

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 PARAMETER-FREE VARIANCE REDUCED ZEROTH- ORDER OPTIMIZATION FOR NONCONVEX PROBLEMS

Anonymous authors

Paper under double-blind review

## ABSTRACT

Zeroth-order optimization has become a vital tool for solving black-box learning problems where explicit gradients are unavailable. However, standard zeroth-order methods typically require careful tuning of algorithmic parameters such as the smoothing parameter and step size, which limits their practicality. In this paper, we propose PF-VRZO(Parameter free variance reduced zeroth-order methods), a novel parameter-free variance-reduced zeroth-order optimization framework for nonconvex finite-sum problems. Our method only requires minimal input information—problem dimension  $d$  and sample size  $n$ —and adaptively adjusts the smoothing and step size parameters during the optimization process. We develop two algorithmic variants based on coordinate-wise and random-direction gradient estimators, respectively. We establish non-asymptotic convergence guarantees showing that PF-VRZO achieves function query complexity of  $\tilde{\mathcal{O}}(d\sqrt{n}\epsilon^{-2})$  for finding stationary points. Additionally, we conduct experiments on nonconvex phase retrieval and distributional robust optimization to validate the effectiveness of our method. To the best of our knowledge, PF-VRZO is the first parameter-free zeroth-order algorithm that incorporates variance reduction techniques tailored specifically for nonconvex optimization problems.

## 1 INTRODUCTION

In the paper, we consider solving the following stochastic nonconvex finite-sum optimization problems.  $f : \mathbb{R}^d \rightarrow \mathbb{R}$

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) \quad (1)$$

where  $f(x)$  and each  $f_i(x)$  are both smooth and possibly nonconvex functions, which captures the standard empirical risk minimization problems in machine learning.

In many important applications, computing explicit gradients is either computationally expensive or infeasible, and only function evaluations are available. Such applications include black-box adversarial attacks on deep neural networks (DNNs) (Papernot et al., 2017; Chen et al., 2017), reinforcement learning (Malik et al., 2018; Kumar et al., 2020), and fine-tuning large-scale models (Malladi et al., 2023). Zeroth-order optimization has thus emerged as a fundamental research direction (Ghadimi & Lan, 2013; Liu et al., 2018b;a; Ji et al., 2019; Lian et al., 2016; Gu et al., 2018), serving as a prototype framework for a wide range of these gradient-free learning tasks. However, a common drawback of standard zeroth-order methods is the introduction of an additional smoothing parameter  $\mu$ . As illustrated in Figure 1, improper tuning of this parameter in practice can lead to suboptimal performance, or even cause the algorithm to diverge.

On the other hand, recent years have seen a growing body of work on parameter-free algorithms (Ivgi et al., 2023; Kreisler et al., 2024; Orabona & Tommasi, 2017; Chen et al., 2022; Defazio & Mishchenko, 2023), particularly in the first-order setting. Several studies have demonstrated that such methods can achieve convergence rates comparable to those of parameter-dependent algorithms, even under nonconvex conditions. We define a parameter-free method as one that does not require prior knowledge of problem-specific parameters such as the smoothness constant  $L$ , the target accuracy  $\epsilon$ , or the total number of iterations  $T$ . This is particularly important in practical applications, where such information is typically unavailable—for example, it is often unclear how many iterations are needed,

054 or how small the gradient or objective value should be for the model to be considered good enough.  
 055 Our expectation for a parameter-free algorithm is that it can be executed with only minimal and  
 056 readily available inputs, such as the sample size  $n$  and the problem dimension  $d$ , and run continuously  
 057 until the model reaches a desirable state—such as sufficiently high test accuracy or low generalization  
 058 error.

059 Although recent works have achieved satisfactory theoretical progress for first-order algorithms,  
 060 research on zeroth-order counterparts remains quite limited. It was not until recently that [Ren &](#)  
 061 [Luo \(2025\)](#) proposed the first parameter-free zeroth-order algorithm. Unfortunately, the theoretical  
 062 guarantee of this method holds only under the assumption that the objective function  $f(x)$  is convex  
 063 and defined over a bounded domain. As acknowledged by the authors, extending this result to the  
 064 nonconvex setting is nontrivial.

065

066

067 Q1. When zeroth-order optimization meets adaptive methods, how does the error introduced  
 068 by inexact gradient estimation accumulate throughout the optimization process, and is such  
 069 error controllable? Can we design an adaptive algorithm that keeps this error within an  
 070 acceptable range?

071

072

073 A1: Based on our results, after  $T$  iterations, the accumulated error is approximately  $\mathcal{O}\left[\frac{1}{T}\left(\sum_{t=0}^{T-1} \mu_t^2 + \right.\right.$   
 074  $\left.\sum_{t=0}^{T-1} \mu_t\right) + \frac{1}{\sqrt{T}}\left(n^{\frac{5}{4}} \sum_{t=0}^{T-1} \mu_t^2 + n^{\frac{1}{2}} \sqrt{\sum_{t=0}^{T-1} 2\mu_t^2}\right)\right]$ . To ensure convergence, it is crucial that con-  
 075 dition  $\sum_{t=0}^{T-1} \mu_t \leq \mathcal{O}(\sqrt{T})$ ,  $\sum_{t=0}^{T-1} \mu_t^2 \leq \mathcal{O}(1)$  holds; otherwise, the algorithm may diverge. This  
 076 observation reveals that the error grows with  $T$ . A natural idea, therefore, is to let the smoothing  
 077 parameter  $\mu$  depend on  $T$ , which would directly guarantee  $\sum_{t=0}^{T-1} \mu_t \leq \mathcal{O}(\sqrt{T})$ ,  $\sum_{t=0}^{T-1} \mu_t^2 \leq \mathcal{O}(1)$ .  
 078 However, this approach conflicts with our goal of designing a parameter-free algorithm, since the  
 079 required number of iterations  $T$  is unknown in advance. To overcome this difficulty while preserving  
 080 the parameter-free property, we introduce a smart adaptive parameter  $\mu_t = \frac{1}{(t+1)\sqrt{nd}}$ , which evolves  
 081 automatically during the optimization process to enforce  $\sum_{t=0}^{T-1} \mu_t \leq \mathcal{O}(\sqrt{T})$ ,  $\sum_{t=0}^{T-1} \mu_t^2 \leq \mathcal{O}(1)$ ,  
 082 without the need for any manually tuned parameters.

083

084

085 Q2: Would the smoothing parameters  $\mu$  that vary with  $t$ , as discussed above, conflict with  
 086 the proof techniques of variance reduction methods? Taking the Spider estimator  $v^t =$   
 087  $\nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{t-1}) + v^{t-1}$  as an example, can we directly replace the terms  $\nabla f_{i_t}(x^t)$   
 088 and  $\nabla f_{i_t}(x^{t-1})$  in the Spider estimator with the zeroth-order estimators  $\bar{\nabla}_{\mu_1} f_{i_t}(x_t)$  and  
 089  $\bar{\nabla}_{\mu_2} f_{i_t}(x_{t-1})$ ? Moreover, can these two zeroth-order estimators be computed directly using  
 090 the adaptive smoothing parameter  $\mu_1 = \mu_2 = \mu_t = \frac{1}{(t+1)\sqrt{nd}}$ ?

091

092

093 A2: We found that directly using the smoothing parameters mentioned above in gradient es-  
 094 timation within variance-reduced methods does not work. This is because the convergence  
 095 proofs for variance reduction often rely on the recursive relation  $\mathbb{E}\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 \leq$   
 096  $\mathbb{E}\|v_{t-1} - \bar{\nabla}_{\mu_{t-1}} f(x_{t-1})\|^2 + (\text{additional terms})$  holding exactly. To ensure this recursive relation,  
 097  $\mathbb{E}[\bar{\nabla}_{\mu_1} f_{i_t}(x_t) - \bar{\nabla}_{\mu_2} f_{i_t}(x_{t-1})] = \bar{\nabla}_{\mu_t} f(x_t) - \bar{\nabla}_{\mu_{t-1}} f(x_{t-1})$  is required. Therefore, simply setting  
 098  $\mu_1 = \mu_2 = \frac{1}{(t+1)\sqrt{nd}}$  does not suffice; a slight modification is needed, where we set  $\mu_1 = \frac{1}{(t+1)\sqrt{nd}}$   
 099 and  $\mu_2 = \frac{1}{(t)\sqrt{nd}}$ .

100

101

102 By addressing the aforementioned challenges, this paper introduces the Parameter-Free Variance-  
 103 Reduced Zeroth-Order (PF-VRZO) method, a novel approach that combines the strengths of adaptive  
 104 algorithms with variance reduction techniques. Our method eliminates the need for manual parameter  
 105 tuning by adaptively adjusting the smoothing parameter and step size during the optimization process.  
 106 Specifically, we propose two variants of PF-VRZO: one based on coordinate-wise gradient estimators  
 107 and another leveraging random-direction estimators.

108  
109 Table 1: Convergence property comparison of the PF-VRZO algorithms for finding an  $\epsilon$ -stationary  
110 point. **C**, **NC**, **S**, and **NS** denote convex, nonconvex, smooth, and non-smooth settings, respectively.  
111 **VR** indicates whether the method is compatible with variance reduction techniques.  $\sigma$  denotes an  
112 upper bound on the variance of stochastic gradients, and  $D_x$  represents the diameter of the domain.  
113 The term “complexity” refers to function query complexity. Here,  $\eta_t$  denotes the step size,  $\mu_t$  the  
114 smoothing parameter,  $c$  a generic constant, and  $T$  the total number of iteration rounds.  $g_t$  refers to the  
115 zeroth-order gradient estimator, while  $v_t$  denotes the SPIDER estimator.\*denotes deterministic case

Method	Problem	VR?	Param-free?	Complexity	$\eta_t$	$\mu_t$
POEM (Ren & Luo)	C-NS	✗	✓	$\tilde{\mathcal{O}}(de^{-2}D_x)$	$\frac{\max_t \{\ x_t - x_0\ \}}{\sum_{s=0}^t \ g_t\ ^2}$	$\frac{d \max_t \{\ x_t - x_0\ \}}{t+1}$
JAGUAR (Veprikov et al.)	NC-S	✓	✗	$^*\mathcal{O}(d\epsilon^{-2})$	$\frac{1}{dL}$	$\mathcal{O}(\frac{\epsilon}{\sqrt{dL}})$
ZO-SGD (Ghadimi & Lan)	NC-S	✗	✗	$\mathcal{O}(\sigma^2\epsilon^{-4})$	$o(\frac{1}{\sqrt{d}} \min\{\frac{1}{L\sqrt{d}}, \frac{c}{\sigma\sqrt{d}}\})$	$o(\frac{c}{d\sqrt{T}})$
ZO-SPIDER-rand (Fang et al.)	NC-S	✓	✗	$\mathcal{O}(d\sqrt{n}\epsilon^{-2})$	$\min\{\frac{ce}{L\ v_t\ }, \frac{c}{L}\}$	$o(\frac{\epsilon}{L\sqrt{d}})$
ZO-SPIDER-coord (Ji et al.)	NC-S	✓	✗	$\mathcal{O}(d\sqrt{n}\epsilon^{-2})$	$\frac{1}{\sqrt{n}L}$	$\frac{1}{\sqrt{TdL}}$
<b>PF-VRZO-coord</b> (Theorem 1)	NC-S	✓	✓	$\mathcal{O}(d\sqrt{n}\epsilon^{-2})$	$\frac{1}{n^{1/4}\sqrt{(n^{1/2} + \sum_{s=0}^t \ v_s\ ^2)}}$	$\frac{1}{(t+1)\sqrt{nd}}$
<b>PF-VRZO-rand</b> (Theorem 2)	NC-S	✓	✓	$\mathcal{O}(d\sqrt{n}\epsilon^{-2})$	$\frac{1}{n^{1/4}\sqrt{d(n^{1/2} + \sum_{s=0}^t \ v_s\ ^2)}}$	$\frac{1}{(t+1)d\sqrt{n}}$

124 The key contributions of this work are as follows:

125

- 126 **A Parameter-Free Zeroth-Order Framework:** We propose PF-VRZO, the first parameter-  
127 free zeroth-order optimization method for nonconvex finite-sum problems. It requires only  
128 minimal inputs—sample size  $n$  and dimension  $d$ , without relying on problem-dependent  
129 parameters such as the smoothness constant or iteration count.
- 130 **Variance Reduction with Adaptive Gradient Estimation:** PF-VRZO incorporates vari-  
131 ance reduction into both coordinate-wise and random-direction zeroth-order estimators,  
132 with adaptive adjustment of smoothing parameters and step sizes, eliminating the need for  
133 manual tuning.
- 134 **Theoretical and Empirical Validation:** We provide convergence guarantees showing that  
135 PF-VRZO achieves a function query complexity  $\tilde{\mathcal{O}}(d\sqrt{n}\epsilon^{-2})$ . Experiments on nonconvex  
136 phase retrieval and distributional robust optimization confirm its comparable performance  
137 compared to existing tuned methods.

## 2 PRELIMINARIES

141 **Remark 1.** By “param-free,” we mean that the method does not require any tunable hyperparam-  
142 eters—no manual adjustment is needed. The algorithm only depends on the dataset size  $n$  and the  
143 model dimension  $d$ , both of which are inherent to the problem setup and readily available before  
144 running the optimization.

145 **Notation** Throughout the paper,  $\|\cdot\|$  denotes the Euclidean norm for vectors,  $\tilde{\mathcal{O}}$  hide the logarithmic  
146 factors, and  $\langle \cdot, \cdot \rangle$  denotes the inner product. We denote by  $d$  the dimension of the problem, and by  $n$   
147 the number of functions in the optimization problem. We use  $f_i(x)$  to denote the  $i$ -th sample function  
148 of  $f(x)$ .

149 **Definition 1** (Smoothness). A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $L$ -smooth if there exists  $L > 0$  such that for  
150 all  $x, y \in \mathbb{R}^d$ :

$$151 \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

152 **Assumption 1** (Lipschitz Gradient). Each function  $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $L$ -smooth such that

$$153 \quad \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|.$$

154 **Assumption 2** (Boundedness). Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be bounded from below by a finite constant  $f^*$ , i.e.,

$$155 \quad f(x_0) - f^* \leq \Delta.$$

156 for the initial solution  $x_0$ .

162 **3 PROPOSED PARAMETER FREE VARIANCE REDUCED ZEROOTH-ORDER**  
 163 **METHODS**  
 164

165 PF-VRZO(coord) method integrates variance reduction with zeroth-order gradient estimation in a  
 166 parameter-free manner. This adaptive structure ensures stable updates and effective convergence,  
 167 even in nonconvex settings.

168 To set the stage for our proposed PF-VRZO algorithm, we first review the fundamentals of zeroth-  
 169 order optimization, followed by a summary of the main techniques introduced in this paper.

171 **3.1 ZEROOTH-ORDER GRADIENT ESTIMATORS**  
 172

173 When the gradient of  $f(x)$  is not directly obtainable, it is often estimated via coordinate-wise methods  
 174 or Gaussian smoothing (Duchi et al., 2015; Gasnikov et al., 2023; Kornowski & Shamir, 2024; Lin  
 175 et al., 2022). In what follows, we first describe the coordinate-wise estimator:

$$176 \quad \bar{\nabla}_\mu f(x) := \sum_{\ell=1}^d \frac{1}{\mu} [f(x + \mu \mathbf{e}_\ell) - f(x)] \mathbf{e}_\ell, \quad (\text{Coord estimator})$$

179 where  $\mathbf{e}_\ell$  is a standard basis vector with 1 at its  $\ell^{\text{th}}$  coordinate, and 0s elsewhere. The error of the  
 180 coordinate-wise gradient estimator is upper bounded as follows, and it approaches zero as  $\mu \rightarrow 0$   
 181 (Gao et al., 2018).

182 Besides the coordinate-wise estimator, the random-direction estimator is another widely used zeroth-  
 183 order method, before introduce random-direction estimator, we first introduce smoothing function  
 184  $f_\mu(x) := \mathbb{E}_{\{w \sim U_b\}}[f(x + \mu w)]$ , where  $U_b$  is a uniform distribution over the unit Euclidean ball,  
 185 following Gao et al. (2018), its gradient can be expressed as  $\nabla f_\mu(x) := \mathbb{E}_{\{\rho \sim U_{S_p}\}} \left[ \frac{n}{\mu} f(x + \mu \rho) \rho \right]$ .  
 186 Here  $U_{S_p}$  is a uniform distribution over the unit Euclidean sphere, and  $\rho \in \mathbb{R}^d$  is a random vector  
 187 sampled from unit Euclidean sphere  $U_{S_p}$ . Now we can define zeroth-order random-direction estimator  
 188  $\hat{\nabla} f(x)$  as follows, which is an unbiased estimator of  $\nabla f_\mu(x)$ :

$$189 \quad \hat{\nabla}_\mu f(x) := \frac{d}{\mu} [f(x + \mu \rho) - f(x)] \rho. \quad (\text{Random-direction estimator})$$

190 Random-direction estimator is an unbiased estimate of the gradient of the smoothing function , i.e,  
 191  $\mathbb{E}[\hat{\nabla}_\mu f(x)] = \nabla f_\mu(x)$ .

192 Both of the aforementioned zeroth-order estimators rely on a fixed smoothing parameter  $\mu$ , whose  
 193 improper tuning may lead to substantially degraded performance, ranging from slow convergence to  
 194 divergence (Figure 1). To overcome this limitation, we develop a framework that integrates three key  
 195 components: variance reduction, adaptive stepsize, and adaptive smoothing parameter. The latter two,  
 196 in particular, set our method apart from existing approaches and enable new convergence guarantees.

200 **3.2 VARIANCE REDUCTION TECHNIQUE**  
 201

202 As a celebrated technique in stochastic optimization, variance reduction has been instrumental in the  
 203 development of algorithms with significantly reduced theoretical complexity, SPIDER(Fang et al.,  
 204 2018) is a variance reduction-typed method with optimal complexity guarantee, which uses large  
 205 batch and small batch alternately to estimate stochastic gradients in a recursive way as follows:

$$206 \quad v_t = \nabla f_B(x^t) - \nabla f_B(x^{t-1}) + v^{t-1}, \quad (\text{SPIDER})$$

207 with clipped step size  $\eta_t = \min\{c_1, \frac{c_2 \epsilon}{\|v_t\|}\}$  , where  $c_1, c_2$  are some constants, and  $\nabla f_B(x) =$   
 208  $\frac{1}{|B|} \sum_{\xi \in B} \nabla f(x)$  with a small batch size  $B$ .

210 **3.3 ADAPTIVE STEPSIZE**  
 211

212 The step size  $\gamma_t$  in PF-VRZO is chosen in a parameter-free and adaptive manner. Specifically, it is set  
 213 as:

$$214 \quad \gamma_t = \frac{1}{n^{1/4} c \sqrt{n^{1/2} + \sum_{s=0}^t \|v_s\|^2}},$$

We set  $c = 1$  when using the coordinate-wise estimator and  $c = \sqrt{d}$  when using the random-direction estimator. This design avoids reliance on unknown constants such as the Lipschitz constant or desired accuracy. By incorporating the accumulated gradient norms, the step size automatically decays, which helps balance exploration and convergence.

### 3.4 ADAPTIVE SMOOTHING PARAMETER

In PF-VRZO, the smoothing parameter  $\mu_t$  plays a critical role in estimating gradients via zeroth-order information. Unlike traditional methods that fix  $\mu$  based on prior knowledge of the target accuracy  $\epsilon$  or total iterations  $T$ , PF-VRZO adaptively sets:

$$\mu_t = \frac{1}{(t+1)\sqrt{nd}} (\text{Coord}), \quad \mu_t = \frac{1}{(t+1)d\sqrt{n}} (\text{Random})$$

which decreases over time. This schedule ensures that early iterations benefit from smoother approximations for stability, while later iterations use finer estimates for improved accuracy. The adaptive design of  $\mu_t$  eliminates the need for manual tuning and allows the algorithm to adjust automatically throughout the optimization process.

### 3.5 PARAMETER-FREE VARIANCE REDUCED ZEROOTH-ORDER METHOD(COORDWISE)

---

#### Algorithm 1 PF-VRZO(coord)

---

```

237 Set  $c = 1$  for coordwise estimator,  $\mu_{-1} = \mu_0$ .
238 for  $t = 0$  to  $T-1$  do
239     Compute  $\mu_t = \frac{1}{(t+1)\sqrt{nd}}$ 
240     if  $t \bmod n = 0$  then
241          $v_t = \bar{\nabla}_{\mu_t} f(x_t)$  {Full zeroth-order gradient computation}
242     else
243         Uniformly sample  $i_t \in \{1, \dots, n\}$ 
244         Compute  $\bar{\nabla}_{\mu_t} f_{i_t}(x_t)$  with  $\mu_t$  and  $\bar{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1})$  with  $\mu_{t-1}$  .
245          $v_t = \bar{\nabla}_{\mu_t} f_{i_t}(x_t) - \bar{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1}) + v_{t-1}$ 
246     end if
247      $\gamma_t = \frac{1}{n^{1/4} c \sqrt{(n^{1/2} + \sum_{s=0}^t \|v_s\|^2)}}$ 
248      $x_{t+1} = x_t - \gamma_t v_t$ 
249 end for

```

---

**Explanation of Algorithm 1:** For the constant  $c$ , we set  $c = 1$  in this algorithm (which uses the coord estimator) and  $c = \sqrt{d}$  in the algorithm with the rand estimator. At each iteration, the algorithm adaptively adjusts the smoothing parameter  $\mu_t = 1/(t+1)\sqrt{nd}$ , allowing finer gradient estimates as optimization progresses. Every  $n$  iterations, a full zeroth-order gradient is computed as mentioned in [Coord estimator](#). For the remaining steps, a variance-reduced estimator  $v_t$  is constructed by combining the current and previous stochastic gradient estimates with  $v_{t-1}$ . The step size  $\gamma_t$  is also adaptively computed based on the historical norm of the gradient estimates, eliminating the need for manual tuning.

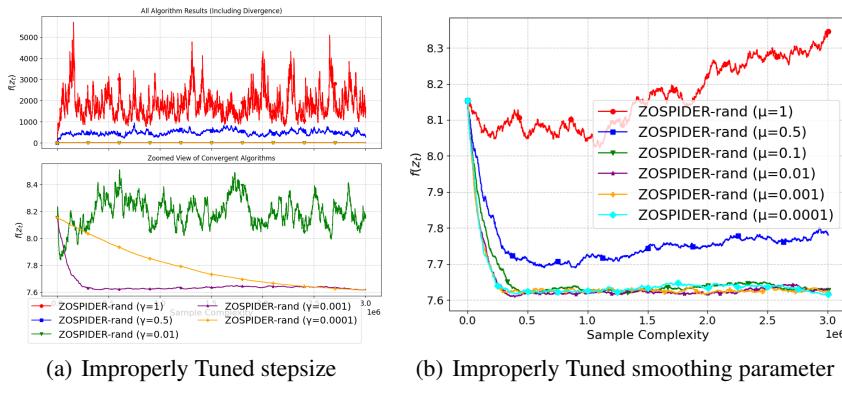
---

To establish the convergence of our method, we divide the analysis into three parts.

$$\frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] \leq \frac{1}{T} \underbrace{\left[ \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|] \right]}_{\text{part I}} + \underbrace{\left[ \sum_{t=0}^{T-1} \mathbb{E}[\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|] \right]}_{\text{part II}} + \underbrace{\left[ \sum_{t=0}^{T-1} \|\bar{\nabla}_{\mu_t} f(x_t) - \nabla f(x_t)\| \right]}_{\text{part III}}.$$

For each of these parts, we now present the corresponding lemmas (the detailed proofs can be found in [Appendix C](#)). Let  $\delta_t := \frac{\sqrt{d}L\mu_t}{2}$  denote the error coefficient of the zeroth-order estimator. Then, we can derive the following results for Algorithm 1. First, we introduce a preliminary bound that will be repeatedly used in the subsequent analysis. [The following lemma provides a bound that is frequently](#)

270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282



283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323

Figure 1: This figure demonstrates the detrimental effects of improper parameter tuning on the optimization process through two subfigures. In (a), an improperly tuned stepsize leads to phenomena such as erratic fluctuations (e.g., the red curve in the upper subplot of (a)) and even non-convergence, while properly tuned stepsizes enable stable convergence (lower subplot of (a)). In (b), an improperly tuned smoothing parameter (e.g.,  $\mu = 1$  in the red curve) causes the optimization process to fail to converge, whereas appropriately tuned values (e.g.,  $\mu = 0.0001$ ) allow for effective convergence. Collectively, these results indicate that improperly tuned parameters can severely impair the optimization process, and in severe cases, even lead to non-convergence.

used in the proof. Although it may seem somewhat large, there is no need to worry because it will appear in logarithmic form in the proof.

**Lemma 1.** *Under assumptions 1 and 2, we have*

$$\sum_{t=0}^{T-1} \|v_t\|^2 \leq \Phi(T) + 1.$$

where  $\Phi(T) := \frac{4TL^2n^{1.5}}{c^2} + (32n^2 + 6) \sum_{t=0}^{T-1} \delta_t^2 + \frac{6L^2T}{nc^2} + 6T \|\nabla f(x_0)\|^2 - 1$ .

Next, we provide an upper bound for each part separately. To facilitate the analysis, we transform the problem of the average gradient into two components: the gradient estimator  $v_t$  (Part I) and the average of gradient estimation errors. The estimation error can be further decomposed into two parts: one is the error incurred by  $v_t$  estimating the zeroth-order estimator  $\bar{\nabla}_{\mu_t} f(x_t)$  (Part II), and the other is the error arising from replacing the true gradient  $\nabla f(x_t)$  with the zeroth-order estimator (Part III). The following lemma aims to provide an upper bound for the SPIDER estimator  $v_t$ . Due to the complexity of this problem, we split the analysis into two lemmas.

**Lemma 2** (Part I(1)). *Under assumptions 1 and 2, we have*

$$\mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t\| \right] \leq n^{1/4} \sqrt{T} \left( 2\Delta c + 2c \sum_{t=0}^{T-1} \gamma_t \delta_t^2 + 1 + \frac{L}{c} \log(\Phi(T)) + c \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 \right] \right).$$

The following lemma provides an upper bound for the last term in Part I(1).

**Lemma 3** (Part I(2)). *Under assumptions 1 and 2, we have*

$$\mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 \right] \leq \frac{2L^2}{c^3} \log(\Phi(T)) + \sum_{t=0}^{T-1} 16n \gamma_t \delta_t^2.$$

For the error incurred by the estimator  $v_t$  in estimating the zeroth-order estimator  $\bar{\nabla}_{\mu_t} f(x_t)$ , we present the following lemma:

**Lemma 4** (Part II). *Under assumptions 1 and 2, we have*

$$\frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\| \right] \leq \frac{Ln^{1/4}}{c\sqrt{T}} \log(\Phi(T)) + \frac{1}{\sqrt{T}} \sqrt{8n \sum_{t=0}^{T-1} \delta_t^2}.$$

324 Based on the properties of the coordinate-wise zeroth-order estimator, we can directly give the upper  
 325 bound for Part III as follows.

326 **Lemma 5** (Part III). *Under assumptions 1 and 2, we have  $\frac{1}{T} \sum_{t=0}^{T-1} \|\bar{\nabla}_{\mu_t} f(x_t) - \nabla_{\mu_t} f(x_t)\| \leq$   
 327  $\frac{1}{T} \sum_{t=0}^{T-1} \delta_t$ .*

329 **Theorem 1** (Converge result of PF-VRZO(coord)). *Under assumptions 1, 2, we can derive the  
 330 following result for Algorithm 1:*

$$\begin{aligned} 331 \quad & \frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] \\ 332 \quad & \leq \frac{n^{1/4}}{\sqrt{T}} \left( 2\Delta \cdot c + 1 + \left( \frac{L}{c} + \frac{L^2}{c^2} \right) \log(\Phi(T)) + \frac{L^2 \pi^2}{24n^{1/4}} + \sqrt{\frac{\pi^2}{24}} \frac{L}{n^{1/4}} + \frac{L^2 \pi^2}{12} + \frac{L}{2} \right) \\ 333 \end{aligned}$$

337 By setting  $c = 1$ , we can find stationary points of  $f(x)$  with  $T = \tilde{\mathcal{O}}(\sqrt{n}\epsilon^{-2})$ .

$$\begin{aligned} 339 \quad & \frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] \leq \underbrace{\frac{1}{T} \left[ \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|] \right]}_{\text{part I}} + \underbrace{\sum_{t=0}^{T-1} \mathbb{E}[\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|]}_{\text{part II}} + \underbrace{\sum_{t=0}^{T-1} \|\bar{\nabla}_{\mu_t} f(x_t) - \nabla f(x_t)\|}_{\text{part III}} \\ 340 \quad & \leq \frac{n^{1/4}}{\sqrt{T}} \left( 2\Delta \cdot c + 1 + \left( \frac{L}{cn^{3/4}} + \frac{L^2}{c^2} \right) \log(\Phi(T)) \right) \\ 341 \quad & + \frac{1}{T} (2c \sum_{t=0}^{T-1} \gamma_t \delta_t^2 + \sum_{t=0}^{T-1} \delta_t) + \frac{1}{\sqrt{T}} \left( n^{\frac{5}{4}} \sum_{t=0}^{T-1} c \gamma_t \delta_t^2 + n^{\frac{1}{2}} \sqrt{\sum_{t=0}^{T-1} 2\delta_t^2} \right). \\ 342 \end{aligned}$$

343 Take  $\delta_t = \frac{L}{2\sqrt{n(t+1)}}$  i.e.  $(\mu_t = \frac{1}{\sqrt{nd}}(t+1))$  then we can give an upper bound of  $\sum_{t=0}^{T-1} \delta_t^2$  and  $\sum_{t=0}^{T-1} \delta_t$   
 344 as follows:

$$\sum_{t=0}^{T-1} \delta_t \leq \frac{L \ln T}{2\sqrt{n}}, \quad \sum_{t=0}^{T-1} \delta_t^2 \leq \frac{L^2 \pi^2}{24n}.$$

354 With some calculations, we can obtain the final result.

355 **Remark 2** (Discussion on the complexity). *Each coordwise estimator zeroth-order gradient estimation requires  $\mathcal{O}(d)$  function evaluations. And since SPIDER consumes, on average,  $\mathcal{O}(1+n/n)$  zeroth-order estimators per iteration, multiplying this by the total number of iterations  $T = \tilde{\mathcal{O}}(\sqrt{n}\epsilon^{-2})$  yields a total function query complexity of #Function =  $\tilde{\mathcal{O}}(d(1 + \frac{n}{n})T) = \tilde{\mathcal{O}}(d\sqrt{n}\epsilon^{-2})$ .*

### 360 3.6 PROPOSED PARAMETER-FREE VARIANCE REDUCED ZEROTH-ORDER 361 METHOD(RANDOM-DIRECTION ESTIMATOR)

363 In contrast to the coordinate-wise approach, which requires  $\mathcal{O}(d)$  function evaluations per random  
 364 estimator, the random method only incurs  $\mathcal{O}(1)$  function evaluations per iteration. Nevertheless, it  
 365 often requires  $d$  times more iterations to achieve comparable accuracy. Therefore, the choice between  
 366 the two methods can be made according to the practitioner's computational budget and application  
 367 requirements. The analysis of the random-direction method follows essentially the same structure as  
 368 that of the coordinate-wise method, although the final results differ slightly.

$$\begin{aligned} 369 \quad & \frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] \leq \underbrace{\frac{1}{T} \left[ \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|] \right]}_{\text{part I}} + \underbrace{\sum_{t=0}^{T-1} \mathbb{E}[\|v_t - \nabla f_{\mu_t}(x_t)\|]}_{\text{part II}} + \underbrace{\sum_{t=0}^{T-1} \|\nabla f_{\mu_t}(x_t) - \nabla f(x_t)\|}_{\text{part III}} \\ 370 \end{aligned}$$

373 The proof of this part follows a similar argument as the coordinate estimator case and is therefore  
 374 omitted. The complete proof can be found in Appendix D.

376 **Remark 3** (Proof Differences between the Random-direction and Coord Methods). *In the coordinate-  
 377 wise method, we provide a bound  $\mathcal{O}(\|x_t - x_{t-1}\|^2) + \mathcal{O}(\mu_t^2 + \mu_{t-1}^2)$  for the quantity  $\|\bar{\nabla}_{\mu_t} f_{i_t}(x_t) -$   
 $\bar{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1})\|^2$ . Although an estimation error exists, the smoothness of the coordinate estimator*

---

**Algorithm 2** PF-VRZO(Random-direction)

---

378 Set  $c = \sqrt{d}$  for random-direction estimator and  $\mu_{-1} = \mu_0$ .  
 379  
 380 **for**  $t = 0$  **to**  $T-1$  **do**  
 381   Compute smoothing parameter  $\mu_t = \frac{1}{(t+1)d\sqrt{n}}$ , smoothing vector  $\rho_t \sim U_B$ .  
 382   **if**  $t \bmod n = 0$  **then**  
 383      $v_t = \hat{\nabla}_{\mu_t} f(x_t)$  {Full zeroth-order gradient computation}  
 384   **else**  
 385     Sample  $i_t \in \{1, \dots, n\}$  uniformly at random  
 386     Compute  $\hat{\nabla}_{\mu_t} f_{i_t}(x_t)$  with parameter  $\mu_t$  and rand vector  $\rho_t$ ,  $\hat{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1})$  with different  
 387     parameter  $\mu_{t-1}$  and the same rand vector  $\rho_t$ .  
 388      $v_t = \hat{\nabla}_{\mu_t} f_{i_t}(x_t) - \hat{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1}) + v_{t-1}$   
 389   **end if**  
 390      $\gamma_t = \frac{1}{n^{1/4}c\sqrt{(n^{1/2} + \sum_{s=0}^t \|v_s\|^2)}}$   
 391      $x_{t+1} = x_t - \gamma_t v_t$   
 392   **end for**  
 393  
 394

395 **Explanation of Algorithm 2** Algorithm 2 shares an overall structure with Algorithm 1, with  
 396 key differences as follows: 1.The zeroth-order estimator employs a random-direction estimator as  
 397 mentioned in [Random-direction estimator](#), where random numbers distributed on the unit sphere  
 398 are generated by first sampling from a  $d$ -dimensional Gaussian distribution and then normalizing  
 399 the sample. 2.We set  $c = \sqrt{d}$  and use a smoothing parameter  $\mu_t = \frac{1}{(t+1)d\sqrt{n}}$ , introducing constant  
 400 differences (involving  $\sqrt{d}$ ) compared to the coordinate-wise variant ,where  $c = 1$  and  $\mu_t = \frac{1}{(t+1)\sqrt{dn}}$

---

402  
 403 *remains roughly of the same order as that of  $f(x)$ . In contrast, for the random-direction method,*  
 404 *we obtain the estimate  $\|\hat{\nabla}_{\mu_t} f(x_t) - \hat{\nabla}_{\mu_{t-1}} f(x_{t-1})\|^2 \leq \mathcal{O}(\mu_t^2 + \mu_{t-1}^2) + \mathcal{O}(d\|x_t - x_{t-1}\|^2)$ , which*  
 405 *suggests—albeit informally—that the smoothness of the random estimator is approximately  $d$  times*  
 406 *larger than that of  $f(x)$ . This distinction is reflected in the conclusions of various parts of the*  
 407 *analysis, and, in particular, it necessitates choosing  $c = \sqrt{d}, \mu_t = \frac{1}{(t+1)d\sqrt{n}}$  in the proof of the*  
 408 *theorem (whereas  $c = 1, \mu_t = \frac{1}{(t+1)\sqrt{dn}}$  suffices in the coordinate-wise case). As a result, the number*  
 409 *of iterations required by the random-direction method is  $d$  times larger than that of the coord method.*

410 **Theorem 2** (Converge result of PF-VRZO(random-direction)). *Under assumptions 1, 2, we can derive*  
 411 *the following result for Algorithm 2:*

$$\begin{aligned} & \frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] \\ & \leq \frac{n^{1/4}}{\sqrt{T}} \left( \Delta \cdot c + 1 + \left( \frac{L\sqrt{d}}{cn^{3/4}} + \frac{L^2 d}{c^2} \right) \log(\phi(T)) + \frac{L^2 \pi^2}{24n^{1/4}} + \frac{L}{n^{1/4}} \sqrt{\frac{\pi^2}{24}} + \frac{L^2 \pi^2}{12} + \frac{L}{2} \right) \end{aligned}$$

412 By setting  $c = \sqrt{d}, \mu_t = \frac{1}{d\sqrt{n}}(t+1)$ , we can find stationary points of  $f(x)$  with  $T = \tilde{\mathcal{O}}(d\sqrt{n}\epsilon^{-2})$ .

413 **Remark 4** (Discussion on the complexity). *Each Random-direction zeroth-order gradient estimation*  
 414 *requires  $\mathcal{O}(1)$  function evaluations. And since SPIDER consumes, on average,  $\mathcal{O}(1 + n/n)$  zeroth-*  
 415 *order estimators per iteration, multiplying this by the total number of iterations  $T = \tilde{\mathcal{O}}(d\sqrt{n}\epsilon^{-2})$*   
 416 *yields a total function query complexity of  $\#Function = \tilde{\mathcal{O}}\left((1 + \frac{n}{n})T\right) = \tilde{\mathcal{O}}(d\sqrt{n}\epsilon^{-2})$ .*

---

## 4 EXPERIMENTS

426 We conduct two experiments to evaluate the effectiveness of our method: the first focuses on Phase  
 427 Retrieval, as shown in Figures 2(a) and 2(b), while the second examines Distributional Robust  
 428 Optimization (DRO), presented in Figures ?? and ???. To validate the performance of our algorithm,  
 429 we compare it with ZO-SPIDER([Ji et al., 2019](#)) and ZO-SGD ([Ghadimi & Lan, 2013](#)), both of which  
 430 rely on manually tuned hyperparameters to ensure convergence, in contrast to our parameter-free

approach. We measure computational cost using both sample complexity and time. Here, sample complexity refers to the total number of function value evaluations. Due to space limitations, we defer the detailed descriptions of the hyperparameter settings to Appendix E. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

#### 4.1 APPLICATION TO NONCONVEX PHASE RETRIEVAL

Phase retrieval is a well-known nonconvex problem in machine learning and signal processing (Miao et al., 1999). Let  $x \in \mathbb{R}^d$  represent the true underlying object, and assume we collect  $m$  intensity measurements, given by  $y_r = |\mathbf{a}_r^\top x|^2$  for  $r = 1, 2, \dots, m$ , where  $\mathbf{a}_r \in \mathbb{R}^d$ . The challenge in phase retrieval lies in recovering the signal by solving the associated nonconvex optimization problem:

$$\min_{z \in \mathbb{R}^d} f(z) := \frac{1}{2m} \sum_{r=1}^m (y_r - |\mathbf{a}_r^\top z|^2)^2. \quad (2)$$

We assess the effectiveness of our algorithms on the nonconvex phase retrieval task defined in (2). As illustrated in Figures 2(a) and 2(b), the proposed PF-VRZO algorithm demonstrates robust performance, notably without requiring manual tuning of algorithmic parameters.

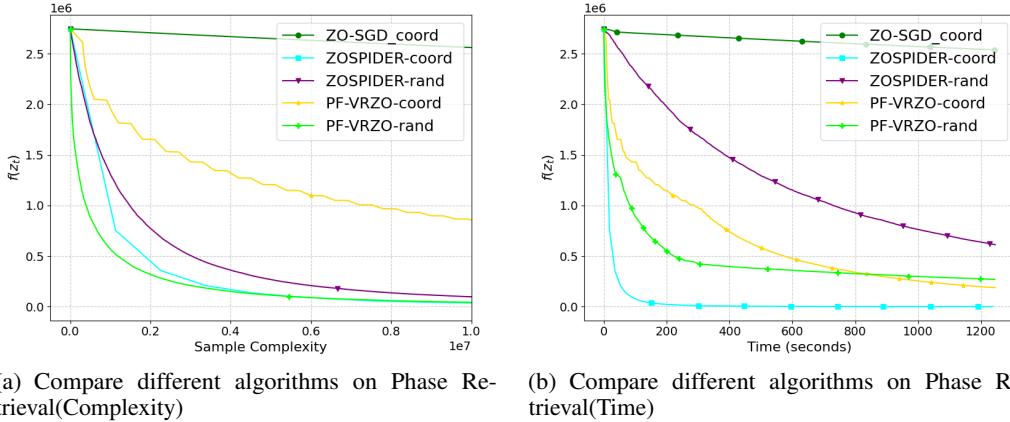


Figure 2: This figure compares the performance of different algorithms on Phase Retrieval through two subfigures. In (a), we evaluate the sample complexity of algorithms including PF-VRZO-coord, PF-VRZO-rand, ZO-SGD-coord, ZOSPIDER-coord, and ZOSPIDER-rand. In (b), we assess their time efficiency. Notably, the proposed PF-VRZO method, even without parameter tuning, demonstrates competitive performance when compared to other algorithms that undergo multiple parameter adjustments, indicating its robustness and effectiveness in Phase Retrieval tasks.

#### 4.2 APPLICATION TO DISTRIBUTIONAL ROBUST OPTIMIZATION

Distributional Robust Optimization (DRO) is a widely used framework for training robust models, Under mild conditions, it aims to solve the following problem:

$$\min_{x \in \mathcal{X}, \eta \in \mathbb{R}} L(x, \eta) := \lambda \mathbb{E}_{\xi \sim P} \psi^* \left( \frac{\ell_\xi(x) - \eta}{\lambda} \right) + \eta \quad (3)$$

We consider the nonconvex DRO problem (3) on three real-world datasets. The Life Expectancy dataset contains 2,413 samples with 20 associated features. The Communities and Crime dataset consists of 1,994 samples and 122 predictive features. The Arcene dataset includes 200 samples with 10,000 high-dimensional features, making it a challenging benchmark for robust optimization. After standard preprocessing steps, including missing value imputation and variable standardization, we retain 70% samples for training, where each input  $x_i \in \mathbb{R}^{34}$  and corresponding target  $y_i \in \mathbb{R}$ . We set the regularization parameter to  $\lambda = 0.01$ , and adopt the  $\chi^2$ -divergence, with the convex conjugate

486 given by  $\psi^*(t) = \frac{1}{4}(t+2)^2 - 1$ . The regularized loss function is defined as:  
 487

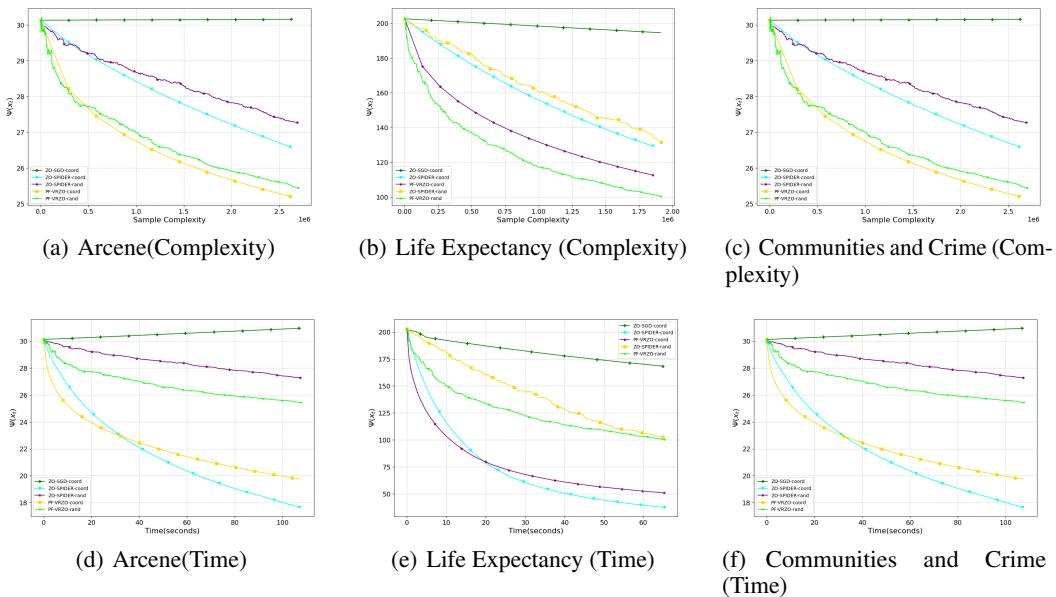
$$488 \quad \ell_\xi(w) = \frac{1}{2}(y_\xi - x_\xi^\top w)^2 + 0.1 \sum_{j=1}^{34} \ln(1 + |w^{(j)}|).$$

$$489$$

$$490$$

491 We initialize  $w_0 \in \mathbb{R}^{34}$  from a Gaussian distribution and set the initial step size  $\eta_0 = 0.1$ .  
 492

493 Based on the experimental results shown in Figures 3, we observe that the proposed PF-VRZO  
 494 method exhibits a brief oscillation in the objective value at the beginning, likely due to insufficient  
 495 accumulated gradient information. However, the method quickly resumes descent and ultimately  
 496 achieves strong performance without the need for any parameter tuning.  
 497



517 Figure 3: This figure evaluates the performance of different algorithms on Distributionally Robust  
 518 Optimization (DRO) tasks across three datasets (Arcene, Life Expectancy, Communities and Crime),  
 519 with results split into two metrics: Sample Complexity (subfigures (a)-(c)): Measures the number  
 520 of samples required for algorithms to converge. Time Efficiency (subfigures (d)-(f)): Measures  
 521 the runtime (in seconds) for algorithms to converge. Across all datasets and metrics, the proposed  
 522 methods (e.g., PF-VRZO variants) demonstrate competitive or superior performance—consistently  
 523 achieving faster convergence. This further validates the effectiveness of the parameter-free design of  
 524 PF-VRZO in practical DRO scenarios.  
 525

## 5 CONCLUSION

529 In this paper, we propose a parameter-free variance-reduced zeroth-order method (PF-VRZO) for  
 530 nonconvex optimization. Our method is based on the SPIDER framework and employs a coordinate-  
 531 wise or random-direction zeroth-order gradient estimator. We establish the convergence of our  
 532 method, demonstrating that it achieves a sample complexity of  $\tilde{\mathcal{O}}(d\sqrt{n}\epsilon^{-2})$  for finding stationary  
 533 points of nonconvex functions. Additionally, we conduct experiments on nonconvex phase retrieval  
 534 and distributionally robust optimization to validate the effectiveness of our method. An interesting  
 535 future direction is to investigate whether the logarithmic,  $L$ -dependent, and  $\Delta$ -dependent terms  
 536 in the complexity bounds are optimal. (Carmon & Hinder, 2024) shows that under the convex-  
 537 but-nonsmooth (C-NS) setting, any adaptive algorithm necessarily suffers from worse complexity.  
 538 However, it remains unclear whether a similar conclusion holds under the nonconvex-smooth (NC-S)  
 539 setting.

540 **6 ETHICS STATEMENT**  
 541

542 Our study focuses on developing a novel optimization algorithm and does not involve human subjects,  
 543 animal experimentation, or the use of sensitive personal data. All experiments are conducted on  
 544 publicly available datasets that are commonly used within the academic community. We adhere to the  
 545 ICLR Code of Ethics, and our work introduces no new privacy or ethical risks beyond those inherent  
 546 in standard academic research on optimization methods.

548 **7 REPRODUCIBILITY STATEMENT**  
 549

550 We have made every effort to ensure the reproducibility of our results. The paper provides detailed  
 551 specifications for our proposed algorithm, PF-VRZO, including its variants and their theoretical foun-  
 552 dations. We have meticulously described our experimental setup, including the specific nonconvex  
 553 problems we studied, the parameters used for all compared algorithms (e.g., learning rates and batch  
 554 sizes for ZO-SGD, PF-VRZO, ZO-SPIDER), and the hardware used.

556 **REFERENCES**  
 557

558 Amit Attia and Tomer Koren. How free is parameter-free stochastic optimization? In *Forty-first*  
 559 *International Conference on Machine Learning*, 2024. URL [https://openreview.net/](https://openreview.net/forum?id=6L4K5jmSJq)  
 560 [forum?id=6L4K5jmSJq](https://openreview.net/forum?id=6L4K5jmSJq).

561 Yair Carmon and Oliver Hinder. The price of adaptivity in stochastic convex optimization. In Shipra  
 562 Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*,  
 563 volume 247 of *Proceedings of Machine Learning Research*, pp. 772–774. PMLR, 30 Jun–03 Jul  
 564 2024. URL <https://proceedings.mlr.press/v247/carmon24a.html>.

565 Keyi Chen, John Langford, and Francesco Orabona. Better parameter-free stochastic optimization  
 566 with ode updates for coin-betting. *Proceedings of the AAAI Conference on Artificial Intelligence*,  
 567 36(6):6239–6247, Jun. 2022. doi: 10.1609/aaai.v36i6.20573. URL <https://ojs.aaai.org/index.php/AAAI/article/view/20573>.

568 Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order  
 569 optimization based black-box attacks to deep neural networks without training substitute models.  
 570 *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017. URL  
 571 <https://api.semanticscholar.org/CorpusID:2179389>.

572 Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation.  
 573 *ArXiv*, abs/2301.07733, 2023. URL <https://api.semanticscholar.org/CorpusID:255999725>.

574 John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning  
 575 and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011. URL <https://api.semanticscholar.org/CorpusID:538820>.

576 John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for  
 577 zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on*  
 578 *Information Theory*, 61(5):2788–2806, 2015. doi: 10.1109/TIT.2015.2409256.

579 Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-  
 580 convex optimization via stochastic path-integrated differential estimator. In S. Bengio,  
 581 H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Ad-  
 582 vances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.,  
 583 2018. URL [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf).

584 Xiang Gao, Bo Jiang, and Shuzhong Zhang. On the information-adaptive variants of the admm:  
 585 An iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, Jul 2018.  
 586 ISSN 1573-7691. doi: 10.1007/s10915-017-0621-6. URL <https://doi.org/10.1007/s10915-017-0621-6>.

594 Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii, Farshed Abdukhakimov, Dmitry Kamzolov,  
 595 Aleksandr Beznosikov, Martin Takáč, Pavel Dvurechensky, and Bin Gu. The power of first-order  
 596 smooth optimization for black-box non-smooth problems, 2023. URL <https://arxiv.org/abs/2201.12289>.

597

598 Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex  
 599 stochastic programming, 2013. URL <https://arxiv.org/abs/1309.5549>.

600

601 Bin Gu, Zhouyuan Huo, Cheng Deng, and Heng Huang. Faster derivative-free stochastic algorithm  
 602 for shared memory machines. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the*  
 603 *35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine*  
 604 *Learning Research*, pp. 1812–1821. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/gu18a.html>.

605

606 Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is sgd’s best friend: A parameter-free dynamic  
 607 step size schedule. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt,  
 608 Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML*  
 609 *2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning*  
 610 *Research*, pp. 14465–14499. PMLR, 2023. URL <https://proceedings.mlr.press/v202/ivgi23a.html>.

611

612 Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms  
 613 and analysis for nonconvex optimization. In *International Conference on Machine Learning*, 2019.  
 614 URL <https://api.semanticscholar.org/CorpusID:174800372>.

615

616 Wei Jiang, Sifan Yang, Yibo Wang, and Lijun Zhang. Adaptive variance reduction for  
 617 stochastic optimization under weaker assumptions. In A. Globerson, L. Mackey, D. Bel-  
 618 grave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural In-  
 619 formation Processing Systems*, volume 37, pp. 22047–22080. Curran Associates, Inc.,  
 620 2024. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/272efd3a6091ceefcbc79f1f3a6fdb4-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/272efd3a6091ceefcbc79f1f3a6fdb4-Paper-Conference.pdf).

621

622 Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive vari-  
 623 ance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Wein-  
 624 berger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Asso-  
 625 ciates, Inc., 2013. URL [https://proceedings.neurips.cc/paper\\_files/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf).

626

627 Ali Kavis, EFSTRATIOS PANTELEIMON SKOULAKIS, Kimon Antonakopoulos, Leello Tadesse  
 628 Dadi, and Volkan Cevher. Adaptive stochastic variance reduction for non-convex finite-sum  
 629 minimization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.),  
 630 *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=k98U0cb0Ig>.

631

632 Ahmed Khaled and Chi Jin. Tuning-free stochastic optimization. In *Forty-first International*  
 633 *Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=A6fmX9QCEa>.

634

635 Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order  
 636 nonsmooth nonconvex stochastic optimization. *J. Mach. Learn. Res.*, 25:122:1–122:14, 2024. URL  
 637 <https://jmlr.org/papers/v25/23-1159.html>.

638

639 Itai Kreisler, Maor Ivgi, Oliver Hinder, and Yair Carmon. Accelerated parameter-free stochastic  
 640 optimization. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference*  
 641 *on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3257–  
 642 3324. PMLR, 30 Jun–03 Jul 2024. URL <https://proceedings.mlr.press/v247/kreisler24a.html>.

643

644 Harshit Kumar, Dionysios S. Kalogerias, George Pappas, and Alejandro Ribeiro. Zeroth-  
 645 order deterministic policy gradient. *ArXiv*, abs/2006.07314, 2020. URL <https://api.semanticscholar.org/CorpusID:219636207>.

646

648 Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, and Ji Liu. A comprehensive  
 649 linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order  
 650 to first-order. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.),  
 651 *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.,  
 652 2016. URL [https://proceedings.neurips.cc/paper\\_files/paper/2016/file/db1915052d15f7815c8b88e879465ale-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2016/file/db1915052d15f7815c8b88e879465ale-Paper.pdf).

653

654 Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic  
 655 and stochastic nonsmooth nonconvex optimization. In S. Koyejo, S. Mohamed,  
 656 A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information  
 657 Processing Systems*, volume 35, pp. 26160–26175. Curran Associates, Inc.,  
 658 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/a78f142aecd481e68c75276756e0a0d91-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/a78f142aecd481e68c75276756e0a0d91-Paper-Conference.pdf).

659

660 L. Liu, Minhao Cheng, Cho-Jui Hsieh, and Dacheng Tao. Stochastic zeroth-order optimiza-  
 661 tion via variance reduction method. *ArXiv*, abs/1805.11811, 2018a. URL <https://api.semanticscholar.org/CorpusID:44117287>.

662

663 Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini.  
 664 Zeroth-order stochastic variance reduction for nonconvex optimization. In S. Bengio,  
 665 H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Ad-  
 666 vances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.,  
 667 2018b. URL [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/ba9a56ce0a9bfa26e8ed9e10b2cc8f46-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/ba9a56ce0a9bfa26e8ed9e10b2cc8f46-Paper.pdf).

668

669 Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L. Bartlett, and Martin J.  
 670 Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic  
 671 systems. *ArXiv*, abs/1812.08305, 2018. URL <https://api.semanticscholar.org/CorpusID:56517260>.

672

673 Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen,  
 674 and Sanjeev Arora. Fine-tuning language models with just forward passes. In A. Oh,  
 675 T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neu-  
 676 ral Information Processing Systems*, volume 36, pp. 53038–53075. Curran Associates, Inc.,  
 677 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/a627810151be4d13f907ac898ff7e948-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/a627810151be4d13f907ac898ff7e948-Paper-Conference.pdf).

678

679 H. B. McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization.  
 680 *ArXiv*, abs/1002.4908, 2010. URL <https://api.semanticscholar.org/CorpusID:13318811>.

681

682 Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. Extending the methodology of  
 683 x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*,  
 684 400(6742):342–344, Jul 1999. ISSN 1476-4687. doi: 10.1038/22498. URL <https://doi.org/10.1038/22498>.

685

686 Francesco Orabona and Tatiana Tommasi. Training deep networks without learning rates through  
 687 coin betting. In *Neural Information Processing Systems*, 2017. URL <https://api.semanticscholar.org/CorpusID:6762437>.

688

689 Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram  
 690 Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM  
 691 on Asia Conference on Computer and Communications Security*, ASIA CCS ’17, pp. 506–519,  
 692 New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349444. doi:  
 693 10.1145/3052973.3053009. URL <https://doi.org/10.1145/3052973.3053009>.

694

695 Kunjie Ren and Luo Luo. A parameter-free and near-optimal zeroth-order algorithm for stochastic  
 696 convex optimization. In *Forty-second International Conference on Machine Learning*, 2025. URL  
 697 <https://openreview.net/forum?id=JVouJNs8Vr>.

698

699 Andrey Veprikov, Alexander Bogdanov, Vladislav Mikhailovich Minashkin, and Aleksandr  
 700 Beznosikov. New aspects of black box conditional gradient: Variance reduction and one  
 701 point feedback. In *International Conference on Computational Optimization*, 2024. URL  
 702 <https://openreview.net/forum?id=xSwFzrDMVK>.

702	CONTENTS	
703		
704		
705	<b>1</b> <b>Introduction</b>	<b>1</b>
706		
707	<b>2</b> <b>Preliminaries</b>	<b>3</b>
708		
709	<b>3</b> <b>Proposed parameter free variance reduced zeroth-order methods</b>	<b>4</b>
710		
711	3.1 Zeroth-order gradient estimators . . . . .	4
712	3.2 Variance reduction technique . . . . .	4
713	3.3 Adaptive stepsize . . . . .	4
714	3.4 Adaptive smoothing parameter . . . . .	5
715	3.5 Parameter-Free Variance Reduced Zeroth-Order method(coordwise) . . . . .	5
716	3.6 Proposed Parameter-Free Variance Reduced Zeroth-Order method(random-direction estimator) . . . . .	7
717		
718		
719		
720		
721	<b>4</b> <b>Experiments</b>	<b>8</b>
722		
723	4.1 Application to Nonconvex Phase Retrieval . . . . .	9
724	4.2 Application to Distributional robust optimization . . . . .	9
725		
726		
727	<b>5</b> <b>Conclusion</b>	<b>10</b>
728		
729	<b>6</b> <b>ETHICS STATEMENT</b>	<b>11</b>
730		
731	<b>7</b> <b>REPRODUCIBILITY STATEMENT</b>	<b>11</b>
732		
733		
734	<b>A</b> <b>Related work</b>	<b>15</b>
735		
736	A.1 Zeroth-order optimization . . . . .	15
737	A.2 Parameter-free optimization . . . . .	15
738		
739	<b>B</b> <b>Useful facts</b>	<b>15</b>
740		
741	<b>C</b> <b>Parameter free variance reduced zeroth-order method(coord)</b>	<b>17</b>
742		
743	C.1 Part I . . . . .	18
744	C.2 Part II . . . . .	21
745	C.3 Part III . . . . .	21
746	C.4 Final proof for coordinate estimator . . . . .	22
747		
748		
749		
750	<b>D</b> <b>Parameter free variance reduced zeroth-order method(random-direction estimator)</b>	<b>23</b>
751		
752	D.1 Part I . . . . .	25
753	D.2 Part II . . . . .	28
754	D.3 Part III . . . . .	29
755	D.4 Final proof for the random-direction estimator . . . . .	29

756	<b>E Hyperparameters Details</b>	<b>29</b>
757	E.1 Phase Retrieval	29
758	E.2 DRO	30
759	E.3 A small experiment to verify the effectiveness of the adaptive smoothing parameter	30
760		
761		
762		

## 763 A RELATED WORK

### 764 A.1 ZEROTH-ORDER OPTIMIZATION

767 The ZO-SGD method was first introduced by (Ghadimi & Lan, 2013), serving as a foundational  
 768 approach in zeroth-order stochastic optimization. To enhance its efficiency, several follow-up  
 769 works (Liu et al., 2018a) proposed accelerated variants, collectively referred to as ZO-SVRG,  
 770 which incorporate the SVRG framework (Johnson & Zhang, 2013). In addition, to further reduce  
 771 the function query complexity, ZO-SPIDER-Coord (Ji et al., 2019) were developed based on the  
 772 stochastic path-integrated differential estimator.

### 773 A.2 PARAMETER-FREE OPTIMIZATION

775 Recent advances in the nonconvex and smooth setting have drawn inspiration from AdaGrad, as  
 776 introduced in the concurrent seminal works (Duchi et al., 2011; McMahan & Streeter, 2010). Building  
 777 on this foundation, (Kavis et al., 2022) were the first to develop a parameter-free algorithm that remains  
 778 compatible with variance reduction techniques. This was later improved by (Jiang et al., 2024),  
 779 who proposed ADA-STORM, reducing the overall complexity by a logarithmic factor. Moreover, a  
 780 series of follow-up studies (Ivgi et al., 2023; Kreisler et al., 2024; Orabona & Tommasi, 2017; Chen  
 781 et al., 2022; Defazio & Mishchenko, 2023) have explored parameter-free methods in various problem  
 782 settings, and other works have investigated the fundamental lower bounds of such algorithms (Khaled  
 783 & Jin, 2024; Attia & Koren, 2024; Carmon & Hinder, 2024).

## 785 B USEFUL FACTS

787 **Lemma B.1** (Jensen’s inequality). *For convex function  $f(x)$  we have*

$$788 f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)],$$

790 two extended versions of Jensen’s inequality are

$$792 \|\mathbb{E}[x]\| \leq \mathbb{E}[\|x\|], \text{ for } x \in \mathbb{R}^d$$

$$793 \left\| \sum_{i=1}^k a_i \right\|^2 \leq k \sum_{i=1}^k \|a_i\|^2, \text{ for } a_i \in \mathbb{R}^d.$$

796 **Lemma B.2** (Young’s inequality). *For any vectors  $a, b \in \mathbb{R}^d$ , and  $\zeta \geq 0$ , the following inequality  
 797 holds:*

$$798 \|a\|^2 \leq (1 + \zeta) \|a - b\|^2 + (1 + \zeta^{-1}) \|b\|^2,$$

800 an extended version of Young’s inequality is

$$801 \langle a, b \rangle \leq \frac{\|a\|^2}{2\zeta} + \frac{\zeta \|b\|^2}{2}.$$

804 **Lemma B.3** (variance decomposition). *For random vector  $x \in \mathbb{R}^d$  and any  $y \in \mathbb{R}^d$ , the variance of  
 805  $x$  can be decomposed as*

$$807 \mathbb{E} [\|x - \mathbb{E}[x]\|^2] = \mathbb{E} [\|x - y\|^2] - \mathbb{E} [\|\mathbb{E}[x] - y\|^2],$$

808 which implies

$$809 \mathbb{E} [\|x - \mathbb{E}[x]\|^2] \leq \mathbb{E} [\|x\|^2].$$

810 **Lemma B.4.** For random variable  $X, Y$ , if  $X, Y$  are independent, and  $\mathbb{E}[X] = 0$  or  $\mathbb{E}[Y] = 0$ , we have  
 811

$$812 \mathbb{E}[\|X - Y\|^2] = \mathbb{E}[\|X\|^2] + \mathbb{E}[\|Y\|^2].$$

814 *Proof.*  
 815

$$816 \mathbb{E}[\|X - Y\|^2] = \mathbb{E}[\|X\|^2 + \|Y\|^2 + 2\mathbb{E}\langle X, Y \rangle] = \mathbb{E}[\|X\|^2] + \mathbb{E}[\|Y\|^2].$$

□

820 **Lemma B.5.** For i.i.d.  $x_1, x_2, x_3 \dots x_n$ , if  $\mathbb{E}[x_i] = x$ ,  $\mathbb{E}[\|x_i - x\|^2] \leq \sigma^2$ , we have  
 821

$$822 \mathbb{E} \left[ \left\| \frac{1}{b} \sum_{i=1}^b x_i - x \right\|^2 \right] \leq \frac{\mathbb{E}[\|x_i\|^2]}{b}.$$

826 *Proof.*  
 827

$$828 \begin{aligned} \mathbb{E} \left[ \left\| \frac{1}{b} \sum_{i=1}^b x_i - x \right\|^2 \right] \\ 829 &= \frac{1}{b^2} \mathbb{E} \left[ \left\| \sum_{i=1}^b (x_i - x) \right\|^2 \right] \\ 830 &= \frac{1}{b^2} \sum_{i=1}^b \mathbb{E}[\|x_i - x\|^2] \\ 831 &= \frac{1}{b} \mathbb{E}[\|x_i - x\|^2] \leq \frac{\mathbb{E}[\|x_i\|^2]}{b}, \end{aligned}$$

832 where the second inequality holds because  $\|a + b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b \rangle$ , and  $\mathbb{E}[\langle x_i - x, x_j - x \rangle] = 0 (j \neq i)$  for iid random variable  $x_i$ .  
 833 □

834 **Lemma B.6** (Sum of Square Roots Inequality). Let  $\alpha_1, \dots, \alpha_T$  be a sequence of non-negative real  
 835 numbers ( $\alpha_t \geq 0$  for all  $t$ ). Then:

$$836 \sqrt{\sum_{t=1}^T \alpha_t} \leq \sum_{t=1}^T \frac{\alpha_t}{\sqrt{\sum_{s=1}^t \alpha_s}}.$$

837 **Lemma B.7** (Logarithmic Sum Bound). For any sequence of non-negative real numbers  $a_1, \dots, a_T$   
 838 with  $a_1 \geq 1$ , we have:

$$839 \sum_{\ell=1}^T \frac{a_\ell}{1 + \sum_{i=1}^\ell a_i} \leq \log \left( \sum_{i=1}^T a_i + 1 \right)$$

840 **Lemma B.8** (Sum of  $\frac{1}{i}$  and  $\frac{1}{i^2}$ ).  
 841

$$842 \sum_{i=1}^{T-1} \frac{1}{i} \leq \log(T).$$

$$843 \sum_{i=1}^{\infty} \frac{1}{i^2} \leq \frac{\pi^2}{6}.$$

864 C PARAMETER FREE VARIANCE REDUCED ZEROOTH-ORDER METHOD(COORD)  
865  
866867 **Algorithm 3** PF-VRZO(coord)

---

868 Set  $c = 1$  for coordwise estimator,  $\mu_{-1} = \mu_0$ .  
869 **for**  $t = 0$  **to**  $T-1$  **do**  
870   Compute  $\mu_t = \frac{1}{(t+1)\sqrt{nd}}$   
871   **if**  $t \bmod n = 0$  **then**  
872      $v_t = \bar{\nabla}_{\mu_t} f(x_t)$  {Full zeroth-order gradient computation}  
873   **else**  
874     Uniformly sample  $i_t \in \{1, \dots, n\}$   
875     Compute  $\bar{\nabla}_{\mu_t} f_{i_t}(x_t)$  with  $\mu_t$  and  $\bar{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1})$  with  $\mu_{t-1}$ .  
876      $v_t = \bar{\nabla}_{\mu_t} f_{i_t}(x_t) - \bar{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1}) + v_{t-1}$   
877   **end if**  
878    $\gamma_t = \frac{1}{n^{1/4}c\sqrt{(n^{1/2} + \sum_{s=0}^t \|v_s\|^2)}}$   
879    $x_{t+1} = x_t - \gamma_t v_t$   
880 **end for**  
881  
882  
883  
884

---

885 Table 2: Meaning of Symbols  
886

Symbol	Meaning
$\gamma_t$	stepsize $\frac{1}{(n^{1/4}c\sqrt{n^{1/2} + \sum_{s=0}^t \ v_s\ ^2})}$ .
$\mu_t$	Smoothing parameter at iteration $t$ .
$v_t$	Spider estimator.
$\bar{\nabla}_{\mu} f(x_t)$	zeroth-order estimator(coord) .
$\delta_t$	$\sqrt{d}L\mu_t/2$ , the estimation error with respect to $\bar{\nabla}f_{\mu}$ .

897 To establish the convergence of our method, we divide the analysis into three parts.  
898

899  
900 
$$\frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] \leq \underbrace{\frac{1}{T} \left[ \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|] \right]}_{\text{part I}} + \underbrace{\sum_{t=0}^{T-1} \mathbb{E}[\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|]}_{\text{part II}} + \underbrace{\sum_{t=0}^{T-1} \|\bar{\nabla}_{\mu_t} f(x_t) - \nabla f(x_t)\|}_{\text{part III}}.$$
901  
902  
903  
904

905 **Lemma C.1** ((Gao et al., 2018)). For  $L$ -smooth function  $f(x)$ , its gradient  $\nabla f(x)$  and its coord  
906 zeroth-order estimator  $\bar{\nabla}_{\mu} f(x)$ , we have  
907

908 
$$\|\bar{\nabla}_{\mu} f(x) - \nabla f(x)\|^2 \leq \delta_t^2.$$
  
909

910 where  $\delta_t := \sqrt{d}L\mu_t/2$ , and  $\mu_t$  is the smoothing parameter at iteration  $t$ .  
911912 **Lemma C.2.** Under assumptions 1 and 2, we can derive the following result for Algorithm 1:  
913

914 
$$\sum_{t=0}^{T-1} \|v_t\|^2 \leq \Phi(T) + 1.$$
  
915

916 where  $\frac{4TL^2n^{1.5}}{c^2} + (32n^2 + 6) \sum_{t=0}^{T-1} \delta_t^2 + \frac{6L^2T}{nc^2} + 6T \|\nabla f(x_0)\|^2 - 1$ . Here the notation  $\Phi(T)$  is  
917 introduced only for brevity, and will be repeatedly used in the subsequent analysis.

918 *Proof.*

$$\begin{aligned}
\|v_t\|^2 &= \left\| \sum_{s=t-t \bmod n+1}^t (\bar{\nabla}_{\mu_s} f_{i_s}(x_s) - \bar{\nabla}_{\mu_{s-1}} f_{i_s}(x_{s-1})) + \bar{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2 \\
&\leq 2 \cdot \left\| \sum_{s=t-t \bmod n+1}^t \bar{\nabla}_{\mu_s} f_{i_s}(x_s) - \bar{\nabla}_{\mu_{s-1}} f_{i_s}(x_{s-1}) \right\|^2 + 2 \cdot \left\| \bar{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2 \\
&\leq 2n \cdot \sum_{s=t-t \bmod n+1}^t \left\| \bar{\nabla}_{\mu_s} f_{i_s}(x_s) - \bar{\nabla}_{\mu_{s-1}} f_{i_s}(x_{s-1}) \right\|^2 + 2 \cdot \left\| \bar{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2 \\
&\stackrel{\text{lem C.1}}{\leq} 2n \cdot \sum_{s=t-t \bmod n+1}^t 2L^2 (x_s - x_{s-1})^2 + 8(\delta_s^2 + \delta_{s-1}^2) + 2 \cdot \left\| \bar{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2 \\
&\leq \frac{4L^2 n^{1.5}}{c^2} + 16n \sum_{s=t-t \bmod n+1}^t (\delta_s^2 + \delta_{s-1}^2) + 2 \cdot \left\| \bar{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2,
\end{aligned}$$

935 in last inequality we use  $\|x_s - x_{s-1}\| = \frac{1}{n^{1/4}c} * \underbrace{\left\| \frac{v_t}{\sqrt{\left(n^{1/2} + \sum_{s=0}^t \|v_s\|^2\right)}} \right\|}_{\leq 1} \leq \frac{1}{n^{1/4}c}$ . then we

939 bound  $\bar{\nabla}_{\mu_t} f(x)$  below:

$$\begin{aligned}
\|\bar{\nabla}_{\mu_t} f(x_t)\| &\leq \|\bar{\nabla}_{\mu_t} f(x_t) - \nabla f(x_t)\| + \|\nabla f(x_t)\| \\
&\leq \delta_t + \|\nabla f(x_t) - \nabla f(x_0)\| + \|\nabla f(x_0)\| \\
&\leq L \|x_t - x_0\| + \delta_t + \|\nabla f(x_0)\| \\
&\leq L \sum_{i=1}^t \|x_i - x_{i-1}\| + \delta_t + \|\nabla f(x_0)\| \\
&\leq \left( \frac{L}{c\sqrt{n}} \right) + \delta_t + \|\nabla f(x_0)\|.
\end{aligned}$$

940 Combining the above results we obtain (Without loss of generality, we set  $\delta_{-1} = \delta_0$ ):

$$\begin{aligned}
\sum_{t=0}^{T-1} \|v_t\|^2 &\leq \sum_{t=0}^{T-1} \left( \frac{4L^2 n^{1.5}}{c^2} + 16n \sum_{s=t-t \bmod n+1}^t (\delta_s^2 + \delta_{s-1}^2) + 2 \cdot \left\| \bar{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2 \right) \\
&\leq \frac{4TL^2 n^{1.5}}{c^2} + 16n \sum_{t=0}^{T-1} \sum_{s=t-t \bmod n+1}^t (\delta_s^2 + \delta_{s-1}^2) + 2 \sum_{t=0}^{T-1} \left\| \bar{\nabla}_{\mu_t} f(x_t) \right\|^2 \\
&\leq \frac{4TL^2 n^{1.5}}{c^2} + 32n^2 \sum_{t=0}^{T-1} \delta_t^2 + 2 \sum_{t=0}^{T-1} \left( \left( \frac{L}{c\sqrt{n}} \right) + \delta_t + \|\nabla f(x_0)\| \right)^2 \\
&\leq \frac{4TL^2 n^{1.5}}{c^2} + (32n^2 + 6) \sum_{t=0}^{T-1} \delta_t^2 + \frac{6L^2 T}{nc^2} + 6T \|\nabla f(x_0)\|^2.
\end{aligned}$$

943 Since this equation will be used repeatedly, we define  $\Phi(T) := \frac{4TL^2 n^{1.5}}{c^2} + (32n^2 + 6) \sum_{t=0}^{T-1} \delta_t^2 +$   
944  $\frac{6L^2 T}{nc^2} + 6T \|\nabla f(x_0)\|^2 - 1$  to simplify the resulting expressions.  $\square$

## 946 C.1 PART I

948 **Lemma C.3** (part I(1)). *Under assumptions 1 and 2, we can derive the following result for Algorithm 949 1:*

$$\mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t\| \right] \leq n^{1/4} \sqrt{T} \left( 2\Delta c + 2c \sum_{t=0}^{T-1} \gamma_t \delta_t^2 + 1 + \frac{L}{c} \log(\Phi(T)) + c \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \|\bar{\nabla}_{\mu_t} f(x_t) - v_t\|^2 \right] \right).$$

972 *Proof.*

$$\begin{aligned}
 974 \quad \mathbb{E}[f(x_{t+1}) \mid \mathcal{F}_t] &\leq \mathbb{E}\left[f(x_t) + \nabla f(x_t)^T(x_{t+1} - x_t) + \frac{L}{2}\|x_t - x_{t+1}\|^2 \mid \mathcal{F}_t\right] \\
 975 \\
 976 \quad &= \mathbb{E}\left[f(x_t) - \gamma_t v_t^T \nabla f(x_t) + \frac{L}{2}\gamma_t^2\|v_t\|^2 \mid \mathcal{F}_t\right] \\
 977 \\
 978 \quad &\leq \mathbb{E}\left[f(x_t) + \frac{\gamma_t}{2}\|v_t - \nabla f(x_t)\|^2 - \frac{\gamma_t}{2}(1 - L\gamma_t)\|v_t\|^2 \mid \mathcal{F}_t\right] \\
 979 \\
 980 \quad &\stackrel{\text{lem C.1}}{\leq} \mathbb{E}\left[f(x_t) + \gamma_t\delta_t^2 + \gamma_t\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 - \frac{\gamma_t}{2}(1 - L\gamma_t)\|v_t\|^2 \mid \mathcal{F}_t\right], \\
 981 \\
 982 \end{aligned}$$

thus we obtain:

$$984 \quad \mathbb{E}[\gamma_t \cdot \|v_t\|^2] \leq 2\mathbb{E}[f(x_t) - f(x_{t+1})] + 2\gamma_t\delta_t^2 + \mathbb{E}[L\gamma_t^2 \cdot \|v_t\|^2] + 2 \cdot \mathbb{E}[\gamma_t \cdot \|\bar{\nabla}_{\mu_t} f(x_t) - v_t\|^2].$$

985 By summing from  $t = 0$  to  $T - 1$  we get:

$$987 \quad \sum_{t=0}^{T-1} \mathbb{E}[\gamma_t \cdot \|v_t\|^2] \leq 2\Delta + 2 \sum_{t=0}^{T-1} \gamma_t\delta_t^2 + \mathbb{E}\left[\sum_{t=0}^{T-1} L\gamma_t^2 \cdot \|v_t\|^2\right] + \mathbb{E}\left[\sum_{t=0}^{T-1} 2\gamma_t \cdot \|\bar{\nabla}_{\mu_t} f(x_t) - v_t\|^2\right].$$

990 Recall that  $\gamma_t = n^{-1/4}c^{-1} \left(n^{1/2} + \sum_{s=0}^t \|v_s\|^2\right)^{-1/2}$  and Lemma B.7 we obtain:

$$\begin{aligned}
 994 \quad \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t \cdot \|v_t\|^2\right] &\leq 2\Delta + 2 \sum_{t=0}^{T-1} \gamma_t\delta_t^2 + \frac{L}{c^2\sqrt{n}} \cdot \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\|v_t\|^2}{\sqrt{n} + \sum_{s=0}^t \|v_s\|^2}\right] \\
 995 \\
 996 \quad &\quad + 2\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t \cdot \|\bar{\nabla}_{\mu_t} f(x_t) - v_t\|^2\right] \\
 997 \\
 998 \quad &\leq 2\Delta + 2 \sum_{t=0}^{T-1} \gamma_t\delta_t^2 + \frac{L}{c^2\sqrt{n}} \log(\Phi(T)) + 2\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t \cdot \|\bar{\nabla}_{\mu_t} f(x_t) - v_t\|^2\right]. \\
 999 \\
 1000 \\
 1001 \\
 1002 \end{aligned}$$

1003 Lower bounding the right-hand side:

$$\begin{aligned}
 1005 \quad \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t \cdot \|v_t\|^2\right] &\geq \mathbb{E}\left[\frac{\sum_{t=0}^{T-1} \|v_t\|^2}{n^{1/4}c\sqrt{n^{1/2} + \sum_{t=0}^{T-1} \|v_t\|^2}}\right] \\
 1006 \\
 1007 \quad &\geq \frac{1}{c} \cdot \mathbb{E}\left[\frac{\sum_{t=0}^{T-1} \|v_t\|^2/\sqrt{n}}{\sqrt{1 + \sum_{t=0}^{T-1} \|v_t\|^2/\sqrt{n}}}\right] \\
 1008 \\
 1009 \quad &\geq \frac{1}{c} \cdot (\mathbb{E}\left[\sqrt{\sum_{t=0}^{T-1} \|v_t\|^2/\sqrt{n}}\right] - 1) \\
 1010 \\
 1011 \quad &\geq \frac{1}{cn^{1/4}\sqrt{T}} \mathbb{E}\left[\sum_{t=0}^{T-1} \|v_t\|\right] - \frac{1}{c}. \\
 1012 \\
 1013 \\
 1014 \\
 1015 \\
 1016 \\
 1017 \\
 1018 \\
 1019 \end{aligned}$$

1020 Combining all results:

$$1022 \quad \mathbb{E}\left[\sum_{t=0}^{T-1} \|v_t\|\right] \leq n^{1/4}\sqrt{T} \left(2\Delta c + 2c \sum_{t=0}^{T-1} \gamma_t\delta_t^2 + 1 + \frac{L}{c\sqrt{n}} \log(\Phi(T)) + c \cdot \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t \cdot \|\bar{\nabla}_{\mu_t} f(x_t) - v_t\|^2\right]\right).$$

1023  $\square$

1026 **Lemma C.4** (part I(2)). *Under assumptions 1 and 2, we can derive the following result for Algorithm*  
 1027 *I:*

$$1028 \quad \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 \right] \leq \frac{2L^2}{c^3} \log(\Phi(T)) + \sum_{t=0}^{T-1} 16n\gamma_t\delta_t^2.$$

1031 *Proof.* Let  $\mathcal{F}_t$  be the sigma-algebra generated by  $\{i_0, \dots, i_t\}$  and  $x_0$ . From the definition of  $\gamma_t$ , it  
 1032 follows that  $\gamma_t \leq \gamma_{t-1}$ ; this condition is imposed to resolve measurability concerns. Consequently,

$$1034 \quad \mathbb{E} [\gamma_t \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 | \mathcal{F}_{t-1}] \leq \mathbb{E} [\gamma_{t-1} \cdot \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 | \mathcal{F}_{t-1}].$$

1035 Hence, our analysis can be reduced to studying  $\mathbb{E} [\gamma_{t-1} \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 | \mathcal{F}_{t-1}]$ .

$$\begin{aligned} 1037 \quad & \mathbb{E} [\gamma_{t-1} \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 | \mathcal{F}_{t-1}] \\ 1038 \quad &= \gamma_{t-1} \mathbb{E} [\| \bar{\nabla}_{\mu_t} f_i(x_t) - \bar{\nabla}_{\mu_{t-1}} f_i(x_{t-1}) - \bar{\nabla}_{\mu_t} f(x_t) + \bar{\nabla}_{\mu_{t-1}} f(x_{t-1}) + (v_{t-1} - \bar{\nabla}_{\mu_{t-1}} f(x_{t-1})) \|^2 | \mathcal{F}_{t-1}] \\ 1039 \quad &= \gamma_{t-1} \mathbb{E} [\| \bar{\nabla}_{\mu_t} f_i(x_t) - \bar{\nabla}_{\mu_{t-1}} f_i(x_{t-1}) - \bar{\nabla}_{\mu_t} f(x_t) + \bar{\nabla}_{\mu_{t-1}} f(x_{t-1}) \|^2 | \mathcal{F}_{t-1}] \\ 1040 \quad & \quad + \gamma_{t-1} \mathbb{E} [\| v_{t-1} - \bar{\nabla}_{\mu_{t-1}} f(x_{t-1}) \|^2 | \mathcal{F}_{t-1}] \\ 1041 \quad &= \gamma_{t-1} \mathbb{E} [\| \bar{\nabla}_{\mu_t} f_i(x_t) - \bar{\nabla}_{\mu_{t-1}} f_i(x_{t-1}) \|^2 | \mathcal{F}_{t-1}] + \gamma_{t-1} \mathbb{E} [\| v_{t-1} - \bar{\nabla}_{\mu_{t-1}} f(x_{t-1}) \|^2 | \mathcal{F}_{t-1}] \\ 1042 \quad &\leq 2L^2 \gamma_{t-1} \mathbb{E} [\|x_t - x_{t-1}\|^2 | \mathcal{F}_{t-1}] + \gamma_{t-1} \mathbb{E} [\|v_{t-1} - \bar{\nabla}_{\mu_{t-1}} f(x_{t-1})\|^2 | \mathcal{F}_{t-1}] + 4\gamma_{t-1}(\delta_t^2 + \delta_{t-1}^2) \\ 1043 \quad &= 2L^2 \gamma_{t-1}^3 \mathbb{E} [\|v_{t-1}\|^2 | \mathcal{F}_{t-1}] + \gamma_{t-1} \mathbb{E} [\|v_{t-1} - \bar{\nabla}_{\mu_{t-1}} f(x_{t-1})\|^2 | \mathcal{F}_{t-1}] + 4\gamma_{t-1}(\delta_t^2 + \delta_{t-1}^2). \end{aligned}$$

1046 We obtain the following by first conditioning on all randomness up to round  $t$ , and then taking the  
 1047 total expectation:

$$1049 \quad \mathbb{E} [\gamma_t \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2] \leq 2L^2 \mathbb{E} [\gamma_{t-1}^3 \|v_{t-1}\|^2] + \mathbb{E} [\gamma_{t-1} \|v_{t-1} - \bar{\nabla}_{\mu_{t-1}} f(x_{t-1})\|^2 + 4\gamma_{t-1}(\delta_t^2 + \delta_{t-1}^2)].$$

1051 Since  $\mathbb{E} [\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|] = 0$  whenever  $t \bmod n = 0$ , it follows that

$$1053 \quad \mathbb{E} [\gamma_t \cdot \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2] \leq 2\mathbb{E} \left[ \sum_{s=t-t \bmod n}^{t-1} L^2 \gamma_s^3 \|v_s\|^2 + 4\gamma_s(\delta_s^2 + \delta_{s-1}^2) \right],$$

1055 which leads to:

$$1056 \quad \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 \right] \leq 2\mathbb{E} \left[ \sum_{t=0}^{T-1} L^2 n \gamma_t^3 \|v_t\|^2 + 8n\gamma_t\delta_t^2 \right],$$

1059 observe that the first term can be bounded by the following terms:

$$\begin{aligned} 1061 \quad & \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t^3 \|v_t\|^2 \right] \\ 1062 \quad &= \frac{1}{c^3} \mathbb{E} \left[ \sum_{t=0}^{T-1} \frac{1}{n^{1/2} + \sum_{s=0}^t \|v_s\|^2} \cdot \frac{\|v_t\|^2}{n^{1/2} + \sum_{s=0}^t \|v_s\|^2} \right] \\ 1063 \quad &\leq \frac{1}{c^3 n} \mathbb{E} \left[ \sum_{t=0}^{T-1} \frac{1}{n^{3/4} \sqrt{n^{1/2}}} \cdot \frac{\|v_t\|^2}{n^{1/2} + \sum_{s=0}^t \|v_s\|^2} \right] \\ 1064 \quad &\leq \frac{1}{c^3 n} \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \frac{\|v_t\|^2}{1 + \sum_{s=0}^t \|v_s\|^2} \right] \\ 1065 \quad &\leq \frac{1}{c^3 n} \log(\Phi(T)), \end{aligned}$$

1074 where the fourth inequality follows by Lemma B.7 and Lemma C.2.

1075 Finally we obtain

$$1077 \quad \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2 \right] \leq \frac{2L^2}{c^3} \log(\Phi(T)) + \sum_{t=0}^{T-1} 16n\gamma_t\delta_t^2.$$

1079  $\square$

1080

## C.2 PART II

1081

1082

**Lemma C.5.** *Under assumptions 1 and 2, we can derive the following result for Algorithm 1:*

1083

1084

1085

1086

1087

1088

*Proof.*

1089

1090

1091

1092

1093

1094

the final inequality holds since  $\left\| \sum_{t=0}^{T-1} a_t \right\|^2$  can be bounded by  $T \cdot \sum_{t=0}^{T-1} \|a_t\|^2$  using Jensen's inequality. By an argument entirely analogous to that of Lemma C.4, we can establish the same result for the estimator  $v_t = \bar{\nabla}_{\mu_t} f_{i_t}(x_t) - \bar{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1}) + v_{t-1}$ :

1095

1096

1097

1098

1099

1100

1101

1102

1103

$$\begin{aligned} \mathbb{E} [\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2] &\leq 4(\delta_t^2 + \delta_{t-1}^2) + 2L^2 \mathbb{E} [\|x_t - x_{t-1}\|^2] + \mathbb{E} [\|v_{t-1} - \bar{\nabla}_{\mu_{t-1}} f(x_{t-1})\|^2] \\ &\leq L^2 \mathbb{E} [\gamma_{t-1}^2 \|v_{t-1}\|^2] + \mathbb{E} [\|v_{t-1} - \bar{\nabla} f(x_{t-1})\|^2] + 4(\delta_t^2 + \delta_{t-1}^2) \\ &= \sum_{\tau=t-(t \bmod n)+1}^{t-1} L^2 \mathbb{E} [\gamma_{\tau}^2 \|v_{\tau}\|^2] + 4(\delta_t^2 + \delta_{t-1}^2), \end{aligned}$$

1104

by a telescoping summation over  $t$  we get that

1105

1106

1107

1108

$$\sum_{t=0}^{T-1} \mathbb{E} [\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2] \leq L^2 n \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t^2 \|v_t\|^2 \right] + 8n \sum_{t=0}^{T-1} \delta_t^2.$$

1109

1110

Now as discussed in Lemma C.4, using the step-size selection  $\gamma_t$  we obatain:

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E} [\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|^2] &\leq L^2 n \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t^2 \|v_t\|^2 \right] + 8n \sum_{t=0}^{T-1} 2\delta_t^2 \\ &= \frac{L^2 \sqrt{n}}{c^2} \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \frac{\|v_t\|^2}{\sqrt{n} + \sum_{s=0}^t \|\bar{\nabla}_s\|^2} \right] + 8n \sum_{t=0}^{T-1} \delta_t^2 \\ &\leq \frac{L^2 \sqrt{n}}{c^2} \log(\Phi(T)) + 8n \sum_{t=0}^{T-1} \delta_t^2 \\ &\leq \frac{L^2 \sqrt{n}}{c^2} \log(\Phi(T)) + 8n \sum_{t=0}^{T-1} 2\delta_t^2, \end{aligned}$$

1124

1125

where lstonequality follows by Lemma B.6 and Lemma C.2. Putting everything together we get

1126

1127

1128

1129

1130

1131

1132

## C.3 PART III

1133

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x_t) - \bar{\nabla}_{\mu_t} f(x_t)\| \leq \frac{1}{T} \sum_{t=0}^{T-1} \delta_t.$$

□

1134 C.4 FINAL PROOF FOR COORDINATE ESTIMATOR  
11351136 **Theorem C.1.** *Under assumptions 1 and 2, based on the previous lemmas C.3, C.4, C.5, we can*  
1137 *derive the following result for Algorithm 1:*

1139 
$$1140 \mathbb{E}\left[\sum_{t=0}^{T-1} \|\nabla f(x_t)\|\right] \leq \frac{n^{1/4}}{\sqrt{T}} \left( 2\Delta \cdot c + 1 + \left(\frac{L}{c} + \frac{L^2}{c^2}\right) \log(\Phi(T)) + \frac{L^2\pi^2}{24n^{1/4}} + \sqrt{\frac{\pi^2}{24}} \frac{L}{n^{1/4}} + \frac{L^2\pi^2}{12} + \frac{L}{2} \right)$$
  
1142

1143 setting  $c = 1$ , we obtain  $T = \tilde{\mathcal{O}}(\sqrt{n}\epsilon^{-2})$ , where the  $\tilde{\mathcal{O}}$  notation hides logarithmic factors.  
11441145  
1146 *Proof.*  
1147

1148 
$$1149 \mathbb{E}\left[\sum_{t=0}^{T-1} \|\nabla f(x_t)\|\right] \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|] + \sum_{t=0}^{T-1} \mathbb{E}[\|v_t - \bar{\nabla}_{\mu_t} f(x_t)\|] + \|\bar{\nabla}_{\mu_t} f(x_t) - \nabla f(x_t)\|$$
  
1150 
$$1151 \leq \frac{n^{1/4}}{\sqrt{T}} \left( 2\Delta \cdot c + 1 + \left(\frac{L}{cn^{3/4}} + \frac{L^2}{c^2}\right) \log(\Phi(T)) \right)$$
  
1152 
$$1153 + \frac{1}{T} (2c \sum_{t=0}^{T-1} \gamma_t \delta_t^2 + \sum_{t=0}^{T-1} \delta_t) + \frac{1}{\sqrt{T}} \left( n^{\frac{5}{4}} \sum_{t=0}^{T-1} c \gamma_t \delta_t^2 + n^{\frac{1}{2}} \sqrt{\sum_{t=0}^{T-1} 2\delta_t^2} \right).$$
  
1154

1155 Due to the fact that  $\gamma_t \leq \frac{1}{cn^{1/4}}$  we obtain:  
1156

1157 
$$1158 \mathbb{E}\left[\sum_{t=0}^{T-1} \|\nabla f(x_t)\|\right] \leq \frac{n^{1/4}}{\sqrt{T}} \left( 2\Delta \cdot c + 1 + \left(\frac{L}{cn^{3/4}} + \frac{L^2}{c^2}\right) \log(\Phi(T)) \right)$$
  
1159 
$$1160 + \frac{1}{T} (2 \sum_{t=0}^{T-1} \frac{1}{n^{1/4}} \delta_t^2 + \sum_{t=0}^{T-1} \delta_t) + \frac{1}{\sqrt{T}} \left( n \sum_{t=0}^{T-1} \delta_t^2 + n^{\frac{1}{2}} \sqrt{\sum_{t=0}^{T-1} 2\delta_t^2} \right).$$
  
1161

1162 Take  $\delta_t = \frac{L}{2\sqrt{n(t+1)}}$  i.e.  $(\mu_t = \frac{1}{\sqrt{nd}}(t+1))$  then :  
1163

1164 
$$1165 \sum_{t=0}^{T-1} \delta_t \leq \frac{L \ln T}{\sqrt{n}}.$$
  
1166

1167 
$$1168 \sum_{t=0}^{T-1} \delta_t^2 < \frac{L^2 \pi^2}{24n}.$$
  
1169

1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
2212  
2213  
2214  
2215  
2216  
2217  
2218  
2219  
2220  
2221  
2222  
2223  
2224  
2225  
2226  
2227  
2228  
2229  
2230  
2231  
2232  
2233  
2234  
2235  
2236  
2237  
2238  
2239  
2240  
2241  
2242  
2243  
2244  
2245  
2246  
2247  
2248  
2249  
2250  
2251  
2252  
2253  
2254  
2255  
2256  
2257  
2258  
2259  
2260  
2261  
2262  
2263  
2264  
2265  
2266  
2267  
2268  
2269  
2270  
2271  
2272  
2273  
2274  
2275  
2276  
2277  
2278  
2279  
2280  
2281  
2282  
2283  
2284  
2285  
2286  
2287  
2288  
2289  
2290  
2291  
2292  
2293  
2294  
2295  
2296  
2297  
2298  
2299  
2300  
2301  
2302  
2303  
2304  
2305  
2306  
2307  
2308  
2309  
2310  
2311  
2312  
2313  
2314  
2315  
2316  
2317  
2318  
2319  
2320  
2321  
2322  
2323  
2324  
2325  
2326  
2327  
2328  
2329  
2330  
2331  
2332  
2333  
2334  
2335  
2336  
2337  
2338  
2339  
2340  
2341  
2342  
2343  
2344  
2345  
2346  
2347  
2348  
2349  
2350  
2351  
2352  
2353  
2354  
2355  
2356  
2357  
2358  
2359  
2360  
2361  
2362  
2363  
2364  
2365  
2366  
2367  
2368  
2369  
2370  
2371  
2372  
2373  
2374  
2375  
2376  
2377  
2378  
2379  
2380  
2381  
2382  
2383  
2384  
2385  
2386  
2387  
2388  
2389  
2390  
2391  
2392  
2393  
2394  
2395  
2396  
2397  
2398  
2399  
2400  
2401  
2402  
2403  
2404  
2405  
2406  
2407  
2408  
2409  
2410  
2411  
2412  
2413  
2414  
2415  
2416  
2417  
2418  
2419  
2420  
2421  
2422  
2423  
2424  
2425  
2426  
2427  
2428  
2429  
2430  
2431  
2432  
2433  
2434  
2435  
2436  
2437  
2438  
2439  
2440  
2441  
2442  
2443  
2444  
2445  
2446  
2447  
2448  
2449  
2450  
2451  
2452  
2453  
2454  
2455  
2456  
2457  
2458  
2459  
2460  
2461  
2462  
2463  
2464  
2465  
2466  
2467  
2468  
2469  
2470  
2471  
2472  
2473  
2474  
2475  
2476  
2477  
2478  
2479  
2480  
2481  
2482  
2483  
2484  
2485  
2486  
2487  
2488  
2489  
2490  
2491  
2492  
2493  
2494  
2495  
2496  
2497  
2498  
2499  
2500  
2501  
2502  
2503  
2504  
2505  
2506  
2507  
2508  
2509  
2510  
2511  
2512  
2513  
2514  
2515  
2516  
2517  
2518  
2519  
2520  
2521  
2522  
2523  
2524  
2525  
2526  
2527  
2528  
2529  
2530  
2531  
2532  
2533  
2534  
2535  
2536  
2537  
2538  
2539  
2540  
2541  
2542  
2543  
2544  
2545  
2546  
2547  
2548  
2549  
2550  
2551  
2552  
2553  
2554  
2555  
2556  
2557  
2558  
2559  
2560  
2561  
2562  
2563  
2564  
2565  
2566  
2567  
2568  
2569  
2570  
2571  
2572  
2573  
2574  
2575  
2576  
2577  
2578  
2579  
2580  
2581  
2582  
2583  
2584  
2585  
2586  
2587  
2588  
2589  
2590  
2591  
2592  
2593  
2594  
2595  
2596  
2597  
2598  
2599  
2600  
2601  
2602  
2603  
2604  
2605  
2606  
2607  
2608  
2609  
2610  
2611  
2612  
2613  
2614  
2615  
2616  
2617  
2618  
2619  
2620  
2621  
2622  
2623  
2624  
2625  
2626  
2627  
2628  
2629  
2630  
2631  
2632  
2633  
2634  
2635  
2636  
2637  
2638  
2639  
2640  
2641  
2642  
2643  
2644  
2645  
2646  
2647  
2648  
2649  
2650  
2651  
2652  
2653  
2654  
2655  
2656  
2657  
2658  
2659  
2660  
2661  
2662  
2663  
2664  
2665  
2666  
2667  
2668  
2669  
2670  
2671  
2672  
2673  
2674  
2675  
2676  
2677  
2678  
2679  
2680  
2681  
2682  
2683  
2684  
2685  
2686  
2687  
2688  
2689  
2690  
2691  
2692  
2693  
2694  
2695  
2696  
2697  
2698  
2699  
2700  
2701  
2702  
2703  
2704  
2705  
2706  
2707  
2708  
2709  
2710  
2711  
2712  
2713  
2714  
2715  
2716  
2717  
2718  
2719  
2720  
2721  
2722  
2723  
2724  
2725  
2726  
2727  
2728  
2729  
2730  
2731  
2732  
2733  
2734  
2735  
2736  
2737  
2738  
2739  
2740  
2741  
2742  
2743  
2744  
2745  
2746  
2747  
2748  
2749  
2750  
2751  
2752  
2753  
2754  
2755  
2756  
2757  
2758  
2759  
2760  
2761  
2762  
2763  
2764  
2765  
2766  
2767  
2768  
2769  
2770  
2771  
2772  
2773  
2774  
2775  
2776  
2777  
2778  
2779  
2780  
2781  
2782  
2783  
2784  
2785  
2786  
2787  
2788  
2789  
2790  
2791  
2792  
2793  
2794  
2795  
2796  
2797  
27

## D PARAMETER FREE VARIANCE REDUCED ZEROTH-ORDER METHOD (RANDOM-DIRECTION ESTIMATOR)

**Algorithm 4** PF-VRZO(Random-direction)

```

1193 Set  $c = \sqrt{d}$  for random-direction estimator and  $\mu_{-1} = \mu_0$ .
1194 for  $t = 0$  to  $T-1$  do
1195   Compute smoothing parameter  $\mu_t = \frac{1}{(t+1)d\sqrt{n}}$ , smoothing vector  $\rho_t \sim U_B$ .
1196   if  $t \bmod n = 0$  then
1197      $v_t = \hat{\nabla}_{\mu_t} f(x_t)$  {Full zeroth-order gradient computation}
1198   else
1199     Sample  $i_t \in \{1, \dots, n\}$  uniformly at random
1200     Compute  $\hat{\nabla}_{\mu_t} f_{i_t}(x_t)$  with parameter  $\mu_t$  and rand vector  $\rho_t$ ,  $\hat{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1})$  with different
1201     parameter  $\mu_{t-1}$  and the same rand vector  $\rho_t$ .
1202      $v_t = \hat{\nabla}_{\mu_t} f_{i_t}(x_t) - \hat{\nabla}_{\mu_{t-1}} f_{i_t}(x_{t-1}) + v_{t-1}$ 
1203   end if
1204    $\gamma_t = \frac{1}{n^{1/4}c\sqrt{(n^{1/2} + \sum_{s=0}^t \|v_s\|^2)}}$ 
1205    $x_{t+1} = x_t - \gamma_t v_t$ 
1206 end for

```

Table 3: Meaning of Symbols

Symbol	Meaning
$\gamma_t$	stepsize $\frac{1}{(n^{1/4}c\sqrt{n^{1/2} + \sum_{s=0}^t \ v_s\ ^2})}.$
$\mu_t$	Smoothing parameter at iteration $t$ .
$\rho_t$	Smoothing vector at iteration $t$ .
$v_t$	Spider operator.
$\nabla_{\mu_t} f(x_t)$	zeroth-order estimator(rand) using smoothing parameter $\mu_t$ and $\rho_t$ .
$\hat{\nabla} f_\mu(\cdot)$	expecation of zeroth-order estimator(rand).
$\Delta_t$	$Ld\mu_t/2$ , the estimation error with respect to $\nabla f_\mu(\cdot)$ .

Following a similar approach as with the coordinate operator, we analyze the convergence of the gradient of  $f(x)$  by dividing it into three parts.

$$\frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] \leq \underbrace{\frac{1}{T} \left[ \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|] \right]}_{\text{part I}} + \underbrace{\sum_{t=0}^{T-1} \mathbb{E}[\|v_t - \nabla f_{\mu_t}(x)\|]}_{\text{part II}} + \underbrace{\sum_{t=0}^{T-1} \|\nabla f_{\mu_t}(x) - \nabla f(x_t)\|}_{\text{part III}}.$$

**Lemma D.1** ([\(Ji et al., 2019\)](#)). Let  $f_\mu(x) = \mathbb{E}_{w \sim U_B}[f(x + \mu w)]$  be a smooth approximation of  $f(x)$ , where  $U_B$  is the uniform distribution over the  $d$ -dimensional unit Euclidean ball  $B$ , and  $\rho \in \mathbb{R}^d$  is a random vector sampled from unit Euclidean sphere  $U_{S_n}$ . Then we have

1.  $|f_\mu(x) - f(x)| \leq \frac{\mu^2 L}{2}$  and  $\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\mu L d}{2}$  for any  $x \in \mathbb{R}^d$ .
2.  $\mathbb{E}\|\hat{\nabla}_\mu f_i(x_1) - \hat{\nabla}_\mu f_i(x_2)\|^2 \leq 3dL^2\|x_1 - x_2\|^2 + \frac{3L^2d^2\mu^2}{2}$  for any  $i$  and any  $x_1, x_2 \in \mathbb{R}^d$ .
3.  $\mathbb{E}_{\rho \sim U_{S_p}} \left[ \|\hat{\nabla} f(x)\|^2 \right] \leq 2d\|\nabla f(x)\|^2 + \frac{L^2\mu^2d^2}{2}$ .

1242    **Lemma D.2.** For random-direction estimator  $\hat{\nabla}_{\mu_t} f_i(x_t) = \frac{d}{\mu_t} [f(x_t + \mu_t \rho_t) - f(x_t)] \rho_t$ ,  $\hat{\nabla}_{\mu_{t-1}} f_i(x_{t-1}) = \frac{d}{\mu_{t-1}} [f(x + \mu_{t-1} \rho_t) - f(x_{t-1})] \rho_t$ , where both estimators use the  
 1243 same random direction  $\rho_t$  sampled from unit Euclidean sphere  $U_{S_p}$  but different smoothing parameters  
 1244  $\mu_t$  and  $\mu_{t-1}$ , we have:  
 1245

$$1246 \quad \|\hat{\nabla}_{\mu_t} f(x_t) - \hat{\nabla}_{\mu_{t-1}} f(x_{t-1})\|^2 \leq \frac{3}{2} (\Delta_t^2 + \Delta_{t-1}^2) + 3dL^2 \|x_t - x_{t-1}\|^2.$$

1247    *Proof.*

$$\begin{aligned} 1248 \quad & \mathbb{E} [\|\hat{\nabla}_{\mu_t} f(x_t) - \hat{\nabla}_{\mu_{t-1}} f(x_t)\|^2] \\ 1249 \quad &= d^2 \mathbb{E} \left[ \left\| \frac{\rho_t}{\mu_t} \left[ f(x_t + \mu_t \rho_t) - f(x_t) - \langle \nabla f(x_t), \rho_t \rangle \right] \rho_t - \frac{\rho_t}{\mu_{t-1}} \left[ f(x_{t-1} + \mu_{t-1} \rho_t) - f(x_{t-1}) - \langle \nabla f(x_{t-1}), \rho_t \rangle \right] \rho_t \right. \right. \\ 1250 \quad & \quad \left. \left. + \left( \langle \nabla f(x_t), \rho_t \rangle - \langle \nabla f(x_{t-1}), \rho_t \rangle \right) \rho_t \right\|^2 \right] \\ 1251 \quad & \leq d^2 \left( \frac{3L^2}{2} (\mu_t^2 + \mu_{t-1}^2) + \mathbb{E} [3\|\langle \nabla f(x_t), \rho_t \rangle - \langle \nabla f(x_{t-1}), \rho_t \rangle\|^2] \right) \\ 1252 \quad & = d^2 \left( \frac{3L^2}{2} (\mu_t^2 + \mu_{t-1}^2) + \mathbb{E} [3\|\langle \nabla f(x_{t-1}) - \nabla f(x_t), \rho_t \rangle\|^2] \right) \quad (\|\rho_t\|^2 = 1) \\ 1253 \quad & \leq d^2 \left( \frac{3L^2}{2} (\mu_t^2 + \mu_{t-1}^2) + \mathbb{E} \left[ \frac{3}{d} \|\nabla f(x_{t-1}) - \nabla f(x_t)\|^2 \right] \right) \quad (\mathbb{E}[\rho_t \rho_t^T] = \frac{1}{d} I_d \text{ (Ji et al., 2019)}) \\ 1254 \quad & \leq \frac{3}{2} (\Delta_t^2 + \Delta_{t-1}^2) + 3dL^2 \|x_t - x_{t-1}\|^2. \\ 1255 \quad & \end{aligned}$$

□

1256    **Lemma D.3.** Under assumptions 1 and 2, we can derive the following result for Algorithm 2

$$1257 \quad \mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t\|^2 \right] \leq \phi(T) + 1.$$

1258    where  $\phi(T) := \frac{6dL^2 n^{1.5}}{c^2} T + \frac{4dL^2 T^3}{n} c^2 + 4dnT \|\nabla f(x_0)\|^2 + (6n^2 + 2) \sum_{t=0}^{T-1} \Delta_t^2 - 1$ . Similar to  
 1259 the coordinate method, the notation  $\phi(T)$  is introduced only for brevity, and will be repeatedly used  
 1260 in the subsequent analysis.

1261    *Proof.*

$$\begin{aligned} 1262 \quad & \mathbb{E} \|v_t\|^2 = \left\| \sum_{s=t-t \bmod n+1}^t \left( \hat{\nabla}_{\mu_s} f_{i_s}(x_s) - \hat{\nabla}_{\mu_{s-1}} f_{i_s}(x_{s-1}) \right) + \hat{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2 \\ 1263 \quad & \leq 2 \mathbb{E} \left\| \sum_{s=t-t \bmod n+1}^t \hat{\nabla}_{\mu_s} f_{i_s}(x_s) - \hat{\nabla}_{\mu_{s-1}} f_{i_s}(x_{s-1}) \right\|^2 + 2 \mathbb{E} \left\| \hat{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2 \\ 1264 \quad & \leq 2n \mathbb{E} \sum_{s=t-t \bmod n+1}^t \left\| \hat{\nabla}_{\mu_s} f_{i_s}(x_s) - \hat{\nabla}_{\mu_{s-1}} f_{i_s}(x_{s-1}) \right\|^2 + 2 \mathbb{E} \left\| \hat{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2 \\ 1265 \quad & \stackrel{\text{lem D.2}}{\leq} 6n \sum_{s=t-t \bmod n+1}^t [dL^2 (x_s - x_{s-1})^2 + \frac{1}{2} (\Delta_s^2 + \Delta_{s-1}^2)] + 2 \mathbb{E} \left\| \hat{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2 \\ 1266 \quad & \leq \frac{6dL^2 n^2}{c^2} + 3n \sum_{s=t-t \bmod n+1}^t (\Delta_s^2 + \Delta_{s-1}^2) + 2 \mathbb{E} \left\| \hat{\nabla}_{\mu_{t-t \bmod n}} f(x_{t-t \bmod n}) \right\|^2, \\ 1267 \quad & \end{aligned}$$

1268    from lemma D.1:

$$\mathbb{E}_{\rho \sim U_{S_p}} [\|\hat{\nabla}_{\mu_t} f(x)\|^2] \leq 2d \|\nabla f(x)\|^2 + \Delta_t^2.$$

1296 Next, we bound  $\|\nabla f(x)\|$  bellow:  
1297

$$\begin{aligned}
1298 \|\nabla f(x_t)\| &= \|\nabla f(x_t) - \nabla f(x_0) + \nabla f(x_0)\| \\
1299 &\leq \|\nabla f(x_t) - \nabla f(x_0)\| + \|\nabla f(x_0)\| \\
1300 &\leq L\|x_t - x_0\| + \|\nabla f(x_0)\| \\
1301 &\leq L\|x_t - x_{t-1}\| + L\|x_{t-1} - x_0\| + \|\nabla f(x_0)\| \\
1302 &\leq L \sum_{i=1}^t \|x_i - x_{i-1}\| + \|\nabla f(x_0)\| \\
1303 &\leq \frac{Lt}{c\sqrt{n}} + \|\nabla f(x_0)\|.
\end{aligned}$$

1308 Combine the above results, we have:  
1309

$$\begin{aligned}
1310 \mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t\|^2 \right] &\leq \sum_{t=0}^{T-1} \left( \frac{6L^2n}{c^2} + 3n \sum_{s=t-t \bmod n+1}^t (\Delta_s^2 + \Delta_{s-1}^2) + \mathbb{E}[2 \|\hat{\nabla} f_\mu(x_{t-t \bmod n})\|^2] \right) \\
1311 &\leq \frac{6dTL^2n}{c^2} + 2d \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 + 3n \sum_{t=0}^{T-1} \sum_{s=t-t \bmod n+1}^t (\Delta_s^2 + \Delta_{s-1}^2) + 2 \sum_{t=0}^{T-1} \Delta_t^2 \\
1312 &\leq \frac{6L^2dn}{c^2} T + 2d \sum_{t=0}^{T-1} \left( \frac{Lt}{\sqrt{nc}} + \|\nabla f(x_0)\| \right)^2 + (6n^2 + 2) \sum_{t=0}^{T-1} \Delta_t^2 \\
1313 &\leq \frac{6dL^2n}{c^2} T + \frac{4dL^2T^3}{n} c^2 + 4dnT \|\nabla f(x_0)\|^2 + (6n^2 + 2) \sum_{t=0}^{T-1} \Delta_t^2.
\end{aligned}$$

1322 Similar to the coordwise method, we define  $\phi(T) := \frac{6dL^2n^{1.5}}{c^2} T + \frac{4dL^2T^3}{n} c^2 + 4dnT \|\nabla f(x_0)\|^2 +$   
1323  $(6n^2 + 2) \sum_{t=0}^{T-1} \Delta_t^2 - 1$  to simplify the resulting expressions.  $\square$   
1324

## 1325 D.1 PART I

1327 **Lemma D.4** (part I(1)). *Under assumptions 1 and 2, we can derive the following result for Algorithm 2*

$$1330 \mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t\| \right] \leq n^{1/4} \sqrt{T} \left( 2\Delta c + 1 + \frac{L}{c} \log(\phi(T)) + c \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \|\nabla f(x_t) - v_t\|^2 \right] \right)$$

1333 *Proof.*

$$\begin{aligned}
1335 \mathbb{E} [f(x_{t+1}) \mid \mathcal{F}_t] &\leq \mathbb{E} \left[ f(x_t) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \|x_t - x_{t+1}\|^2 \mid \mathcal{F}_t \right] \\
1336 &= \mathbb{E} \left[ f(x_t) - \gamma_t v_t^T \nabla f(x_t) + \frac{L}{2} \gamma_t^2 \|v_t\|^2 \mid \mathcal{F}_t \right] \\
1337 &\leq \mathbb{E} \left[ f(x_t) + 2\gamma_t \|v_t - \nabla f(x_t)\|^2 - \frac{\gamma_t}{2} (1 - L\gamma_t) \|v_t\|^2 \mid \mathcal{F}_t \right],
\end{aligned}$$

1342 which leads to:

$$1343 \mathbb{E} [\gamma_t \cdot \|v_t\|^2] \leq 2\mathbb{E} [f(x_t) - f(x_{t+1})] + \mathbb{E} [L\gamma_t^2 \cdot \|v_t\|^2] + 2c \cdot \mathbb{E} [\gamma_t \cdot \|\nabla f(x_t) - v_t\|^2].$$

1345 By summing from  $t = 0$  to  $T - 1$  we get:

$$1348 \sum_{t=0}^{T-1} \mathbb{E} [\gamma_t \cdot \|v_t\|^2] \leq 2\Delta + \mathbb{E} \left[ \sum_{t=0}^{T-1} L\gamma_t^2 \cdot \|v_t\|^2 \right] + \mathbb{E} \left[ \sum_{t=0}^{T-1} 2\gamma_t \cdot \|\nabla f(x_t) - v_t\|^2 \right].$$

1350

Recall that  $\gamma_t = n^{-1/4}c^{-1} \left( n^{1/2} + \sum_{s=0}^t \|v_s\|^2 \right)^{-1/2}$ :

1351

1352

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|v_t\|^2 \right] &\leq 2\Delta + \frac{L}{c^2} \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \frac{\|v_t\|^2}{\sqrt{n} + \sum_{s=0}^t \|v_s\|^2} \right] + \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|\nabla f(x_t) - v_t\|^2 \right] \\ &\leq 2\Delta + \frac{L}{\sqrt{nc^2}} \log(\phi(T)) + \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|\nabla f(x_t) - v_t\|^2 \right]. \end{aligned}$$

1353

1354

1355

Lower bounding the right-hand side:

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

Combining all results:

1377

1378

1379

1380

1381

1382

1383

1384

1385

**Lemma D.5** (part I(2)). *Under assumptions 1 and 2, we can derive the following result for Algorithm 2*

1386

1387

1388

1389

1390

1391

1392

1393

*Proof.* Let  $\mathcal{F}_t$  be the sigma-algebra generated by  $\{i_0, \dots, i_t\}$  and  $x_0$ . From the definition of  $\gamma_t$ , it follows that  $\gamma_t \leq \gamma_{t-1}$ ; this condition is imposed to resolve measurability concerns. Consequently,

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

$$\mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|v_t - \nabla f(x_t)\|^2 \right] \leq \frac{6dL^2}{c^3} \log(\phi(T)) + \sum_{t=0}^{T-1} (3n + 8)n\gamma_t\Delta_t^2.$$

□

$$\mathbb{E} [\gamma_t \|v_t - \nabla f(x_t)\|^2 | \mathcal{F}_{t-1}] \leq \mathbb{E} [\gamma_{t-1} \|v_t - \nabla f(x_t)\|^2 | \mathcal{F}_{t-1}],$$

Hence, our analysis can be reduced to studying  $\mathbb{E} [\gamma_{t-1} \|v_t - \nabla f(x_t)\|^2 | \mathcal{F}_{t-1}]$ .

$$\begin{aligned} \mathbb{E} [\gamma_{t-1} \|v_t - \nabla f(x_t)\|^2 | \mathcal{F}_{t-1}] \\ \leq 2\mathbb{E} [\gamma_{t-1} \|v_t - \nabla f_{\mu_t}(x_t)\|^2 | \mathcal{F}_{t-1}] + 2\mathbb{E} [\gamma_{t-1} \|\nabla f(x_t) - \nabla f_{\mu_t}(x_t)\|^2 | \mathcal{F}_{t-1}]. \end{aligned}$$

1404 As established in Lemma D.1, the second term can be bounded by  $4\gamma_{t-1}\Delta_t^2$ . In the following, we  
 1405 focus on the analysis of the first term.  
 1406

$$\begin{aligned}
 & \mathbb{E} [\gamma_{t-1}\|v_t - \nabla f_{\mu_t}(x_t)\|^2 \mid \mathcal{F}_{t-1}] \\
 &= \gamma_{t-1}\mathbb{E} [\|\hat{\nabla}_{\mu_t}f_{i_t}(x_t) - \hat{\nabla}_{\mu_{t-1}}f_{i_t}(x_{t-1}) - \nabla f_{\mu_t}(x_t) + \nabla f_{\mu_{t-1}}(x_{t-1}) + (v_{t-1} - \nabla f_{\mu_{t-1}}(x_{t-1}))\|^2 \mid \mathcal{F}_{t-1}] \\
 &= \gamma_{t-1}\mathbb{E} [\|\hat{\nabla}_{\mu_t}f_{i_t}(x_t) - \hat{\nabla}_{\mu_{t-1}}f_{i_t}(x_{t-1}) - \nabla f_{\mu_{t-1}}(x_{t-1}) + (v_{t-1} - \nabla f_{\mu_{t-1}}(x_{t-1}))\|^2 \mid \mathcal{F}_{t-1}] \\
 &\quad + \gamma_{t-1}\mathbb{E} [\|v_{t-1} - \nabla f(x_{t-1})\|^2 \mid \mathcal{F}_{t-1}] \\
 &\leq \gamma_{t-1}\mathbb{E} [\|\hat{\nabla}_{\mu_t}f_{i_t}(x_t) - \hat{\nabla}_{\mu_{t-1}}f_{i_t}(x_{t-1})\|^2 \mid \mathcal{F}_{t-1}] + \gamma_{t-1}\mathbb{E} [\|v_{t-1} - \nabla f_{\mu_{t-1}}(x_{t-1})\|^2 \mid \mathcal{F}_{t-1}] \\
 &\stackrel{\text{Lem D.2}}{\leq} 3dL^2\gamma_{t-1}\mathbb{E} [\|x_t - x_{t-1}\|^2 \mid \mathcal{F}_{t-1}] + \gamma_{t-1}\mathbb{E} [\|v_{t-1} - \nabla f_{\mu_{t-1}}(x_{t-1})\|^2 \mid \mathcal{F}_{t-1}] + \frac{3\gamma_{t-1}}{2}(\Delta_t^2 + \Delta_{t-1}^2) \\
 &= 3dL^2\gamma_{t-1}^3\mathbb{E} [\|v_{t-1}\|^2 \mid \mathcal{F}_{t-1}] + \gamma_{t-1}\mathbb{E} [\|v_{t-1} - \nabla f_{\mu_{t-1}}(x_{t-1})\|^2 \mid \mathcal{F}_{t-1}] + \frac{3\gamma_{t-1}}{2}(\Delta_t^2 + \Delta_{t-1}^2).
 \end{aligned} \tag{4}$$

1420 We obtain the following by first conditioning on all randomness up to round  $t$ , and then taking the  
 1421 total expectation:

$$\mathbb{E} [\gamma_t\|v_t - \nabla f_{\mu_t}(x_t)\|^2] \leq \mathbb{E} [\gamma_{t-1}\|v_{t-1} - \nabla f_{\mu_{t-1}}(x_{t-1})\|^2] + 3dL^2\mathbb{E} [\gamma_{t-1}^3\|v_{t-1}\|^2] + \frac{3\gamma_{t-1}}{2}(\Delta_t^2 + \Delta_{t-1}^2).$$

1425 Since  $\mathbb{E} [\gamma_t \cdot \|v_t - \nabla f_{\mu_t}(x_t)\|^2] \leq \gamma_{t-1}\mathbb{E} [\|v_t - \nabla f_{\mu_t}(x_t)\|^2] = 0$  whenever  $t \bmod n = 0$ , it  
 1426 follows that

$$\mathbb{E} [\gamma_t \cdot \|v_t - \nabla f_{\mu_t}(x_t)\|^2] \leq \mathbb{E} \left[ \sum_{s=t-t \bmod n}^{t-1} 3dL^2\gamma_s^3\|v_s\|^2 + \frac{3\gamma_s}{2}(\Delta_s^2 + \Delta_{s+1}^2) \right].$$

1431 Combine the above results we obtain:

$$\begin{aligned}
 & \mathbb{E} [\gamma_{t-1}\|v_t - \nabla f(x_t)\|^2 \mid \mathcal{F}_{t-1}] \\
 &\stackrel{\text{Lem D.1}}{\leq} 2\mathbb{E} [\gamma_{t-1}\|v_t - \nabla f_{\mu_t}(x_t)\|^2 \mid \mathcal{F}_{t-1}] + 2\mathbb{E} [\gamma_{t-1}\|\nabla f(x_t) - \nabla f_{\mu_t}(x_t)\|^2 \mid \mathcal{F}_{t-1}] \\
 &\leq \mathbb{E} \left[ \sum_{s=t-t \bmod n}^{t-1} 6dL^2\gamma_s^3\|v_s\|^2 + 3\gamma_{s-1}(\Delta_s^2 + \Delta_{s+1}^2) \right] + 8\gamma_{t-1}\Delta_t^2,
 \end{aligned}$$

1439 summing over  $t$  from 0 to  $T-1$  we get that

$$\mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|v_t - \nabla f(x_t)\|^2 \right] \leq \mathbb{E} \left[ \sum_{t=0}^{T-1} 6dL^2n\gamma_t^3\|v_t\|^2 + (3n+8)\gamma_t\Delta_t^2 \right].$$

1444 Observe that the first term can be bounded by the following terms:

$$\begin{aligned}
 & \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t^3\|v_t\|^2 \right] \\
 &= \frac{1}{c^3} \mathbb{E} \left[ \sum_{t=0}^{T-1} \frac{1}{n^{1/2} + \sum_{s=0}^t \|v_s\|^2} \cdot \frac{\|v_t\|^2}{n^{1/2} + \sum_{s=0}^t \|v_s\|^2} \right] \\
 &\leq \frac{1}{c^3n} \mathbb{E} \left[ \sum_{t=0}^{T-1} \frac{1}{n^{3/4}\sqrt{n^{1/2}}} \cdot \frac{\|v_t\|^2}{n^{1/2} + \sum_{s=0}^t \|v_s\|^2} \right] \\
 &\leq \frac{1}{c^3n} \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \frac{\|v_t\|^2}{1 + \sum_{s=0}^t \|v_s\|^2} \right] \\
 &\leq \frac{1}{c^3n} \log(\phi(T)),
 \end{aligned}$$

1458 where the fourth inequality follows by Lemma B.7 and Lemma D.3.  
 1459  
 1460 Finally we obtain

$$1461 \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t \cdot \|v_t - \nabla f(x_t)\|^2 \right] \leq \frac{6dL^2}{c^3} \log(\phi(T)) + \sum_{t=0}^{T-1} (3n+8)n\gamma_t \Delta_t^2.$$

1464 □  
 1465

## 1466 D.2 PART II

1467 **Lemma D.6.** *Under assumptions 1 and 2, we can derive the following result for Algorithm 2*

$$1469 \mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t - \nabla f_{\mu_t}(x_t)\| \right] \leq \frac{6\sqrt{d}Ln^{1/4}}{c\sqrt{T}} \log(\phi(T)) + \frac{1}{\sqrt{T}} \sqrt{(3n+8) \sum_{t=0}^{T-1} 2\Delta_t^2}.$$

1472  
 1473 *Proof.*

$$1474 \mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t - \nabla f_{\mu_t}(x_t)\| \right] \leq \sqrt{T} \cdot \sqrt{\mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t - \nabla f_{\mu_t}(x_t)\|^2 \right]},$$

1475 where the inequality follows by the fact that  $\|\sum_{t=0}^{T-1} y_t\|^2 \leq T \cdot \sum_{t=0}^{T-1} \|y_t\|^2$ . For the same reason  
 1476 with equation 4, we obtain:

$$1477 \mathbb{E} [\|v_t - \nabla f_{\mu_t}(x_t)\|^2] \leq \mathbb{E} \left[ \sum_{s=t-t \bmod n}^{t-1} 6dL^2\gamma_s^2\|v_s\|^2 + 3(\Delta_s^2 + \Delta_{s+1}^2) \right] + 8\Delta_t^2,$$

1478 by a telescoping summation over  $t$  we get that

$$1479 \sum_{t=0}^{T-1} \mathbb{E} [\|v_t - \nabla f_{\mu_t}(x_t)\|^2] \leq \mathbb{E} \left[ \sum_{t=0}^{T-1} 6dL^2n\gamma_t^2\|v_t\|^2 + (3n+8)\Delta_t^2 \right].$$

1480 Using the step-size selection  $\gamma_t$  we can provide a bound on the total variance  $\mathbb{E} [\|v_t - \nabla f_{\mu_t}(x_t)\|^2]$ :

$$1481 \begin{aligned} & \sum_{t=0}^{T-1} \mathbb{E} [\|v_t - \nabla f_{\mu_t}(x_t)\|^2] \\ & \leq 6dL^2n \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma_t^2\|v_t\|^2 \right] + (3n+8) \sum_{t=0}^{T-1} 2\Delta_t^2 \\ & = \frac{6dL^2\sqrt{n}}{c^2} \cdot \mathbb{E} \left[ \sum_{t=0}^{T-1} \frac{\|v_t\|^2}{\sqrt{n} + \sum_{s=0}^t \|v_s\|^2} \right] + (3n+8) \sum_{t=0}^{T-1} \Delta_t^2 \\ & \leq \frac{6dL^2\sqrt{n}}{c^2} \log \left( 1 + \mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t\|^2 \right] \right) + (3n+8) \sum_{t=0}^{T-1} \Delta_t^2 \\ & \leq \frac{6dL^2\sqrt{n}}{c^2} \log(\phi(T)) + (3n+8) \sum_{t=0}^{T-1} 2\Delta_t^2, \end{aligned}$$

1482 where last inequality follows by Lemma B.7 and Lemma D.3. Putting everything together we get

$$1483 \mathbb{E} \left[ \sum_{t=0}^{T-1} \|v_t - \nabla f_{\mu_t}(x_t)\| \right] \leq \frac{6\sqrt{d}Ln^{1/4}}{c\sqrt{T}} \log(\phi(T)) + \frac{1}{\sqrt{T}} \sqrt{(3n+8) \sum_{t=0}^{T-1} 2\Delta_t^2}.$$

1484 □  
 1485

1512 D.3 PART III  
1513

1514  $\sum_{t=0}^{T-1} \|\nabla f_{\mu_t}(x) - \nabla f(x_t)\| \leq \frac{1}{T} \sum_{t=0}^{T-1} \Delta_t$   
1515

1516 D.4 FINAL PROOF FOR THE RANDOM-DIRECTION ESTIMATOR  
15171518 **Theorem D.1.** *Under assumptions 1 and 2, based on the previous lemmas D.4, D.5, D.6, we can*  
1519 *derive the following result for Algorithm 2:*

1520  $\frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] \leq \frac{n^{1/4}}{\sqrt{T}} \left( \Delta \cdot c + 1 + \left( \frac{L\sqrt{d}}{cn^{3/4}} + \frac{L^2 d}{c^2} \right) \log(\phi(T)) + \frac{L^2 \pi^2}{24n^{1/4}} + \frac{L}{n^{1/4}} \sqrt{\frac{\pi^2}{24}} + \frac{L^2 \pi^2}{12} + \frac{L}{2} \right),$   
1521

1522 *setting  $c = \sqrt{d}$ , we obtain  $T = \tilde{\mathcal{O}}(d\sqrt{n}\epsilon^{-2})$ , where the  $\tilde{\mathcal{O}}$  notation hides logarithmic factors.*  
15231524 *Proof.*  
1525

1526 
$$\begin{aligned} \frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] &\leq \frac{1}{T} \underbrace{\left[ \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|] \right]}_{\text{part I}} + \underbrace{\sum_{t=0}^{T-1} \mathbb{E}[\|v_t - \nabla f_{\mu_t}(x)\|]}_{\text{part II}} + \underbrace{\sum_{t=0}^{T-1} \|\nabla f_{\mu_t}(x) - \nabla f(x_t)\|}_{\text{part III}} \\ &\leq \frac{n^{1/4}}{\sqrt{T}} \left( 2\Delta \cdot c + 1 + \left( \frac{L\sqrt{d}}{cn^{3/4}} + \frac{dL^2}{c^2} \right) \log(\phi(T)) \right) \\ &\quad + \frac{1}{T} \left( 2c \sum_{t=0}^{T-1} \gamma_t \Delta_t^2 + \sum_{t=0}^{T-1} \Delta_t \right) + \frac{1}{\sqrt{T}} \left( n^{5/4} \sum_{t=0}^{T-1} c\gamma_t \Delta_t^2 + n^{1/2} \sqrt{\sum_{t=0}^{T-1} 2\Delta_t^2} \right), \end{aligned}$$
  
1527

1528 take  $\Delta_t = \frac{L}{2\sqrt{n(t+1)}}$  i.e.  $(\mu_t = \frac{1}{d\sqrt{n(t+1)}})$ , from we obtain:  
1529

1530 
$$\sum_{t=0}^{T-1} \Delta_t \leq \frac{L \ln T}{2\sqrt{n}}.$$
  
1531

1532 
$$\sum_{t=0}^{T-1} \Delta_t^2 < \frac{L^2 \pi^2}{24n}.$$
  
1533

1534 *combing the above results we obtain:*  
1535

1536 
$$\frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|\nabla f(x_t)\| \right] \leq \frac{n^{1/4}}{\sqrt{T}} \left( \Delta \cdot c + 1 + \left( \frac{L\sqrt{d}}{cn^{3/4}} + \frac{L^2 d}{c^2} \right) \log(\phi(T)) + \frac{L^2 \pi^2}{24n^{1/4}} + \frac{L}{n^{1/4}} \sqrt{\frac{\pi^2}{24}} + \frac{L^2 \pi^2}{12} + \frac{L}{2} \right),$$
  
1537

1538 *setting  $c = \sqrt{d}$ , we obtain  $T = \tilde{\mathcal{O}}(d\sqrt{n}\epsilon^{-2})$ , where the  $\tilde{\mathcal{O}}$  notation hides logarithmic factors.  $\square$*   
15391540 E HYPERPARAMETERS DETAILS  
15411542 E.1 PHASE RETRIEVAL  
15431544 We choose the problem dimension to be  $d = 100$  and the sample size to be  $n = 3000$ . The  
1545 measurement vectors  $a_r \in \mathbb{R}^d$  and the true parameter  $z \in \mathbb{R}^d$  are generated element-wise from  
1546 a Gaussian distribution  $\mathcal{N}(0, 0.5)$ . For the initialization,  $z_0 \in \mathbb{R}^d$  is drawn element-wise from  
1547  $\mathcal{N}(5, 0.5)$ . The measurements are then constructed as  $y_i = |a_r^T z|^2 + m_i$  for  $i = 1, \dots, n$ , where the  
1548 noise term  $m_i$  is sampled from  $\mathcal{N}(0, 4^2)$ , representing additive Gaussian noise.  
15491550 We set the parameters for ZO-SGD with a learning rate of  $\gamma = 2 \times 10^{-8}$  and a batch size of  $\sqrt{n}$ . For  
1551 ZO-SPIDER-coord and ZO-SPIDER-rand, we set the learning rate to  $\gamma = 10^{-7}$ , the epoch size to  
1552  $q = n$ , and the batch sizes to  $B = n$  and  $B' = 1$ . For the proposed PF-VRZO method, we similarly  
1553 set the epoch size to  $q = n$ , and choose  $B = n$  and  $B' = 1$  for both the coord and random-direction  
1554 estimators.  
1555

1566  
1567

E.2 DRO

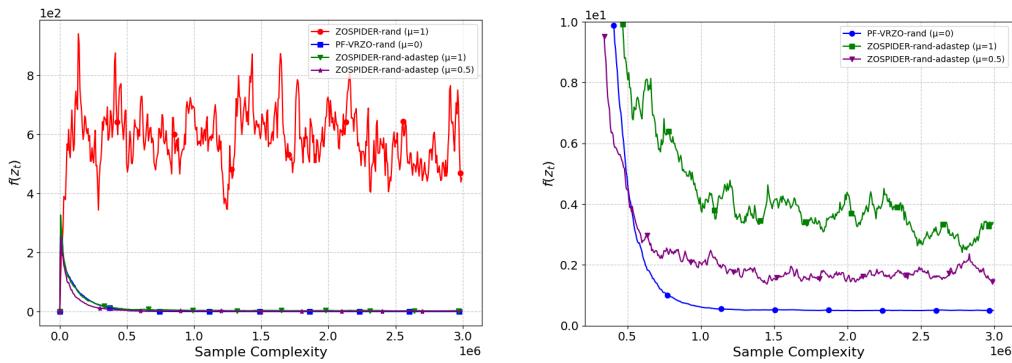
We set the parameters for ZO-SGD with a learning rate  $\gamma = 1 \times 10^{-8}$  and a batch size of  $\sqrt{n}$ . For ZO-SPIDER-coord and ZO-SPIDER-rand, the learning rates are set to  $\gamma = 10^{-6}$  and  $\gamma = 10^{-8}$ , respectively. Both methods use an epoch size of  $q = \sqrt{n}$ , with batch sizes  $B = n$  and  $B' = \sqrt{n}$ . For the proposed PF-VRZO method, we also set the epoch size to  $q = \sqrt{n}$ , and choose  $B = n$  and  $B' = \sqrt{n}$  for both the coord and random-direction estimators. *We remark that the setting  $q = n$ ,  $B = n$ , and  $B' = 1$  is also valid, although it yields slightly worse empirical performance in this experiment.*

1575  
1576  
1577

### E.3 A SMALL EXPERIMENT TO VERIFY THE EFFECTIVENESS OF THE ADAPTIVE SMOOTHING PARAMETER

1578  
1579  
1580  
1581  
1582  
1583

This is a small experiment designed to demonstrate the effectiveness of our adaptive smoothing parameter. We conducted an ablation experiment (placed at the end of the appendix due to page limits) based on the Nonconvex Phase Retrieval setup in the main text. We compare the following four variants: 1. Original ZO-SPIDER, using step size  $\gamma = 0.001$  and  $\mu = 1$ . 2. ZO-SPIDER-adastep, adaptive step size but fixed  $\mu = 1$ . 3. ZO-SPIDER-adastep, adaptive step size but fixed  $\mu = 0.5$ . 4. Our parameter-free PF-VRZO (adaptive step size + adaptive  $\mu_t$ ).

1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596

(a) Comparison of four algorithms: The original ZO-SPIDER (red curve) exhibits severe divergence (function value exceeds 600), which obscures the performance of the other three algorithms (with smaller function values).

(b) Zoomed view of the region where  $f(z) < 10$  in (a): This magnification clarifies the convergence behaviors of the three algorithms with smaller function values, while our PF-VRZO (blue curve) achieves full optimization.

1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610

Figure 4: The original ZO-SPIDER (Group 1) diverges drastically under this parameter setting, with function values surging beyond 600. - Groups 2 and 3 (ZO-SPIDER-adastep) outperform Group 1, yet their function values stagnate (plateauing around 4 and 2, respectively) and fail to decrease further. This aligns with our theoretical analysis: since the fixed  $\mu$  does not diminish with  $T$ , estimator noise accumulates to a point that halts progress. - The  $\mu = 0.5$  variant plateaus later than  $\mu = 1$ —a result consistent with the observation that a smaller fixed  $\mu$  delays (but does not resolve) the stagnation issue. Our PF-VRZO (Group 4), which employs an adaptive  $\mu_t$ , achieves complete optimization successfully.

1611  
1612  
1613  
1614  
1615  
1616  
1617

From the experimental results, we highlight: 1. Adaptive step sizes generally improve convergence behavior. 2. Our adaptive smoothing parameter  $\mu_t$  works synergistically with adaptive step sizes. From our theoretical analysis, a fixed  $\mu$  cannot shrink as  $T$  grows, so the zeroth-order estimator noise eventually fails to meet the increasingly stringent accuracy requirement in later stages of training, causing the algorithm to stall. In contrast, our adaptive  $\mu_t$  avoids this issue by design and ensures stable convergence.

1618  
1619

$$f(\bar{x}_t) - f(x_*) \leq \frac{1}{\sum_{k=0}^{t-1} \bar{r}_k} \sum_{k=0}^{t-1} \bar{r}_k (f(x_k) - f(x_*)).$$