Zeroth-Order Katyusha: An Accelerated Derivative-Free Method for Composite Convex Optimization

Silan Zhang and Yujie Tang*

Abstract

We investigate accelerated zeroth-order algorithms for smooth composite convex optimization problems. While for unconstrained optimization, existing methods that merge 2-point zeroth-order gradient estimators with first-order frameworks usually lead to satisfactory performance, for constrained/composite problems, there is still a gap in the complexity bound that is related to the non-vanishing variance of the 2-point gradient estimator near an optimal point. To bridge this gap, we propose the Zeroth-Order Loopless Katyusha (*ZO-L-Katyusha*) algorithm, leveraging the variance reduction as well as acceleration techniques from the first-order loopless Katyusha algorithm. We show that *ZO-L-Katyusha* is able to achieve accelerated linear convergence for compositve smooth and strongly convex problems, and has the same oracle complexity as the unconstrained case. Moreover, the number of function queries to construct a zeroth-order gradient estimator in *ZO-L-Katyusha* can be made to be $O(1)$ on average. These results suggest that *ZO-L-Katyusha* provides a promising approach towards bridging the gap in the complexity bound for zeroth-order composite optimization.

1 Introduction

Zeroth-order (ZO) or derivative-free optimization problems have been the subject of intense studies recently. They have emerged from various practical situations where explicit gradient information of the objective function can be very difficult or even impossible to compute, and only function evaluations are accessible to the decision maker. Zeroth-order optimization methods have seen wide applications in multiple disciplines, including signal processing [1], parameter tuning for machine learning [2], black-box adversarial attacks on deep neural networks [3], model-free control [4], etc.

Various types of zeroth-order algorithms have been proposed to address the challenge of lacking gradient information in the optimization procedure. In this paper, we particularly focus on the

*This work was supported by the National Natural Science Foundation of China through grant 72301008. S. Zhang and Y. Tang are with the College of Engineering, Peking University, Beijing, China. Email: zhangsilan@stu.pku.edu.cn, yujietang@pku.edu.cn

line of methods initiated by [5] and further developed by [6–8] and other related works. Basically, in this line of methods, a zeroth-order gradient estimator is constructed based on function values acquired at a few randomly explored points, which will then be combined with first-order optimization frameworks, leading to a zeroth-order optimization algorithm. For instance, [6] shows that, by combining a two-point Gaussian random gradient estimator with the vanilla gradient descent and Nesterov's accelerated gradient descent iterations, we can obtain a zeroth-order optimization algorithm for deterministic unconstrained smooth convex problems with an oracle complexity of $O(d/\epsilon)$ and $O(d/\sqrt{\epsilon})$, respectively, where d denotes the problem dimension. Similar techniques have also been adapted to the stochastic settings and nonconvex settings [6–9]. Recently, [10] proposed a zeroth-order stochastic approximation algorithms based on the Frank-Wolfe method, focusing particularly on handling constraints, high dimensionality and saddle-point avoiding. [11] proposed a one-point feedback scheme that queries the function value once per iteration but achieves comparable performance with two-point zeroth-order algorithms. [12] and [13] studied zeroth-order algorithms for stochastic weakly convex optimization. [14] investigated how to escape saddle points for zeroth-order algorithms using two-point gradient estimators. In sum, for *unconstrained* problems, existing works have shown that we are generally able to design zeroth-order algorithms with oracle complexities at most $O(d)$ times greater than their first-order counterparts.

However, for *constrained* deterministic optimization problems, we notice a gap in terms of oracle complexity between zeroth-order and first order methods. In [15], the authors proposed a zeroth-order Frank-Wolf algorithm with an oracle complexity of $O(d/\epsilon)$, but their method employs a gradient estimator whose computation requires $O(d)$ function queries, making it less flexible for problems of high dimensions. The recent work [16] proposed an inexact preconditioned zeroth-order algorithm for nonconvex composite optimization problems, based on a $2d$-point gradient estimator and also estimation of second-order information using finite differences. Classical two-point gradient estimators, on the other hand, can only achieve a suboptimal oracle complexity $O(d/\epsilon^2)$ for both Frank-Wolf [17] and projected gradient methods [7]. It has been suspected that the major cause of this gap is the non-vanishing variance of two-point gradient estimators as the iterate approaches an optimal point on the boundary, which could slow down convergence and lead to higher complexities [18]. Moreover, zeroth-order methods that apply acceleration techniques to constrained problems are scanty, and to the best of our knowledge, none of them has achieved a theoretical oracle complexity that can match the first-order counterpart.

The non-vanishing variance of the gradient estimator suggests the adaption of variance reduction as a possible approach to close the aforementioned gap in zeroth-order optimization. Variance reduction techniques have been widely used in machine learning problems to further enhance the performance of stochastic-gradient-descent-type methods. Algorithms such as SAG[19], SAGA[20], SVRG[21], SARAH[22] and SPIDER[23] have emerged to solve finite-sum problems that are prevalent in the machine learning community. Along this line, some variants incorporating first-order acceleration techniques have also been proposed, including ASVRG [24], Katyusha [25] and its loopless counterpart L-Katyusha [26]. We noticed that some existing works [23, 27, 28] have investigated combining variance reduction techniques with zeroth-order optimization. However, [23, 27]

considers unconstrained nonconvex problems, and [28] considers composite nonconvex problems. Moreover, these works concentrate more on the finite-sum problems or stochastic problems for machine learning tasks. The power of variance reduction techniques on reducing the variance of zeroth-order gradient estimators in deterministic constrained settings is still underexplored.

1.1 Our Contributions

In this paper, we develop an accelerated proximal zeroth-order algorithm for composite optimization problems with strongly convex objectives. The proposed algorithm leverages both variance reduction and acceleration techniques, and achieves superior convergence rate compared to existing zeroth-order algorithms. Specifically, our contributions can be summarized as follows:

- We propose *ZO-L-Katyusha*, an accelerated proximal zeroth-order algorithm, to minimize a strongly convex composite objective function. One advantage of *ZO-L-Katyusha* is that, with properly chosen algorithmic parameters, the algorithm only needs $O(1)$ function queries per iteration on average. Together with the loopless structure, this makes it more flexible than algorithms based on $O(d)$-point finite difference gradient estimators, especially when the dimension of the problem d is high. On the other hand, *ZO-L-Katyusha* also allows using purely $O(d)$-point gradient estimators, making the algorithm flexible and applicable to a wide range of practical scenarios.
- We prove that the oracle complexity of *ZO-L-Katyusha* can be upper bounded by $O(d\sqrt{\kappa}\ln(1/\epsilon))$, where κ is the reciprocal condition number of the objective function. This is significantly better than the state-of-art zeroth-order algorithms for *constrained* or *composite* strongly convex optimization problems. To the best of our knowledge, this is the first accelerated zeroth-order algorithm that can match the performance of first order algorithms up to a factor of variable dimension under constrained settings.

1.2 Notations

Throughout this paper, we denote the standard inner product in $\mathbb{R}^n$ by $\langle x, y\rangle = x^T y$, and denote the Euclidean norm by $\|x\| = \sqrt{x^T x}$. We use e_i to denote the vector that has only one non-zero entry 1 at its i-th coordinate, and use I_d to denote the $d \times d$ identity matrix. The unit sphere in $\mathbb{R}^d$ will be denoted by $\mathbb{S}_d \coloneqq \{x \in \mathbb{R}^d : \|x\| = 1\}$. For any set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality.

2 Problem Formulation

In this paper, we consider finding a solution to the following optimization problem:

$$\min_{x\in\mathbb{R}^d} \; F(x) \stackrel{\text{def}}{=} f(x) + \psi(x). \tag{1}$$

Here $f : \mathbb{R}^d \to \mathbb{R}$ is a convex and differentiable function, and $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex and lower semicontinuous but possibly non-differentiable. We impose the following assumptions on the information that can be accessed by the decision maker:

- The decision maker may only obtain the value of the function f by querying a black-box, or a zeroth-order oracle, which returns the value $f(x)$ whenever a point $x \in \mathbb{R}^d$ is fed into the black-box as its input. The decision maker does not have access to the derivatives of any order of f.
- The decision maker is able to efficiently evaluate the following proximal operator

$$\text{prox}_{\eta\psi}(x) \coloneqq \arg\min_{y \in \mathbb{R}^d} \left\{ \frac{1}{2}\|x - y\|^2 + \eta\psi(y) \right\}$$

  for any $\eta > 0$.

The first assumption, as we have shown in Section 1, is relevant to many problems in areas such as signal processing and machine learning, and necessitates the development of a zeroth-order optimization algorithm. The second assumption is standard in composite optimization and has been imposed in relevant studies such as [13,28], and allows us to incorporate simple nonsmoothness into our problem settings and algorithm design. For example, if we take ψ to the indicator function of a closed and convex set $\mathcal{X} \subseteq \mathbb{R}^d$:

$$\psi(x) = I_{\mathcal{X}}(x) = \begin{cases} 0, & \text{if } x \in \mathcal{X}, \\ +\infty, & \text{otherwise,} \end{cases}$$

Then the main problem (1) is equivalent to the constrained optimization problem $\min_{x \in \mathcal{X}} f(x)$, and the proximal operator reduces to the projection operator onto the convex set $\mathcal{X}$. We see that our problem formulation includes constrained optimization as a special case.

We also make the following technical assumptions that will facilitate convergence analysis of our algorithm:

Assumption 1. 1. The function f is L-smooth on $\mathbb{R}^d$ for some $L > 0$, i.e.,

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

2. There exist some $\mu_f \geq 0$ and $\mu_\psi \geq 0$ such that for all $x, y \in R^d$,

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu_f}{2}\|x - y\|^2,$$
$$\psi(x) \geq \psi(y) + \langle v, x - y \rangle + \frac{\mu_\psi}{2}\|x - y\|^2,$$

   where v is any subgradient of ψ at y.

3. $\mu \coloneqq \mu_f + \mu_\psi > 0$, so that F is μ-strongly convex.

The strongly convex setting will be the main focus of this paper, and investigation of the non-strongly convex setting has been our ongoing work. But we expect that the algorithm design techniques (presented in Section 3) can also be adapted to non-strongly convex problems.

2.1 Gaps in the Algorithm Design

When $\psi = 0$, i.e., the main problem (1) is an unconstrained smooth convex problem, we can find zeroth-order algorithms in existing literature that can solve (1) with satisfactory performance. For example, [6, 7, 9] investigated the following 2-point zeroth-order gradient estimator:

$$\mathsf{G}_f^{(2)}(x; u, \beta) = d \cdot \frac{f(x + \beta u) - f(x)}{\beta} u, \tag{2}$$

where $\beta > 0$ is a parameter called the *smoothing radius*, and u is a *random perturbation direction* either sampled from the normal distribution $\mathcal{N}(0, d^{-1} I_d)$ or the uniform distribution on the unit sphere $\mathcal{U}(\mathbb{S}_d)$. It has been shown that for unconstrained problems, by combining the gradient estimator (2) with certain first-order frameworks, we can obtain zeroth-order algorithms with convergence guarantees that are similar to the first-order counterparts. Particularly, [6] showed that, if we replace the true gradient in Nesterov's accelerated gradient descent by the 2-point gradient estimator (2), as long as the algorithmic parameters are adjusted accordingly, the resulting algorithm can successfully solve an unconstrained L-smooth and μ-strongly convex problem with oracle complexity upper bounded by

$$O\left(d\sqrt{\frac{L}{\mu}} \ln \frac{1}{\epsilon}\right),$$

which agrees with the first-order Nesterov's accelerated gradient descent method except for an $O(d)$ factor.

On the other hand, if we shift our focus to constrained or composite convex problems, we soon find that straightforward combination of the 2-point gradient estimator $\mathsf{G}_f^{(2)}(x; u, \beta)$ with a first-order optimization scheme will result in slower convergence and higher complexity than in the unconstrained setting; see Section 5.2 for a numerical example. The recent study [18] suggests that slower convergence of such algorithms is related to the fact that

$$\mathbb{E}\left[\left\|\mathsf{G}_f^{(2)}(x; u, \beta) - \nabla f(x)\right\|^2\right] \sim d\|\nabla f(x)\|^2$$

assuming $\beta > 0$ sufficiently small; here the expectation is taken with respect to $u \sim \mathcal{N}(0, I_d)$ or $u \sim \mathcal{U}(\mathbb{S}_d)$. For unconstrained smooth problems, as the algorithm's iterate x^k approaches an optimal point, we have $\nabla f(x^k) \to 0$, and consequently $\mathbb{E}\left[\left\|\mathsf{G}_f^{(2)}(x^k; u, \beta) - \nabla f(x^k)\right\|^2\right] \approx 0$, indicating that the randomness of $\mathsf{G}_f^{(2)}(x^k; u, \beta)$ will have a diminishing effect and thus the algorithm's behavior resembles a deterministic first-order method. However, for constrained or composite problems, the gradient at an optimal point may not vanish. As a result, as the iterate x^k approaches an optimal

point, the quantity $\mathbb{E}\left[\left\|\mathsf{G}_f^{(2)}(x^k;u,\beta)-\nabla f(x^k)\right\|^2\right]$ in general will not be close to zero, indicating that the algorithm's behavior resembles a stochastic first-order algorithm, which has strictly inferior convergence behavior.

There are also existing works (including [15,16,27], etc.) that propose the following $(d+1)$-point gradient estimator (or its variants) for solving constrained convex problems:

$$\sum_{i=1}^{d}\frac{f(x+\beta e_i)-f(x)}{2\beta}e_i. \tag{3}$$

However, each computation of the gradient estimator (3) then requires $O(d)$ evaluations of the function f, which may not be desired when the problem dimension d is high, as argued by [14]. Besides, for constrained composite convex optimization problems, convergence results for accelerated zeroth-order algorithm using $O(d)$-point gradient estimators are yet to be established.

The above discussion demonstrates that there still exists a gap between constrained/composite zeroth-order methods and unconstrained zeroth-order methods, especially if we insist on employing gradient estimators that can be constructed by only $O(1)$ function queries. On the other hand, the above analysis also suggests that the variance of the zeroth-order gradient estimator plays a key role in the algorithm's convergence behavior, which inspires us that we may consider incorporating variance reduction techniques into zeroth-order optimization, to bridge the aforementioned gap. Later, we shall see that, our proposed algorithm provides a promising approach towards bridging this gap: It achieves accelerated linear convergence for composite smooth and strongly convex problems, while the construction of each gradient estimator requires $O(1)$ function queries *on average*.

3 Our Algorithm

In this section, we present our proposed algorithm, the Zeroth-Order loopless Katyusha (*ZO-L-Katyusha*) algorithm, while is presented in Algorithm 1.

The algorithm generally follows the framework of loopless Katyusha [26]: We keep a reference point w^k at which we evaluate a relatively accurate estimation of the gradient $\hat{\nabla}f(w^k)$, and use $\hat{\nabla}f(w^k)$ to adjust the subsequent estimated gradients. The reference point w^k will be updated in a random fashion as proposed by [26]. However, the major difference is that we use variance reduction for handling the randomness of the two-point zeroth-order gradient estimator, rather than deadling with the finite-sum problem structure. Specifically, we set the adjusted gradient estimator to be

$$g^k=\hat{\nabla}_{\mathcal{S}_k}f(x^k)-\frac{1}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k}d\left\langle\hat{\nabla}f(w^k),u\right\rangle u+\hat{\nabla}f(w^k), \tag{4}$$

Algorithm 1: ZO-L-Katyusha

Parameters: Stepsize parameters $\theta \in (0,1)$, $M > 0$; batch size $|\mathcal{S}| \leq d$; smoothing radius $\beta > 0$; probability $p \in (0,1]$

Initialization: $\sigma = \mu_f/M$, $\eta = 1/(3\theta)$, $y^0 = z^0 = w^0 \in \mathbb{R}^d$

for $k = 0, 1, 2, \ldots$ **do**

$x^k = \theta z^k + \frac{1}{2} w^k + (\frac{1}{2} - \theta) y^k$

Generate a set of random vectors $\mathcal{S}_k$ by

- **Option I**: Drawing $|\mathcal{S}|$ samples uniformly from $\{e_1, e_2, ..., e_d\}$ without replacement
- **Option II**: Drawing $|\mathcal{S}|$ samples independently from $\mathcal{U}(\mathbb{S}_d)$, the uniform distribution over the unit sphere

$g^k = \hat{\nabla}_{\mathcal{S}_k} f(x^k) - \frac{1}{|\mathcal{S}_k|} \sum_{u \in \mathcal{S}_k} \langle \hat{\nabla} f(w^k), u \rangle u + \hat{\nabla} f(w^k)$

$z^{k+1} = \mathrm{prox}_{\frac{\eta}{(1+\eta\sigma)M}\psi} \left(\frac{\eta\sigma x^k + z^k - \frac{\eta}{M} g^k}{1 + \eta\sigma} \right)$

$y^{k+1} = x^k + \theta(z^{k+1} - z^k)$

$w^{k+1} = \begin{cases} y_k, & \text{with probability } p \\ w_k, & \text{with probability } 1 - p \end{cases}$

end

where we denote

$$\hat{\nabla}_{\mathcal{S}} f(x) = \frac{1}{|\mathcal{S}|} \sum_{u \in \mathcal{S}} d \cdot \frac{f(x + \beta u) - f(x)}{\beta} u, \tag{5}$$

$$\hat{\nabla} f(x) = \sum_{i=1}^{d} \frac{f(x + \beta e_i) - f(x)}{\beta} e_i. \tag{6}$$

In (4), each $\mathcal{S}_k$ is a randomly generated set of identically distributed perturbation directions for $|\mathcal{S}_k|$ 2-point gradient estimators, which will be averaged to form $\hat{\nabla}_{\mathcal{S}_k} f(x^k)$ as shown in (5). For simplicity, we employ the same batch size (which we denote by $|\mathcal{S}|$) for every $\mathcal{S}_k$. Unlike variance reduction algorithms for finite-sum problems, here we choose to draw $|\mathcal{S}_k|$ random directions, rather than drawing individual loss functions. The reference gradient $\hat{\nabla} f(w^k)$, on the other hand, is generated by the $(d+1)$-point gradient estimator, giving sufficiently accurate approximation to $\nabla f(w^k)$ when the smoothing radius β is chosen properly.

We offer two sampling strategies for the set $\mathcal{S}_k$, as presented in Algorithm 1: The first option is to draw S vectors from $\{e_1, e_2, \ldots, e_d\}$ *without replacement*, i.e., we choose the $|\mathcal{S}|$ random perturbation directions to be distinct and parallel to the coordinate axes; as an extreme case, when $|\mathcal{S}| = d$, we get a $(d+1)$-point gradient estimator. The second option is to draw $|\mathcal{S}|$ vectors independently from $\mathcal{U}(\mathbb{S}_d)$, the uniform distribution on the unit sphere. We shall see that these two

sampling strategies lead to largely similar performance guarantees when $|\mathcal{S}| \ll d$.

Algorithm 1 is a relatively general zeroth-order algorithm with several tunable parameters. In Section 4, we will show how these parameters should generally be chosen to guarantee convergence. Here, we discuss some special choices of the parameters:

- Suppose we set $|\mathcal{S}| = 1$ and $p = (\gamma d)^{-1}$ for some $\gamma > 0$, then on average we update w^k after approximately γd iterations. Since each reference gradient $\hat{\nabla} f(w^k)$ requires $d + 1$ function queries, while each $\hat{\nabla}_{\mathcal{S}_k} f(x^k)$ requires 2 function queries, we see that the averaged number of function queries needed per zeroth-order gradient estimation is

$$\frac{(d+1) \times 1 + 2 \times (\gamma d)}{1 + \gamma d} = O(1).$$

  We shall see in Section 4 that such choice of $|\mathcal{S}|$ and p guarantees convergence as long as other parameters are selected accordingly. This justifies our previous claim that, in *ZO-L-Katyusha*, the number of function queries for the construction of each gradient estimator can be made to be $O(1)$ on average.

- If we use the first sampling strategy, and set the mini-batch size $|\mathcal{S}| = d$ and probability $p = 1$, then *ZO-L-Katyusha* reduces to an accelerated zeroth-order algorithm using purely $(d + 1)$-point gradient estimators, with $w^{k+1} = y^k$ and $g^k = \hat{\nabla} f(x^k)$. We can see that *ZO-L-Katyusha* provides abundant flexibility which allows the algorithm to be applicable to a wide range of practical scenarios.

4 Convergence Analysis

In this section, we present our theoretical results on the convergence rate of the proposed *ZO-L-Katyusha* algorithm. We first present a main theorem that covers general sampling strategies and parameter choices. Then we provide corollaries on the algorithm's oracle complexities for some special cases. The proof of the main theorem will be postponed to Section 4.1.

Theorem 1. *Let $x^* \in \mathbb{R}^d$ be the unique optimal solution to* (1)*. Denote*

$$A = \begin{cases} \max\left\{\dfrac{4d(d-|\mathcal{S}|)}{(d-1)|\mathcal{S}|}, 1\right\}, & \textit{Option I is used,} \\ \dfrac{4d}{|\mathcal{S}|}, & \textit{Option II is used,} \end{cases}$$

and suppose in ZO-L-Katyusha*, we set $M = (A + 1)L/3$. Denote*

$$\Psi^k := \frac{\mu + 3\theta M}{2}\|z^k - x^*\|^2 + \frac{1}{\theta}(F(y^k) - F(x^*)) + \frac{1+\theta}{2p\theta}(F(w^k) - F(x^*)).$$

Then for the corresponding sequence generated by ZO-L-Katyusha, we have

$$\mathbb{E}\big[\Psi^k\big] \le (1-\Delta)^k \Psi^0 + \frac{\beta^2 d^2 L}{\Delta}\left(\frac{L}{d\mu} + \frac{1}{A\theta}\right),$$

where

$$\Delta = \min\left\{\frac{\mu}{2\mu + 6\theta M}, \frac{\theta}{2}, \frac{p\theta}{1+\theta}\right\}.$$

The convergence guarantees presented in Theorem 1 is quite general, and applies to arbitrary choices of $\theta \in (0,1)$, $|\mathcal{S}| \le d$, $\beta > 0$ and $p \in (0,1]$. Next we provide corollaries for three special cases, with θ and q fixed so that we can get concrete results on the complexity of our algorithm.

Corollary 1 (Mini-batch, Option I)**.** *Suppose $\mathcal{S}_k$ is drawn uniformly from $\{e_1, e_2, \ldots, e_d\}$ without replacement, with $|\mathcal{S}| \le \sqrt{d}$. Given arbitrary $\epsilon > 0$, let the parameters of* ZO-L-Katyusha *satisfy*

$$M = \frac{4d(d-|\mathcal{S}|)L}{3(d-1)|\mathcal{S}|} + \frac{L}{3}, \qquad \theta = \min\left\{\sqrt{\frac{d\mu}{M}}, \frac{1}{2}\right\}, \qquad \beta = O\left(\sqrt{\frac{\mu\epsilon}{d^{\frac{3}{2}}L^2}}\right), \qquad p = \frac{1}{d}.$$

Then we achieve $\mathbb{E}\big[F(w^k) - F(x^*)\big] \le \epsilon$ for

$$k \ge O\left(d\sqrt{\frac{L}{\mu|\mathcal{S}|}}\,\ln\frac{1}{\epsilon}\right).$$

Corollary 2 (Mini-batch, Option II)**.** *Suppose each vector in $\mathcal{S}_k$ is drawn independently from $\mathcal{U}(\mathbb{S}_d)$, the uniform distribution on the unit sphere, with $|\mathcal{S}| \le \sqrt{d}$. Given arbitrary $\epsilon > 0$, let the parameters of* ZO-L-Katyusha *satisfy*

$$M = \frac{4dL}{|\mathcal{S}|} + \frac{L}{3}, \qquad \theta = \min\left\{\sqrt{\frac{d\mu}{M}}, \frac{1}{2}\right\}, \qquad \beta = O\left(\sqrt{\frac{\mu\epsilon}{d^{\frac{3}{2}}L^2}}\right), \qquad p = \frac{1}{d}.$$

Then we achieve $\mathbb{E}\big[F(w^k) - F(x^*)\big] \le \epsilon$ for

$$k \ge O\left(d\sqrt{\frac{L}{\mu|\mathcal{S}|}}\,\ln\frac{1}{\epsilon}\right).$$

Corollary 3 (Full batch, Option I)**.** *Suppose we choose $\mathcal{S}_k = \{e_1, e_2, \ldots, e_d\}$ for every $k \ge 0$. Given arbitrary $\epsilon > 0$, let the parameters of* ZO-L-Katyusha *satisfy*

$$M = \frac{2L}{3}, \qquad \theta = \min\left\{\sqrt{\frac{\mu}{M}}, \frac{1}{2}\right\}, \qquad \beta = O\left(\sqrt{\frac{\mu\epsilon}{d^2L^2}}\right), \qquad p = 1.$$

Then we achieve $\mathbb{E}\big[F(w^k) - F(x^*)\big] \leq \epsilon$ *for*

$$k \geq O\left(\sqrt{\frac{L}{\mu}}\,\ln\frac{1}{\epsilon}\right).$$

We provide further discussions on these results:

1. In the mini-batch setups (analyzed in Corollaries 1 and 2), if we choose $|\mathcal{S}| = 1$, then as discussed at the end of Section 3, the averaged number of zeroth-order oracle calls (i.e., function queries) for each calculation of a gradient estimator is $O(1)$. Thus, the oracle complexity of ZO-L-Katyusha can be upper bounded by $O\Big(d\sqrt{L/\mu}\ln(1/\epsilon)\Big)$.
2. In the full-batch setup with Option I (analyzed in Corollary 3), since the number of zeroth-order oracle calls for each gradient estimator is $O(d)$, the oracle complexity of ZO-L-Katyusha is also upper bounded by $O\Big(d\sqrt{L/\mu}\ln(1/\epsilon)\Big)$, the same as the mini-batch setups with $|\mathcal{S}| = 1$.

In sum, regardless of whether we choose the mini-batch setup or the full-batch setup, the algorithm enjoys the same oracle complexity bound. Since first-order Nesterov's accelerated gradient descent has an oracle complexity upper bounded by $O\left(\sqrt{L/\mu}\ln(1/\epsilon)\right)$ for smooth and strongly convex optimization, we see that our algorithm's complexity bound accords with that of first-order accelerated methods apart from an $O(d)$ factor. In other words, our algorithm provides a promising approach towards bridging the gap mentioned and discussed in Section 2.1.

4.1 Proof Outline

Here we provide the outline of the proof of Theorem 1. Due to space limitations, we postpone the proofs of some technical lemmas to the Appendix. Throughout the proof, we let $\mathcal{F}_k$ denote the filtration generated by the family of random vectors $\{z^t, w^t, y^t : t \leq k\}$.

We first present some lemmas that characterize the bias and mean square error of the adjusted gradient estimator with respect to the true gradient.

Lemma 1. *Suppose Option I is adopted as the sampling strategy for* $\mathcal{S}_k$*. Then*

$$\|\mathbb{E}[g^k \,|\, \mathcal{F}_k] - \nabla f(x^k)\|^2 \leq \frac{1}{4}L^2\beta^2 d, \tag{7}$$

and

$$\mathbb{E}[\|g^k - \nabla f(x^k)\|^2 \,|\, \mathcal{F}_k] \leq \frac{4d(d-|\mathcal{S}|)L}{(d-1)|\mathcal{S}|}\cdot\big(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle\big) + 2L^2\beta^2 d^2. \tag{8}$$

Lemma 2. *Suppose Option II is adopted as the sampling strategy for* $\mathcal{S}_k$*. Then*

$$\|\mathbb{E}[g^k \,|\, \mathcal{F}_k] - \nabla f(x^k)\|^2 \leq L^2\beta^2, \tag{9}$$

and

$$\mathbb{E}[\|g^k - \nabla f(x^k)\|^2 \,|\, \mathcal{F}_k] \leq \frac{4dL}{|\mathcal{S}|} \cdot (f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle) + 2L^2\beta^2 d^2. \tag{10}$$

The next lemma provides a necessary technical result for establishing convergence.

Lemma 3. *We have*

$$\begin{aligned} &\frac{1}{\theta}\big(f(y^{k+1}) - f(x^k)\big) - \frac{\eta}{2(M - L\eta\theta)}\|g^k - \nabla f(x^k)\|^2 \\ &\leq \frac{M}{2\eta}\|z^{k+1} - z^k\|^2 + \langle g^k, z^{k+1} - z^k \rangle. \end{aligned} \tag{11}$$

Now, for notational simplicity, we introduce the following quantities:

$$\begin{aligned} &\mathcal{Z}^k = \frac{M + \eta\mu}{2\eta}\|z^k - x^*\|^2, \qquad \mathcal{Y}^k = \frac{1}{\theta}(F(y^k) - F^*), \\ &\mathcal{W}^k = \frac{1+\theta}{2p\theta}(F(w^k) - F^*), \end{aligned}$$

so that the stochastic Lyapunov function Φ^k can be written as $\Phi^k = \mathcal{Z}^k + \mathcal{Y}^k + \mathcal{W}^k$. Since w^{k+1} equals w^k with probability $1 - p$ and equals y^k with probability p, it's not hard to check that the following identity holds:

$$\mathbb{E}[\mathcal{W}^{k+1} \,|\, \mathcal{F}_k] = (1-p)\mathcal{W}^k + \frac{1+\theta}{2}\mathcal{Y}^k, \qquad \forall k \geq 0. \tag{12}$$

Using this set of notations, we proceed to establish the following lemma, which is another necessary technical result for our proof.

Lemma 4. *We have*

$$\begin{aligned} &\langle g^k, x^* - z^{k+1} \rangle + \frac{\mu_f}{2}\|x^k - x^*\|^2 + \frac{M\mathcal{Z}^k}{M + \eta\mu} \\ &\geq \frac{M}{2\eta}\|z^k - z^{k+1}\|^2 + \mathcal{Z}^{k+1} + \psi(z^{k+1}) - \psi(x^*). \end{aligned} \tag{13}$$

Now we are ready to prove our main theorem. First of all, we use the μ_f-strong convexity of f to obtain

$$\begin{aligned} f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle + \frac{\mu_f}{2}\|x^k - x^*\|^2 \\ &= f(x^k) + \frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle \nabla f(x^k), x^* - z^k \rangle + \langle \nabla f(x^k), z^k - x^k \rangle \\ &= f(x^k) + \frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle \nabla f(x^k), x^* - z^k \rangle \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2\theta}\langle\nabla f(x^k),x^k-w^k\rangle+\frac{1-2\theta}{2\theta}\langle\nabla f(x^k),x^k-y^k\rangle\\
&\geq f(x^k)+\frac{\mu_f}{2}\|x^k-x^*\|^2+\langle\nabla f(x^k),x^*-z^k\rangle\\
&\quad+\frac{1}{2\theta}\langle\nabla f(x^k),x^k-w^k\rangle+\frac{1-2\theta}{2\theta}(f(x^k)-f(y^k)),
\end{aligned}
$$

where in the third step we used Line 2 in Algorithm 1, and in the last step we used the convexity of f to get $\langle\nabla f(x^k),x^k-y^k\rangle\geq f(x^k)-f(y^k)$. Then, we notice that

$$
\langle\nabla f(x^k),x^*-z^k\rangle=-\langle\mathbb{E}[g^k\,|\,\mathcal{F}_k]-\nabla f(x^k),x^*-z^k\rangle+\langle\mathbb{E}[g^k\,|\,\mathcal{F}_k],x^*-z^k\rangle,
$$

and using the inequality $\langle a,b\rangle\geq-\frac{1}{2\epsilon}\|a\|^2-\frac{\epsilon}{2}\|b\|^2$ for any $\epsilon>0$, we have

$$
\begin{aligned}
\langle\mathbb{E}[g^k\,|\,\mathcal{F}_k]-\nabla f(x^k),x^*-z^k\rangle&\geq-\frac{1}{\mu}\left\|\mathbb{E}[g^k\,|\,\mathcal{F}_k]-\nabla f(x^k)\right\|^2-\frac{\mu}{4}\left\|x^*-z^k\right\|^2\\
&\geq-\frac{\beta^2L^2d}{\mu}-\frac{\eta\mu}{2(M+\eta\mu)}\mathcal{Z}^k,
\end{aligned}
$$

where we used Lemmas 1 and 2 to bound the term $\left\|\mathbb{E}[g^k\,|\,\mathcal{F}_k]-\nabla f(x^k)\right\|^2$, and also used the definition of $\mathcal{Z}^k$. Therefore

$$
\begin{aligned}
f(x^*)\geq f(x^k)&+\frac{1}{2\theta}\langle\nabla f(x^k),x^k-w^k\rangle-\frac{\eta\mu}{2(M+\eta\mu)}\mathcal{Z}^k\\
&-\frac{\beta^2L^2d}{\mu}+\frac{1-2\theta}{2\theta}(f(x^k)-f(y^k))+\frac{\mu_f}{2}\|x^k-x^*\|^2+\langle\mathbb{E}[g^k\,|\,\mathcal{F}_k],x^*-z^k\rangle.
\end{aligned}
$$

Next, we note that the inequality (13) implies

$$
\begin{aligned}
&\frac{\mu_f}{2}\|x^k-x^*\|^2+\langle\mathbb{E}[g^k\,|\,\mathcal{F}_k],x^*-z^k\rangle\\
&=\mathbb{E}\Big[\langle g^k,x^*-z^{k+1}\rangle+\frac{\mu_f}{2}\|x^k-x^*\|^2\,\Big|\,\mathcal{F}_k\Big]+\mathbb{E}\big[\langle g^k,z^{k+1}-z^k\rangle\,\big|\,\mathcal{F}_k\big]\\
&\geq\mathbb{E}\left[\frac{M}{2\eta}\|z^k-z^{k+1}\|^2+\mathcal{Z}^{k+1}+\psi(z^{k+1})-\psi(x^*)\,\middle|\,\mathcal{F}_k\right]+\mathbb{E}\left[\langle g^k,z^{k+1}-z^k\rangle-\frac{M\mathcal{Z}^k}{M+\eta\mu}\,\middle|\,\mathcal{F}_k\right],
\end{aligned}
$$

and (11) together with Lemmas 1 and 2 leads to

$$
\begin{aligned}
&\mathbb{E}\left[\frac{M}{2\eta}\|z^k-z^{k+1}\|^2+\langle g^k,z^{k+1}-z^k\rangle\,\middle|\,\mathcal{F}_k\right]\\
&\geq\mathbb{E}\left[\frac{f(y^{k+1})-f(x^k)}{\theta}-\frac{\eta}{2(M-L\eta\theta)}\|g^k-\nabla f(x^k)\|^2\,\middle|\,\mathcal{F}_k\right]\\
&\geq\frac{1}{\theta}\mathbb{E}[f(y^{k+1})-f(x^k)\,|\,\mathcal{F}_k]-\frac{\eta}{2(M-L\eta\theta)}\cdot2L^2\beta^2d^2
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{\eta AL}{2(M-L\eta\theta)}(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k\rangle) \\
&= \frac{1}{\theta}\mathbb{E}[f(y^{k+1}) - f(x^k)\,|\,\mathcal{F}_k] - \frac{L\beta^2 d^2}{A\theta} - \frac{1}{2\theta}(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k\rangle),
\end{aligned}
$$

where in the last step we used $\eta = 1/(3\theta)$ and $M = (A+1)L/3$. Consequently,

$$
\begin{aligned}
&\frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle \mathbb{E}[g^k\,|\,\mathcal{F}_k], x^* - z^k\rangle \\
&\geq \frac{1}{\theta}\mathbb{E}[f(y^{k+1}) - f(x^k)\,|\,\mathcal{F}_k] - \frac{L\beta^2 d^2}{A\theta} - \frac{1}{2\theta}(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k\rangle) \\
&\quad + \mathbb{E}\left[\mathcal{Z}^{k+1} + \psi(z^{k+1}) - \psi(x^*) - \frac{M\mathcal{Z}^k}{M+\eta\mu}\,\middle|\,\mathcal{F}_k\right].
\end{aligned}
$$

Summarizing the previous results, we can show that

$$
\begin{aligned}
&\mathbb{E}\big[f(x^*) - \psi(z^{k+1}) + \psi(x^*) - \mathcal{Z}^{k+1}\,\big|\,\mathcal{F}_k\big] \\
&\geq f(x^k) + \frac{1-2\theta}{2\theta}(f(x^k) - f(y^k)) - \frac{M\mathcal{Z}^k}{M+\eta\mu} + \frac{1}{\theta}\mathbb{E}[f(y^{k+1}) - f(x^k)\,|\,\mathcal{F}_k] \\
&\quad - \frac{1}{2\theta}(f(w^k) - f(x^k)) - \frac{\beta^2 L^2 d}{\mu} - \frac{\eta\mu}{2(M+\eta\mu)}\mathcal{Z}^k - \frac{L\beta^2 d^2}{A\theta} \\
&= -\frac{M+\eta\mu/2}{M+\eta\mu}\mathcal{Z}^k - \frac{1-2\theta}{2\theta}f(y^k) + \frac{1}{\theta}\mathbb{E}[f(y^{k+1})\,|\,\mathcal{F}_k] - \frac{1}{2\theta}f(w^k) - \frac{\beta^2 L^2 d}{\mu} - \frac{L\beta^2 d^2}{A\theta}.
\end{aligned}
$$

Moreover, since ψ is convex and

$$
y^{k+1} = x^k + \theta(z^{k+1} - z^k) = \theta z^{k+1} + \frac{1}{2}w^k + \left(\frac{1}{2} - \theta\right)y^k,
$$

by Jensen's inequality, we have

$$
\psi(z^{k+1}) \geq \frac{1}{\theta}\psi(y^{k+1}) - \frac{1}{2\theta}\psi(w^k) - \frac{1-2\theta}{2\theta}\psi(y^k).
$$

Hence, we arrive at

$$
\begin{aligned}
f(x^*) \geq\ & \mathbb{E}[\mathcal{Z}^{k+1}\,|\,\mathcal{F}_k] - \frac{M+\eta\mu/2}{M+\eta\mu}\mathcal{Z}^k - \frac{1-2\theta}{2\theta}F(y^k) \\
&+ \frac{1}{\theta}\mathbb{E}[F(y^{k+1})\,|\,\mathcal{F}_k] - \frac{F(w^k)}{2\theta} - \psi(x^*) - C(\beta),
\end{aligned}
$$

where we denote

$$
C(\beta) \coloneqq \beta^2 d^2 L\left(\frac{L}{d\mu} + \frac{1}{A\theta}\right).
$$

After arranging the terms using (12), we get

$$\begin{aligned}\mathbb{E}\big[\mathcal{Z}^{k+1}+\mathcal{Y}^{k+1}+\mathcal{W}^{k+1}\,\big|\,\mathcal{F}_k\big] &\le \frac{M+\eta\mu/2}{M+\eta\mu}\mathcal{Z}^k+\left(\frac{1}{2}-\theta\right)\mathcal{Y}^k+\frac{p}{1+\theta}\mathcal{W}^k\\ &\quad+(1-p)\mathcal{W}^k+\frac{1+\theta}{2}\mathcal{Y}^k+C(\beta)\\ &= \left(1-\frac{\eta\mu/2}{M+\eta\mu}\right)\mathcal{Z}^k+\left(1-\frac{\theta}{2}\right)\mathcal{Y}^k\\ &\quad+\left(1-\frac{p\theta}{1+\theta}\right)\mathcal{W}^k+C(\beta)\\ &\le (1-\Delta)(\mathcal{Z}^k+\mathcal{Y}^k+\mathcal{W}^k)+C(\beta),\end{aligned}$$

where we used the definition of the quantity Δ in the last step. In other words,

$$\mathbb{E}\left[\Phi^{k+1}-\frac{1}{\Delta}C(\beta)\,\middle|\,\mathcal{F}_k\right]\le(1-\Delta)\left(\Phi^k-\frac{1}{\Delta}C(\beta)\right).$$

The conclusion of the theorem is now evident.

5 Numerical experiments

5.1 Algorithmic Performance

In this section, we consider the following constrained optimization problem:

$$\min_{x\in\mathcal{X}}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln(1+\exp(-b_ia_i^Tx))+\frac{\mu}{2}\|x\|^2\right\},$$

which takes the form of an L2 regularized logistic regression problem with constraints. Here we let $\mathcal{X}\subseteq\mathbb{R}^d$ be a closed box, and set $d=40, n=30, \mu=0.02$ for the test case. The data set $\{(a_i,b_i)\}_{i=1}^n$ is manually synthesized. This problem can be formulated as a special case of (1) where

$$f(x)=\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-b_ia_i^Tx)),$$

is smooth and convex, and

$$\psi(x)=\frac{\mu}{2}\|x\|^2+I_{\mathcal{X}}(x)$$

is strongly convex. Note that here we only use this problem for the purpose of comparison of the algorithms' convergence behavior; a more practical approach would also need to exploit its finite-sum structure.

We test the 2-point and $(d+1)$-point version of *ZO-L-Katyusha* (i.e., the mini-batch setup and

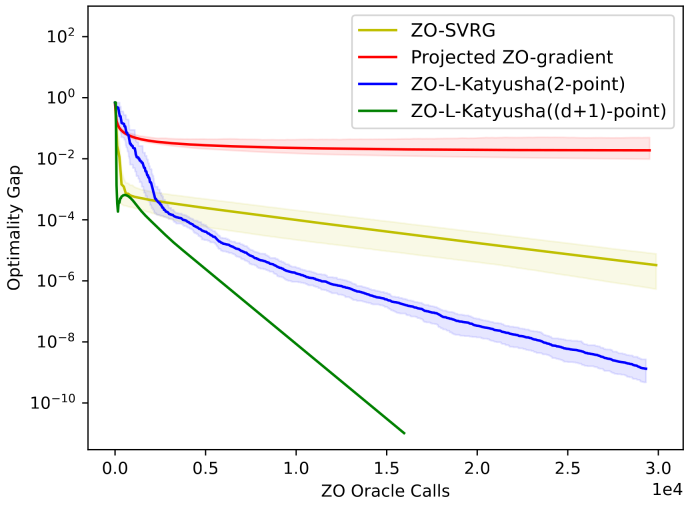

Figure 1: The performance of 2-point and $(d+1)$-point ZO-Katyusha

the full-batch setup) against existing algorithms. The results are shown in Figure 1. Here *ZO-SVRG* is adapted from [27], with some minor adaptations to fit our specific problem formulation, and *Projected ZO-gradient* is the standard stochastic projected zeroth-order gradient descent algorithm, to be found in [7]. Each solid curve in the Figure represents the average of 50 random trials for each algorithm, and the corresponding light shade represents the 5% to 95% quantile interval among these trials.

From Figure 1, some observations can be made. First, *Projected ZO-gradient* performs the worst, with the curve almost flattened out in the latter half of the iterations, this may be due to the diminishing stepsize policy introduced to deal with the non-vanishing variance of the gradient estimator, which leads to a sub-optimal $O(d/\epsilon^2)$ function query complexity. While with variance reduction techniques, all the other algorithms can employ a constant stepsize. Second, our *ZO-Katyusha* algorithm performs better than *ZO-SVRG*, while these two algorithms use similar gradient estimators, acceleration enables the former to converge even faster.

5.2 A Note on Accelerated Zeroth-Order Method

It is already mentioned that simply plugging a 2-point ZO gradient estimator into existing first order projected accelerated methods, like the one in [29], does not result in fast convergence as one may expect. To demonstrate this, we test the performance of the following iterations:

$$\begin{aligned} x^{k+1} &= P_{\mathcal{X}}\left(y_k - \frac{1}{dL} g^k\right), \\ y^{k+1} &= x^{k+1} - \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}(x^{k+1} - x^k), \end{aligned} \tag{14}$$

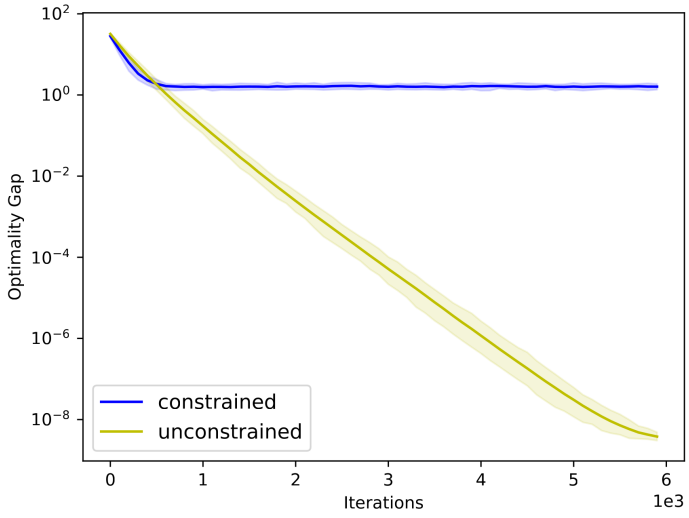

Figure 2: Performance of the algorithm (14) under constrained and unconstrained cases.

where $P_{\mathcal{X}}$ denotes projection onto $\mathcal{X}$, and g^k is generated by the simple 2-point gradient estimator (3). For the numerical test, we set f to be a strongly convex quadratic function, and set $\mathcal{X}$ to be i) $\mathbb{R}^d$ (the unconstrained case), and ii) a high-dimensional box (the constrained case), respectively. We run the above algorithm 50 times over the test cases.

The results are shown in Figure 2. We see that while in the unconstrained case, this algorithm converges quite fast, in the constrained case, the algorithm does not exhibit a clear trend of convergence. This partially illustrates the gap mentioned in Section 2.1, and suggests that some variance reduction techniques may be necessary if we want to design accelerated zeroth-order algorithms based on 2-point gradient estimators.

6 Conclusion

We proposed the Zeroth-Order Loopless Katyusha (*ZO-L-Katyusha*) algorithm for composite strongly convex optimization, which is able to achieve accelerated linear convergence while only requiring $O(1)$ function queries per iteration. Some future directions include: i) analysis of zeroth-order Katyusha methods for non-strongly convex problems; ii) developing accelerated zeroth-order algorithms for constrained problems using purely $O(1)$-point gradient estimators; iii) extending our algorithm to incorporate finite-sum structures.

References

[1] S. Liu, P.-Y. Chen, B. Kailkhura, G. Zhang, A. O. Hero, III, and P. K. Varshney, "A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications," *IEEE Signal Processing Magazine*, vol. 37, no. 5, pp. 43–54, 2020.

[2] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[3] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[4] Z. Ren, A. Zhong, and N. Li, "LQR with tracking: A zeroth-order approach and its global convergence," in *2021 American Control Conference (ACC)*, 2021, pp. 2562–2568.

[5] A. S. Nemirovskij and D. B. Yudin, *Problem complexity and method efficiency in optimization*. John Wiley & Sons, 1983.

[6] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, pp. 527–566, 2017.

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

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

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

[10] K. Balasubramanian and S. Ghadimi, "Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points," *Foundations of Computational Mathematics*, vol. 22, no. 1, pp. 35–76, 2022.

[11] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "A new one-point residual-feedback oracle for black-box learning and control," *Automatica*, vol. 136, p. 110006, 2022.

[12] V. Kungurtsev and F. Rinaldi, "A zeroth order method for stochastic weakly convex optimization," *Computational Optimization and Applications*, vol. 80, no. 3, pp. 731–753, 2021.

[13] S. Pougkakiotis and D. Kalogerias, "A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization," *SIAM Journal on Scientific Computing*, vol. 45, no. 5, pp. A2679–A2702, 2023.

[14] Z. Ren, Y. Tang, and N. Li, "Escaping saddle points in zeroth-order optimization: The power of two-point estimators," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 28 914–28 975.

[15] A. K. Sahu, M. Zaheer, and S. Kar, "Towards gradient free and projection free stochastic optimization," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 89, 2019, pp. 3468–3477.

[16] S. Liu, L. Wang, N. Xiao, and X. Liu, "An inexact preconditioned zeroth-order proximal method for composite optimization," *arXiv preprint arXiv:2401.03565*, 2024.

[17] K. Balasubramanian and S. Ghadimi, "Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[18] R. Jin, Y. Tang, and J. Song, "Zeroth-order feedback-based optimization for distributed demand response," *arXiv preprint arXiv:2311.00372*, 2023.

[19] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, pp. 83–112, 2017.

[20] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," *Advances in neural information processing systems*, vol. 27, 2014.

[21] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in neural information processing systems*, vol. 26, 2013.

[22] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," in *International conference on machine learning*. PMLR, 2017, pp. 2613–2621.

[23] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," *Advances in neural information processing systems*, vol. 31, 2018.

[24] F. Shang, L. Jiao, K. Zhou, J. Cheng, Y. Ren, and Y. Jin, "ASVRG: accelerated proximal SVRG," in *Asian Conference on Machine Learning*. PMLR, 2018, pp. 815–830.

[25] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *Journal of Machine Learning Research*, vol. 18, no. 221, pp. 1–51, 2018.

[26] D. Kovalev, S. Horváth, and P. Richtárik, "Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop," in *Algorithmic Learning Theory*. PMLR, 2020, pp. 451–467.

[27] K. Ji, Z. Wang, Y. Zhou, and Y. Liang, "Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization," in *International conference on machine learning*. PMLR, 2019, pp. 3100–3109.

[28] F. Huang, B. Gu, Z. Huo, S. Chen, and H. Huang, "Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 1503–1510.

[29] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Springer, 2018.

[30] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. L. Bartlett, and M. J. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," *Journal of Machine Learning Research*, vol. 21, no. 21, pp. 1–51, 2020.

A Technical Proofs

Throughout this material, we agree on the following notation conventions:

$$\hat{\nabla}_u f(x) = d\frac{f(x+\beta u) - f(x)}{\beta} u, \qquad \nabla_u f(x) = d\langle \nabla f(x), u\rangle u,$$

for $u \in \mathbb{R}^d$, and

$$\nabla_{\mathcal{S}} f(x) = \frac{1}{|\mathcal{S}|}\sum_{u\in\mathcal{S}} \nabla_u f(x)$$

for $\mathcal{S} = \{u_1, u_2, ..., u_{|\mathcal{S}|}\}, u_i \in \mathbb{R}^d$. With this notation, we have

$$\hat{\nabla} f(x) = \frac{1}{d}\sum_{i=1}^{d} \hat{\nabla}_{e_i} f(x), \qquad \nabla f(x) = \frac{1}{d}\sum_{i=1}^{d} \nabla_{e_i} f(x).$$

Recall that the function f is assumed to be convex and L-smooth, and $\beta > 0$ represents the smoothing radius.

Note that

$$\mathbb{E}\big[g^k \,\big|\, \mathcal{F}_k\big] = \mathbb{E}\left[\hat{\nabla}_{\mathcal{S}_k} f(x^k) - \frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T \hat{\nabla} f(w^k) + \hat{\nabla} f(w^k), \,\middle|\, \mathcal{F}_k\right]$$

and since $\mathbb{E}\big[\sum_{u\in\mathcal{S}_k} uu^T \,\big|\, \mathcal{F}_k\big] = \frac{|\mathcal{S}_k|}{d} I_d$ holds for both Option I and Option II in Algorithm 1 (the proof for Option II can be found in, e.g., Lemma 5 of [27]), we see that

$$\mathbb{E}\big[g^k \,\big|\, \mathcal{F}_k\big] = \mathbb{E}\Big[\hat{\nabla}_{\mathcal{S}_k} f(x^k) \,\Big|\, \mathcal{F}_k\Big].$$

To proceed, we introduce the following auxiliary lemmas:

Lemma 5. *Suppose $u \in \mathbb{S}_d$. Then for any $x \in \mathbb{R}^d$, we have*

$$\|\hat{\nabla}_u f(x) - \nabla_u f(x)\| \leq \frac{dL\beta}{2}. \tag{15}$$

Proof. We have

$$\begin{aligned}
\|\hat{\nabla}_u f(x) - \nabla_u f(x)\| &= d\left|\frac{f(x+\beta u) - f(x)}{\beta} - \langle \nabla f(x), u\rangle\right| \\
&= \frac{d}{\beta}\left|\int_0^\beta \left(\langle \nabla f(x+\gamma u), u\rangle - \langle \nabla f(x), u\rangle\right) d\gamma\right| \\
&\leq \frac{d}{\beta}\int_0^\beta \left|\langle \nabla f(x+\gamma u), u\rangle - \langle \nabla f(x), u\rangle\right| d\gamma \\
&\leq \frac{d}{\beta}\int_0^\beta \|\nabla f(x+\gamma u) - \nabla f(x)\|\|u\| d\gamma \\
&\leq \frac{d}{\beta}\int_0^\beta L\gamma\|u\|^2 d\gamma = \frac{dL\beta}{2},
\end{aligned}$$

where in the second step we used the fundamental theorem of calculus, and in the fifth step we used the L-smoothness of f. □

Lemma 6 ([30, Lemma 14])**.** *Suppose $u \sim \mathcal{U}(\mathbb{S}_d)$. Then for any $x \in \mathbb{R}^d$, we have*

$$\left\|\mathbb{E}\left[\hat{\nabla}_u f(x)\right] - \nabla f(x)\right\| \leq \beta L.$$

A.1 Proof of Lemma 1

Note that for Option I, we have $\mathbb{E}\left[\hat{\nabla}_{\mathcal{S}_k} f(x^k) \,\middle|\, \mathcal{F}_k\right] = \hat{\nabla} f(x^k)$, and therefore

$$\begin{aligned}
\left\|\mathbb{E}\left[g^k \,\middle|\, \mathcal{F}_k\right] - \nabla f(x^k)\right\|^2 &= \left\|\hat{\nabla} f(x^k) - \nabla f(x^k)\right\|^2 \\
&= \frac{1}{d^2}\sum_{i=1}^d \left\|\hat{\nabla}_{e_i} f(x^k) - \nabla_{e_i} f(x^k)\right\|^2 \overset{(15)}{\leq} \frac{dL^2\beta^2}{4}.
\end{aligned}$$

Then, for the proof of (8), we have

$$\begin{aligned}
&\left\|g^k - \nabla f(x^k)\right\|^2 \\
&= \left\|\hat{\nabla}_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(x^k) - \left(\frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T - I_d\right)\left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right)\right. \\
&\qquad \left. + \nabla_{\mathcal{S}_k} f(x^k) - \left(\frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T - I_d\right)\nabla f(w^k) - \nabla f(x^k)\right\|^2
\end{aligned}$$

$$\leq 4\left\|\hat{\nabla}_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(x^k)\right\|^2 + 4\left\|\left(\frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T - I_d\right)\left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right)\right\|^2$$
$$+ 2\left\|\nabla_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(w^k) + \nabla f(w^k) - \nabla f(x^k)\right\|^2. \tag{16}$$

For the first term on the right-hand side, notice that all elements in $\mathcal{S}_k$ are orthogonal for Option I, and so

$$\left\|\hat{\nabla}_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(x^k)\right\|^2 = \sum_{u\in\mathcal{S}_k}\left\|\frac{1}{|\mathcal{S}_k|}\left(\hat{\nabla}_u f(x^k) - \nabla_u f(x^k)\right)\right\|^2 \leq \frac{\beta^2 L^2 d^2}{4|\mathcal{S}_k|}.$$

Then for the second term, we notice that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(\frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T - I_d\right)\left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right)\right\|^2 \middle| \mathcal{F}_k\right] \\
&= \left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right)^T \mathbb{E}\left[\left(\frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T - I_d\right)^2 \middle| \mathcal{F}_k\right]\left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right) \\
&= \left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right)^T \mathbb{E}\left[\left(\frac{d^2}{|\mathcal{S}_k|^2} - 2\frac{d}{|\mathcal{S}_k|}\right)\sum_{u\in\mathcal{S}_k} uu^T + I_d \middle| \mathcal{F}_k\right]\left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right) \\
&= \frac{d - |\mathcal{S}_k|}{|\mathcal{S}_k|}\left\|\hat{\nabla} f(w^k) - \nabla f(w^k)\right\|^2 \leq \frac{d(d-|\mathcal{S}_k|)L^2\beta^2}{4|\mathcal{S}_k|},
\end{aligned}$$

where in the second step we used $\left(\sum_{u\in\mathcal{S}_k} uu^T\right)^2 = \sum_{u\in\mathcal{S}_k} uu^T$, and in the last step we used

$$\left\|\hat{\nabla} f(w^k) - \nabla f(w^k)\right\|^2 = \frac{1}{d^2}\sum_{i=1}^d \left\|\hat{\nabla}_{e_i} f(w^k) - \nabla_{e_i} f(w^k)\right\|^2 \overset{(15)}{\leq} \frac{dL^2\beta^2}{4}.$$

As a result, we get

$$\begin{aligned}
&\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2 \middle| \mathcal{F}_k\right] \\
&\leq 2\,\mathbb{E}\left[\left\|\nabla_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(w^k) + \nabla f(w^k) - \nabla f(x^k)\right\|^2 \middle| \mathcal{F}_k\right] + \frac{2d^2 - |\mathcal{S}_k|d}{|\mathcal{S}_k|}\beta^2 L^2.
\end{aligned}$$

Now to simplify notation, given $\mathcal{S}_k$, we let

$$b_i = \nabla_{e_i} f(x^k) - \nabla_{e_i} f(w^k) - (\nabla f(x^k) - \nabla f(w^k)), \qquad \iota_i = \begin{cases} 1, & \text{if } e_i \in \mathcal{S}_k, \\ 0, & \text{otherwise.} \end{cases}$$

Then it is obvious that $\mathbb{E}\big[\iota_i^2 \,\big|\, \mathcal{F}_k\big] = \frac{|\mathcal{S}_k|}{d}$ and $\mathbb{E}[\iota_i\iota_j \,|\, \mathcal{F}_k] = \frac{|\mathcal{S}_k|(|\mathcal{S}_k|-1)}{d(d-1)}$ for $i \neq j$. Therefore

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\nabla_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(w^k) + \nabla f(w^k) - \nabla f(x^k)\big\|^2 \,\Big|\, \mathcal{F}_k\Big] \\
&= \frac{1}{|\mathcal{S}_k|^2}\mathbb{E}\left[\left\|\sum_{i=1}^d \iota_i b_i\right\|^2 \,\middle|\, \mathcal{F}_k\right] \\
&= \frac{1}{|\mathcal{S}_k|^2}\left(\sum_{i=1}^d \mathbb{E}\Big[\iota_i^2 \|b_i\|^2 \,\Big|\, \mathcal{F}_k\Big] + \sum_{i\neq j}\mathbb{E}[\iota_i\iota_j\langle b_i, b_j\rangle \,|\, \mathcal{F}_k]\right) \\
&= \frac{1}{|\mathcal{S}_k|^2}\left(\left(\frac{|\mathcal{S}_k|}{d} - \frac{|\mathcal{S}_k|(|\mathcal{S}_k|-1)}{d(d-1)}\right)\sum_{i=1}^d \|b_i\|^2 + \frac{|\mathcal{S}_k|(|\mathcal{S}_k|-1)}{d(d-1)}\left\|\sum_{i=1}^d b_i\right\|^2\right). \\
&= \frac{d-|\mathcal{S}_k|}{|\mathcal{S}_k|d(d-1)}\sum_{i=1}^d \|b_i\|^2,
\end{aligned}
$$

where we used $\sum_{i=1}^d b_i = 0$. Then since $\sum_{i=1}^d \left\|a_i - \frac{1}{d}\sum_{j=1}^d a_j\right\|^2 \leq \sum_{i=1}^d \|a_i\|^2$, we have

$$
\begin{aligned}
\sum_{i=1}^d \|b_i\|^2 &= \sum_{i=1}^d \left\|\nabla_{e_i} f(x^k) - \nabla_{e_i} f(w^k) - \frac{1}{d}\sum_{j=1}^d \big(\nabla_{e_j} f(x^k) - \nabla_{e_j} f(w^k)\big)\right\|^2 \\
&\leq \sum_{i=1}^d \big\|\nabla_{e_i} f(x^k) - \nabla_{e_i} f(w^k)\big\|^2 = d^2\|\nabla f(x^k) - \nabla f(w^k)\|^2 \\
&\leq 2Ld^2(f(w^k) - f(x^k) + \langle \nabla f(x^k), w^k - x^k\rangle),
\end{aligned}
$$

where the last step follows from the L-smoothness of f. Summarizing these results, we get

$$
\begin{aligned}
&\mathbb{E}\big[\|g^k - \nabla f(x^k)\|^2 \,\big|\, \mathcal{F}_k\big] \\
&\leq \frac{4d(d-|\mathcal{S}|)L}{(d-1)|\mathcal{S}|}(f(w^k) - f(x^k) + \langle \nabla f(x^k), w^k - x^k\rangle) + \frac{2d^2 - |\mathcal{S}|d}{|\mathcal{S}|}\beta^2 L^2,
\end{aligned}
$$

and by $\frac{2d^2-|\mathcal{S}|d}{|\mathcal{S}|} \leq 2d^2$ we conclude the proof.

A.2 Proof of Lemma 2

In this part, we shall denote $\mathcal{S}_k = \{u_1^k, \ldots, u_{|\mathcal{S}|}^k\}$.

First, note that for Option II, $u_1^k, \ldots, u_{|\mathcal{S}|}^k$ are i.i.d. and follows the distribution $\mathcal{U}(\mathbb{S}_d)$, we have

$$
\big\|\mathbb{E}\big[g^k \,\big|\, \mathcal{F}_k\big] - \nabla f(x^k)\big\|^2 = \left\|\mathbb{E}\Big[\hat{\nabla}_{u_1^k} f(x^k) \,\Big|\, \mathcal{F}_k\Big] - \nabla f(x)\right\|^2 \leq \beta^2 L^2,
$$

where the last step follows from Lemma 6.

Then, to prove (10), we notice that the bound (16) still holds for Option II:

$$\begin{aligned}
&\left\|g^k - \nabla f(x^k)\right\|^2 \\
&\leq 4\left\|\hat{\nabla}_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(x^k)\right\|^2 + 4\left\|\left(\frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T - I_d\right)\left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right)\right\|^2 \\
&\quad + 2\left\|\nabla_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(w^k) + \nabla f(w^k) - \nabla f(x^k)\right\|^2.
\end{aligned}$$

This time, for the first term on the right-hand side, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\nabla}_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(x^k)\right\|^2 \middle| \mathcal{F}_k\right] &= \mathbb{E}\left[\left\|\frac{1}{|\mathcal{S}_k|}\sum_i \left(\hat{\nabla}_{u_i^k} f(x^k) - \nabla_{u_i^k} f(x^k)\right)\right\|^2 \middle| \mathcal{F}_k\right] \\
&\leq \mathbb{E}\left[\left|\frac{1}{|\mathcal{S}_k|}\sum_i \left\|\hat{\nabla}_{u_i^k} f(x^k) - \nabla_{u_i^k} f(x^k)\right\|\right|^2 \middle| \mathcal{F}_k\right] \\
&\leq \frac{\beta^2 L^2 d^2}{4},
\end{aligned}$$

where the last inequality follows from the bound (15); in order to bound the second term, we notice that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T - I_d\right)^2 \middle| \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\frac{d^2}{|\mathcal{S}_k|^2}\sum_{i,j} u_i^k (u_i^k)^T u_j^k (u_j^k)^T - 2\frac{d}{|\mathcal{S}_k|}\sum_i u_i^k (u_i^k)^T + I_d \middle| \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\frac{d^2}{|\mathcal{S}_k|^2}\left(\sum_i u_i^k (u_i^k)^T u_i^k (u_i^k)^T + \sum_{i\neq j} u_i^k (u_i^k)^T u_j^k (u_j^k)^T\right) - 2\frac{d}{|\mathcal{S}_k|}\sum_i u_i^k (u_i^k)^T + I_d \middle| \mathcal{F}_k\right] \\
&= \frac{d^2}{|\mathcal{S}_k|^2}\left(\sum_i \mathbb{E}\left[u_i^k (u_i^k)^T \,\middle|\, \mathcal{F}_k\right] + \sum_{i\neq j} \mathbb{E}\left[u_i^k (u_i^k)^T \,\middle|\, \mathcal{F}_k\right]\cdot \mathbb{E}\left[u_j^k (u_j^k)^T \,\middle|\, \mathcal{F}_k\right]\right) \\
&\quad - 2\frac{d}{|\mathcal{S}_k|}\sum_i \mathbb{E}\left[u_i^k (u_i^k)^T \,\middle|\, \mathcal{F}_k\right] + I_d \\
&= \frac{d-1}{|\mathcal{S}_k|} I_d,
\end{aligned}$$

where we used $\mathbb{E}\big[u_i(u_i^k)^T \,\big|\, \mathcal{F}_k\big] = \frac{1}{d} I_d$, the proof of which can be seen in Lemma 5 of [27]. Therefore

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(\frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T - I_d\right)\left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right)\right\|^2 \,\middle|\, \mathcal{F}_k\right] \\
&= \left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right)^T \mathbb{E}\left[\left(\frac{d}{|\mathcal{S}_k|}\sum_{u\in\mathcal{S}_k} uu^T - I_d\right)^2 \,\middle|\, \mathcal{F}_k\right]\left(\hat{\nabla} f(w^k) - \nabla f(w^k)\right) \\
&= \frac{d-1}{|\mathcal{S}_k|}\left\|\hat{\nabla} f(w^k) - \nabla f(w^k)\right\|^2 \\
&= \frac{(d-1)}{|\mathcal{S}_k|}\frac{1}{d^2}\sum_{i=1}^{d} \|\hat{\nabla}_{e_i} f(w^k) - \nabla e_i f(w^k)\|^2 \\
&\leq \frac{d(d-1)}{|\mathcal{S}_k|}\beta^2 L^2.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\mathbb{E}\Big[\left\|g^k - \nabla f(x^k)\right\|^2 \,\Big|\, \mathcal{F}_k\Big] \\
&\leq 2\mathbb{E}\Big[\left\|\nabla_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(w^k) + \nabla f(w^k) - \nabla f(x^k)\right\|^2 \,\Big|\, \mathcal{F}_k\Big] + \left(\frac{d(d-1)}{|\mathcal{S}_k|} + d^2\right) L^2\beta^2
\end{aligned} \tag{17}$$

Next, because $u_i^k, i = 1, 2, \ldots, |\mathcal{S}|$ are i.i.d. and sampled from $\mathcal{U}(\mathbb{S}_d)$, it follows that

$$\begin{aligned}
&\mathbb{E}\Big[\left\|\nabla_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(w^k) + \nabla f(w^k) - \nabla f(x^k)\right\|^2 \,\Big|\, \mathcal{F}_k\Big] \\
&= \frac{1}{|\mathcal{S}_k|}\mathbb{E}\left[\left\|\nabla_{u_1^k} f(x^k) - \nabla_{u_1^k} f(w^k) + \nabla f(w^k) - \nabla f(x^k)\right\|^2 \,\middle|\, \mathcal{F}_k\right] \\
&\leq \frac{1}{|\mathcal{S}_k|}\mathbb{E}\left[\left\|\nabla_{u_1^k} f(x^k) - \nabla_{u_1^k} f(w^k)\right\|^2 \,\middle|\, \mathcal{F}_k\right] \\
&= \frac{d^2}{|\mathcal{S}_k|}\mathbb{E}\Big[\left\|\langle \nabla f(x^k), u_1^k\rangle u_1^k - \langle \nabla f(w^k), u_1^k\rangle u_1^k\right\|^2 \,\Big|\, \mathcal{F}_k\Big] \\
&= \frac{d^2}{|\mathcal{S}_k|}\mathbb{E}\Big[\left(\langle \nabla f(x^k) - \nabla f(w^k), u_1^k\rangle\right)^2 \,\Big|\, \mathcal{F}_k\Big].
\end{aligned}$$

By using $\mathbb{E}\Big[u_1^k (u_1^k)^T \,\Big|\, \mathcal{F}_k\Big] = \frac{1}{d} I_d$, we get

$$\begin{aligned}
&\mathbb{E}\Big[\left\|\nabla_{\mathcal{S}_k} f(x^k) - \nabla_{\mathcal{S}_k} f(w^k) + \nabla f(w^k) - \nabla f(x^k)\right\|^2 \,\Big|\, \mathcal{F}_k\Big] \\
&\leq \frac{d}{|\mathcal{S}_k|}\mathbb{E}\Big[\left\|\nabla f(x^k) - \nabla f(w^k)\right\|^2 \,\Big|\, \mathcal{F}_k\Big] \\
&\leq \frac{2dL}{|\mathcal{S}_k|}(f(w^k) - f(x^k) + \langle \nabla f(x^k), w^k - x^k\rangle).
\end{aligned} \tag{18}$$

Plugging (18) back into (17) and using $\frac{d(d-1)}{|\mathcal{S}_k|} + d^2 \leq 2d^2$ concludes the proof.

A.3 Proof of Lemma 3

We have

$$\begin{aligned}
&\frac{M}{2\eta}\left\|z^{k+1}-z^k\right\|^2 + \left\langle g^k, z^{k+1}-z^k\right\rangle \\
&= \frac{1}{\theta}\left(\frac{M}{2\eta\theta}\left\|y^{k+1}-x^k\right\|^2 + \left\langle g^k, y^{k+1}-x^k\right\rangle\right) \\
&= \frac{1}{\theta}\left(\frac{L}{2}\left\|y^{k+1}-x^k\right\|^2 + \left\langle \nabla f(x^k), y^{k+1}-x^k\right\rangle + \left(\frac{M}{2\eta\theta}-\frac{L}{2}\right)\left\|y^{k+1}-x^k\right\|^2\right) \\
&\quad + \frac{1}{\theta}\left(\left\langle g^k - \nabla f(x^k), y^{k+1}-x^k\right\rangle\right) \\
&\geq \frac{1}{\theta}\left(f(y^{k+1}) - f(x^k) + \left(\frac{M}{2\eta\theta}-\frac{L}{2}\right)\left\|y^{k+1}-x^k\right\|^2 + \left\langle g^k - \nabla f(x^k), y^{k+1}-x^k\right\rangle\right) \\
&\geq \frac{1}{\theta}\left(f(y^{k+1}) - f(x^k) - \frac{\eta\theta}{2(M-L\eta\theta)}\|g^k - \nabla f(x^k)\|^2\right).
\end{aligned}$$

Here the first step comes from the update rule for y_{k+1} in *ZO-L-Katyusha*; we used the smoothness of f in the third step; in the last step we used $\langle a, b\rangle \geq -\frac{\|a\|^2}{2\epsilon} - \frac{\|b\|^2\epsilon}{2}$ for any $\epsilon > 0$.

A.4 Proof of Lemma 4

From the algorithm we have

$$z^{k+1} = \mathrm{prox}_{\frac{\eta}{(1+\eta\sigma)M}\psi}\left(\frac{\eta\sigma x^k + z^k - \frac{\eta}{M}g^k}{1+\eta\sigma}\right),$$

and so by the first-order optimality condition for minimizing convex functions, we can find $\xi^k \in \partial\psi(z^{k+1})$ such that

$$z^{k+1} - \frac{1}{1+\eta\sigma}\left(\eta\sigma x^k + z^k - \frac{\eta}{M}g^k\right) + \frac{\eta}{(1+\eta\sigma)M}\xi^k = 0.$$

Together with $\mu_f = M\sigma$, we obtain

$$g^k = \frac{M}{\eta}(z^k - z^{k+1}) + \mu_f(x^k - z^{k+1}) - \xi^k.$$

Therefore, we have

$$\langle g^k, z^{k+1} - x^*\rangle = \mu_f\langle x^k - z^{k+1}, z^{k+1} - x^*\rangle + \frac{M}{\eta}\langle z^k - z^{k+1}, z^{k+1} - x^*\rangle - \langle \xi^k, z^{k+1} - x^*\rangle$$

$$
\begin{aligned}
&= \frac{\mu_f}{2}\left(\|x^k - x^*\|^2 - \|x^k - z^{k+1}\|^2 - \|z^{k+1} - x^*\|^2\right) \\
&\quad + \frac{M}{2\eta}\left(\|z^k - x^*\|^2 - \|z^k - z^{k+1}\|^2 - \|z^{k+1} - x^*\|^2\right) - \langle \xi^k, z^{k+1} - x^* \rangle \\
&\leq \frac{\mu_f}{2}\|x^k - x^*\|^2 + \frac{M}{2\eta}\left(\|z^k - x^*\|^2 - \left(1 + \frac{\eta\mu_f}{M}\right)\|z^{k+1} - x^*\|^2\right) \\
&\quad - \frac{M}{2\eta}\|z^k - z^{k+1}\|^2 + \psi(x^*) - \psi(z^{k+1}) - \frac{\mu_\psi}{2}\|z^{k+1} - x^*\|^2 \\
&= \frac{\mu_f}{2}\|x^k - x^*\|^2 + \frac{M}{2\eta}\left(\|z^k - x^*\|^2 - \left(1 + \frac{\eta\mu}{M}\right)\|z^{k+1} - x^*\|^2\right) \\
&\qquad - \frac{M}{2\eta}\|z^k - z^{k+1}\|^2 + \psi(x^*) - \psi(z^{k+1}),
\end{aligned}
$$

where the third step follows from the fact that ψ is μ_ψ-strongly convex, and in the last equality we used $\mu = \mu_\psi + \mu_f$.

Since $\mathcal{Z}_k = \frac{M+\eta\mu}{2\eta}\|z^k - x^*\|^2$, rearranging terms yields the desired result.