

# Boosting One-Point Derivative-Free Online Optimization via Residual Feedback

Yan Zhang<sup>1b</sup>, Student Member, IEEE, Yi Zhou<sup>1b</sup>, Kaiyi Ji<sup>1b</sup>, Yi Shen<sup>1b</sup>, Student Member, IEEE, and Michael M. Zavlanos<sup>1b</sup>, Senior Member, IEEE

**Abstract**—Zeroth-order optimization (ZO) typically relies on two-point feedback to estimate the gradient of the objective function. Nevertheless, two-point feedback cannot be used for online optimization with time-varying objective functions, where only a single query of the function value is possible at each time step. In this work, we propose a new one-point feedback method for online optimization that estimates the gradient using the residual between two feedback points at consecutive time instants. Moreover, we develop regret bounds for ZO with residual feedback for constrained convex and unconstrained nonconvex online optimization problems. Specifically, for both deterministic and stochastic problems and for both Lipschitz and smooth objective functions, we show that using residual feedback can produce gradient estimates with much smaller variance compared to conventional one-point feedback methods. As a result, our regret bounds are much tighter compared to existing regret bounds for ZO with conventional one-point feedback, which suggests that ZO with residual feedback can better track the optimizer of online optimization problems. In addition, our regret bounds rely on weaker assumptions than those used in conventional one-point feedback methods. Numerical experiments show that ZO with residual feedback significantly outperforms existing one-point feedback methods.

**Index Terms**—Online optimization, regret analysis, zeroth-order optimization (ZO).

## I. INTRODUCTION

Zeroth-order optimization (ZO) algorithms have been widely used to solve online optimization problems where first or second-order information (i.e., gradient or Hessian information) is unavailable at each time instant. Such problems arise, e.g., in online learning and involve adversarial training [1] and reinforcement learning [2], [3] among others. The goal is to minimize a sequence of time-varying objective functions  $\{f_t(x)\}_{t=1:T}$ , where the value  $f_t(x_t)$  is revealed to the agent after an action  $x_t$  is selected and is used to adapt the agent's future strategy. Since the objective functions are not known a priori, the quality of an online decision is measured using notions of regret, that

compare the total cost incurred by an online decision to the cost of the fixed or varying optimal decision that a clairvoyant agent could select.

Perhaps the most popular zeroth-order gradient estimator is the two-point estimator that has been extensively studied in [4], [5], [6], [7], [8], [9], and [10]. This estimator queries the function value  $f_t(x)$  at two different decision variables at each time step, and uses the difference in the two function values to estimate the desired gradient (two-point feedback), i.e.,

$$\tilde{g}_t^{(2)}(x) = \frac{d}{\delta} (f_t(x + \delta u) - f_t(x)) u \quad (1)$$

where  $\delta > 0$  is a parameter and  $u$  is uniformly sampled from a unit sphere  $\mathbb{S}^d$  in space  $\mathbb{R}^d$ . Although this two-point estimator produces gradient estimates with low variance that improve the convergence speed of ZO, it can only be used for online optimization when the same objective function can be queried multiple times, e.g., for online learning with incoming data streams in [11]. When the objective function changes by nature or by an adversary with every new query, one-point estimators can be used instead that query the objective function  $f_t(x)$  only once at each time instant (one-point feedback), i.e.,

$$\tilde{g}_t^{(1)}(x) = \frac{d}{\delta} f_t(x + \delta u) u. \quad (2)$$

One-point feedback was first proposed and analyzed in [12] for convex online optimization problems. Subsequently, the authors in [13] and [14] showed that the regret of convex online optimization methods using one-point gradient estimation can be improved if the objective functions are assumed to be smooth and self-concordant regularization is used. More recently, Gasnikov et al. [15] developed regret bounds for ZO with one-point feedback also for stochastic convex problems. On the other hand, Hazan et al. [16] characterized the convergence of one-point zeroth-order methods for static stochastic nonconvex optimization problems. However, as shown in these studies, one-point feedback produces gradient estimates with large variance which results in increased regret. In addition, the regret analysis for ZO with one-point feedback usually requires the strong assumption that the function value is uniformly upper bounded over time, so this method cannot be used for practical nonstationary optimization problems. In [17], the authors studied a two-point oracle that evaluates different objective functions at the two queries. The analysis is based on a strong convexity assumption and cannot be extended to general convex and nonconvex problems.

**Contributions:** In this article, we propose a novel one-point gradient estimator for the zeroth-order online optimization and develop new regret bounds to study its performance. Our proposed estimator uses the residual between two consecutive feedback points to estimate the gradient and, therefore, we refer to it as residual feedback. We show that, for both deterministic and stochastic problems, using residual feedback produces gradient estimates with lower variance compared to those produced using the conventional one-point feedback proposed

Manuscript received 8 December 2022; revised 11 November 2023; accepted 9 March 2024. Date of publication 28 March 2024; date of current version 29 August 2024. This work was supported in part by AFOSR under Award #FA9550-19-1-0169 and in part by NSF under Award CNS-1932011. Recommended by Associate Editor G. Notarstefano. (Yan Zhang and Yi Zhou contributed equally to this work.) (Corresponding author: Yan Zhang.)

Yan Zhang, Yi Shen, and Michael M. Zavlanos are with the Department of Mechanical Engineering and Material Science, Duke University, Durham, NC 27705 USA (e-mail: yan.zhang2@duke.edu; yi.shen478@duke.edu; michael.zavlanos@duke.edu).

Yi Zhou is with the Department of Electrical and Computer Engineering, The University of Utah, Salt Lake City, UT 84112 USA (e-mail: yi.zhou@utah.edu).

Kaiyi Ji is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: ji.367@osu.edu).

Digital Object Identifier 10.1109/TAC.2024.3382358

in [12] and [15]. As a result, we obtain tighter regret bounds both for constrained convex and unconstrained nonconvex problems, especially when the value of the objective function is large. Moreover, our regret analysis relies on weaker assumptions compared to those for ZO with conventional one-point feedback. Finally, we present numerical experiments that demonstrate that ZO with residual feedback significantly outperforms the conventional one-point method in its ability to track the time-varying optimizers of online learning problems. To the best of our knowledge, this is the first time a one-point zeroth-order method is theoretically studied for nonconvex online optimization problems. It is also the first time that a one-point gradient estimator demonstrates comparable empirical performance to that of the two-point method.

*Related Work:* Online optimization problems are only one instance of optimization problems that ZO methods have been used to solve. For example, Balasubramanian and Ghadimi [18] applied ZO to solve a set-constrained optimization problem where the projection onto the constraint set is nontrivial. The authors in [19] and [20] applied a variance-reduced technique and acceleration methods to achieve better convergence speed in ZO. Wang et al. [21] improved the dependence of the iteration complexity on the dimension of the problem under an additional sparsity assumption on the gradient of the objective function. The authors in [22] and [23] applied zeroth-order oracles to distributed optimization problems when only bandit feedbacks are available at each local agents. Our proposed residual feedback oracle can be used to solve such optimization problems as well. Also related is the work in [24], in which the authors considered nonconvex online bandit optimization problems with a single query at each time step. However, this method employs the exploration and exploitation bandit learning framework and the proposed analysis is restricted to a special class of nonconvex objective functions. The authors in [25], [26], and [27] studied online bandit algorithms using ellipsoid methods. In particular, these methods induce heavy computation per step and achieve regret bounds that have bad dependence on the problem dimension. As a comparison, our one-point method is computation light and achieves regret bounds that have better dependence on the problem dimension. A similar one-point oracle was proposed in [28] and [29] for static convex optimization problems but the analysis cannot be extended to the online optimization setting.

## II. PRELIMINARIES AND RESIDUAL FEEDBACK

In this section, we provide basic definitions and results on ZO that will be needed in the subsequent analysis. We also define the residual feedback gradient estimator to solve online optimization problems with unknown gradient information. Consider the following online bandit optimization problem:

$$\min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x) \quad (\text{P})$$

where  $\mathcal{X} \subset \mathbb{R}^d$  is a convex set and  $\{f_t\}_t$  is a sequence of objective functions that are unknown to the agent a priori. Specifically, we assume that at any time  $t$ , first the agent makes a decision  $x_t$  and then the value of the objective function  $f_t$  at  $x_t$  is revealed. We also assume that the derivatives of the objective functions are unavailable. Therefore, the agent needs to use a zeroth-order oracle to estimate the derivative information. The goal is to determine an online decision  $x_t$  (or a sequence of time-varying decisions) with cost that is as close as possible to the cost of a fixed decision (or a sequence of varying optimal decisions) that a clairvoyant agent could select, which is measured by notions of regret.

First, we define the class of Lipschitz and smooth objective functions we are concerned with. Consider the set  $\mathcal{X}_\delta := \{z : z = x + \delta u, \text{ for any } x \in \mathcal{X} \text{ and } u \in \mathbb{U}^d\}$ , where  $\mathbb{U}^d$  represents the unit sphere in space  $\mathbb{R}^d$ .

**Definition 2.1 (Lipschitz functions):** The class of Lipschitz-continuous functions  $C^{0,0}$  satisfies: for any  $f \in C^{0,0}$ ,  $|f(x) - f(y)| \leq L_0 \|x - y\| \quad \forall x, y \in \mathcal{X}_\delta$ , where  $L_0 > 0$  is the Lipschitz parameter over set  $\mathcal{X}_\delta$ . The class of smooth functions  $C^{1,1}$  satisfies: for any  $f \in C^{1,1}$ ,  $\|\nabla f(x) - \nabla f(y)\| \leq L_1 \|x - y\| \quad \forall x, y \in \mathcal{X}_\delta$ , where  $L_1 > 0$  is the smoothness parameter over set  $\mathcal{X}_\delta$ .

The key idea in ZO is to estimate the unknown first-order gradient of the objective function  $f$  using zeroth-order oracles that perturb the objective function around the current point along all directions uniformly. The ability of these oracles to correctly estimate the gradient is typically analyzed using the smoothed version of the function  $f$  defined as  $f_\delta(x) := \mathbb{E}_{u \sim \mathbb{U}^d} [f(x + \delta u)]$ , where the coordinates of the vector  $u$  are uniformly sampled from a unit ball  $\mathbb{U}^d$  in space  $\mathbb{R}^d$ . Note that the objective function  $f_t$  is defined over the larger domain  $\mathcal{X}_\delta$  rather than  $\mathcal{X}$ , since the objective function  $f_t$  can be evaluated outside the set  $\mathcal{X}$  during iterations. On the other hand, the smoothed function  $f_{\delta,t}$  is defined over the set  $\mathcal{X}$ . We have the following results bounding the approximation errors of the function  $f_\delta(x)$ .

**Lemma 2.2:** Consider a function  $f$  and its smoothed version  $f_\delta$ . It holds that

$$|f_\delta(x) - f(x)| \leq \begin{cases} \delta L_0, & \text{if } f \in C^{0,0} \\ \delta^2 L_1, & \text{if } f \in C^{1,1} \end{cases}$$

$$\text{and } \|\nabla f_\delta(x) - \nabla f(x)\| \leq \delta L_1 d, \text{ if } f \in C^{1,1}.$$

The smoothed function  $f_\delta(x)$  also satisfies the following amenable property.

**Lemma 2.3:** If  $f \in C^{0,0}$  is  $L_0$ -Lipschitz, then  $f_\delta \in C^{1,1}$  with Lipschitz constant  $L_{1,\delta} = d\delta^{-1}L_0$ .

The proofs of the above lemmas are included in [30].

**Definition 2.4: (Objective functions)** We call the sequence of objective functions  $\{f_0, f_1, \dots, f_t\}$  naturally nonstationary when the objective function  $f_t$  is selected based on the agent's past decisions  $\{x_0 + \delta u_0, x_1 + \delta u_1, \dots, x_{t-1} + \delta u_{t-1}\}$  and does not depend on its decision  $x_t + \delta u_t$ . The same sequence of objective functions is called adversarially nonstationary if the selection of  $f_t$  depends also on the agent's current decision  $x_t + \delta u_t$ . In addition, at each time step  $t$ , the objective function  $f_t$  is bounded by a constant  $f_t^*$  from below.

In this article we consider both natural and adversarial objective function sequences, as defined in Definition 2.4. Natural nonstationary learning problems arise, for example, in reinforcement learning, when the environment changes because of the natural shift in the noise distribution of the agent dynamics and reward functions. On the other hand, in multiagent games, if an agent plays against an adversarial agent who selects its policy based on the first agent's policy at time  $t$ , then the first agent faces an adversarial nonstationary environment. In such problems where the system evolves from  $f_t$  to  $f_{t+1}$ , two-point feedback (1) cannot be used to estimate the unknown gradient of  $f_t$  as it requires two different evaluations of  $f_t$  at two different decisions  $x_t$  and  $x_t + \delta u_t$  at the same time, which is not possible since  $f_t$  changes after one decision variable is evaluated. Instead, a more practical approach is to use the one-point feedback scheme (2) in [15]. However, the gradient estimates produced by the one-point feedback method in (2) have large variance that leads to large regret and, therefore, poor ability to track the optimizer of the online problem. To address this limitation, in this article we propose a novel one-point gradient estimator, which we call a one-point residual feedback estimator, that has reduced variance and

is defined as

$$\tilde{g}_t(x_t) := \frac{d}{\delta} (f_t(x_t + \delta u_t) - f_{t-1}(x_{t-1} + \delta u_{t-1})) u_t \quad (3)$$

where  $u_{t-1}, u_t \sim \mathcal{U}^d$  are independent random vectors. To elaborate, the proposed residual feedback estimator in (3) queries  $f_t$  at a single perturbed point  $x_t + \delta u_t$ , and then subtracts the value  $f_{t-1}(x_{t-1} + \delta u_{t-1})$  obtained from the previous iteration. Next, we discuss some basic properties of this new estimator. We first show that this estimator provides an unbiased gradient estimate of the smoothed function  $f_{\delta,t}$ .

**Lemma 2.5:** The residual feedback estimator satisfies  $\mathbb{E}[\tilde{g}_t(x_t)] = \nabla f_{\delta,t}(x_t)$  for all  $x_t \in \mathcal{X}$  and  $t$ .

*Proof:* The proof follows from the fact that  $\frac{d}{\delta} f_t(x_t + \delta u_t) u_t$  is an unbiased estimator of  $\nabla f_{\delta,t}(x_t)$  according to [12] and  $u_t$  has zero mean and is independent from  $u_{t-1}, x_{t-1}$ . ■

In this article, we consider the following ZO projected gradient update with residual feedback

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta \tilde{g}_t(x_t)) \quad (4)$$

where  $\eta$  is the learning rate and  $\Pi_{\mathcal{X}}$  is the projection operator onto the constrained set  $\mathcal{X}$ . For unconstrained problems, let  $\mathcal{X} = \mathbb{R}^d$ . The following result bounds the second moment of the gradient estimate generated by using residual feedback.

**Lemma 2.6 (Second moment):** Assume that  $f_t \in C^{0,0}$  with Lipschitz constant  $L_0$  for all time  $t$ . Then, under the ZO update rule in (4), the second moment of the residual feedback (3) satisfies

$$\mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \leq \frac{4d^2L_0^2\eta^2}{\delta^2} \mathbb{E}[\|\tilde{g}_{t-1}(x_{t-1})\|^2] + D_t \quad (5)$$

where  $D_t := 16d^2L_0^2 + \frac{2d^2}{\delta^2} \mathbb{E}[(f_t(x_{t-1} + \delta u_{t-1}) - f_{t-1}(x_{t-1} + \delta u_{t-1}))^2]$ .

The proof of the above lemma can be found in [30]. The above lemma shows that the second moment of the gradient estimates obtained using residual feedback forms a contraction with perturbation term  $D_t$ , provided that we choose  $\eta$  and  $\delta$  such that the contracting rate satisfies  $\alpha = 4d^2L_0^2\eta^2\delta^{-2} < 1$ . As we show later in the analysis, this contraction property leads to gradient estimates with small variances that allow to reduce the regret of the online ZO algorithm (4).

### III. ZO WITH RESIDUAL FEEDBACK FOR CONVEX ONLINE OPTIMIZATION

In this section, we consider the online bandit problem (P), where the sequence of functions  $\{f_t\}_{t=0:T-1}$  are all convex and the constraint set is compact. In particular, we are interested in analyzing the static regret of algorithm (4) defined as

$$R_T := \mathbb{E} \left[ \sum_{t=0}^{T-1} f_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x) \right]. \quad (6)$$

Denote  $\mathcal{X}^* = \operatorname{argmin}_{x \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x)$  the set of all optimal points and let  $x^* = \operatorname{argmin}_{x \in \mathcal{X}^*} \|x\|$ . Let  $x_0$  be a given initial point and define  $R = \|x_0 - x^*\|$ .

First, we make the following assumption on the nonstationarity of the online learning problem.

**Assumption 3.1 (Bounded variation):** There exists  $V_f > 0$  such that for all  $t$  and all  $x \in \mathcal{X}$ ,  $|f_t(x) - f_{t-1}(x)| \leq V_f$ .

Assumption 3.1 states that the variation of the objective function between two consecutive time instants is uniformly bounded over time. We note that this assumption is weaker than the assumption that the objective function is uniformly bounded, i.e.,  $|f_t(x)| \leq B \quad \forall t, x$ , which is used in the analysis of ZO with conventional one-point feedback in [12] and [15].

For any sequence of objective functions, natural or adversarial, as defined in Definition 2.4, the following result characterizes the regret of ZO with residual feedback when the objective function  $f_t$  is convex and Lipschitz.

**Theorem 3.2 (Regret for Convex Lipschitz  $f_t$ ):** Let Assumption 3.1 hold. Assume that  $f_t \in C^{0,0}$  is convex with Lipschitz constant  $L_0$  over the compact set  $\mathcal{X}_\delta$  for all  $t$ . Run ZO with residual feedback with  $\eta = \frac{1}{2\sqrt{2}dL_0T^{\frac{1}{4}}}$  and  $\delta = \sqrt{dT}^{-\frac{1}{4}}$ . Then, we have that

$$R_T \leq \sqrt{2}dL_0R^2T^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{2\sqrt{2}dL_0T^{\frac{3}{4}}} + 4\sqrt{2}d^{\frac{3}{2}}L_0T^{\frac{1}{4}} + 2\sqrt{d}L_0T^{\frac{3}{4}} + \frac{\sqrt{d}V_f^2}{\sqrt{2}L_0}T^{\frac{3}{4}}.$$

The proof can be found in Appendix A. Next, we present the regret of ZO with residual feedback when the objective function  $f_t$  is convex and smooth. Since  $f_t$  is defined on a compact set  $\mathcal{X}_\delta$ , if  $f_t$  has a Lipschitz gradient then it is also Lipschitz with a constant  $L_0$  over the set  $\mathcal{X}_\delta$ . As before, the sequence of objective functions can be either natural or adversarial, as per Definition 2.4.

**Theorem 3.3 (Regret for Convex Smooth  $f_t$ ):** Let Assumption 3.1 hold. Assume that  $f_t(x) \in C^{1,1}$  is convex and smooth with constant  $L_1$  over the compact set  $\mathcal{X}_\delta$  for all  $t$ . Run ZO with residual feedback with  $\eta = \frac{1}{2\sqrt{2}L_0d^{\frac{2}{3}}T^{\frac{1}{3}}}$  and  $\delta = \frac{1}{d^{\frac{1}{6}}T^{\frac{1}{6}}}$ . Then, we have that

$$R_T \leq \sqrt{2}L_0d^{\frac{2}{3}}R^2T^{\frac{2}{3}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{2\sqrt{2}L_0d^{\frac{2}{3}}T^{\frac{2}{3}}} + 8\sqrt{2}L_0\frac{(d+4)^2}{d^{\frac{2}{3}}}T^{\frac{1}{3}} + 2d^{\frac{2}{3}}L_1T^{\frac{2}{3}} + \frac{\sqrt{2}}{L_0}d^{\frac{2}{3}}V_f^2T^{\frac{2}{3}}.$$

The proof can be found in Appendix B. According to Theorems 3.2 and 3.3, using the proposed residual feedback gradient estimator, the regret of the one-point ZO no longer depends on the uniform bound of the function value, which can be very large in practice. Instead, our regret only relies on how fast the function varies over time.

### IV. ZO WITH RESIDUAL FEEDBACK FOR NONCONVEX ONLINE OPTIMIZATION

In this section, we analyze the regret of ZO with residual feedback for the unconstrained online bandit problem (P), where the objective functions  $\{f_t\}_{t=0:T-1}$  are nonconvex. To the best of our knowledge, this is the first time that a one-point zeroth-order method is studied for nonconvex online optimization. Throughout this section, we make the following assumption on the objective functions.

**Assumption 4.1:** There exist  $W_T, \tilde{W}_T > 0$  such that for any sequence  $\{x_t\}_{t=1}^T$  the following conditions hold.

- 1)  $\sum_{t=1}^T (f_t(x_t) - f_{t-1}(x_t)) \leq W_T$ .
- 2)  $\sum_{t=1}^T (f_t(x_t) - f_{t-1}(x_t))^2 \leq \tilde{W}_T$ .

The above two conditions in Assumption 4.1 measure the accumulated first-order and second-order function variations. Such variations are called the regularity measures in online nonstationary learning problems as in [10] and [31].

First, we consider the case where  $\{f_t\}_t$  are nonconvex and Lipschitz continuous functions. Since the objective function  $f_t$  is not necessarily differentiable,  $\nabla f(t)$  may not exist. Therefore, we define the regret as the accumulated gradient of the smoothed function, i.e.,  $R_{g,\delta}^T := \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2]$ , inspired by the study of zeroth-order oracles in static nonsmooth optimization problems in [8]. In addition, similar to [8], we require that the smoothed function  $f_{\delta,t}$  is close to the original function  $f_t$  such that  $|f_{\delta,t}(x) - f_t(x)| \leq \epsilon_f$  for all  $t$ . To satisfy this

condition, we need to choose  $\delta \leq (L_0)^{-1}\epsilon_f$  according to Lemma 2.2. Then, we can show the following regret bound for ZO with residual feedback, when the objective functions are either natural or adversarial, as per Definition 2.4.

**Theorem 4.2 (Nonconvex Lipschitz  $f_t$ ):** Let Assumptions 4.1 hold. Assume that  $f_t \in C^{0,0}$  with Lipschitz constant  $L_0$  and that  $f_t$  is bounded below by  $f_t^*$  for all  $t$ . Run ZO with residual feedback with  $\eta = \epsilon_f^{\frac{3}{2}}(2\sqrt{2}L_0^2 d^{\frac{3}{2}} T^{\frac{1}{2}})^{-1}$  and  $\delta = \epsilon_f L_0^{-1}$ . Then, we have that

$$R_{g,\delta}^T \leq 2\sqrt{2}L_0^2 (\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T) d^{\frac{3}{2}} \epsilon_f^{-\frac{3}{2}} T^{\frac{1}{2}} \\ + 4\sqrt{2}L_0^2 \epsilon_f^{\frac{1}{2}} d^{\frac{3}{2}} T^{\frac{1}{2}} + \frac{L_0^2 d^{\frac{3}{2}} \widetilde{W}_T}{\sqrt{2} \epsilon_f^{\frac{3}{2}} T^{\frac{1}{2}}} + \frac{\epsilon_f^{\frac{1}{2}} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{2\sqrt{2}dT}.$$

Asymptotically,  $R_{g,\delta}^T = \mathcal{O}(d^{\frac{3}{2}} L_0^2 \epsilon_f^{-\frac{3}{2}} (W_T + \widetilde{W}_T T^{-1}) T^{\frac{1}{2}})$ .

The proof can be found in Appendix C. Theorem 4.2 implies that the regret bound satisfies  $R_{g,\delta}^T/T \rightarrow 0$  whenever  $W_T = o(T^{\frac{1}{2}} \epsilon_f^{\frac{3}{2}})$  and  $\widetilde{W}_T = o(T^{\frac{3}{2}} \epsilon_f^{\frac{3}{2}})$ . In particular, if the bounded variation Assumption 4.1 holds, then we have  $\widetilde{W}_T \leq \mathcal{O}(\text{TV}_f^2)$ , and it suffices to let  $T^{-\frac{1}{2}} \epsilon_f^{-\frac{3}{2}} = o(1)$ .

Next, we assume that the objective functions  $f_t$  in (P) are nonconvex and smooth and define the regret  $R_g^T := \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_t(x_t)\|^2]$ . Specifically, we provide the following regret bound for ZO with residual-feedback for natural or adversarial objective functions, as per Definition 2.4.

**Theorem 4.3 (Nonconvex smooth  $f_t$ ):** Let Assumptions 4.1 hold. Assume that  $f_t \in C^{0,0} \cap C^{1,1}$  with Lipschitz constant  $L_0$  and smoothness constant  $L_1$  and that  $f_t$  is bounded below by  $f_t^*$  for all  $t$ . Run ZO with residual feedback for  $T$  iterations with  $\eta = (2\sqrt{2}L_0 d^{\frac{4}{3}} T^{\frac{1}{2}})^{-1}$  and  $\delta = (d^{\frac{1}{3}} T^{\frac{1}{2}})^{-1}$ . Then,

$$R_g^T \leq 4\sqrt{2}L_0 (\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T) d^{\frac{4}{3}} T^{\frac{1}{2}} + 2L_1^2 d^{\frac{4}{3}} T^{\frac{1}{2}} \\ + 8\sqrt{2}L_1 L_0 d^{\frac{4}{3}} T^{\frac{1}{2}} + \frac{\sqrt{2}L_1 d^{\frac{4}{3}} \widetilde{W}_T}{L_0} + \frac{L_1 \mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{\sqrt{2}L_0 d^{\frac{4}{3}} T^{\frac{1}{2}}}.$$

Asymptotically,  $R_g^T = \mathcal{O}(d^{\frac{4}{3}} L_0 W_T T^{\frac{1}{2}} + d^{\frac{4}{3}} L_1 L_0^{-1} \widetilde{W}_T)$ .

The proof can be found in Appendix D. Theorem 4.3 implies that the regret bound satisfies  $R_g^T/T \rightarrow 0$  whenever  $W_T = o(T^{\frac{1}{2}})$  and  $\widetilde{W}_T = o(T)$ . We note that these requirements on  $W_T$  and  $\widetilde{W}_T$  are weaker than those in the case of nonsmooth problems, as they do not rely on the small parameter  $\epsilon_f$ .

## V. ZO WITH RESIDUAL FEEDBACK FOR STOCHASTIC ONLINE OPTIMIZATION

Our proposed residual feedback gradient estimator can be also extended to solve stochastic online bandit problems. Since the regret analysis is similar to that for deterministic online problems presented before, we only introduce the key technical lemmas and comment on the differences in the proof. Consider the stochastic online bandit problems

$$\min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} \mathbb{E}[F_t(x; \xi_t)], \text{ where } \mathbb{E}[F_t(x; \xi_t)] = f_t(x) \quad \forall t$$

where  $\xi_t$  denotes a certain noise that is independent of  $x$ . Different from the deterministic online problems discussed before, the agent here can only query noisy evaluations of the objective function. To solve the above problem, we propose the following stochastic residual feedback:

$$\tilde{g}_t(x_t) := \frac{du_t}{\delta} (F_t(x_t + \delta u_t; \xi_t) - F_{t-1}(x_{t-1} + \delta u_{t-1}; \xi_{t-1})) \quad (7)$$

where  $\xi_{t-1}$  and  $\xi_t$  are independent random sample noises at consecutive iterations  $t-1$  and  $t$ , respectively, and  $u_t$  and  $u_{t-1} \in \mathbb{US}^d$  are random search directions sampled by the user. Since the noisy function value  $F(x; \xi_t)$  is an unbiased estimate of the objective function  $f_t(x)$ , it is straightforward to show that (7) is an unbiased gradient estimate of the function  $f_{\delta,t}(x)$ . To analyze the regret of ZO with stochastic residual feedback, we first consider the convex case and make the following assumption on the variation of the stochastic objective functions.

**Assumption 5.1: (Bounded stochastic variation)** There exists  $V_{f,\xi} > 0$  such that for all  $t$  and  $x_{t-1} \in \mathcal{X}_\delta$ ,  $\mathbb{E}[(F_t(x_{t-1}, \xi_t) - F_{t-1}(x_{t-1}, \xi_{t-1}))^2] \leq V_{f,\xi}^2$ , where the expectation is taken over the evaluation noises  $\xi_t$  and  $\xi_{t-1}$ .

The above assumption generalizes Assumption 3.1 to stochastic problems. The bound  $V_{f,\xi}^2$  controls both the variation of function and the variation due to stochastic sampling. The following lemma characterizes the second moment of the stochastic residual feedback gradient estimates. Its proof can be found in [30].

**Lemma 5.2:** Assume  $F(x, \xi) \in C^{0,0}$  with Lipschitz constant  $L_0$  for all  $\xi$  and  $x \in \mathcal{X}_\delta$ . Then, under the ZO update rule, we have that

$$\mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \leq \frac{4d^2 L_0^2 \eta^2}{\delta^2} \mathbb{E}[\|\tilde{g}_t(x_{t-1})\|^2] + D_{t,\xi}$$

where  $D_{t,\xi} := 16L_0^2 d^2 + \frac{2d^2}{\delta^2} \mathbb{E}[(F_t(x_{t-1} + \delta u_{t-1}, \xi_t) - F_{t-1}(x_{t-1} + \delta u_{t-1}, \xi_{t-1}))^2]$ .

Observe that the above second moment bound is very similar to that in Lemma 2.6, and the only difference is the perturbation term. Consequently, ZO with stochastic residual feedback achieves almost the same regret bounds as those in Theorems 3.2 and 3.3, and one simply needs to replace  $V_f$  by  $V_{f,\xi}$ . For nonconvex problems, we adopt the following assumption that generalizes Assumption 4.1.

**Assumption 5.3:** There exists  $W_T, \widetilde{W}_{T,\xi} > 0$  such that for any sequence  $\{x_t\}_{t=1}^T$  the following two conditions hold.

- 1)  $\sum_{t=1}^T (f_{\delta,t}(x_t) - f_{\delta,t-1}(x_t)) \leq W_T$ .
- 2)  $\sum_{t=1}^T \mathbb{E}[(F_t(x_{t-1}, \xi_t) - F_{t-1}(x_{t-1}, \xi_{t-1}))^2] \leq \widetilde{W}_{T,\xi}$ , where the expectation is taken over evaluation noises  $\xi_t$  and  $\xi_{t-1}$ .

Then, following similar steps as those in the proofs of Theorems 4.2 and 4.3, we can obtain similar regret bounds for ZO with stochastic residual feedback (simply replace  $W_T$  and  $\widetilde{W}_T$  in Theorems 4.2 and 4.3 by  $W_{T,\xi}$  and  $\widetilde{W}_{T,\xi}$ , respectively).

## VI. NUMERICAL EXPERIMENTS

In this section, we compare the performance of ZO with one-point, two-point, and residual feedback in solving nonstationary resource allocation problems, where either the reward or transition functions are varying over episodes. Specifically, we consider a multistage resource allocation problem with time-varying sensitivity to the lack of resource supply. Specifically, 16 agents are located on a  $4 \times 4$  grid. During episode  $t$ , at step  $k$ , agent  $i$  stores  $m_i(k)$  amount of resources and has a demand for resources in the amount of  $d_i(k)$ . Also, agent  $i$  decides to send a fraction of resources  $a_{ij}(k) \in [0, 1]$  to its neighbors  $j \in \mathcal{N}_i$  on the grid. The local amount of resources and demands of agent  $i$  evolve as  $m_i(k+1) = m_i(k) - \sum_{j \in \mathcal{N}_i} a_{ij}(k)m_i(k) + \sum_{j \in \mathcal{N}_i} a_{ji}(k)m_j(k) - d_i(k)$  and  $d_i(k) = \psi_i \sin(\omega_i k + \phi_i) + w_{i,k}$ , where  $w_{i,k}$  is the noise in the demand. At each step  $k$ , agent  $i$  receives a local cost  $r_{i,t}(k)$ , such that  $r_{i,t}(k) = 0$  when  $m_i(k) \geq 0$  and  $r_{i,t}(k) = \zeta_t m_i(k)^2$  when  $m_i(k) < 0$ , where  $\zeta_t$  represents the varying sensitivity of the agents to the lack of supply during episode  $t$ . Let agent  $i$  make its decisions according to a parameterized policy function  $\pi_{i,t}(o_i; \theta_{i,t}) : \mathcal{O}_i \rightarrow [0, 1]^{|\mathcal{N}_i|}$ , where  $\theta_{i,t}$  is the parameter of the policy function  $\pi_{i,t}$  at episode  $t$ ,  $o_i \in \mathcal{O}_i$  denotes agent  $i$ 's local observation. Specifically, we let  $o_i(k) = [m_i(k), d_i(k)]^T$ . Our goal is to track the

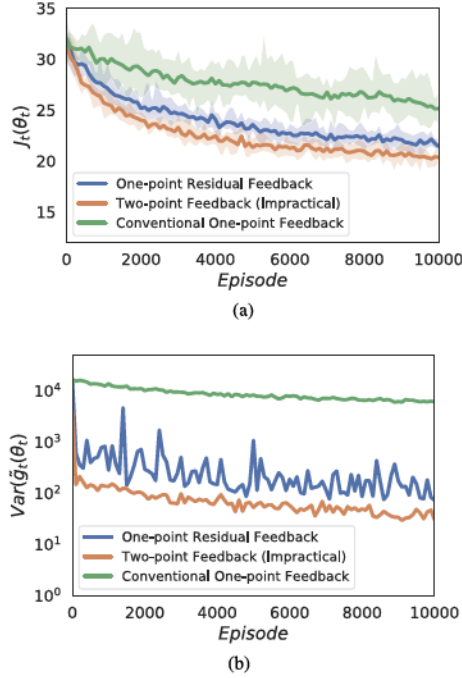


Fig. 1. Comparative results of ZO with the proposed one-point residual feedback (3) (black), the two-point oracle in [7] (orange) and the conventional one-point oracle in [15] (green) for the nonstationary resource allocation problem. (a) Varying cost  $J_t(\theta_t)$  achieved using three different oracles. (b) Variance of the gradient estimates at agent 1 returned by the three methods. The two point method (orange) is infeasible to use in practice and is presented here to serve as a simulation benchmark.

time-varying optimal policy so that the accumulated cost over the grid  $J_t(\theta_t) = \sum_{i=1}^{16} \sum_{k=0}^H \gamma^k r_{i,t}(k)$  during each episode is maintained at a low level, where  $\theta_t = [\dots, \theta_{i,t}, \dots]$  is the policy parameter,  $H$  is the problem horizon at each episode, and  $\gamma$  is the discount factor.

All experiments are conducted using Matlab R2019a on Ubuntu 18.04 with the AMD Ryzen 2700X 8-core processor and 16 GB 2133 MHz memory. The policy function  $\pi_{i,t}(o_i; \theta_{i,t})$  is parameterized as:  $a_{ij} = \exp(z_{ij}) / \sum_j \exp(z_{ij})$ , where  $z_{ij} = \sum_{p=1}^9 \psi_p(o_i) \theta_{ij}(p)$  and  $\theta_i = [\dots, \theta_{ij}, \dots]^T$  and the episode index  $t$  is omitted for notational simplicity. Specifically, the feature function  $\psi_p(o_i)$  is selected as  $\psi_p(o_i) = \|o_i - c_p\|^2$ , where  $c_p$  is the parameter of the  $p$ th feature function. Effectively, the agents need to make decisions on 64 actions, and each action is decided by nine parameters. Therefore, the problem dimension is  $d = 576$ . The discount factor is selected as  $\gamma = 0.75$  and the length of the horizon is  $H = 30$ . The time-varying sensitivity parameter  $\zeta_{i,t}$  is generated as follows: let  $\zeta_{i,0} = 1$  and  $\zeta_{i,t+1} = \zeta_{i,t} + 0.1P_t$ , where  $P_t$  is a random number uniformly sampled from  $[-1, 1]$ .

In Fig. 1(a), we present the cost  $J_t(\theta_t)$  achieved during each episode after ten trials of ZO with residual-feedback, one-point, and two-point feedback which, as before, is impossible to use in practice for this nonstationary problem either. It can be seen that ZO with our proposed residual-feedback achieves a cost  $J_t(\theta_t)$  that is as low as the cost achieved by the two-point feedback in this nonstationary environment. In particular, ZO with both residual and two-point feedback performs much better than ZO with conventional one-point feedback. Fig. 1(b) also compares the estimated variance of the gradient estimates returned by these feedback schemes. It can be seen that the variance of the gradient estimates returned by the residual feedback oracle is comparable to that of the gradient estimates returned by the two-point oracle and is much smaller than that returned by the conventional one-point oracle.

## VII. CONCLUSION

In this article, we proposed a novel one-point residual feedback oracle for zeroth-order online optimization, which estimates the gradient of the time-varying objective function using a single query of the function value at each time instant. For both deterministic and stochastic problems, we showed that ZO with the proposed residual feedback estimator achieves much lower regret than that of ZO with conventional one-point feedback for convex online optimization problems. In addition, we provided regret bounds for ZO with residual feedback for nonconvex online optimization problems. To the best of our knowledge, this is the first time that a one-point zeroth-order method is theoretically studied for nonconvex online problems. Numerical experiments on a nonstationary reinforcement learning problem were conducted and the proposed residual-feedback estimator was shown to significantly outperform the conventional one-point method.

## APPENDIX

### A. Proof of Theorem 3.2

Note that  $f_{\delta,t}(x)$  is convex for all  $t$ , we then conclude that

$$f_{\delta,t}(x_t) - f_{\delta,t}(x) \leq \langle \nabla f_{\delta,t}(x_t), x_t - x \rangle, \text{ for all } x \in \mathcal{X}. \quad (8)$$

Adding and subtracting  $\tilde{g}_t(x_t)$  after  $\nabla f_{\delta,t}(x_t)$  in above inequality, and taking expectation over  $u_t$  on both sides, we obtain that

$$\mathbb{E}[f_{\delta,t}(x_t) - f_{\delta,t}(x)] \leq \mathbb{E}[\langle \tilde{g}_t(x_t), x_t - x \rangle]. \quad (9)$$

Since  $x_{t+1} = \Pi_{\mathcal{X}}[x_t - \eta \tilde{g}_t(x_t)]$ , for any  $x \in \mathcal{X}$  we have that

$$\begin{aligned} \|x_{t+1} - x\|^2 &= \|\Pi_{\mathcal{X}}[x_t - \eta \tilde{g}_t(x_t)] - \Pi_{\mathcal{X}}[x]\|^2 \\ &\leq \|x_t - \eta \tilde{g}_t(x_t) - x\|^2 \\ &= \|x_t - x\|^2 - 2\eta \langle \tilde{g}_t(x_t), x_t - x \rangle + \eta^2 \|\tilde{g}_t(x_t)\|^2. \end{aligned} \quad (10)$$

Rearranging the above inequality yields that

$$\begin{aligned} \langle \tilde{g}_t(x_t), x_t - x \rangle &\leq \frac{1}{2\eta} (\|x_t - x\|^2 - \|x_{t+1} - x\|^2) + \frac{\eta}{2} \|\tilde{g}_t(x_t)\|^2. \end{aligned} \quad (11)$$

Taking expectation on both sides of the above inequality over  $u_t$ , using inequality (9), and telescoping the resulting bound from  $t = 0$  to  $T$ , we obtain that

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=0}^T f_{\delta,t}(x_t) - \sum_{t=0}^T f_{\delta,t}(x) \right] &\leq \frac{1}{2\eta} \|x_0 - x\|^2 + \frac{\eta}{2} \mathbb{E} \left[ \sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2 \right]. \end{aligned}$$

Since  $f_t(x) \in C^{0,0}$ , we know that  $|f_{\delta,t}(x) - f_t(x)| \leq \delta L_0$ . Therefore, we obtain from the above inequality that

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x) \right] &= \mathbb{E} \left[ \sum_{t=0}^T f_{\delta,t}(x_t) - \sum_{t=0}^T f_{\delta,t}(x) \right] \\ &\quad + \mathbb{E} \left[ \sum_{t=0}^T (f_t(x_t) - f_{\delta,t}(x_t)) - \sum_{t=0}^T (f_t(x) - f_{\delta,t}(x)) \right] \\ &\leq \frac{1}{2\eta} \|x_0 - x\|^2 + \frac{\eta}{2} \mathbb{E} \left[ \sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2 \right] + 2L_0 \delta T. \end{aligned} \quad (12)$$

On the other hand, telescoping the second moment bound in (5) over  $t = 1, 2, \dots, T$ , adding  $\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]$  on both sides, adding  $\frac{4d^2L_0^2\eta^2}{\delta^2}\mathbb{E}[\|\tilde{g}_T(x_T)\|^2]$  to the right-hand side and using Assumption 3.1, we obtain that

$$\begin{aligned} & \mathbb{E} \left[ \sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2 \right] \\ & \leq \frac{1}{1-\alpha} \mathbb{E} [\|\tilde{g}_0(x_0)\|^2] + \frac{16}{1-\alpha} d^2 L_0^2 T + \frac{2d^2 V_f^2}{1-\alpha} \frac{1}{\delta^2} T \end{aligned} \quad (13)$$

where  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2}$ . Substituting the above bound into (12) yields that

$$\begin{aligned} & \mathbb{E} \left[ \sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x) \right] \leq \frac{\eta}{2(1-\alpha)} \mathbb{E} [\|\tilde{g}_0(x_0)\|^2] \\ & + \frac{1}{2\eta} \|x_0 - x\|^2 + \frac{8}{1-\alpha} L_0^2 d^2 \eta T + 2L_0 \delta T + \frac{d^2 V_f^2}{1-\alpha} \frac{\eta}{\delta^2} T. \end{aligned}$$

Since the above inequality holds for all  $x \in \mathcal{X}$ , we can replace  $x$  with  $x^*$ . When the upper bound on  $\|x_0 - x^*\| \leq R$  is known, let  $\eta = \frac{R^{\frac{1}{2}}}{2\sqrt{2}dL_0T^{\frac{1}{4}}}$  and  $\delta = \frac{\sqrt{dR}}{T^{\frac{1}{4}}}$ , so that  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2} = \frac{R^2}{2T} \leq \frac{1}{2}$ , when  $T \geq R^2$ . Then, we obtain that

$$\begin{aligned} & \mathbb{E} \left[ \sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x^*) \right] \\ & \leq \sqrt{2}dL_0\sqrt{RT}^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] R^{\frac{3}{2}}}{2\sqrt{2}dL_0T^{\frac{3}{4}}} + 4\sqrt{2}d^{\frac{3}{2}}L_0R^{\frac{3}{2}}T^{\frac{1}{4}} \\ & + 2L_0\sqrt{dRT}^{\frac{3}{4}} + \frac{\sqrt{dRV_f^2}}{\sqrt{2}L_0}T^{\frac{3}{4}}. \end{aligned} \quad (14)$$

When  $R$  is unknown, let  $\eta = \frac{1}{2\sqrt{2}dL_0T^{\frac{1}{4}}}$  and  $\delta = \frac{\sqrt{d}}{T^{\frac{1}{4}}}$ , so that  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2} = \frac{1}{2T} \leq \frac{1}{2}$ . Then, we obtain that

$$\begin{aligned} & \mathbb{E} \left[ \sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x^*) \right] \\ & \leq \sqrt{2}dL_0R^2T^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{2\sqrt{2}dL_0T^{\frac{3}{4}}} + 4\sqrt{2}d^{\frac{3}{2}}L_0T^{\frac{1}{4}} \\ & + 2\sqrt{d}L_0T^{\frac{3}{4}} + \frac{\sqrt{dV_f^2}}{\sqrt{2}L_0}T^{\frac{3}{4}}. \end{aligned} \quad (15)$$

### B. Proof of Theorem 3.3

As discussed above Theorem 3.3, there exists a constant  $L_0$  with which  $f_t$  is Lipschitz over the compact set  $\mathcal{X}_\delta$ . Since  $f_t(x) \in C^{1,1}$ , we know that  $|f_{\delta,t}(x) - f_t(x)| \leq \delta^2 L_1$ . Following the same proof logic as that for proving (12), we obtain that  $\mathbb{E}[\sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x)] \leq \frac{1}{2\eta} \|x_0 - x\|^2 + \frac{\eta}{2} \mathbb{E}[\sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2] + 2L_1\delta^2 T$ . Substituting the bound in (13) into the previous inequality, we obtain that  $\mathbb{E}[\sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x)] \leq \frac{1}{2\eta} \|x_0 - x\|^2 + \frac{\eta}{2(1-\alpha)} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \frac{8}{1-\alpha} L_0^2 d^2 \eta T + 2L_1\delta^2 T + \frac{d^2 V_f^2}{1-\alpha} \frac{\eta}{\delta^2} T$ . Since this inequality holds for all  $x \in \mathcal{X}$ , we can replace  $x$  with  $x^*$ .

Assuming the bound  $\|x_0 - x^*\| \leq R$  is known, let  $\eta = \frac{R^{\frac{1}{2}}}{2\sqrt{2}dL_0T^{\frac{1}{4}}}$  and  $\delta = \frac{\frac{1}{2}R^{\frac{1}{2}}}{T^{\frac{1}{4}}}$  so that  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2} = \frac{R^2}{2T} \leq \frac{1}{2}$  when  $T \geq R^2$ .

Plugging these parameters into above inequality, we finally obtain that

$$\begin{aligned} & \mathbb{E} \left[ \sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x) \right] \\ & \leq \sqrt{2}L_0 d^{\frac{3}{2}} R^{\frac{3}{2}} T^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] R^{\frac{3}{2}}}{2\sqrt{2}L_0 d^{\frac{3}{2}} T^{\frac{3}{4}}} + 4\sqrt{2}L_0 d^{\frac{3}{2}} R^{\frac{3}{2}} T^{\frac{1}{4}} \\ & + 2L_1 d^{\frac{3}{2}} R^{\frac{3}{2}} T^{\frac{3}{4}} + (\sqrt{2}L_0)^{-1} d^{\frac{3}{2}} R^{\frac{3}{2}} V_f^2 T^{\frac{3}{4}}. \end{aligned} \quad (16)$$

When the bound  $\|x_0 - x^*\| \leq R$  is unknown. Choose  $\eta = \frac{1}{2\sqrt{2}L_0 d^{\frac{3}{2}} T^{\frac{1}{4}}}$  and  $\delta = \frac{1}{d^{\frac{1}{2}} T^{\frac{1}{4}}}$  so that  $\alpha = \frac{4dL_0^2\eta^2}{\delta^2} = \frac{1}{2T} \leq \frac{1}{2}$ . Plugging these parameters into above inequality, we finally obtain that

$$\begin{aligned} & \mathbb{E} \left[ \sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x) \right] \\ & \leq \sqrt{2}L_0 d^{\frac{3}{2}} \|x_0 - x\|^2 T^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{2\sqrt{2}L_0 d^{\frac{3}{2}} T^{\frac{3}{4}}} + 8\sqrt{2}L_0 \frac{(d+4)^2}{d^{\frac{3}{2}}} T^{\frac{1}{4}} \\ & + 2d^{\frac{3}{2}} L_1 T^{\frac{3}{4}} + \frac{\sqrt{2}}{L_0} d^{\frac{3}{2}} V_f^2 T^{\frac{3}{4}}. \end{aligned} \quad (17)$$

The proof is complete.  $\blacksquare$

### C. Proof of Theorem 4.2

Note that  $f_t(x) \in C^{0,0}$ . According to Lemma 2.2,  $f_{\delta,t}(x)$  has  $L_{1,\delta}$ -Lipschitz continuous gradient with  $L_{1,\delta} = \frac{d}{\delta} L_0$ . Furthermore, according to [32, Lemma 1.2.3], we have the following inequality:

$$\begin{aligned} & f_{\delta,t}(x_{t+1}) \\ & \leq f_{\delta,t}(x_t) + \langle \nabla f_{\delta,t}(x_t), x_{t+1} - x_t \rangle + \frac{L_{1,\delta}}{2} \|x_{t+1} - x_t\|^2 \\ & = f_{\delta,t}(x_t) - \eta \langle \nabla f_{\delta,t}(x_t), \tilde{g}_t(x_t) \rangle + \frac{L_{1,\delta}\eta^2}{2} \|\tilde{g}_t(x_t)\|^2 \\ & + \frac{L_{1,\delta}\eta^2}{2} \|\tilde{g}_t(x_t)\|^2 \end{aligned} \quad (18)$$

where  $\Delta_t = \tilde{g}_t(x_t) - \nabla f_{\delta,t}(x_t)$ . According to Lemma 2.5, we know that  $\mathbb{E}_{u_t}[\tilde{g}_t(x_t)] = \nabla f_{\delta,t}(x_t)$ . Therefore, taking expectation over  $u_t$  conditional on  $x_t$  on both sides of inequality (18) and rearranging terms, we obtain that

$$\begin{aligned} & \eta \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] \\ & \leq \mathbb{E}[f_{\delta,t}(x_t)] - \mathbb{E}[f_{\delta,t}(x_{t+1})] + \frac{L_{1,\delta}\eta^2}{2} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \\ & \leq \mathbb{E}[f_{\delta,t}(x_t)] - \mathbb{E}[f_{\delta,t+1}(x_{t+1})] + \frac{L_{1,\delta}\eta^2}{2} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \\ & + \mathbb{E}[f_{\delta,t+1}(x_{t+1})] - \mathbb{E}[f_{\delta,t}(x_{t+1})] \end{aligned} \quad (19)$$

where the expectation is conditional on  $x_t$ . Then, we can further condition both sides of (19) on  $x_0$  without changing the sign of inequality, and then apply the tower rule of conditional expectation to make the expectation in (19) become full expectation. Telescoping the above inequality over  $t = 0, \dots, T-1$  and dividing both sides by  $\eta$ , we obtain that

$$\begin{aligned} & \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] \leq \frac{L_{1,\delta}\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \\ & + \frac{\mathbb{E}[f_{\delta,0}(x_0)] - \mathbb{E}[f_{\delta,T}(x_T)] + W_T}{\eta} \\ & \leq \frac{\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T}{\eta} + \frac{L_{1,\delta}\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \end{aligned} \quad (20)$$

where  $f_{\delta,T}^*$  is the lower bound of the smoothed function  $f_{\delta,T}(x)$ .  $f_{\delta,T}^*$  must exist because we assume that the original function  $f_t(x)$  is lower bounded and the smoothed function has a bounded distance from  $f_t(x)$  due to Lemma 2.2 for all  $t$ . The first inequality holds by Assumption 4.1; the bounded accumulated variation of the function  $f$  implies that the smoothed function  $f_\delta$  can also be bounded at the same level. This is due to the definition of the smoothed objective function  $f_{\delta,t}$ . Specifically, we have that  $\sum_{t=1}^T (f_{\delta,t}(x_t) - f_{\delta,t-1}(x_t)) = \mathbb{E}_{v_t \in \mathbb{B}} [\sum_{t=1}^T (f_t(x_t + \delta v_t) - f_{t-1}(x_t + \delta v_t))] \leq \mathbb{E}_{v_t \in \mathbb{B}} [W_T] = W_T$ .

Next, we derive the bound on  $\sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2]$ . Summing the bound in (5) from  $t = 1, \dots, T$ , adding  $\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]$  on both sides, and adding  $\frac{4d^2L_0^2\eta^2}{\delta^2} \mathbb{E}[\|\tilde{g}_T(x_T)\|^2]$  to the right-hand side, according to Assumption 4.1, we obtain that

$$\begin{aligned} & \mathbb{E} \left[ \sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2 \right] \\ & \leq \frac{1}{1-\alpha} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \frac{16}{1-\alpha} L_0^2 d^2 T + \frac{2d^2}{1-\alpha} \frac{\tilde{W}_T}{\delta^2}. \end{aligned} \quad (21)$$

Substituting this bound into the inequality (20), we obtain that

$$\begin{aligned} & \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] \\ & \leq \frac{\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^*}{\eta} + \frac{W_T}{\eta} + \frac{dL_0\eta}{2\delta} \frac{1}{1-\alpha} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2] \\ & \quad + \frac{dL_0\eta}{2\delta} \frac{16}{1-\alpha} L_0^2 d^2 T + \frac{dL_0\eta}{2\delta} \frac{2d^2}{1-\alpha} \frac{\tilde{W}_T}{\delta^2}. \end{aligned}$$

To fulfill the requirement that  $|f_t(x) - f_{\delta,t}(x)| \leq \epsilon_f$ , we set the exploration parameter  $\delta = \frac{\epsilon_f}{L_0}$ . In addition, let the stepsize be  $\eta = \frac{\epsilon_f^{1.5}}{2\sqrt{2}L_0^2 d^{1.5} T^{\frac{1}{2}}}$ . Then, we have that  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2} = \frac{\epsilon_f}{2dT} \leq \frac{1}{2}$  when  $T \geq \frac{\epsilon_f}{d}$ . Therefore, we have that  $\frac{1}{1-\alpha} \leq 2$ . Substituting this bound and the choices of  $\eta$  and  $\delta$  into the bound above, we finally obtain that  $\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] \leq 2\sqrt{2}L_0^2 (\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T) \frac{d^{1.5}}{\epsilon_f^{1.5}} T^{\frac{1}{2}} + \frac{\epsilon_f^{\frac{1}{2}} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{2\sqrt{2}dT} + 4\sqrt{2}L_0^2 \epsilon_f^{\frac{1}{2}} d^{1.5} T^{\frac{1}{2}} + \frac{L_0^2 d^{1.5} \tilde{W}_T}{\sqrt{2} \epsilon_f^{1.5} T^{\frac{1}{2}}}$ .

The proof is complete.  $\blacksquare$

#### D. Proof of Theorem 4.3

Note that when  $f_t \in C^{1,1}$  with Lipschitz constant  $L_1$ , the smoothed function  $f_{\delta,t} \in C^{1,1}$  with Lipschitz constant  $L_1$ . Therefore, following the proof of Theorem 4.2 but replacing  $L_{1,\delta}$  with  $L_1$ , we obtain that

$$\begin{aligned} & \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] \\ & \leq \frac{\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^*}{\eta} + \frac{L_1\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] + \frac{W_T}{\eta}. \end{aligned} \quad (22)$$

Since  $f_t \in C^{1,1}$ , according to Lemma 2.2, we have that  $\|\nabla f_{\delta,t}(x) - \nabla f_t(x)\| \leq dL_1\delta$ . Furthermore, we have that

$$\begin{aligned} & \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_t(x_t)\|^2] = \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_t(x_t) - \nabla f_{\delta,t}(x_t) + \nabla f_{\delta,t}(x_t)\|^2] \\ & \leq 2 \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_t(x_t) - \nabla f_{\delta,t}(x_t)\|^2] + 2 \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2]. \end{aligned}$$

Substituting the bound in (21) into (22) and using the bound above, we obtain that  $\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_t(x_t)\|^2] \leq 2 \frac{\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^*}{\eta} + 2 \frac{W_T}{\eta} + \frac{L_1}{1-\alpha} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2] \eta + \frac{16L_1}{1-\alpha} L_0^2 d^2 \eta T + \frac{2d^2L_1\tilde{W}_T}{1-\alpha} \frac{\eta}{\delta^2} + 2d^2L_1^2\delta^2T$ . Choose  $\eta = \frac{1}{2\sqrt{2}L_0d^{\frac{3}{2}}T^{\frac{1}{2}}}$  and  $\delta = \frac{1}{d^{\frac{1}{2}}T^{\frac{1}{2}}}$ . Then,  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2} = \frac{1}{2\sqrt{T}} \leq \frac{1}{2}$  and  $\frac{1}{1-\alpha} \leq 2$ . Substituting these results into the previous inequality, we finally obtain that  $\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_t(x_t)\|^2] \leq 4\sqrt{2}L_0(\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T) d^{\frac{3}{2}} T^{\frac{1}{2}} + \frac{L_1\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{\sqrt{2}L_0d^{\frac{3}{2}}T^{\frac{1}{2}}} + 8\sqrt{2}L_1L_0d^{\frac{3}{2}}T^{\frac{1}{2}} + \frac{\sqrt{2}L_1}{L_0} d^{\frac{3}{2}} \tilde{W}_T + 2L_1^2 d^{\frac{3}{2}} T^{\frac{1}{2}}$ . The proof is complete.  $\blacksquare$

#### REFERENCES

- [1] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, 2017, pp. 15–26.
- [2] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1467–1476.
- [3] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," in *Proc. 22nd Int. Conf. Artif. Intell. Statist. PMLR*, 2019, pp. 2916–2925.
- [4] A. Agarwal, O. Dekel, and L. Xiao, "Optimal algorithms for online convex optimization with multi-point bandit feedback," in *Proc. Annu. Conf. Learn. Theory*, 2010, pp. 28–40.
- [5] S. Ghadimi and G. Lan, "Stochastic first- and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2341–2368, 2013.
- [6] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2788–2806, May 2015.
- [7] F. Bach and V. Perchet, "Highly-smooth zero-th order online optimization," in *Proc. Conf. Learn. Theory*, 2016, pp. 257–283.
- [8] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations Comput. Math.*, vol. 17, no. 2, pp. 527–566, 2017.
- [9] X. Gao, X. Li, and S. Zhang, "Online learning with non-convex losses and non-stationary regret," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 235–243.
- [10] A. Roy, K. Balasubramanian, S. Ghadimi, and P. Mohapatra, "Stochastic zeroth-order optimization under nonstationarity and nonconvexity," *J. Mach. Learn. Res.*, vol. 23, no. 64, pp. 1–47, 2022.
- [11] E. Dall'Anese, A. Simonetto, S. Becker, and L. Madden, "Optimization and learning with information streams: Time-varying algorithms and applications," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 71–83, May 2020.
- [12] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proc. 17th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2005, pp. 385–394.
- [13] A. Saha and A. Tewari, "Improved regret guarantees for online smooth convex optimization with bandit feedback," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 636–642.
- [14] O. Dekel, R. Eldan, and T. Koren, "Bandit smooth convex optimization: Improving the bias-variance tradeoff," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2926–2934.
- [15] A. V. Gaskov, E. A. Kryukova, A. A. Lagunovskaya, I. N. Usmanova, and F. A. Fedorenko, "Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case," *Automat. Remote Control*, vol. 78, no. 2, pp. 224–234, 2017.
- [16] E. Hazan, K. Y. Levy, and S. Shalev-Shwartz, "On graduated optimization for stochastic non-convex problems," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1833–1841.
- [17] I. Shames, D. Selvaratnam, and J. H. Manton, "Online optimization using zeroth order oracles," *IEEE Control Syst. Lett.*, vol. 4, no. 1, pp. 31–36, Jan. 2020.
- [18] K. Balasubramanian and S. Ghadimi, "Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3455–3464.

- [19] E. Gorbunov, P. Dvurechensky, and A. Gasnikov, "An accelerated method for derivative-free smooth stochastic convex optimization," *SIAM J. Optim.*, vol. 32, no. 2, pp. 1210–1238, 2022.
- [20] K. Ji, Z. Wang, Y. Zhou, and Y. Liang, "Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization," in *Proc. Int. Conf. Mach. Learn. PMLR*, 2019, pp. 3100–3109.
- [21] Y. Wang, S. Du, S. Balakrishnan, and A. Singh, "Stochastic zeroth-order optimization in high dimensions," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 1356–1365.
- [22] D. Hajinezhad and M. M. Zavlanos, "Gradient-free multi-agent nonconvex nonsmooth optimization," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 4939–4944.
- [23] Y. Tang and N. Li, "Distributed zero-order algorithms for nonconvex multi-agent optimization," in *Proc. 57th Annu. Allerton Conf. Commun. Control Comput.*, 2019, pp. 781–786.
- [24] L. Zhang, T. Yang, R. Jin, and Z.-H. Zhou, "Online bandit learning for a special class of non-convex losses," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 3158–3164.
- [25] A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin, "Stochastic convex optimization with bandit feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1035–1043.
- [26] E. Hazan and Y. Li, "An optimal algorithm for bandit convex optimization," 2016, *arXiv:1603.04350*.
- [27] S. Bubeck, Y. T. Lee, and R. Eldan, "Kernel-based methods for bandit convex optimization," in *Proc. 49th Annu. ACM SIGACT Symp. Theory Comput.*, 2017, pp. 72–85.
- [28] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "A new one-point residual-feedback oracle for black-box learning and control," *Automatica*, vol. 136, p. 110006, 2022.
- [29] O. Bilenne, P. Mertikopoulos, and E.-V. Belmega, "Fast optimization with zeroth-order feedback in distributed, multi-user MIMO systems," *IEEE Trans. Signal Process.*, vol. 68, pp. 6085–6100, 2020.
- [30] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "Boosting one-point derivative-free online optimization via residual feedback," 2020, *arXiv:2010.07378*.
- [31] P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou, "Dynamic regret of convex and smooth functions," in *proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12510–12520.
- [32] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87. Berlin, Germany: Springer Science & Business Media, 2013.