Double Variance Reduction: A Smoothing Trick for Composite Optimization Problems without First-Order Gradient

Hao Di [1 2] **Haishan Ye** [1 3] **Yueling Zhang** [4] **Xiangyu Chang** [1] **Guang Dai** [3] **Ivor W. Tsang** [5 6]

Abstract

Variance reduction techniques are designed to decrease the sampling variance, thereby accelerating convergence rates of first-order (FO) and zeroth-order (ZO) optimization methods. However, in composite optimization problems, ZO methods encounter an additional variance called the coordinate-wise variance, which stems from the random gradient estimation. To reduce this variance, prior works require estimating all partial derivatives, essentially approximating FO information. This approach demands $\mathcal{O}(d)$ function evaluations (d is the dimension size), which incurs substantial computational costs and is prohibitive in high-dimensional scenarios. This paper proposes the Zeroth-order Proximal Double Variance Reduction (`ZPDVR`) method, which utilizes the averaging trick to reduce both sampling and coordinate-wise variances. Compared to prior methods, `ZPDVR` relies solely on random gradient estimates, calls the stochastic zeroth-order oracle (SZO) in expectation $\mathcal{O}(1)$ times per iteration, and achieves the optimal $\mathcal{O}(d(n+\kappa)\log(\frac{1}{\epsilon}))$ SZO query complexity in the strongly convex and smooth setting, where κ represents the condition number and ϵ is the desired accuracy. Empirical results validate `ZPDVR`'s linear convergence and demonstrate its superior performance over other related methods.

[1]Center for Intelligent Decision-Making and Machine Learning, School of Management, Xi'an Jiaotong University, China. [2]This work was completed during the internship at SGIT AI Lab, State Grid Corporation of China. [3]SGIT AI Lab, State Grid Corporation of China. [4]International Business School, Beijing Foreign Studies University, Beijing, China. [5]CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore [6]College of Computing and Data Science, NTU, Singapore.. Correspondence to: Haishan Ye <yehaishan@xjtu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

This paper considers the following composite finite-sum optimization problem,

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & F(x) = f(x) + \psi(x), \\ \text{with} \quad & f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \end{aligned} \tag{1}$$

where $f_i(x) : \mathbb{R}^d \to \mathbb{R}$ is a μ-strongly convex and L-smooth function , and $\psi(x) : \mathbb{R}^d \to \mathbb{R}$ is a convex and non-smooth function, such as the ℓ_1 regularization. This formulation is prevalent in several critical applications, such as signal compression (Jenatton et al., 2011), image recovery (Chen et al., 2021), sparse model training (Beck & Teboulle, 2009; Yun et al., 2021). In this study, we assume that only the zeroth-order (ZO) oracle (i.e., the value of function $f(x)$) is available. Under this condition, we explore the ZO variance reduction method to achieve the linear convergence rate for the above composite optimization problem (1) with only random gradient estimates.

ZO optimization methods approximate the directional derivative information by utilizing either the function value or the difference in function values, and employ this estimate to minimize the objective function (Liu et al., 2020). Due to their exclusive reliance on function evaluations, ZO optimization methods simplify algorithm implementation (Nesterov & Spokoiny, 2017; Liu et al., 2020) and are increasingly employed in scenarios where gradient information is either inaccessible or impractical, such as in neural network black-box attacks and robotic stiffness control (Tu et al., 2019; Li et al., 2023). Recently, Malladi et al. (2023) have capitalized on the benefits of ZO optimization methods for fine-tuning large language models, achieving a comparable performance to first-order (FO) methods while significantly reducing memory usage. Furthermore, Chen et al. (2023a) employ ZO methods to train the ResNet-20 model from scratch, illustrating the practicality of these methods in large-scale machine learning applications.

To decrease the variance resulting from the random sampling and accelerate the convergence rate of ZO optimization methods for finite-sum optimization problems (i.e., $\min_{x\in\mathbb{R}^d}\sum_{i=1}^{n} f_i(x)$), researchers have introduced various

Table 1. Summary of SZO query complexity and expected SZO queries per iteration of different methods. FO and ZO represent first-order and zeroth-order methods, respectively. PGD is the proximal gradient descent method.

	Methods	SZO query complexity	Expected SZO queries per iteration
FO	PGD	$\mathcal{O}(dn\kappa\log(\frac{1}{\epsilon}))$	$\mathcal{O}(nd)$
	PSVRG (Xiao & Zhang, 2014)	$\mathcal{O}(d(n+\kappa)\log(\frac{1}{\epsilon}))$	$\mathcal{O}(d)$
ZO	SEGA (Hanzely et al., 2018)	$\mathcal{O}(dn\kappa\log(\frac{1}{\epsilon}))$	$\mathcal{O}(n)$
	ZPDVR (ours)	$\mathcal{O}(d(n+\kappa)\log(\frac{1}{\epsilon}))$	$\mathcal{O}(1)$

variance reduction techniques tailored for diverse settings, encompassing non-convex and non-convex non-smooth problems (Fang et al., 2018; Ji et al., 2019; Chen et al., 2023b; Kazemi & Wang, 2023), spanning from centralized to decentralized scenarios (Yi et al., 2022; Lin et al., 2023). In addition to the sampling variance, ZO gradient estimates also suffer from high inherent gradient approximation variance, which is $\mathcal{O}(d)$ times the value of gradient norm $\|\nabla f(x)\|^2$ (Nesterov & Spokoiny, 2017; Liu et al., 2018). This variance stems from the random directional gradient estimation in ZO methods and can be eliminated by estimating each partial derivative along the coordinate directions instead of random directions. Consequently, we refer to this variance as the *coordinate-wise variance* throughout this paper. For the unconstrained finite-sum optimization problem, the FO optimality condition indicates that the coordinate-wise variance will automatically diminish as $x \to x^*$.

However, in the context of the composite optimization problem (1) as well as the constrained optimization problem, a crucial characteristic is that $\|\nabla f(x^*)\|^2$ is generally not equal to 0 at the optimal point x^*. Therefore, the coordinate-wise variance persists even after reaching the optimal point x^*. Consequently, this non-vanishing variance induced by the random gradient estimate inhibits the convergence to the optimal point x^*, which leads to inferior performance. Prior ZO variance reduction methods for the composite optimization problem (Liu et al., 2018; Kazemi & Wang, 2023) usually adopt stochastic variance reduction gradient (SVRG) techniques (Johnson & Zhang, 2013; Nguyen et al., 2017; Fang et al., 2018) to eliminate the inherent sampling variance and approximate the gradient by computing all partial gradient estimations to circumvent the coordinate-wise variance. Deriving all partial gradient estimations requires $\Omega(d)$ function evaluations to make a step, which is prohibitive for high-dimensional problems and hinders the application of ZO variance reduction methods.

In this work, we employ the averaging trick and introduce a novel gradient estimator for the gradient $\nabla f(x)$, utilizing only random gradient estimates. This proposed estimator calculates the difference between the random gradient estimate and the vector value of the current estimator along the same direction and uses the average scheme to refine estimation errors. We demonstrate that this gradient estimator can progressively reduce the coordinate-wise variance. Furthermore, by combining the proposed estimator with SVRG techniques, we present the **Z**eroth-order **P**roximal **D**ouble **V**ariance **R**eudction method, `ZPDVR`, which can reduce the inherent variances — both the random sampling variance and the coordinate-wise variance, in the ZO composite optimization problem, using only the random gradient estimate. We also conduct a comprehensive analysis of `ZPDVR` and prove that the stochastic zeroth-order oracle (SZO) query complexity (the number of function estimations) of `ZPDVR` is $\mathcal{O}(d(n+\kappa)\log(\frac{1}{\epsilon}))$ and it only calls $\mathcal{O}(1)$ SZO in expectation per iteration. As shown in Table 1, compared to other related FO and ZO methods, `ZPDVR` obtains the lowest SZO query complexity and exhibits the fewest SZO calls in expectation per iteration, which underscores its computational efficiency. The main contributions of this work are summarized as follows:

- We propose a novel ZO variance reduction method for the composite optimization problem (1) using only random gradient estimates, `ZPDVR`, which reduce the inherent variances induced by random sampling and random gradient estimation. `ZPDVR` calls the SZO $\mathcal{O}(1)$ times in expectation per iteration, which demonstrates efficiency in updating parameters per iteration and exhibits applicability in high-dimensional scenarios.

- We conduct a convergence analysis of `ZPDVR`, and derive the $\mathcal{O}(d(n+\kappa)\log\frac{1}{\epsilon})$ SZO query complexity for `ZPDVR`, which aligns with the best-know SZO query complexity of SVRG methods and requires fewer SZO function evaluations than other related FO and ZO methods. The optimal convergence property further demonstrates the computational efficiency of `ZPDVR`.

Notation. Let x^* denote the optimal point of the function F, specifically $x^* = \arg\min_{x\in\mathbb{R}^d} F(x)$, and $[n]$ represent the set $\{1,\ldots,n\}$. For the μ-strongly convex and L-smooth function f_i, $i \in [n]$, there exist two constants $\mu, L > 0$ and

$\mu \leq L$ such that

$$f_i(x_1) \geq f_i(x_2) + \langle \nabla f(x_2), x_1 - x_2 \rangle + \frac{\mu}{2}\|x_1 - x_2\|^2,$$
$$f_i(x_1) \leq f_i(x_2) + \langle \nabla f(x_2), x_1 - x_2 \rangle + \frac{L}{2}\|x_1 - x_2\|^2,$$

for $\forall x_1, x_2 \in \mathbb{R}^d$. For the convex function ψ, the following inequality holds for $\forall x_1, x_2 \in \mathbb{R}^d$

$$\psi(x_1) \geq \psi(x_2) + \langle \partial\psi(x_2), x_1 - x_2 \rangle ,$$

where $\partial\psi(x)$ is the subgradient of ψ at the point x. The proximal operator, denoted as $\text{prox}_{\eta\psi}$, is defined as $\text{prox}_{\eta\psi}(x) = \arg\min_{x' \in \mathbb{R}^d} \psi(x) + \frac{\eta}{2}\|x - x'\|^2$. $\kappa = \frac{L}{\mu}$ is the condition number of the function f, and η denotes the constant learning rate. If $f(x) = \mathcal{O}(g(x))$, it implies there is a positive constant $M > 0$ such that $|f(x)| \leq Mg(x)$. $f(x) = \Omega(g(x))$ indicates that there is a positive constant C such that $C|g(x)| \leq |f(x)|$.

2. Related Work

Variance Reduction. Variance reduction methods leverage the control variate technique (Rubinstein & Marcus, 1985) to reduce inherent sampling variance in stochastic gradient descent (SGD). Specifically, for the strongly convex and smooth finite-sum optimization problem, FO variance reduction methods attain the best-know $\mathcal{O}((\kappa + n)\log(\frac{1}{\epsilon}))$ gradient evaluations (Gower et al., 2020) , resulting in a significant acceleration of $\mathcal{O}(n\kappa\log(\frac{1}{\epsilon}))$ in SGD. The classic SVRG method (Johnson & Zhang, 2013) adopts a double-loop structure, maintaining a snapshot of model parameters to compute the full gradient in the outer loop and constructing an unbiased gradient estimate in the inner loop. Allen-Zhu & Hazan (2016) and Reddi et al. (2016) extend SVRG into the non-convex setting and derive an improved query complexity than gradient descent (GD) by a factor of $\mathcal{O}(n^{\frac{1}{3}})$. Subsequent work (Fang et al., 2018) leverages the stochastic path-integrated differential estimator, proposing SPIDER with $\mathcal{O}(n^{\frac{1}{2}}\epsilon)$ query complexity for non-convex finite-sum optimization problems. Diverging from the double-loop structure, Kovalev et al. (2020) and Li et al. (2021) propose loop-less SVRG methods that execute the coin-flip process independently of prior knowledge about the condition number. In cases where only a gradient sketch is available, Hanzely et al. (2018) and Gower et al. (2021) develop gradient estimators via a sketch-and-project process, progressively diminishing the estimation variance.

Zeroth-order Optimization. ZO optimization methods leverage the function evaluation to approximate the gradient and utilize this estimate to minimize the objective function. Nesterov & Spokoiny (2017) introduce the Gaussian smooth technique and achieve the $\mathcal{O}(dn\kappa\log(\frac{1}{\epsilon}))$ SZO query complexity for ZO-GD in the strongly convex and smooth scenario. The convergence properties of ZO-SGD for non-convex problems are analyzed by Ghadimi & Lan (2013). Ye et al. (2018) integrate Hessian information into gradient estimation and propose a Hessian-aware ZO optimization method. Berahas et al. (2022) derive bounds on the number of samples and the sampling radius to ensure the convergence of ZO-GD with different ZO gradient estimators. Compared to the FO methods, the query complexity of ZO methods exhibits a dependence on the dimension size d (Liu et al., 2020).

Zeroth-order Variance Reduction. Combining with SVRG, Liu et al. (2018) introduce the ZO-SVRG method for non-convex finite-sum problems. Subsequently, Ji et al. (2019) improve the SZO query complexity of ZO-SVRG, surpassing both ZO-GD and ZO-SGD in the non-convex finite-sum setting and aligning with the results of FO SVRG methods. For the composite optimization problem (1), Huang et al. (2019) present the convergence property of ZO proximal SVRG (ZPSVRG) method for non-convex non-smooth problems, and demonstrate that only utilizing the random gradient estimate incurs additional bias and the FO gradient information is necessary. Kazemi & Wang (2023) improve the iterative complexity of ZPSVRG for the non-convex non-smooth problem.

3. Methodology

This section presents details of `ZPDVR`. We first introduce the random stochastic gradient estimator in ZO methods and demonstrate the undiminished variance when employing it in composite optimization problems. Then, we delineate how the proposed method `ZPDVR` can reduce the two types of variances.

3.1. Gradient Estimate in Zeroth-Order Optimization

When only function evaluations are available, here, we utilize the Gaussian smoothing technique (Nesterov & Spokoiny, 2017) to derive the decent direction. Specifically, for the smoothing constant v and the random vector $u \sim \mathcal{N}(0, I_d)$, the directional derivative of f_i in the direction u for the smooth function f_i, $i \in [n]$, can be estimated as:

$$\hat{\nabla} f_i(x, u, v) = \frac{f_i(x + vu) - f_i(x)}{v} u. \tag{2}$$

Similarly, we define

$$\hat{\nabla} f(x, u, v) = \frac{1}{n}\sum_{i=1}^{n} \frac{f_i(x + vu) - f_i(x)}{v} u, \tag{3}$$

as the approximation of the full directional gradient.

Since the smoothing constant v is fixed, for simplicity, we

leave out v in these gradient estimations and set

$$\hat{\nabla} f_i(x,u) := \hat{\nabla} f_i(x,u,v),$$

and

$$\hat{\nabla} f(x,u) := \hat{\nabla} f(x,u,v),$$

in the rest of the paper. Under the smoothness condition, the following lemma and corollary show that these estimators defined in Eq.(2) and Eq.(3) are nearly unbiased in expectation. The proof is deferred to Appendix B.

Lemma 3.1. *Let the random vector u drawn from the multivariate Gaussian distribution $\mathcal{N}(0, I_d)$. For the L-smooth function f_i and any $x \in \mathbb{R}^d$, $i \in [n]$, the estimator in Eq.(2) satisfies:*

$$\hat{\nabla} f_i(x,u) = uu^\top \nabla f_i(x) + \frac{Lv}{2} s_i(x,u)\|u\|^2 u, \quad (4)$$

and its expectation w.r.t. u is

$$\mathbb{E}_u[\hat{\nabla} f_i(x,u)] = \nabla f_i(x) + \frac{Lv}{2}\tau_i(x,u), \quad (5)$$

where $s_i(x,u)$ is a function of x and u within the range of $[0,1]$, and $\tau_i(x,u)$ is the error term with $\|\tau_i\| \le (d+3)^{\frac{3}{2}}$.

Furthermore, the expected norm of the estimation error between $\hat{\nabla} f_i(x,u)$ and $\nabla f(x)$ is bounded as

$$\begin{aligned} &\mathbb{E}_u[\|\hat{\nabla} f_i(x,u) - \nabla f(x)\|^2] \\ &\le 3(d+1)\|\nabla f(x)\|^2 + \frac{3L^2v^2}{4}(d+6)^3 + \\ &\quad 3(d+2)\|\nabla f_i(x) - \nabla f(x)\|^2. \end{aligned} \quad (6)$$

In Eq.(4), $s_i(x,u)$ measures the curvature scaled by L along the specified direction u at the given point x. By taking the maximum value of $s_i(x,u)$ (i.e., $s_i(x,u) := 1$), we derive the upper bound of the distance between estimator $\hat{\nabla} f_i(x,u)$ and the full gradient $\nabla f(x)$ in Eq.(6). This bound comprises three components: the norm of the true gradient $\nabla f(x)$; the trivial perturbation; and the last error $\|\nabla f_i(x) - \nabla f(x)\|^2$ induced by the random sampling of i. Similarly, the following corollary establishes the nearly unbiased and bounded variance properties of $\hat{\nabla} f(x,u)$.

Corollary 3.2. *Let the random vector u drawn from the multivariate Gaussian distribution $\mathcal{N}(0, I_d)$. For the L-smooth function f and any $x \in \mathbb{R}^d$, $i \in [n]$, the full gradient estimator $\hat{\nabla} f(x,u)$ in Eq.(3) satisfies:*

$$\hat{\nabla} f(x,u) = uu^\top \nabla f(x) + \frac{Lv}{2} s(x,u)\|u\|^2 u, \quad (7)$$

with the expectation

$$\mathbb{E}_u[\hat{\nabla} f(x,u)] = \nabla f(x) + \frac{Lv}{2}\tau(x,u), \quad (8)$$

where $s(x,u) = \frac{1}{n}\sum_{i=1}^n s_i(x,u)$ and $\|\tau\| \le (d+3)^{\frac{3}{2}}$.

Moreover, the expected norm of the estimation error between $\hat{\nabla} f(x,u)$ and $\nabla f(x)$ is bounded as

$$\begin{aligned} &\mathbb{E}_u[\|\hat{\nabla} f(x,u) - \nabla f(x)\|^2] \\ &\le \frac{L^2v^2}{2}(d+6)^3 + 2(d+1)\|\nabla f(x)\|^2. \end{aligned} \quad (9)$$

From Eq.(5) and Eq.(8), when the smoothing constant v is sufficiently small, the following approximations hold: $\mathbb{E}[\hat{\nabla} f_i(x,u)] \approx \nabla f_i(x)$ and $\mathbb{E}[\hat{\nabla} f(x,u)] \approx \nabla f(x)$. Besides, Eq.(6) and Eq.(9) indicate that both the expected norm of the estimation errors $\mathbb{E}[\|\hat{\nabla} f_i(x,u) - \nabla f(x)\|^2]$ and $\mathbb{E}[\|\hat{\nabla} f(x,u) - \nabla f(x)\|^2]$ depend on $\|\nabla f(x)\|^2$. The presence of the norm of full gradient $\|\nabla f(x)\|^2$ introduces the additional noise, which requires $\|\nabla f(x)\|^2$ to gradually approach 0 to obtain the precise estimation of $\nabla f(x)$ and guarantee convergence.

However, according to the FO optimality condition, for the objective function in Eq.(1), the following relation holds at the optimal point x^*:

$$0 \in \nabla f(x^*) + \partial\psi(x^*),$$

which implies that $\|\nabla f(x^*)\|^2 \neq 0$ in the general case. Therefore, as characterized by Eq.(9) (Eq.(6)), after excluding the trivial perturbation introduced by v, the variance of the estimator $\hat{\nabla} f(x,u)$ ($\hat{\nabla} f_i(x,u)$) persists as a non-zero value and does not diminish. Analogous to the impact of the sampling noise in SGD, with a constant learning rate η, this significantly compromises the convergence performance.

Remark 3.3. In contrast to the random direction in Eq.(7), researchers (Lian et al., 2016; Ji et al., 2019) utilize the finite-difference technique (Kiefer & Wolfowitz, 1952) to approximate the gradient as follows,

$$\sum_{l=1}^{d} \frac{f(x+ve_l) - f(x)}{v} e_l,$$

where $e_l \in \mathbb{R}^d$ is the l-th standard basis vector. Compared to the random gradient estimate in Eq.(7), the above gradient estimate leverages all partial gradient information to approximate the gradient, and its variance is only influenced by the trivial perturbation arising from the smooth constant v. In fact, it can be regarded as a first-order method and can find the correct solution when $\nabla f(x^*) \neq 0$. However, it requires to take $\Omega(d)$ function evaluations, which is impractical in high-dimensional problems.

3.2. Coordinate-Wise Variance Reduction in Zeroth-Order Optimization

Since $\|\nabla f(x^*)\|^2 \neq 0$, simply extending the SVRG methods (Johnson & Zhang, 2013; Xiao & Zhang, 2014) to

composite as well as constrained optimization problems can not eliminate the overall variance - such methods only reduce the variance induced by the stochastic sampling $\|\nabla f_i(x) - \nabla f(x)\|^2$ in Eq.(6) and omit the coordinate-wise variance. Although driving all partial derivative information can mitigate the coordinate-wise variance (Huang et al., 2019; Kazemi & Wang, 2023), the computation is prohibitive for high-dimensional problems. To reduce the overall variance while preserving computational efficiency, we propose a novel gradient estimator, utilizing only the random gradient estimate throughout the iterative process.

According to Eq.(7), when v is sufficient small, we have that

$$\hat{\nabla} f(x, u) \approx u^\top \nabla f(x) u, \tag{10}$$

which is an approximation of the directional derivative in the direction u. Here, we leverage the directional derivative estimation and build a gradient estimator $\tilde{h}$ by the averaging trick, i.e.,

$$\tilde{h}_{k+1} = \tilde{h}_k + \frac{1}{d+2}(\hat{\nabla} f(x_k, u) - uu^\top \tilde{h}_k). \tag{11}$$

The estimator $\tilde{h}$ employs the difference between the estimated directional derivative $\hat{\nabla} f(x_k, u)$ and the value of h in the same direction u to refine the gradient estimation error. Moreover, the coefficient $\frac{1}{d+2}$ further smoothens this correction across all dimensions. In Section 4, we will demonstrate that by incrementally refining the gradient estimated error at disparate directions, as $x \to x^*$, the estimator $\tilde{h}$ can gradually converge to $\nabla f(x^*)$.

Remark 3.4. We can also connect the estimator $\tilde{h}$ with the sketch-and-project method (Hanzely et al., 2018). Since $u^\top \nabla f(x)$ in Eq.(10) is the gradient sketch, according to Hanzely et al. (2018), $\tilde{h}_{k+1}$ can be obtained by solving the following optimization problem,

$$\begin{aligned} \min_{\tilde{h}} \quad & \|\tilde{h} - \tilde{h}_k\|^2, \\ \text{s.t.} \quad & u^\top h = u^\top \nabla f(x_k). \end{aligned}$$

By solving the Lagrangian function $\|h - h_k\|^2 + \lambda u^\top (\tilde{h} - \nabla f(x_k))$, we have the closed-form solution:

$$h^{k+1} = h_k + \frac{uu^\top (\nabla f(x_k) - \tilde{h}_k)}{\|u\|^2}.$$

Compared to the above update formula, Eq.(11) simplifies the computation of the norm $\|u\|^2$ by approximating it with its expectation ($\mathbb{E}[\|u\|^2] = d$, see Lemma A.4 in Appendix A) plus a constant. Therefore, from the view of the sketch-and-projection method, Eq.(11) essentially requires that the sketches of the estimator $\tilde{h}$ and gradient $\nabla f(x)$ are identical. Hence, when the projections of the gradient $\nabla f(x)$ and the estimator $\tilde{h}$ align along numerous directions, $\tilde{h}$ gradually becomes identical to the gradient.

In Algorithm 1, we present the proposed algorithm ZPDVR, utilizing the proposed estimator $\tilde{h}$ and SVRG techniques to reduce both the sampling variance and coordinate-wise variance in the composite optimization problem Eq.(1). In ZPDVR, like SVRG, we maintain a reference point w_k and pass over the full data to compute the gradient estimate $\tilde{\nabla} f(w_k)$ at the reference point,

$$\tilde{\nabla} f(w_k) = \tilde{h}_k + \hat{\nabla} f(w_k, u) - uu^\top \tilde{h}_k, \tag{12}$$

where we leverage the directional gradient $\hat{\nabla} f(w_k, u)$ to refine the gradient estimator $\tilde{h}_k$ in the direction u.

Furthermore, with the conserved $\tilde{\nabla} f(w_k)$, we can derive the stochastic gradient estimate g_k,

$$g_k = \hat{\nabla} f_i(x_k, u_k) - \hat{\nabla} f_i(w_k, u_k) + \tilde{\nabla} f(w_k), \tag{13}$$

which reduces the sampling variance. Then, the update of x_{k+1} is

$$x_{k+1} = \text{prox}_{\eta\psi}(x - \eta g_k). \tag{14}$$

Notably, with Eq.(12) and Eq.(13), we demonstrate that g_k is also a nearly unbiased estimation of $\nabla f(x_k)$ in the following Lemma. The proof is deferred to Appendix B.

Lemma 3.5. *For the L-smooth function f_i, $i \in [n]$, the expected value of g_k defined in Eq.(13) is*

$$\mathbb{E}_{u,u_k,i}[g_k] = \nabla f(x_k) + \frac{Lv}{2}\tau_{i,k}, \tag{15}$$

where $\tau_{i,k} = \mathbb{E}_{u,u_k,i}[(s_i(x_k, u_k) - s_i(w_k, u_k))\|u_k\|^2 u_k + s(w_k, u)\|u\|^2 u]$ with the norm $\|\tau_{i,k}\| \le 2(d+3)^{\frac{3}{2}}$.

In Algorithm 1, instead of the classic double loop structure, we adopt a loopless scheme (Kovalev et al., 2020; Qian et al., 2021): with a small probability p, we update the reference point w_{k+1} with the value of x_k and take a full pass over the n data points to refine the estimate $\tilde{h}_{k+1}$ in the direction u; with the probability $1 - p$, we remain the previous w_k and $\tilde{h}_k$ and utilize them in the next iteration.

Furthermore, with the probability p to perform a full pass over the dataset, ZPDVR has to query the function value $\mathcal{O}(pn)$ times in the expectation at each iteration. When $p = \frac{1}{n}$, the proposed method ZPDVR recovers the complexity of the FO-SVRG methods. Compared to the complexity $\mathcal{O}(d)$ in other ZO variance reduction methods (Huang et al., 2019; Kazemi & Wang, 2023), ZPDVR is independence of the dimension size d, which makes it efficient and practical for high-dimensional problems. Besides, in ZPDVR, directions u and u_k drawn in Eq.(12) and Eq.(13) are independent. Hence, in Algorithm 1, except for $\tilde{\nabla} f(w_k)$, we also have to save u to update $\tilde{h}_{k+1}$ when $z_k \le p$.

Algorithm 1 Zeroth-order Proximal Doubly Variance Reduction

Input: x_0, $w_0 = x_0$, $\tilde{h}_0$, the probability p, and the learning rate η.

for $k = 0, \ldots, K-1$ **do**
  if $k = 0$ **or** $w_k \neq w_{k-1}$ **then**
    Sample a new vector $u \sim \mathcal{N}(0, I)$ and save it
    $\tilde{\nabla} f(w_k) = \tilde{h}_k + \hat{\nabla} f(w_k, u) - uu^\top \tilde{h}_k$
  else
    $\tilde{\nabla} f(w_k) = \tilde{\nabla} f(w_{k-1})$
  end if
  Sample $u_k \sim \mathcal{N}(0, I)$ and pick $i \in \{1, \ldots, n\}$ uniformly at random
  $g_k = \hat{\nabla} f_i(x_k, u_k) - \hat{\nabla} f_i(w_k, u_k) + \tilde{\nabla} f(w_k)$
  $x_{k+1} = \text{prox}_{\eta\psi}(x_k - \eta g_k)$
  Draw z_k from the uniform distribution $U[0, 1]$
  if $z_k \leq p$ **then**
    $w_{k+1} = x_k$
    $\tilde{h}_{k+1} = \tilde{h}_k + \frac{1}{d+2}(\hat{\nabla} f(x_k, u) - uu^\top \tilde{h}_k)$
  else
    $w_{k+1} = w_k$
    $\tilde{h}_{k+1} = \tilde{h}_k$
  end if
end for

Output: x_K

Remark 3.6. Note that, in Algorithm 1, the update of the reference point w and the gradient estimate $\tilde{h}$ are bound together. Since w is updated with the probability p, the estimator $\tilde{h}$ is not required to be updated at every iteration, which differs significantly from (Hanzely et al., 2018). We demonstrate that sporadically updating the estimator $\tilde{h}$ can reduce the coordinate-wise variance, and ZPDVR is not a trivial extension. Moreover, according to the update of w in Algorithm 1 and Eq.(11), $\tilde{h}_k$ is actually estimating the gradient at the reference point, $\nabla f(w_k)$. Then, when $w \to x^*$, we can obtain that $\tilde{h} \to \nabla f(x^*)$.

4. Convergence Analysis

This section will establish the convergence property of the proposed method ZPVR. Initially, we define the following Lyapunov function Ψ:

$$\Psi(x_k) = \|x_k - x^*\|^2 + \alpha\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{\beta}{n}\sum_{i=1}^{n}\|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2, \tag{16}$$

where both α and β are positive real values.

The Lyapunov function in Eq.(16) comprises three distinct terms: the first term quantifies the distance between the current point x_k and the optimal point x^*; the second term, $\|\tilde{h}_k - \nabla f(x^*)\|^2$, delineates the disparity between the gradient estimator $\tilde{h}_k$ and the gradient at x^*; and the final term serves as an approximation for the difference between the reference point w_k and x^* in the strongly convex setting. During the training process, if the value of the Lyapunov function (16) converges to 0, we have that both x_k and w_k attain the optimal point x^* and the gradient estimator $\tilde{h}_k$ approaches $\nabla f(x^*)$. According to the definition of the Laypunov function, it involves successive iterations with respect to $\|x_k - x^*\|^2$, $\|\tilde{h}_k - \nabla f(x^*)\|^2$, and $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2$. Next, we shall establish recursive formulations for these components in the subsequent three lemmas, respectively. The proofs are provided in Appendix B.

Lemma 4.1. *For the μ-strongly and L smooth function f_i, $i \in [n]$, and the convex function $\psi(x)$, according to Algorithm 1, we have*

$$\begin{aligned}&\mathbb{E}[\|x_{k+1} - x^*\|^2] \\ &\leq (1 - \frac{\mu}{2}\eta)\|x_k - x^*\|^2 - 2\eta\Big(f(x_k) - f(x^*) - \\ &\quad \langle \nabla f(x^*), x_k - x^*\rangle\Big) + \frac{2(d+3)^3 L^2 v^2 \eta}{\mu} + \\ &\quad \eta^2 \mathbb{E}[\|g_k - \nabla f(x^*)\|^2].\end{aligned} \tag{17}$$

In Lemma 4.1, the convexity of f indicates the positive value of $f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^*\rangle$. Except for the ZO perturbation, the iteration of x involves the estimation gap between the stochastic gradient estimation g_k and $\nabla f(x^*)$, i.e., $\mathbb{E}[\|g_k - \nabla f(x^*)\|^2]$, which will be bounded in the later analysis.

Lemma 4.2. *For the L smooth function f_i, $i \in [n]$, following Algorithm 1, we have*

$$\begin{aligned}&\mathbb{E}[\|\tilde{h}_{k+1} - \nabla f(x^*)\|^2] \\ &\leq (1 - \frac{p}{2(d+2)})\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{5(d+6)^3 L^2 v^2 p}{2(d+2)} + \\ &\quad \frac{8Lp}{3(d+2)}(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^*\rangle).\end{aligned} \tag{18}$$

Lemma 4.3. *For the L smooth function f_i, $i \in [n]$, the recursive formulation of $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2$ is as follows:*

$$\begin{aligned}&\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w_{k+1}) - \nabla f_i(x^*)\|^2] \\ &\leq \frac{(1-p)}{n}\sum_{i=1}^{n}\|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + \\ &\quad 2Lp(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^*\rangle).\end{aligned} \tag{19}$$

According to Lemma 4.2 and Lemma 4.3, $\mathbb{E}[\|\tilde{h}_k - \nabla f(x^*)\|^2]$ and $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2]$ is re-

duced by $\mathcal{O}(1-\frac{p}{d})$ and $\mathcal{O}(1-p)$, respectively. Furthermore, all recursive formulations of these three components involve the term $f(x_k)-f(x^*)-\langle\nabla f(x^*),x_k-x^*\rangle$. In the subsequent lemma, we will establish a connection between the term $\|g_k-\nabla f(x^*)\|^2$ introduced in Lemma 4.1 and $\|\tilde{h}_k-\nabla f(x^*)\|^2$, driving the bound for the former item.

Lemma 4.4. *For the L smooth function f_i, $i\in[m]$, the following inequality holds*

$$\begin{aligned}&\mathbb{E}[\|g_k-\nabla f(x^*)\|^2]\\ \le\ & 2(2d+3)\|\tilde{h}_k-\nabla f(x^*)\|^2+4L^2v^2(d+6)^3+\\ &\frac{4(2d+3)}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2+\\ &4(d+3)L(f(x_k)-f(x^*)-\langle\nabla f(x^*),x_k-x^*\rangle).\end{aligned}\tag{20}$$

From Lemma 4.4, the bound of $\mathbb{E}[\|g_k-\nabla f(x^*)\|^2]$ involves $\tilde{h}_k-\nabla f(x^*)$, $f(x_k)-f(x^*)-\langle\nabla f(x^*),x_k-x^*\rangle$, and $\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2$. Hence, setting $p=\frac{1}{n}$ and combing with Lemma 4.1, Lemma 4.2, and Lemma 4.3, we can obtain the whole recursive formula of the Lyapunov function in the following corollary:

Corollary 4.5. *For the L smooth and μ-strongly convex function f_i, $i\in[n]$, and the convex function $\psi(x)$, when $p=\frac{1}{n}$, the recursive formula of the Lyapunov function Ψ is following:*

$$\begin{aligned}&\mathbb{E}[\Psi(x_{k+1})]\\ \le\ &(1-\frac{\mu}{2}\eta)\|x_k-x^*\|^2+\alpha'\|\tilde{h}_k-\nabla f(x^*)\|^2+\\ &\beta'\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2-\gamma C+\Delta,\end{aligned}\tag{21}$$

where $\alpha'=\alpha(1-\frac{1}{2n(d+2)}+\frac{2(2d+3)\eta^2}{\alpha})$, $\beta'=\beta(1-\frac{1}{n}+\frac{4(2d+3)\eta^2}{\beta})$, $\gamma=2\eta-4(d+3)L\eta^2-\frac{8L}{3n(d+2)}\alpha-\frac{2L}{n}\beta$, $C=f(x_k)-f(x^*)-\langle\nabla f(x^*),x_k-x^*\rangle$, and $\Delta=\frac{2(d+3)^3L^2v^2\eta}{\mu}+4L^2v^2(d+6)^3\eta^2+\frac{5(d+6)^3L^2v^2}{2n(d+2)}\alpha$.

According to Corollary 4.5, by carefully selection η, α, and β, we can guarantee that $\gamma\le 0$ and $0<(1-\frac{\mu\eta}{2}),\alpha',\beta'<1$. Hence, we can drive the convergence rate of the Lyapunov function, which is presented in the following theorem.

Theorem 4.6. *Let $p=\frac{1}{n}$, $\alpha=8n(d+2)(2d+3)\eta^2$, $\beta=8n(2d+3)\eta^2$, and $\eta=\frac{1}{(40d+63)L}$. Given the L smooth and u-strongly convex function f_i, $i\in[n]$, and the convex function $\psi(x)$, for $k\in 0,\dots,K-1$, the Lyapunov function satisfies*

$$\begin{aligned}&\mathbb{E}[\Psi(x_{k+1})]\\ \le\ &\max\{1-\frac{1}{\kappa(80d+126)},1-\frac{1}{4n(d+2)}\}\Psi(x_k)+\delta,\end{aligned}\tag{22}$$

Table 2. Summary of data sets and regularization coefficients (λ_1 and λ_2) used in our experiments.

Data Set	n	d	λ_1	λ_2	v
a9a	32561	123	10^{-4}	10^{-4}	10^{-3}
w8a	49749	300	10^{-5}	10^{-4}	10^{-3}
covtype	581012	54	10^{-4}	10^{-5}	10^{-3}
gisette	6000	5000	10^{-4}	10^{-4}	10^{-3}

where $\delta=\frac{2(d+3)^3v^2\kappa}{40d+63}+\frac{8(5d+8)(d+6)^3v^2}{(40d+63)^2}$.

According to Theorem 4.6, we can telescope the inequality (22) from $k=0$ for $K-1$, and then find the SZO query complexity and the value of the smooth constant v to get the ϵ accuracy solution.

Corollary 4.7. *Let $p=\frac{1}{n}$, $\alpha=8n(d+2)(2d+3)\eta^2$, $\beta=8n(2d+3)\eta^2$, and $\eta=\frac{1}{(40d+63)L}$. Given the L smooth and u-strongly convex function f_i, $i\in[n]$, and the convex function $\psi(x)$, we have*

$$\mathbb{E}[\Psi(x_K)]\le(1-\theta)^K\Psi(x_0)+\sigma,\tag{23}$$

where $\theta=\frac{1}{\kappa(80d+126)+4n(d+2)}$, and $\sigma=[(80d+126)+4n(d+2)]\kappa(\kappa+1)(d+6)^2v^2$.

From Corollary 4.7, in order to get ϵ accuracy for the Lyapunov function as well as F, the smooth constant v need to sufficiently small such that $v=\mathcal{O}(\sqrt{\frac{\epsilon}{\kappa^2d^3n}})$. We can also derive the $\mathcal{O}(d(n+\kappa)\log(\frac{1}{\epsilon}))$ SZO query complexity of ZPDVR, which aligns with the best-known results of SVRG methods. The optimal SZO query complexity demonstrates the efficiency of ZPDVR.

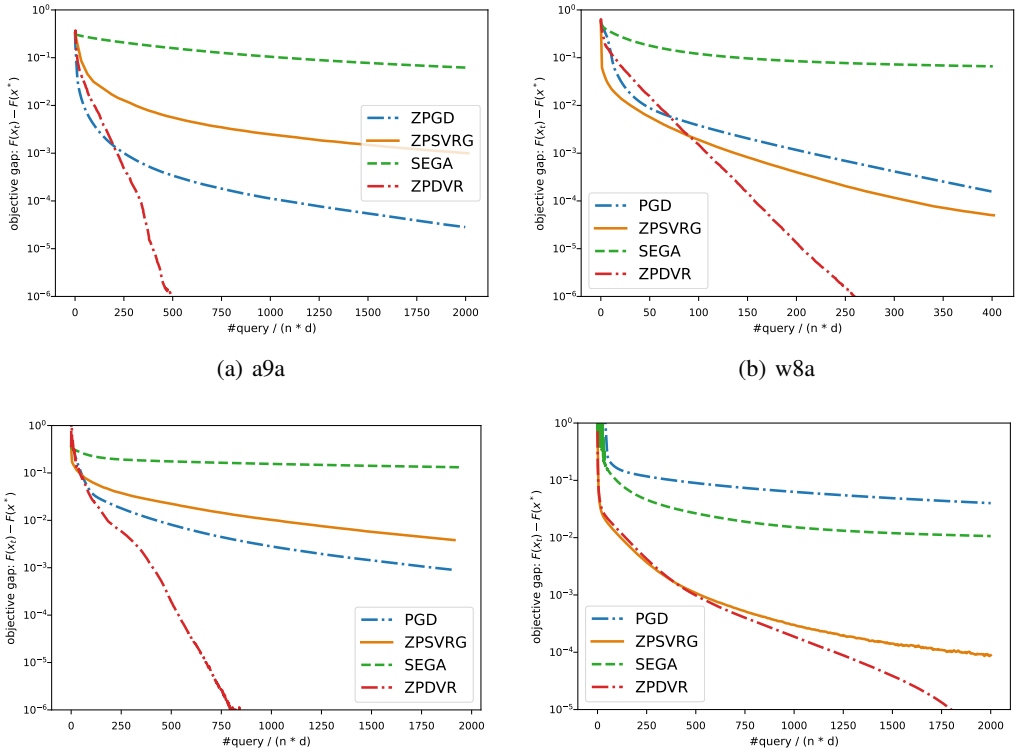
5. Experiment

In this section, we conduct several numerical experiments to demonstrate the convergence propriety of the proposed ZPDVR and compare its performance against other related FO and ZO methods. Here, we focus on a binary logistic classification problem. Given the data set $\mathcal{D}=\{(z_i,y_i)\}_{i=1}^n$, where $z_i\in\mathbb{R}^d$ is the feature vector and $y_i\in\{-1,1\}$ is the label, the strongly convex objective function $F(x)$ is the regularized cross entropy loss, i.e.,

$$\min_{x\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^n\log(1+\exp(-y_ix^\top z_i)+\frac{\lambda_2}{2}\|x\|^2+\lambda_1\|x\|_1,$$

with $f_i(x)=\log(1+\exp(-y_ix^\top z_i))+\frac{\lambda_2}{2}\|x\|^2$ and $\psi(x)=\lambda\|x\|_1$, for $i\in[n]$.

Specifically, we choose three related FO and ZO methods as baselines: Proximal Gradient Descent (PGD), ZPSVRG, and SEGA. In our experiment, ZPSVRG carries out the ZO version of proximal variance reduction method (Xiao & Zhang,



(a) a9a (b) w8a

(c) covtype (d) gisette

Figure 1. Comparison of different zeroth-order methods for the loss residual $F(x) - F(x^*)$ versus the number of SZO. The y axis is on a logarithmic scale and the x label is the number of SZO divided by $n * d$.

2014) with the random gradient estimate defined in Eq.(3), which has the non-vanishing coordinate-wise variance.

In Table 2, we present the regularization coefficients and the smooth constant used in four binary classification data sets from LIBSVM website[1]. To enhance implementation efficiency, ZPSVRG, ZPDVR, and SEGA are executed in the batch version. The grid search is performed to identify optimal hyperparameters for each method. The detailed information of the hyperparameter tuning procedure is provided in Appendix C.

As shown in Figure 1, we compare the objective gap $F(x) - F(x^*)$ of ZPDVR with other methods in terms of SZO query complexity. Across all data sets, ZPDVR achieves the best performance and presents the linear convergence rate as demonstrated in Theorem 4.6. Besides, PGD and SEGA also obtain the linear convergence rate, albeit at a slower pace than ZPDVR. While due to the non-vanishing coordinate-wise variance in ZPSVRG, we observe that the function value of ZPSVRG can not decrease to the specified precision. For instance, in Figure 1(d), ZPSVRG stagnates in the neighborhood of $F(x^*)$ with the radius 10^{-4}. These results emphasize the linear convergence rate and double variance reduction properties of ZPDVR analyzed in Section 4 for the composite problem (1).

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

6. Conclusion

This paper introduces a novel ZO variance reduction method, ZPDVR, designed to diminish both the inherent sampling variance and the coordinate-wise variance in composite optimization problems. Unlike prior method, ZPDVR obviates the need for $\Omega(d)$ function evaluations to approximate FO information. It instead only requires $\mathcal{O}(1)$ SZO query in expectation per iteration, which enhances the computational efficiency, particularly for high-dimensional problems. Additionally, we establish the linear convergence of ZPDVR and show that its SZO query complexity is $\mathcal{O}(d(n+\kappa)\log(\frac{1}{\epsilon}))$, paralleling the optimal results achieved by SVRG methods. Moreover, ZPDVR is adaptable to constrained optimization scenarios, including black-box optimization problems. Empirical results validate both the convergence properties and the superior performance of ZPDVR.

Acknowledge

This work was supported by the National Natural Science Foundation of China under Grant 12101491, the National Natural Science Foundation for Outstanding Young Scholars of China under Grant 72122018, the MOE Project of Key Research Institute of Humanities and Social Sciences No. 22JJD110001, and A*star Centre for Frontier AI Research.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pp. 699–707. PMLR, 2016.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Berahas, A. S., Cao, L., Choromanski, K., and Scheinberg, K. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, 22(2):507–560, 2022.

Chen, A., Zhang, Y., Jia, J., Diffenderfer, J., Liu, J., Parasyris, K., Zhang, Y., Zhang, Z., Kailkhura, B., and Liu, S. Deepzero: Scaling up zeroth-order optimization for deep model training. *arXiv preprint arXiv:2310.02025*, 2023a.

Chen, L., Xu, J., and Luo, L. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In *International Conference on Machine Learning*, volume 50, pp. 5219–5233, 2023b.

Chen, Z., Guo, W., Feng, Y., Li, Y., Zhao, C., Ren, Y., and Shao, L. Deep-learned regularization and proximal operator for image compressive sensing. *IEEE Transactions on Image Processing*, 30:7112–7126, 2021.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, pp. 689–699, 2018.

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

Gower, R. M., Schmidt, M., Bach, F., and Richtárik, P. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: Variance reduction via jacobian sketching. *Mathematical Programming*, 188:135–192, 2021.

Hanzely, F., Mishchenko, K., and Richtárik, P. Sega: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2082–2093, 2018.

Huang, F., Gu, B., Huo, Z., Chen, S., and Huang, H. Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1503–1510, 2019.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

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

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, pp. 315–323, 2013.

Kazemi, E. and Wang, L. Efficient zeroth-order proximal stochastic method for nonconvex nonsmooth black-box problems. *Machine Learning*, pp. 1–24, 2023.

Kiefer, J. and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.

Kovalev, D., Horváth, S., and Richtárik, P. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pp. 451–467. PMLR, 2020.

Li, J., Balasubramanian, K., and Ma, S. Stochastic zeroth-order riemannian derivative estimation and optimization. *Mathematics of Operations Research*, 48(2):1183–1211, 2023.

Li, Z., Bao, H., Zhang, X., and Richtárik, P. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pp. 6286–6295. PMLR, 2021.

Lian, X., Zhang, H., Hsieh, C.-J., Huang, Y., and Liu, J. A comprehensive linear speedup analysis for asynchronous

stochastic parallel optimization from zeroth-order to first-order. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3054–3062. Curran Associates, Inc., 2016.

Lin, Z., Xia, J., Deng, Q., and Luo, L. Decentralized gradient-free methods for stochastic non-smooth non-convex optimization. *arXiv preprint arXiv: 2310.11973*, 2023.

Liu, S., Kailkhura, B., Chen, P.-Y., Ting, P., Chang, S., and Amini, L. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 3727–3737, 2018.

Liu, S., Chen, P.-Y., Kailkhura, B., Zhang, G., Hero III, A. O., and Varshney, P. K. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., and Arora, S. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36, 2023.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

Qian, X., Qu, Z., and Richtárik, P. L-svrg and l-katyusha with arbitrary sampling. *The Journal of Machine Learning Research*, 22(1):4991–5039, 2021.

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

Rubinstein, R. Y. and Marcus, R. Efficiency of multivariate control variates in monte carlo simulation. *Operations Research*, 33(3):661–677, 1985.

Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 742–749, 2019.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Ye, H., Huang, Z., Fang, C., Li, C. J., and Zhang, T. Hessian-aware zeroth-order optimization for black-box adversarial attack. *arXiv preprint arXiv:1812.11377*, 2018.

Yi, X., Zhang, S., Yang, T., and Johansson, K. H. Zeroth-order algorithms for stochastic distributed nonconvex optimization. *Automatica*, 142:110353, 2022.

Yun, J., Lozano, A. C., and Yang, E. Adaptive proximal gradient methods for structured neural networks. *Advances in Neural Information Processing Systems*, 34: 24365–24378, 2021.

A. Some useful Lemmas

Lemma A.1 (Cauchy–Schwarz inequality)**.** *For all vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, we have*

$$\|\sum_{i=1}^{n} x_i\|^2 \leq n \sum_{i=1}^{n} \|x_i\|^2. \tag{24}$$

Lemma A.2 (Young's inequality)**.** *For any two vectors $x, y \in \mathbb{R}^d$, we have*

$$\langle x, y \rangle \leq \frac{a\|x\|^2}{2} + \frac{\|y\|^2}{2a}, \tag{25}$$

where a is a positive real number.

Lemma A.3 (Theorem 2.1.5 of Nesterov et al. (2018))**.** *For a L-smooth function f defined in the domain $\mathcal{X}$, $\forall x, y \in \mathcal{X}$, we have*

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2. \tag{26}$$

Lemma A.4. *Let A and B be two symmetric matrices. Random vector u has the Gaussian distribution, i.e., $u \sim \mathcal{N}(0, I_d)$. Then, we have*

$$\mathbb{E}_u[u^\top A u \cdot u^\top B u] = (\operatorname{tr} A)(\operatorname{tr} B) + 2(\operatorname{tr} AB). \tag{27}$$

Corollary A.5. *Let v be a any vector in $\mathbb{R}^d$. For the random vector u with the Gaussian distribution, i.e., $u \sim \mathcal{N}(0, I_d)$, we have*

$$\mathbb{E}_u[\|uu^\top v\|^2] = (d+2)\|v\|^2. \tag{28}$$

Proof.

$$\begin{aligned}
\mathbb{E}_u[\|uu^\top v\|^2] &= \mathbb{E}_u[\operatorname{tr}(v^\top uu^\top uu^\top v)] \\
&= \mathbb{E}_u[\operatorname{tr}(u^\top uu^\top vv^\top u)] \\
&= \mathbb{E}_u[u^\top I u \cdot u^\top vv^\top u] \\
&\overset{(27)}{=} \operatorname{tr}(I)\operatorname{tr}(vv^\top) + 2\operatorname{tr}(vv^\top) \\
&= (d+2)\operatorname{tr}(vv^\top) \\
&= (d+2)\|v\|^2.
\end{aligned}$$

□

Lemma A.6 (Lemma 1 of Nesterov & Spokoiny (2017))**.** *Let the random vector $u \sim \mathcal{N}(0, I_d)$. we have*

$$\begin{aligned}
\mathbb{E}[\|u\|^q] &\leq d^{\frac{q}{2}} && q \in [0, 2], \\
d^{\frac{q}{2}} \leq \mathbb{E}[\|u\|^q] &\leq (q+d)^{\frac{q}{2}} && q \geq 2.
\end{aligned} \tag{29}$$

B. Missing Proofs

Proof of Lemma 3.1. For the L-smooth function f_i, we have the following Taylor expansion,

$$f_i(x + vu) = f_i(x) + v\langle \nabla f_i(x), u \rangle + \frac{v^2}{2} u^\top \nabla^2 f_i(x') u, \tag{30}$$

where $x' \in (x, x + vu)$. Plugging it into Eq.(2), we have

$$\begin{aligned}
\hat{\nabla} f_i(x, u) &= u\langle u, \nabla f_i(x) \rangle + \frac{v}{2} u^\top \nabla^2 f_i(x') uu \\
&= uu^\top \nabla f_i(x) + \frac{Lv}{2} s_i(x, u)\|u\|^2 u,
\end{aligned} \tag{31}$$

where the last equality employs the fact that $0 \preceq \nabla^2 f_i(x') \preceq L$ for any accessible x', and the function $s_i(x,u)$ is confined to the range $[0,1]$.

Taking the expectation w.r.t. u for $\hat{\nabla} f_i(x)$, we have

$$\begin{aligned}\mathbb{E}[\hat{\nabla} f_i(x,u)] &= \mathbb{E}[uu^\top]\nabla f_i(x) + \frac{Lv}{2}\mathbb{E}[s_i(x,u)\|u\|^2 u] \\ &= \nabla f_i(x) + \frac{Lv}{2}\mathbb{E}[s_i(x,u)\|u\|^2 u].\end{aligned}$$

Since $\left\|\mathbb{E}[s_i(x,u)\|u\|^2 u]\right\| \le \mathbb{E}\left[\left\|s_i(x,u)\|u\|^2 u\right\|\right] \le \mathbb{E}\left[\left\|\|u\|^2 u\right\|\right] = \mathbb{E}[\|u\|^3]$, with Eq.(29) $\mathbb{E}[\|u\|^3] \le (d+3)^{\frac{3}{2}}$, we then have $\left\|\mathbb{E}[s_i(x,u)\|u\|^2 u]\right\| \le (d+3)^{\frac{3}{2}}$.

For the expected norm, we have

$$\begin{aligned}&\mathbb{E}[\|\hat{\nabla} f_i(x,u) - \nabla f(x)\|^2] \\ \overset{(4)}{=}\ & \mathbb{E}[\|uu^\top \nabla f_i(x) + \frac{Lv}{2} s_i(x,u)\|u\|^2 u - \nabla f(x)\|^2] \\ =\ & \mathbb{E}\left[\left\|uu^\top \nabla f_i(x) - uu^\top \nabla f(x) + uu^\top \nabla f(x) - \nabla f(x) + \frac{Lv}{2} s_i(x,u)\|u\|^2 u\right\|^2\right] \\ \overset{(24)}{\le}\ & 3\mathbb{E}[\|uu^\top(\nabla f_i(x) - \nabla f(x)\|^2] + 3\mathbb{E}[\|(uu^\top - I)\nabla f(x)\|^2] + \frac{3L^2v^2}{4}\mathbb{E}[\|u\|^6] \\ \overset{(28)}{\le}\ & 3(d+2)\|\nabla f_i(x) - \nabla f(x)\|^2 + 3(d+1)\|\nabla f(x)\|^2 + \frac{3L^2v^2}{4}\mathbb{E}[\|u\|^6].\end{aligned}$$

Combining with Lemma A.6, we conclude the result in Lemma 3.1. □

Proof of Corollary 3.2. According to Lemma 3.1 and the definition of $\hat{\nabla} f(x,u)$, we have that

$$\hat{\nabla} f(x,u) = uu^\top \nabla f(x) + \frac{Lv}{2} s(x,u)\|u\|^2 u,$$

where $s(x,u) = \frac{1}{n}\sum_{i=1}^n s_i(x,u)$.

Hence, taking the expectation w.r.t. u, we also derive

$$\begin{aligned}\mathbb{E}[\hat{\nabla} f(x,u)] &= \mathbb{E}[uu^\top]\nabla f(x) + \frac{Lv}{2} s(x,u)\mathbb{E}[\|u\|^2 u] \\ &= \nabla f_i(x) + \frac{Lv}{2}\mathbb{E}[s(x,u)\|u\|^2 u],\end{aligned}$$

where $\left\|\mathbb{E}[s(x,u)\|u\|^2 u]\right\| \le (d+3)^{\frac{3}{2}}$

For the expected norm, we have

$$\begin{aligned}&\mathbb{E}[\|\hat{\nabla} f(x,u) - \nabla f(x)\|^2] \\ \overset{(7)}{=}\ & \mathbb{E}[\|uu^\top \nabla f(x) + \frac{Lv}{2} s(x,u)\|u\|^2 u - \nabla f(x)\|^2] \\ \overset{(24)}{\le}\ & 2\mathbb{E}[\|(uu^\top - I)\nabla f(x)\|^2] + \frac{L^2v^2}{2}\mathbb{E}[\|u\|^6] \\ \overset{(28)}{\le}\ & 2(d+1)\|\nabla f(x)\|^2 + \frac{L^2v^2}{2}\mathbb{E}[\|u\|^6],\end{aligned}$$

which implies the result. □

Proof of Lemma 3.5. From Eq.(13) and Eq.(12), we have

$$\begin{aligned}
&\mathbb{E}_{u,u_k,i}[g_k] \\
&= \mathbb{E}_{u,u_k,i}[\hat{\nabla} f_i(x_k,u_k) - \hat{\nabla} f_i(w_k,u_k) + \tilde{h}_k + \hat{\nabla} f(w_k,u) - uu^\top \tilde{h}_k] \\
&\overset{(4),(7)}{=} \mathbb{E}_{u,u_k,i}\Big[u_k u_k^\top \nabla f_i(x_k) + \frac{Lv}{2} s_i(x_k,u_k)\|u_k\|^2 u_k - u_k u_k^\top \nabla f_i(w_k) - \frac{Lv}{2} s_i(w_k,u_k)\|u_k\|^2 u_k + \\
&\qquad \tilde{h}_k + uu^\top \nabla f(w_k) + \frac{Lv}{2} s(w_k,u)\|u\|^2 u - uu^\top \tilde{h}_k\Big] \\
&= \nabla f(x_k) + \frac{Lv}{2}\mathbb{E}_{u,u_k,i}\Big[(s_i(x_k,u_k) - s_i(w_k,u_k))\|u_k\|^2 u_k + s(w_k,u)\|u\|^2 u\Big].
\end{aligned}$$

Let $\tau_{i,k} = \mathbb{E}_{u,u_k,i}[(s_i(x_k,u_k) - s_i(w_k,u_k))\|u_k\|^2 u_k + s(w_k,u)\|u\|^2 u]$. Then, the upper bound for the norm of $\tau_{i,k}$ is

$$\begin{aligned}
\|\tau_{i,k}\| &= \Big\|\mathbb{E}_{u,u_k,i}[(s_i(x_k,u_k) - s_i(w_k,u_k))\|u_k\|^2 u_k + s(w_k,u)\|u\|^2 u]\Big\| \\
&\le \Big\|\mathbb{E}[(s_i(x_k,u_k) - s_i(w_k,u_k))\|u_k\|^2 u_k]\Big\| + \Big\|\mathbb{E}[s(w_k,u)\|u\|^2 u]\Big\| \\
&\le \mathbb{E}\Big[\Big\|(s_i(x_k,u_k) - s_i(w_k,u_k))\|u_k\|^2 u_k\Big\|\Big] + \mathbb{E}\Big[\Big\|s(w_k,u)\|u\|^2 u\Big\|\Big] \\
&\le \mathbb{E}[\|u_k\|^3] + \mathbb{E}[\|u\|^3].
\end{aligned}$$

Finally, combining with Lemma A.6, we can conclude the result. □

Proof of Lemma 4.1. According to the convexity of $\psi(x)$, the optimal point x^* satisfies

$$x^* = \mathrm{prox}_{\eta\psi}(x^* - \eta\nabla f(x^*)).$$

Hence, based on Eq.(14), we have

$$\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x^*\|^2] \\
&= \mathbb{E}[\mathrm{prox}_{\eta\psi}(x_k - \eta g_k) - \mathrm{prox}_{\eta\psi}(x^* - \eta\nabla f(x^*))] \\
&\le \mathbb{E}[\|x_k - x^* - \eta(g_k - \nabla f(x^*))\|^2] \\
&= \|x_k - x^*\|^2 - 2\eta\mathbb{E}[\langle x_k - x^*, g_k - \nabla f(x^*)\rangle] + \eta^2\mathbb{E}[\|g_k - \nabla f(x^*)\|^2] \\
&\overset{(15)}{=} \|x_k - x^*\|^2 - 2\eta\langle x_k - x^*, \nabla f(x_k) - \nabla f(x^*)\rangle - Lv\eta\langle x_k - x^*, \tau_{i,k}\rangle + \eta^2\mathbb{E}[\|g_k - \nabla f(x^*)\|^2] \\
&\le (1-\eta\mu)\|x_k - x^*\|^2 - 2\eta(f(x_k) - f(x^*) - \langle\nabla f(x^*), x_k - x^*\rangle) - Lv\eta\langle x_k - x^*, \tau_{i,k}\rangle + \\
&\qquad \eta^2\mathbb{E}[\|g_k - \nabla f(x^*)\|^2] \\
&\overset{(25)}{\le} (1 - \frac{\mu}{2}\eta)\|x_k - x^*\|^2 - 2\eta(f(x_k) - f(x^*) - \langle\nabla f(x^*), x_k - x^*\rangle) + \frac{L^2 v^2 \eta}{2\mu}\|\tau_{i,k}\|^2 + \eta^2\mathbb{E}[\|g_k - \nabla f(x^*)\|^2].
\end{aligned}$$

Coupling with Lemma 3.5, it yields the final iterative formula. □

Proof of Lemma 4.2. According to the update of $\tilde{h}$, we obtain

$$\begin{aligned}
&\mathbb{E}[\|\tilde{h}_{k+1} - \nabla f(x^*)\|^2] \\
&= p\mathbb{E}[\|\tilde{h}_k + \frac{1}{d+2}(\hat{\nabla} f(x_k,u) - uu^\top \tilde{h}_k) - \nabla f(x^*)\|^2] + (1-p)\|\tilde{h}_k - \nabla f(x^*)\|^2 \\
&\overset{(3)}{=} p\underbrace{\mathbb{E}\Big[\Big\|\tilde{h}_k + \frac{1}{d+2}(uu^\top \nabla f(x_k) + \frac{Lv}{2} s(x_k,u)\|u\|^2 u - uu^\top \tilde{h}_k) - \nabla f(x^*)\Big\|^2\Big]}_{A_1} + (1-p)\|\tilde{h}_k - \nabla f(x^*)\|^2.
\end{aligned}$$

For A_1, we have

$$
\begin{aligned}
A_1 &= \|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{2}{d+2}\mathbb{E}\Big[\Big\langle \tilde{h}_k - \nabla f(x^*), uu^\top(\nabla f(x_k) - \tilde{h}_k) + \frac{Lv}{2}s(x_k,u)\|u\|^2 u\Big\rangle\Big] + \\
&\quad \underbrace{\frac{1}{(d+2)^2}\mathbb{E}\Big[\Big\|uu^\top(\nabla f(x_k) - \tilde{h}_k) + \frac{Lv}{2}s(x_k,u)\|u\|^2 u\Big\|^2\Big]}_{A_2} \\
&= \|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{2}{d+2}\Big\langle \tilde{h}_k - \nabla f(x^*), \nabla f(x_k) - \nabla f(x^*) + \nabla f(x^*) - \tilde{h}_k\Big\rangle + \\
&\quad \frac{Lv}{d+2}\mathbb{E}\Big[\Big\langle \tilde{h}_k - \nabla f(x^*), s(x_k,u)\|u\|^2 u\Big\rangle\Big] + A_2 \\
&= (1 - \frac{2}{d+2})\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{2}{d+2}\Big\langle \tilde{h}_k - \nabla f(x^*), \nabla f(x_k) - \nabla f(x^*)\Big\rangle + \\
&\quad \frac{Lv}{d+2}\mathbb{E}\Big[\Big\langle \tilde{h}_k - \nabla f(x^*), s(x_k,u)\|u\|^2 u\Big\rangle\Big] + A_2 \\
&\overset{(25)}{\leq} (1 - \frac{2}{d+2})\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{2}{d+2}\Big\langle \tilde{h}_k - \nabla f(x^*), \nabla f(x_k) - \nabla f(x^*)\Big\rangle + \\
&\quad \frac{1}{6(d+2)}\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{3L^2v^2}{2(d+2)}\mathbb{E}[\|u\|^6] + A_2 \\
&= (1 - \frac{11}{6(d+2)})\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{2}{d+2}\Big\langle \tilde{h}_k - \nabla f(x^*), \nabla f(x_k) - \nabla f(x^*)\Big\rangle + \frac{3L^2v^2}{2(d+2)}\mathbb{E}[\|u\|^6] + A_2.
\end{aligned}
$$

For A_2, we have

$$
\begin{aligned}
A_2 &= \frac{1}{(d+2)^2}\mathbb{E}\Big[\Big\|uu^\top(\nabla f(x_k) - \tilde{h}_k) + \frac{Lv}{2}s(x_k,u)\|u\|^2 u\Big\|^2\Big] \\
&\overset{(25)}{\leq} \frac{1}{(d+2)^2}\mathbb{E}[\frac{7}{6}\|uu^\top(\nabla f(x_k) - \tilde{h}_k)\|^2 + \frac{7L^2v^2}{4}s^2(x_k,u)\|u\|^6] \\
&\overset{(28)}{\leq} \frac{7}{6(d+2)}\|\nabla f(x_k) - \tilde{h}_k\|^2 + \frac{7L^2v^2}{4(d+2)^2}\mathbb{E}[\|u\|^6] \\
&= \frac{7}{6(d+2)}\Big[\|\nabla f(x_k) - \nabla f(x^*)\|^2 + 2\Big\langle \nabla f(x^*) - \tilde{h}_k, \nabla f(x_k) - \nabla f(x^*)\Big\rangle + \|\tilde{h}_k - \nabla f(x^*)\|^2\Big] + \\
&\quad \frac{7L^2v^2}{4(d+2)^2}\mathbb{E}[\|u\|^6].
\end{aligned}
$$

Plugging A_2 into A_1, we obtain

$$
\begin{aligned}
A_1 &\leq (1 - \frac{2}{3(d+2)})\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{1}{3(d+2)}\Big\langle \nabla f(x^*) - \tilde{h}_k, \nabla f(x_k) - \nabla f(x^*)\Big\rangle + \\
&\quad \frac{7}{6(d+2)}\|\nabla f(x_k) - \nabla f(x^*)\|^2 + \frac{L^2v^2}{2(d+2)}\mathbb{E}[\|u\|^6](3 + \frac{7}{2(d+2)}) \\
&\overset{(25)}{\leq} 1 - \frac{1}{2(d+2)}\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{4}{3(d+2)}\|\nabla f(x_k) - \nabla f(x^*)\|^2 + \frac{L^2v^2}{2(d+2)}\mathbb{E}[\|u\|^6](3 + \frac{7}{2(d+2)}) \\
&\overset{(26),(29)}{\leq} (1 - \frac{1}{2(d+2)})\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{8L}{3(d+2)}(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^*\rangle) + \frac{5(d+6)^3L^2v^2}{2(d+2)},
\end{aligned}
$$

where we also use the fact that $\frac{7}{2(d+2)} < 2$ for $d \geq 1$ in the last inequality.

Hence, plugging A_1 into $\mathbb{E}[\|\tilde{h}_{k+1} - \nabla f(x^*)\|^2]$, we obtain that

$$\begin{aligned}
&\mathbb{E}[\|\tilde{h}_{k+1} - \nabla f(x^*)\|^2] \\
&\leq p\Big[(1 - \frac{1}{2(d+2)})\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{8L}{3(d+2)}(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle) + \frac{5(d+6)^3 L^2 v^2}{2(d+2)}\Big] + \\
&\quad (1-p)\|\tilde{h}_k - \nabla f(x^*)\|^2 \\
&= (1 - \frac{p}{2(d+2)})\|\tilde{h}_k - \nabla f(x^*)\|^2 + \frac{8Lp}{3(d+2)}(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle) + \frac{5(d+6)^3 L^2 v^2 p}{2(d+2)},
\end{aligned}$$

which concludes the result. □

Proof of Lemma 4.3. According to the update of w, we have the following

$$\begin{aligned}
&\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(w_{k+1}) - \nabla f_i(x^*)\|^2] \\
&= \frac{1-p}{n}\sum_{i=1}^n \|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + \frac{p}{n}\sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f(x^*)\|^2 \\
&\overset{(26)}{\leq} \frac{1-p}{n}\sum_{i=1}^n \|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + \frac{2Lp}{n}\sum_{i=1}^n (f_i(x_k) - f_i(x^*) - \langle \nabla f_i(x^*), x_k - x^* \rangle) \\
&= \frac{1-p}{n}\sum_{i=1}^n \|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + 2Lp(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle).
\end{aligned}$$

□

Proof of Lema 4.4. Taking the expectation w.r.t. u, u_k, and i, we have

$$\begin{aligned}
&\mathbb{E}[\|g_k - \nabla f(x^*)\|^2] \\
&\overset{(12),(13)}{=} \mathbb{E}\Big[\|\hat{\nabla} f_i(x_k, u_k) - \hat{\nabla} f_i(w_k, u_k) + \tilde{h}_k + \hat{\nabla} f(w_k, u) - uu^\top \tilde{h}_k - \nabla f(x^*)\|^2\Big] \\
&= \|\tilde{h}_k - \nabla f(x^*)\|^2 + 2\mathbb{E}\Big[\Big\langle \tilde{h}_k - \nabla f(x^*), \hat{\nabla} f_i(x_k, u_k) - \hat{\nabla} f_i(w_k, u_k) + \hat{\nabla} f(w_k, u) - uu^\top \tilde{h}_k \Big\rangle\Big] + \\
&\quad \underbrace{\mathbb{E}\Big[\|\hat{\nabla} f_i(x_k, u_k) - \hat{\nabla} f_i(w_k, u_k) + \hat{\nabla} f(w_k, u) - uu^\top \tilde{h}_k\|^2\Big]}_{B_1} \\
&\overset{(15)}{=} \|\tilde{h}_k - \nabla f(x^*)\|^2 + 2\Big\langle \tilde{h}_k - \nabla f(x^*), \nabla f(x_k) - \nabla f(x^*) + \nabla f(x^*) - \tilde{h}_k \Big\rangle + Lv\Big\langle \tilde{h}_k - \nabla f(x^*), \tau_{i,k} \Big\rangle + B_1 \\
&= -\|\tilde{h}_k - \nabla f(x^*)\|^2 + 2\Big\langle \tilde{h}_k - \nabla f(x^*), \nabla f(x_k) - \nabla f(x^*) \Big\rangle + Lv\Big\langle \tilde{h}_k - \nabla f(x^*), \tau_{i,k} \Big\rangle + B_1.
\end{aligned}$$

For B_1, with Eq.(2) and Eq.(3), we have

$$\begin{aligned}
B_1 &= \mathbb{E}[\|\hat{\nabla} f_i(x_k, u_k) - \hat{\nabla} f_i(w_k, u_k) + \hat{\nabla} f(w_k, u) - uu^\top \tilde{h}_k\|^2] \\
&\overset{(2),(3)}{=} \mathbb{E}\Big[\Big\|u_k u_k^\top \nabla f_i(x_k) + \frac{Lv}{2} s_i(x_k, u_k)\|u_k\|^2 u_k - u_k u_k^\top \nabla f_i(w_k) - \frac{Lv}{2} s_i(w_k, u_k)\|u_k\|^2 u_k + uu^\top \nabla f(w_k) + \\
&\quad \frac{Lv}{2} s(w_k, u)\|u\|^2 u - uu^\top \tilde{h}_k\Big\|^2\Big] \\
&\overset{(25)}{\leq} 2\underbrace{\mathbb{E}\Big[\Big\|u_k u_k^\top \nabla f_i(x_k) - u_k u_k^\top \nabla f_i(w_k) + uu^\top \nabla f(w_k) - uu^\top \tilde{h}_k\Big\|^2\Big]}_{B_2} + \\
&\quad 2\underbrace{\mathbb{E}\Big[\Big\|\frac{Lv}{2}(s_i(x_k, u_k) - s_i(w_k, u_k))\|u_k\|^2 u_k + \frac{Lv}{2} s(w_k, u)\|u\|^2 u\Big\|^2\Big]}_{B_3}.
\end{aligned}$$

For B_2, we have that

$$\begin{aligned}
B_2 &\overset{(28)}{=} (d+2)\mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x^*) + \nabla f_i(x^*) - \nabla f_i(w_k)\|^2] + (d+2)\|\nabla f(w_k) - \tilde{h}_k\|^2 + \\
&\quad 2\left\langle \nabla f(x_k) - \nabla f(w_k), \nabla f(w_k) - \tilde{h}_k \right\rangle \\
&\overset{(25)}{\leq} \frac{2(d+2)}{n}\sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f_i(x^*)\|^2 + \frac{2(d+2)}{n}\sum_{i=1}^n \|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + (d+2)\|\nabla f(w_k) - \tilde{h}_k\|^2 + \\
&\quad 2\left\langle \nabla f(x_k) - \nabla f(w_k), \nabla f(w_k) - \tilde{h}_k \right\rangle \\
&= \frac{2(d+2)}{n}\sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f_i(x^*)\|^2 + \frac{2(d+2)}{n}\sum_{i=1}^n \|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + (d+2)\|\nabla f(w_k) - \nabla f(x^*)\|^2 + \\
&\quad (d+2)\|\nabla f(x^*) - \tilde{h}_k\|^2 + 2(d+2)\left\langle \nabla f(w_k) - \nabla f(x^*), \nabla f(x^*) - \tilde{h}_k \right\rangle + \\
&\quad 2\left\langle \nabla f(x_k) - \nabla f(w_k), \nabla f(w_k) - \tilde{h}_k \right\rangle.
\end{aligned}$$

Since

$$\begin{aligned}
&2\left\langle \nabla f(x_k) - \nabla f(w_k), \nabla f(w_k) - \tilde{h}_k \right\rangle \\
&= 2\left\langle \nabla f(x_k) - \nabla f(x^*), \nabla f(w_k) - \nabla f(x^*) \right\rangle + 2\left\langle \nabla f(x_k) - \nabla f(x^*), \nabla f(x^*) - \tilde{h}_k \right\rangle - 2\|f(w_k) - f(x^*)\|^2 + \\
&\quad 2\left\langle \nabla f(x^*) - \nabla f(w_k), \nabla f(x^*) - \tilde{h}_k \right\rangle,
\end{aligned}$$

plugging it into B_2, we obtain that

$$\begin{aligned}
B_2 \leq\ & \frac{2(d+2)}{n}\sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f_i(x^*)\|^2 + \frac{2(d+2)}{n}\sum_{i=1}^n \|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + d\|\nabla f(w_k) - \nabla f(x^*)\|^2 + \\
& (d+2)\|\nabla f(x^*) - \tilde{h}_k\|^2 + 2(d+1)\left\langle \nabla f(w_k) - \nabla f(x^*), \nabla f(x^*) - \tilde{h}_k \right\rangle + \\
& 2\left\langle \nabla f(x_k) - \nabla f(x^*), \nabla f(w_k) - \nabla f(x^*) \right\rangle + 2\left\langle \nabla f(x_k) - \nabla f(x^*), \nabla f(x^*) - \tilde{h}_k \right\rangle.
\end{aligned}$$

With Lemma A.1 and A.2, we have

$$\begin{aligned}
2(d+1)\left\langle \nabla f(w_k) - \nabla f(x^*), \nabla f(x^*) - \tilde{h}_k \right\rangle &\overset{(25)}{\leq} (d+1)\|\nabla f(w_k) - \nabla f(x^*)\|^2 + (d+1)\|\nabla f(x^*) - \tilde{h}_k\|^2, \\
2\left\langle \nabla f(x_k) - \nabla f(x^*), \nabla f(w_k) - \nabla f(x^*) \right\rangle &\overset{(25)}{\leq} \|\nabla f(x_k) - \nabla f(x^*)\|^2 + \|\nabla f(w_k) - \nabla f(x^*)\|^2, \\
\|\nabla f(w_k) - \nabla f(x^*)\|^2 &\overset{(24)}{\leq} \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(w_k) - \nabla f(x^*)\|^2.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
B_2 \leq\ & \frac{2(d+2)}{n}\sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f_i(x^*)\|^2 + \frac{2(2d+3)}{n}\sum_{i=1}^n \|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + \\
& (2d+3)\|\tilde{h}_k - \nabla f(x^*)\|^2 + \|\nabla f(x_k) - \nabla f(x^*)\|^2 + 2\left\langle \nabla f(x_k) - \nabla f(x^*), \nabla f(x^*) - \tilde{h}_k \right\rangle \\
\overset{(24)}{\leq}\ & \frac{2d+5}{n}\sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f_i(x^*)\|^2 + \frac{2(2d+3)}{n}\sum_{i=1}^n \|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + \\
& (2d+3)\|\tilde{h}_k - \nabla f(x^*)\|^2 + 2\left\langle \nabla f(x_k) - \nabla f(x^*), \nabla f(x^*) - \tilde{h}_k \right\rangle.
\end{aligned}$$

For B_3, we have

$$\begin{aligned}B_3 &= \frac{L^2v^2}{4}\mathbb{E}\Big[\Big\|(s_i(x_k,u_k)-s_i(w_k,u_k))\|u_k\|^2u_k + s(w_k,u)\|u\|^2u\Big\|^2\Big]\\ &\overset{(25)}{\leq} \frac{L^2v^2}{2}(\mathbb{E}[\|u_k\|^6]+\mathbb{E}[\|u\|^6])\\ &\overset{(29)}{\leq} L^2v^2(d+6)^3.\end{aligned}$$

Plugging B_2 and B_3 into B_1, we obtain

$$\begin{aligned}B_1 \leq\ & \frac{2(2d+5)}{n}\sum_{i=1}^n\|\nabla f_i(x_k)-\nabla f_i(x^*)\|^2 + \frac{4(2d+3)}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2 +\\ & 2(2d+3)\|\tilde{h}_k-\nabla f(x^*)\|^2 + 4\Big\langle\nabla f(x_k)-\nabla f(x^*),\nabla f(x^*)-\tilde{h}_k\Big\rangle + 2L^2v^2(d+6)^3.\end{aligned}$$

Plugging B_1 into the original inequality, we have

$$\begin{aligned}&\mathbb{E}[\|g_k-\nabla f(x^*)\|^2]\\ \leq\ & (4d+5)\|\tilde{h}_k-\nabla f(x^*)\|^2 + 2\Big\langle\nabla f(x_k)-\nabla f(x^*),\nabla f(x^*)-\tilde{h}_k\Big\rangle + Lv\Big\langle\tilde{h}_k-\nabla f(x^*),\tau_{i,k}\Big\rangle +\\ & \frac{2(2d+5)}{n}\sum_{i=1}^n\|\nabla f_i(x_k)-\nabla f_i(x^*)\|^2 + \frac{4(2d+3)}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2 + 2L^2v^2(d+6)^3\\ \overset{(25),(24)}{\leq}\ & 2(2d+3)\|\tilde{h}_k-\nabla f(x^*)\|^2 + \frac{L^2v^2}{2}\|\tau_{i,k}\|^2 + \frac{4(d+3)}{n}\sum_{i=1}^n\|\nabla f_i(x_k)-\nabla f_i(x^*)\|^2 +\\ & \frac{4(2d+3)}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2 + 2L^2v^2(d+6)^3\\ \overset{(26)}{\leq}\ & 2(2d+3)\|\tilde{h}_k-\nabla f(x^*)\|^2 + \frac{L^2v^2}{2}\|\tau_{i,k}\|^2 + 4(d+3)L(f(x_k)-f(x^*)-\langle\nabla f(x^*),x_k-x^*\rangle) +\\ & \frac{4(2d+3)}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2 + 2L^2v^2(d+6)^3,\end{aligned}$$

where we can derive the result with the bound of $\mathbb{E}[\|\tau_{i,k}\|^2]$ in Lemma 3.5. □

Proof of Corollary 4.5. Let $C = f(x_k)-f(x^*)-\langle\nabla f(x^*),x_k-x^*\rangle$. With Lemma 4.1, 4.2, 4.3, and $p=\frac{1}{n}$, we have

$$\begin{aligned}&\mathbb{E}[\Psi(x_{k+1})]\\ \overset{(17),(18),(19)}{\leq}\ & (1-\frac{\mu}{2}\eta)\|x_k-x^*\|^2 - 2\eta C + \frac{2(d+3)^3L^2v^2\eta}{\mu} + \eta^2\mathbb{E}[\|g_k-\nabla f(x^*)\|^2] + \alpha\Big[\frac{5(d+6)^3L^2v^2}{2n(d+2)} +\\ & (1-\frac{1}{2n(d+2)})\|\tilde{h}_k-\nabla f(x^*)\|^2 + \frac{8L}{3n(d+2)}C\Big] + \beta\Big[(1-\frac{1}{n})\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2 + \frac{2L}{n}C\Big].\end{aligned}$$

Plugging Lemma 4.4 into the above inequality, we achieve

$$\begin{aligned}&\mathbb{E}[\psi(x_{k+1})]\\ \overset{(20)}{\leq}\ & (1-\frac{\mu}{2}\eta)\|x_k-x^*\|^2 - 2\eta C + \frac{2(d+3)^3L^2v^2\eta}{\mu} + \eta^2\Big[2(2d+3)\|\tilde{h}_k-\nabla f(x^*)\|^2 + 4L^2v^2(d+6)^3 +\\ & 4(d+3)LC + \frac{4(2d+3)}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2\Big] + \alpha\Big[(1-\frac{1}{2n(d+2)})\|\tilde{h}_k-\nabla f(x^*)\|^2 + \frac{8L}{3n(d+2)}C +\\ & \frac{5(d+6)^3L^2v^2}{2n(d+2)}\Big] + \beta\Big[(1-\frac{1}{n})\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(w_k)-\nabla f_i(x^*)\|^2 + \frac{2L}{n}C\Big].\end{aligned}$$

Rearranging the preceding inequality, we obtain

$$\begin{aligned}
&\mathbb{E}[\psi(x_{k+1})] \\
&\leq (1-\frac{\mu}{2}\eta)\|x_k - x^*\|^2 + \alpha(1-\frac{1}{2n(d+2)}+\frac{2(2d+3)\eta^2}{\alpha})\|\tilde{h}_k - \nabla f(x^*)\|^2 + \\
&\quad \beta(1-\frac{1}{n}+\frac{4(2d+3)\eta^2}{\beta})\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 - (2\eta - 4(d+3)L\eta^2 - \frac{8L}{3n(d+2)}\alpha - \frac{2L}{n}\beta)C + \\
&\quad \frac{2(d+3)^3L^2v^2\eta}{\mu} + 4L^2v^2(d+6)^3\eta^2 + \frac{5(d+6)^3L^2v^2}{2n(d+2)}\alpha,
\end{aligned}$$

which concludes the result. □

Proof of Theorem 4.6. Plugging the values of η, α, β in to Corollary 4.5, we obtain that

$$\begin{aligned}
&\mathbb{E}[\Psi(x_{k+1})] \\
&\leq (1-\frac{1}{(80d+126)\kappa})\|x_k - x^*\|^2 + (1-\frac{1}{4n(d+2)})\alpha\|\tilde{h}_k - \nabla f(x^*)\|^2 + \\
&\quad \beta(1-\frac{1}{2n})\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 - \lambda C + \frac{2(d+3)^3v^2\kappa}{40d+63} + \frac{8(5d+8)(d+6)^3v^2}{(40d+63)^2},
\end{aligned}$$

where $\lambda = \frac{2}{(40d+63)L} - \frac{\frac{188}{3}d+100}{(40d+63)^2L}$, i.e., $\lambda > 0$. Then, we can eliminate the term C in the above inequality and obtain,

$$\begin{aligned}
&\mathbb{E}[\Psi(x_{k+1})] \\
&\leq (1-\frac{1}{(80d+126)\kappa})\|x_k - x^*\|^2 + (1-\frac{1}{4n(d+2)})\alpha\|\tilde{h}_k - \nabla f(x^*)\|^2 + \\
&\quad \beta(1-\frac{1}{2n})\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w_k) - \nabla f_i(x^*)\|^2 + \delta, \\
&\leq \max\{1-\frac{1}{\kappa(80d+126)}, 1-\frac{1}{4n(d+2)}\}\Psi(x_k) + \delta,
\end{aligned}$$

where $\delta = \frac{2(d+3)^3v^2\kappa}{40d+63} + \frac{8(5d+8)(d+6)^3v^2}{(40d+63)^2}$. □

Proof of Corollary 4.7. Since $\theta = \frac{1}{\kappa(80d+126)+4n(d+2)} \leq \min\{\frac{1}{\kappa(80d+126)}, \frac{1}{4n(d+2)}\}$, based on Theorem 4.6, we have that

$$\mathbb{E}[\Psi(x_{k+1})] \leq (1-\theta)\Psi(x_k) + \delta.$$

Telescoping it from $k = 0$ for $K-1$, we obtain that

$$\begin{aligned}
\mathbb{E}[\Psi(x_K)] &\leq (1-\theta)^K\Psi(x_0) + \sum_{k=0}^{k-1}(1-\theta)^k\delta \\
&= (1-\theta)^K\Psi(x_0) + \frac{1-(1-\theta)^K}{\theta}\delta \\
&\leq (1-\theta)^K\Psi(x_0) + \frac{\delta}{\theta}.
\end{aligned}$$

Furthermore, since $\delta \leq (\kappa+1)d^2v^2$, we conclude the result. □

C. Hyperparameter Tuning

We employ the grid search method to determine the optimal hyperparameters for ZPDVR and other baseline methods. The total number of SZO function evaluations, and the batch size of samples and directions are fixed and equal for all methods. Apart from the learning rate η, the probability ρ in ZPDVR and the number of the inner loop m in ZPSVRG are also tuned parameters. The hyperparameter search fields for each method in the four data sets are as follows:

- a9a. The learning rate η is ranged from $\{1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}, 5\times10^{-1}, 1, 2, 5, 10\}$ for PGD, $\{1\times10^{-1}, 5\times10^{-1}, 1, 5, 10\}$ for ZPDVR, $\{1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}, 5\times10^{-1}, 1\}$ for ZPSVRG, $\{1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}, 5\times10^{-1}, 1, 5, 10\}$ for SEGA. The probability ρ is ranged from $\{1\times10^{-2}, 2\times10^{-2}, 3\times10^{-2}, 4\times10^{-2}, 5\times10^{-2}\}$ and the number of inner loop m is selected from $\{10, 50, 100, 500, 1000, 5000\}$;
- w8a. The learning rate η is ranged from $\{1\times10^{-1}, 5\times10^{-1}, 1, 2, 5, 10\}$ for PGD, $\{1\times10^{-3}, 7\times10^{-3}, 1\times10^{-2}, 7\times10^{-2}, 1\times10^{-1}\}$ for ZPDVR, $\{1\times10^{-3}, 5\times10^{-3}, 7\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 7\times10^{-2}, 1\times10^{-1}\}$ for ZPSVRG, $\{1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}, 5\times10^{-1}, 1, 5, 10\}$ for SEGA. The probability ρ is ranged from $\{5\times10^{-4}, 1\times10^{-3}, 7\times10^{-3}\}$ and the number of inner loop m is selected from $\{10, 50, 100, 500, 1000\}$;
- covtype. The Learning rate η is ranged from $\{1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}, 5\times10^{-1}, 1, 2, 3, 4, 5, 10\}$ for PGD, $\{1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}, 5\times10^{-1}, 1\}$ for ZPDVR, $\{1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 2\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}\}$ for ZPSVRG, $\{1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}, 5\times10^{-1}, 1, 5, 10\}$ for SEGA. The number of inner loop m is selected from $\{10, 50, 100, 500, 1000\}$. In this dataset, we periodically update w and $\tilde{h}$ for ZPDVR (update these parameters after completing a full pass over the data set). This procedure omits the randomness caused by ρ and is equivalent to set $\rho = \frac{B}{n}$, where B is the batch size of samples;
- gisette. The learning rate η is range from $\{1\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}\}$ for PGD, $\{1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}\}$ for ZPDVR, $\{1\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}\}$ for ZSVRG, $\{1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}, 5\times10^{-2}, 1\times10^{-1}, 5\times10^{-1}, 1, 5, 10\}$ for SEGA. The number of inner loop m is selected from $\{10, 20, 50, 100\}$. Similarly, we also periodically update w and $\tilde{h}$ for ZPDVR.