb{E}\left[f(\tilde{\mathbf{x}}^{s-1})-f(\mathbf{x}^*)\right]+\frac{1}{2}\|\mathbf{x}^{s-1}-\mathbf{x}^*\|^2\\
&+\frac{\beta}{\alpha}\sum_{k=1}^{m_s}\Gamma_{k-1}\cdot\frac{3}{4}\mu^2Ld+\frac{\beta}{\alpha}\sum_{k=1}^{m_s}\Gamma_{k-1}\cdot(2-\alpha_s)L\sqrt{d}Z\mu
\end{aligned}\tag{109}$$

Applying this inequality iteratively for $s=1,...,S$, we obtain

$$\begin{aligned}
&\frac{\beta}{\alpha}\left(1-\alpha-p+p\sum_{k=1}^{m_s}\Gamma_{k-1}\right)\mathbb{E}\left[f(\tilde{\mathbf{x}}^S)-f(\mathbf{x}^*)\right]+\frac{1}{2}\mathbb{E}\left[\|\mathbf{x}^S-\mathbf{x}^*\|^2\right]\\
\leq&\left(\frac{1}{\Gamma_{m_s}}\right)^S\left[\frac{\beta}{\alpha}\left(1-\alpha-p+p\sum_{k=1}^{m_s}\Gamma_{k-1}\right)\mathbb{E}\left[f(\tilde{\mathbf{x}}^0)-f(\mathbf{x}^*)\right]+\frac{1}{2}\|\mathbf{x}^0-\mathbf{x}^*\|^2\right]\\
&+\sum_{s=1}^{S}\left(\frac{1}{\Gamma_{m_s}}\right)^s\left[\frac{\beta}{\alpha}\sum_{k=1}^{m_s}\Gamma_{k-1}\cdot\frac{3}{4}\mu^2Ld+\frac{\beta}{\alpha}\sum_{k=1}^{m_s}\Gamma_{k-1}\cdot(2-\alpha_s)L\sqrt{d}Z\mu\right]
\end{aligned}\tag{110}$$

Note that, since

$$\frac{\beta}{\alpha}\left(1-\alpha-p+p\sum_{k=1}^{m_s}\Gamma_{k-1}\right)\geq\frac{\beta p}{\alpha}\sum_{k=1}^{m_s}\Gamma_{k-1}\geq\frac{\beta p m_s}{\alpha}=\frac{\beta p m_s}{n}$$

and $p=\frac{1}{2}$, the inequality above implies

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{\mathbf{x}}^S)-f(\mathbf{x}^*)\right]\\
\leq&\left(\frac{1}{\Gamma_{m_s}}\right)^S\left[\mathbb{E}\left[f(\tilde{\mathbf{x}}^0)-f(\mathbf{x}^*)\right]+\frac{\alpha}{\beta n}\mathbb{E}\left[\|\mathbf{x}^0-\mathbf{x}^*\|^2\right]\right]+\sum_{s=1}^{S}\left(\frac{1}{\Gamma_{m_s}}\right)^s\left[\frac{3}{2}\mu^2Ld+(4-2\alpha_s)L\sqrt{d}Z\mu\right]\\
\leq&\left(\frac{1}{\Gamma_{m_s}}\right)^S\left[\mathbb{E}\left[f(\tilde{\mathbf{x}}^0)-f(\mathbf{x}^*)\right]+\frac{\alpha}{\beta n}\mathbb{E}\left[\|\mathbf{x}^0-\mathbf{x}^*\|^2\right]\right]+\frac{S}{\Gamma_{m_s}}\left[\frac{3}{2}\mu^2Ld+(4-2\alpha_s)L\sqrt{d}Z\mu\right]
\end{aligned}\tag{111}$$

Since $\frac{\alpha}{\beta}=12L\alpha^2=n\gamma$, from the strong convexity of f we have

$$\frac{\alpha}{\beta n}\mathbb{E}\left[\|\mathbf{x}^0-\mathbf{x}^*\|^2\right]=\gamma\mathbb{E}\left[\|\mathbf{x}^0-\mathbf{x}^*\|^2\right]\leq 2\mathbb{E}\left[f(\mathbf{x}^0)-f(\mathbf{x}^*)\right]$$

Note that $\mathbf{x}^0=\tilde{\mathbf{x}}^0$, combining the above inequality with (111), we have

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{\mathbf{x}}^S)-f(\mathbf{x}^*)\right]\\
\leq&\left(\frac{1}{\Gamma_{m_s}}\right)^S 3\mathbb{E}\left[f(\tilde{\mathbf{x}}^0)-f(\mathbf{x}^*)\right]+\frac{S}{\Gamma_{m_s}}\left[\frac{3}{2}\mu^2Ld+(4-2\alpha_s)L\sqrt{d}Z\mu\right]\\
=&\left(1+\sqrt{\frac{\gamma}{12nL}}\right)^{-nS}3\mathbb{E}\left[f(\tilde{\mathbf{x}}^0)-f(\mathbf{x}^*)\right]+\frac{S}{\Gamma_n}\left[\frac{3}{2}\mu^2Ld+(4-2\alpha_s)L\sqrt{d}Z\mu\right]
\end{aligned}\tag{112}$$

Then we complete the proof. ■

Theorem 7 *Suppose Assumptions 1, 4 and 5 hold. Set*

$$m_s = n, \quad p_s = \frac{1}{2}, \quad \alpha_s = \min\left\{\sqrt{\frac{n\gamma}{12L}}, \frac{1}{2}\right\}, \quad \beta_s = \frac{1}{12L\alpha_s}$$

$$\Gamma_k = (1+\gamma\beta_s)^k, \qquad \theta_k = \begin{cases} \Gamma_{k-1} - (1-\alpha_s-p_s)\Gamma_k, & k \leq m_s - 1 \\ \Gamma_{k-1}, & k = m_s \end{cases}$$

We obtain

$$\begin{aligned} &\mathbb{E}\left[f(\tilde{\mathbf{x}}^S) - f(\mathbf{x}^*)\right] \\ &\leq \begin{cases} \left(\frac{1}{4}\right)^S \frac{4}{3}\mathbb{E}\left[f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*)\right] + \frac{1}{2}\mu^2 Ld + \frac{4}{3}L\sqrt{d}Z\mu, & n \geq \frac{18L}{\gamma} \\ \left(1+\sqrt{\frac{\gamma}{12nL}}\right)^{-nS} 3\mathbb{E}\left[f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*)\right] + \frac{S}{\Gamma_n}\left[\frac{3}{2}\mu^2 Ld + (4-2\alpha_s)L\sqrt{d}Z\mu\right], & n < \frac{18L}{\gamma} \end{cases} \end{aligned}$$

Proof The result can be directly derived from Lemma 8 and 9. ■

Appendix F. Additional Experiment Results

In this section, we provide our choice of parameters and more experiments results of *AdaptRdct-C* (ZO-SVRG/ZO-SAGA/ZO-Varag) and *AdaptRdct-NC* (ZO-SVRG/ZO-SAGA/ZO-Varag) under different parameter settings.

Choice of Parameters

The parameters of the algorithms are set according to their convergence analyses. To be specific, under convex setting, the parameters of ZO-SVRG and ZO-SAGA are set according to Theorem 4 and 6 respectively, the parameters of *AdaptRdct-C* (ZO-SVRG/ZO-SAGA/ZO-Varag) are set according to Theorem 3, 5 and 7 respectively. The parameters of ZO-Varag are set according to (Chen et al., 2020, Theorem 6). The parameters of ZO-SPIDER-Coord are set according to (Ji et al., 2019, Appendix, Theorem 7).

Under non-convex setting, the parameters of ZO-SVRG are set according to (Liu et al., 2018, Corollary 1). The parameters of *AdaptRdct-NC* (ZO-SVRG/ZO-SAGA/ZO-Varag) are set according to Theorem 3, 5 and 7 respectively. The parameters of ZO-SPIDER-Coord are set according to (Ji et al., 2019, Corollary 3).

In the experiment of generation of black-Box adversarial examples, we set a mini-batch size of 10 for all the algorithms. We conduct grid search for the regularization parameter σ on {1e-2, 1e-1, 1} for algorithms with our reduction frameworks. For all the algorithms, we conduct grid search for the Lipschitz constant L on {1, 1e1, 1e2, 1e3}. Note that L is related to the choice of learning rate. The smoothing parameter μ is set to $1/d$.

In the experiment of logistic regression, we set a mini-batch size of 64 for all the algorithms. We conduct grid search for the regularization parameter γ_0 and σ on {1e-4, 5e-4,

1e-3, 5e-3, 1e-2} for algorithms with our reduction frameworks under convex and non-convex settings, respectively. For all the algorithms, we conduct grid search for the Lipschitz constant L on {1e-2, 5e-2, 1e-1, 5e-1, 1, 5, 1e1, 5e1, 1e2}. The smoothing parameter μ is set to $1/d$.

Table 3: Choices of parameters for the experiment of generation of black-Box adversarial examples.

Algorithms	Cifar-10		Fmnist		Mnist	
	σ	L	σ	L	σ	L
ZO-SVRG	-	1	-	1	-	1
ZO-SPIDER-Coord	-	1	-	1	-	1
AdaptRdct-NC (ZO-SVRG)	1	1	1	1	1	1
AdaptRdct-NC (ZO-SAGA)	1e-1	1	1	1	1	1
AdaptRdct-NC (ZO-Varag)	1e-2	1	1e-2	1	1e-1	1

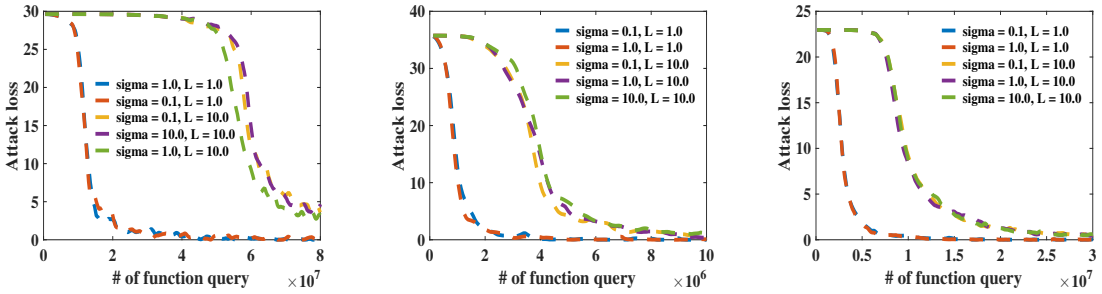
Table 4: Choices of parameters for the experiment of *convex* logistic regression.

Algorithms	German		Ijcnn1		Mushrooms	
	γ_0	L	γ_0	L	γ_0	L
ZO-SVRG	-	5e-2	-	5e-2	-	5e-2
ZO-SAGA	-	5e-2	-	1e-1	-	1e-1
ZO-Varag	-	5e-2	-	1e-2	-	1e-1
ZO-SPIDER-Coord	-	1	-	1e-1	-	1
AdaptRdct-NC (ZO-SVRG)	1e-3	1e-1	1e-4	1e-2	1e-4	1e-2
AdaptRdct-NC (ZO-SAGA)	5e-3	5e-2	1e-4	5e-2	1e-4	1e-2
AdaptRdct-NC (ZO-Varag)	1e-3	1e-1	5e-4	5e-2	5e-4	5e-2

Table 5: Choices of parameters for the experiment of *non-convex* logistic regression.

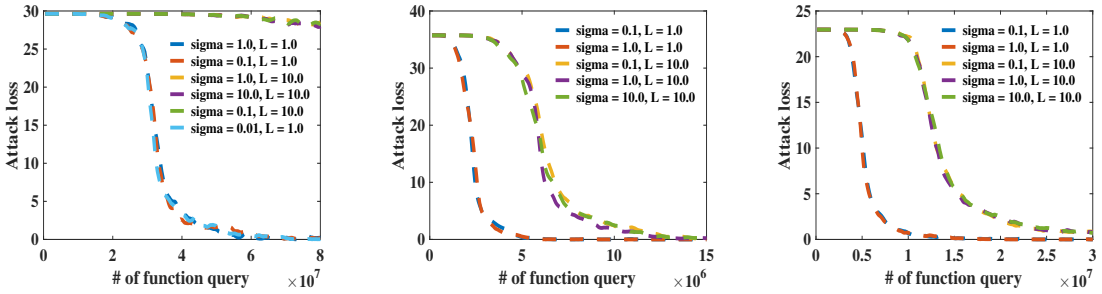
Algorithms	German		Ijcnn1		Mushrooms	
	σ	L	σ	L	σ	L
ZO-SVRG	-	1	-	1	-	1e-2
ZO-SPIDER-Coord	-	1e2	-	1e1	-	1e1
AdaptRdct-NC (ZO-SVRG)	5e-2	5	1e-2	5	5e-4	5e-2
AdaptRdct-NC (ZO-SAGA)	1e-2	1e1	1e-2	5	5e-4	5e-2
AdaptRdct-NC (ZO-Varag)	5e-4	5e-1	1e-4	1e-2	5e-4	5e-2

More Results on Generation of Black-Box Adversarial Examples



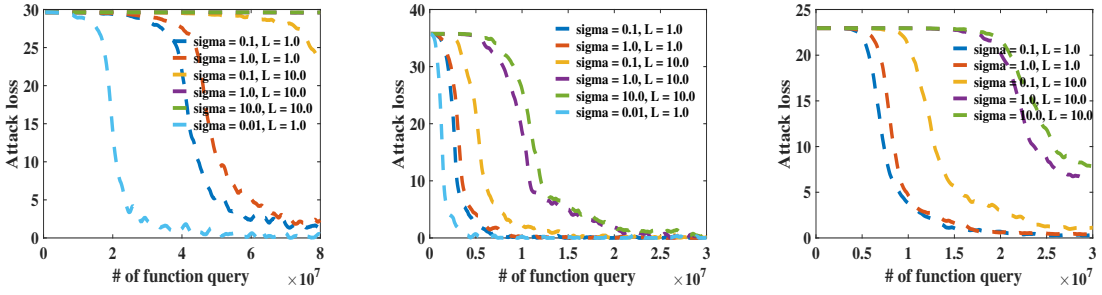
(a) Cifar-10 (b) Fmnist (c) Mnist

Figure 3: *AdaptRdct-NC* (ZO-SVRG) running under different parameter settings



(a) Cifar-10 (b) Fmnist (c) Mnist

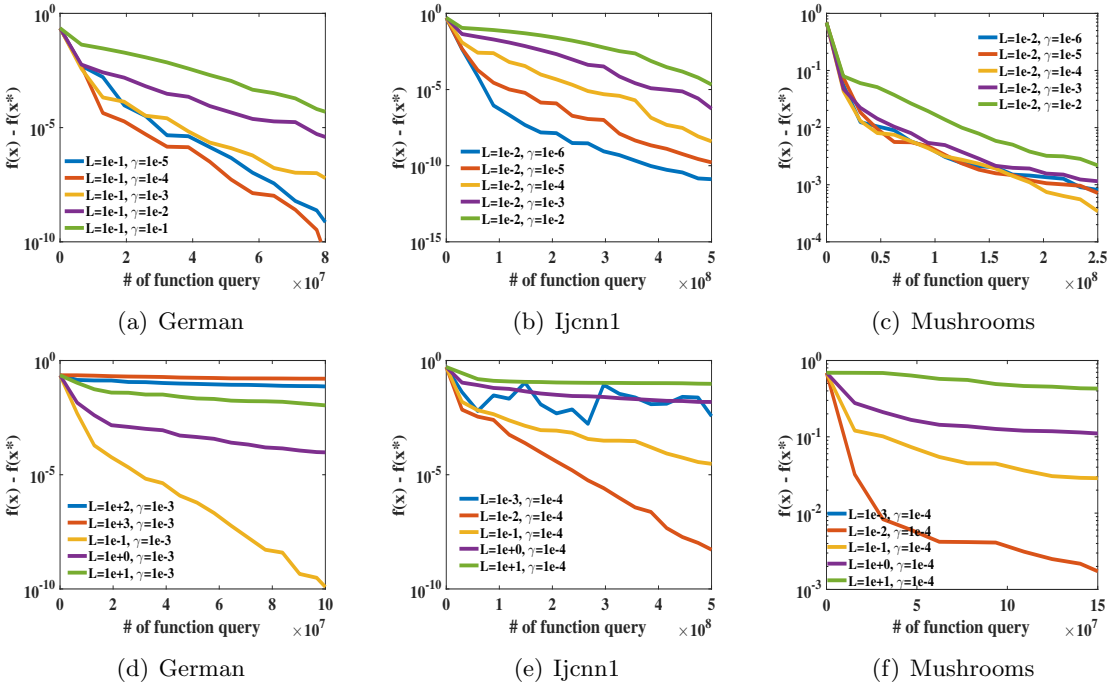
Figure 4: *AdaptRdct-NC* (ZO-SAGA) running under different parameter settings



(a) Cifar-10 (b) Fmnist (c) Mnist

Figure 5: *AdaptRdct-NC* (ZO-Varag) running under different parameter settings

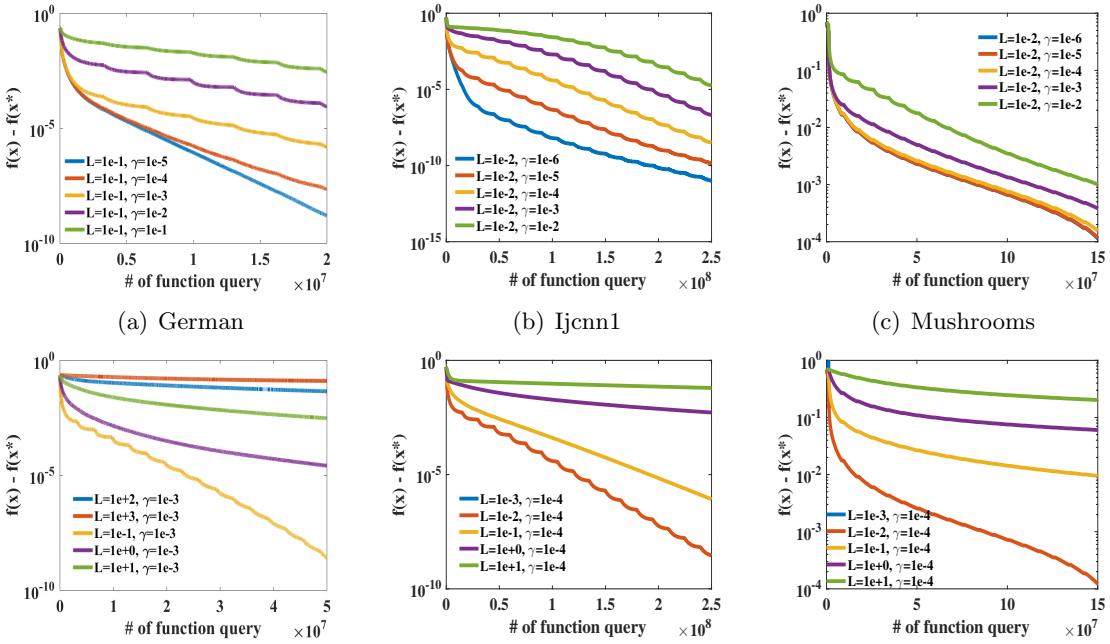
More Results on Convex Logistic Regression



(a) German (b) Ijcnn1 (c) Mushrooms

(d) German (e) Ijcnn1 (f) Mushrooms

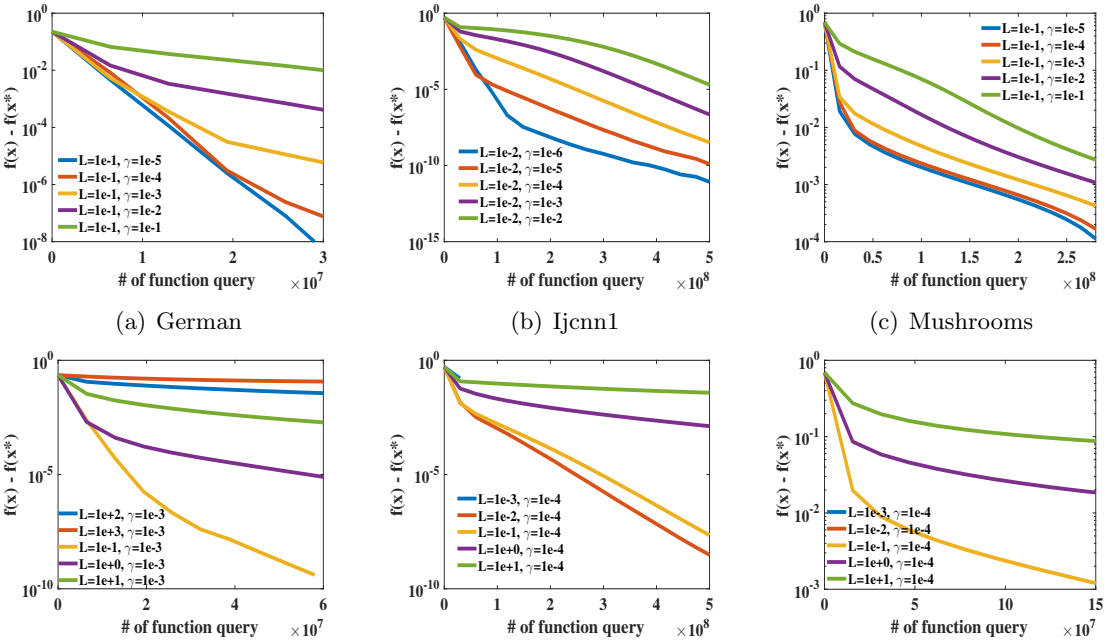
Figure 6: *AdaptRdct-C* (ZO-SVRG) running under different parameter settings



(a) German (b) Ijcnn1 (c) Mushrooms

(d) German (e) Ijcnn1 (f) Mushrooms

Figure 7: *AdaptRdct-C* (ZO-SAGA) running under different parameter settings

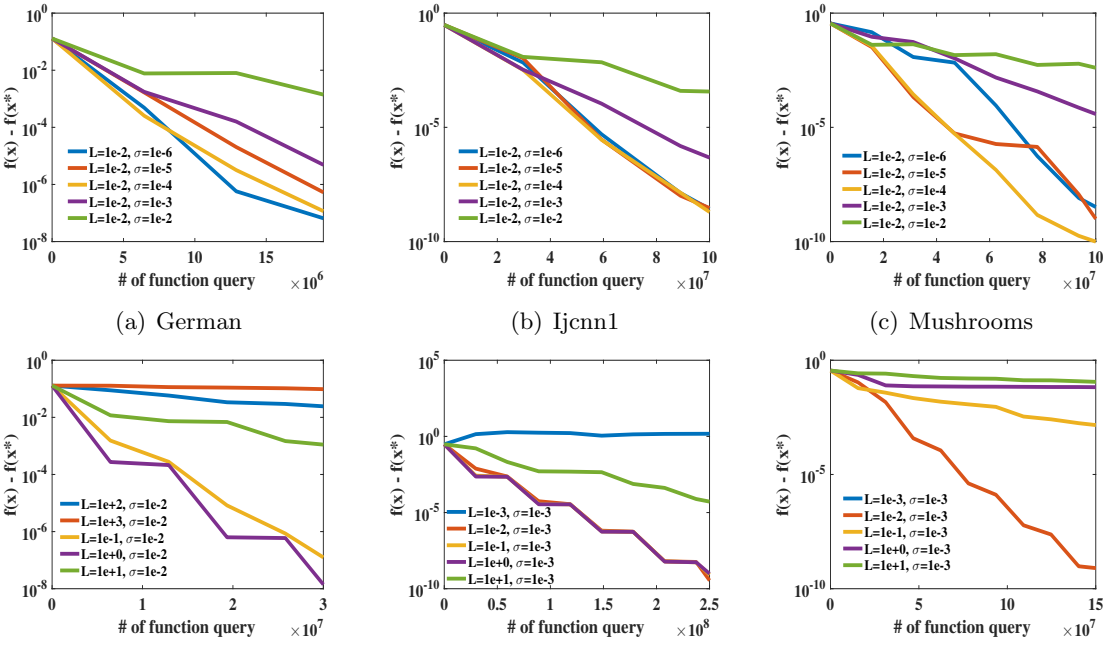


(a) German (b) Ijcnn1 (c) Mushrooms

(d) German (e) Ijcnn1 (f) Mushrooms

Figure 8: *AdaptRdct-C* (ZO-Varag) running under different parameter settings

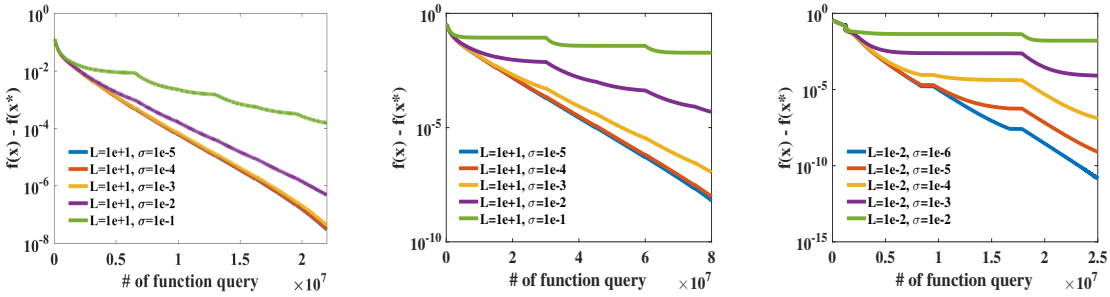
More Results on Non-convex Logistic Regression



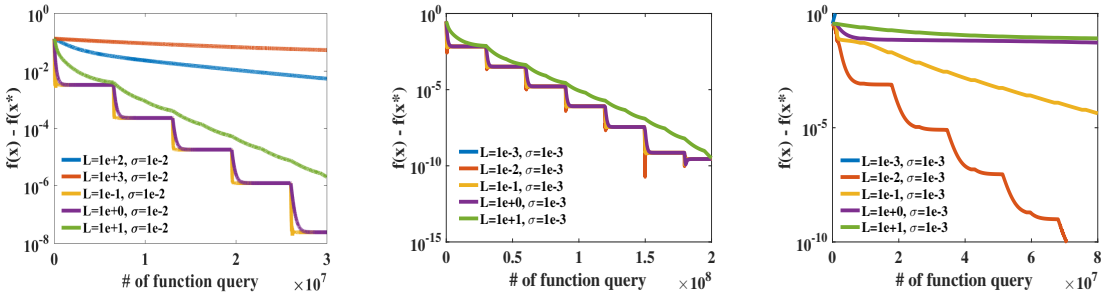
(a) German (b) Ijcnn1 (c) Mushrooms

(d) German (e) Ijcnn1 (f) Mushrooms

Figure 9: *AdaptRdct-NC* (ZO-SVRG) running under different parameter settings

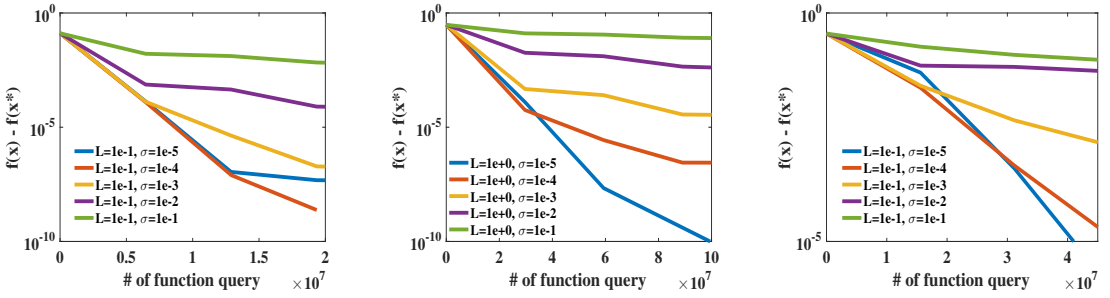


(a) German (b) Ijcnn1 (c) Mushrooms

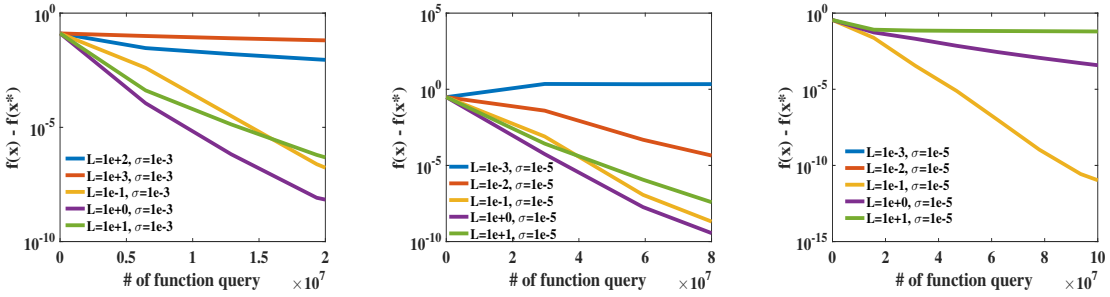


(d) German (e) Ijcnn1 (f) Mushrooms

Figure 10: *AdaptRdct-NC* (ZO-SAGA) running under different parameter settings



(a) German (b) Ijcnn1 (c) Mushrooms



(d) German (e) Ijcnn1 (f) Mushrooms

Figure 11: *AdaptRdct-NC* (ZO-Varag) running under different parameter settings

References

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

Zeyuan Allen-Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. In *Advances in Neural Information Processing Systems*, pages 1614–1622, 2016.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Yuwen Chen, Antonio Orvieto, and Aurelien Lucchi. An accelerated dfo algorithm for finite-sum convex functions. In *International Conference on Machine Learning*, pages 1681–1690. PMLR, 2020.

Zaiyi Chen, Zhuoning Yuan, Jinfeng Yi, Bowen Zhou, Enhong Chen, and Tianbao Yang. Universal stagewise learning for non-convex problems with convergence on averaged solutions. *arXiv preprint arXiv:1808.06296*, 2018.

Zaiyi Chen, Yi Xu, Haoyuan Hu, and Tianbao Yang. Katalyst: Boosting convex katayusha for non-convex problems with a large condition number. In *International Conference on Machine Learning*, pages 1102–1111. PMLR, 2019.

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

Bryan Conroy and Paul Sajda. Fast, exact model selection and permutation testing for l2-regularized logistic regression. In *Artificial Intelligence and Statistics*, pages 246–254, 2012.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *arXiv preprint arXiv:1407.0202*, 2014.

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

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

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

Roy Frostig, Rong Ge, Sham M Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on learning theory*, pages 728–763. PMLR, 2015.

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

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

Bin Gu, Zhouyuan Huo, Cheng Deng, and Heng Huang. Faster derivative-free stochastic algorithm for shared memory machines. In *International Conference on Machine Learning*, pages 1812–1821, 2018a.

Bin Gu, Zhouyuan Huo, and Heng Huang. Asynchronous doubly stochastic group regularized learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1791–1800. PMLR, 2018b.

Bin Gu, Wenhan Xian, and Heng Huang. Asynchronous stochastic frank-wolfe algorithms for nonconvex optimization. In *28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, 2019.

Bin Gu, Zhiyuan Dang, Xiang Li, and Heng Huang. Federated doubly stochastic kernel learning for vertically partitioned data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2483–2493, 2020a.

Bin Gu, Wenhan Xian, Zhouyuan Huo, Cheng Deng, and Heng Huang. A unified q-memorization framework for asynchronous stochastic optimization. *Journal of Machine Learning Research*, 21(190):1–53, 2020b.

Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

Feihu Huang, Bin Gu, Zhouyuan Huo, Songcan Chen, and Heng Huang. Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1503–1510, 2019.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. *arXiv preprint arXiv:1910.12166*, 2019.

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

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

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. *arXiv preprint arXiv:1905.12412*, 2019.

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

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18 (1):7854–7907, 2018.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *arXiv preprint arXiv:1805.10367*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

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

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

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Xia Shen, Moudud Alam, Freddy Fikse, and Lars Rönnegård. A novel generalized ridge regression method for quantitative genetics. *Genetics*, 193(4):1255–1268, 2013.

Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019.

Qingsong Zhang, Bin Gu, Cheng Deng, and Heng Huang. Secure bilevel asynchronous vertical federated learning with backward updating. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10896–10904, 2021.