A New One-Point Residual-Feedback Oracle For Black-Box Learning and Control

Yan Zhang* [a], Yi Zhou* [b], Kaiyi Ji [c], Michael M. Zavlanos [a]

[a] *Mechanical Enginerring and Material Science, Duke University, Durham, NC 27708 USA*

[b] *Electrical and Computer Engineering, The University of Utah, Salt Lake City, UT 84112 USA*

[c] *Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA*

Abstract

Zeroth-order optimization (ZO) algorithms have been recently used to solve black-box or simulation-based learning and control problems, where the gradient of the objective function cannot be easily computed but can be approximated using the objective function values. Many existing ZO algorithms adopt two-point feedback schemes due to their fast convergence rate compared to one-point feedback schemes. However, two-point schemes require two evaluations of the objective function at each iteration, which can be impractical in applications where the data are not all available a priori, e.g., in online optimization. In this paper, we propose a novel one-point feedback scheme that queries the function value once at each iteration and estimates the gradient using the residual between two consecutive points. When optimizing a deterministic Lipschitz function, we show that the query complexity of ZO with the proposed one-point residual feedback matches that of ZO with the existing two-point schemes. Moreover, the query complexity of the proposed algorithm can be improved when the objective function has Lipschitz gradient. Then, for stochastic bandit optimization problems where only noisy objective function values are given, we show that ZO with one-point residual feedback achieves the same convergence rate as that of two-point scheme with uncontrollable data samples. We demonstrate the effectiveness of the proposed one-point residual feedback via extensive numerical experiments.

Key words: Zeroth-Order Optimization, Residual-Feedback

1 Introduction

Zeroth-order optimization algorithms have been widely-used to solve control and machine learning problems where first or second order information (i.e., gradient or Hessian information) is unavailable, e.g., controlling complex systems whose dynamics can not be modeled explicitly but can only be given by high-fidelity simulators Ghadimi & Lan (2013), adversarial training Chen et al. (2017), reinforcement learning Fazel et al. (2018); Malik et al. (2018) and human-in-the-loop control Luo et al. (2020). In these problems, the goal is to solve the following generic optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{P}$$

where $x \in \mathbb{R}^d$ corresponds to the parameters and f denotes the total loss. Using zeroth-order information, i.e., function evaluations, first-order gradients can be estimated to solve the problem (P).

Existing zeroth-order optimization (ZO) algorithms can be divided into two categories, namely, ZO with one-point feedback and ZO with two-point feedback. Flaxman et al. (2005) was among the first to propose a ZO algorithm with one-point feedback, that queries one function value at each iteration to estimate the gradient. The corresponding one-point gradient estimator $\widetilde{\nabla} f(x)$ takes the form [2]

$$\text{(One-point feedback):}\ \widetilde{\nabla} f(x) = \frac{u}{\delta} f(x+\delta u), \tag{1}$$

Email addresses: yan.zhang2@duke.edu (Yan Zhang*), yi.zhou@utah.edu (Yi Zhou*), ji.367@osu.edu (Kaiyi Ji), michael.zavlanos@duke.edu (Michael M. Zavlanos).

[1] *Equal Contribution. This work is supported in part by AFOSR under award #FA9550-19-1-0169 and by NSF under award CNS-1932011.

[2] In Flaxman et al. (2005), the estimator is $\widetilde{\nabla} f(x) = \frac{du}{\delta} f(x+\delta u)$ where $x \in \mathbb{R}^d$ and u is uniformly sampled from a unit sphere in $\mathbb{R}^d$. In this paper, we follow Nesterov & Spokoiny (2017) and sample u from the standard normal distribution.

Table 1
Iteration Complexity of Zeroth-order Methods with One-point, Two-point and Proposed Feedback Schemes

Complexity [3]		Convex $C^{0,0}$	Convex $C^{1,1}$	Nonconvex $C^{0,0}$	Nonconvex $C^{1,1}$
One-point	Gasnikov et al. (2017)	$d^2\epsilon^{-4}$	$d^2\epsilon^{-3}$	–	–
Two-point	Duchi et al. (2015)	$d\log(d)\epsilon^{-2}$	$d\epsilon^{-2}$	–	–
	Shamir (2017)	$d\epsilon^{-2}$	–	–	–
	Nesterov & Spokoiny (2017)	$d^2\epsilon^{-2}$	$d\epsilon^{-1}$	$d^3\epsilon_f^{-1}\epsilon^{-2}$	$d\epsilon^{-1}$
	Bach & Perchet (2016)	–	$d^2\epsilon^{-3}$ (UN)	–	–
Residual One-point	Deterministic	$d^2\epsilon^{-2}$	$d^3\epsilon^{-1.5}$	$d^4\epsilon_f^{-1}\epsilon^{-2}$	$d^3\epsilon^{-1.5}$
	Stochastic	$d^2\epsilon^{-4}$	$d^2\epsilon^{-3}$	$d^3\epsilon_f^{-3}\epsilon^{-2}$	$d^4\epsilon^{-3}$

where δ is an exploration parameter and $u \in \mathbb{R}^d$ is sampled from the standard normal distribution elementwise. In particular, Flaxman et al. (2005) showed that the above one-point gradient estimator has a large estimation variance and the resulting ZO algorithm achieves a convergence rate of at most $\mathcal{O}(T^{-\frac{1}{4}})$, where T is the number of iterations, which is much slower than that of gradient descent algorithms used to solve problem (P). Assuming smoothness and relying on self-concordant regularization, Dekel et al. (2015); Saha & Tewari (2011) further improved this convergence speed. However, the gap in the iteration complexity between ZO algorithms with one-point feedback and gradient-based methods remained. In order to reduce the large estimation variance of the above one-point gradient estimator, Agarwal et al. (2010); Nesterov & Spokoiny (2017); Shamir (2017) introduced the following two-point gradient estimators

$$\text{(Two-point feedback):}\ \widetilde{\nabla} f(x) = \frac{u}{\delta}\big(f(x+\delta u) - f(x)\big), \quad \text{or}\ \frac{u}{2\delta}\big(f(x+\delta u) - f(x-\delta u)\big), \tag{2}$$

that have lower estimation variance and showed that ZO with these two-point feedbacks achieves a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$ (or $\mathcal{O}(\frac{1}{T})$ when the problem is smooth), which is order-wise much faster than the convergence rate achieved by ZO algorithms with one-point feedback. Therefore, as also pointed out in Larson et al. (2019), a fundamental question we seek to answer in this paper is:

- *(Q1): Does there exist a one-point feedback for which zeroth-order optimization can achieve the same query complexity as that of two-point feedback methods?*

The literature discussed above focuses on deterministic optimization problems (P). Nevertheless, in practice, many problems involve randomness in the environment and parameters, giving rise to the following stochastic optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) = \mathbb{E}_\xi[F(x,\xi)], \tag{Q}$$

where only a noisy function evaluation $F(x,\xi)$ with a random data sample ξ is available. ZO algorithms have also been developed to solve the above problem (Q), e.g., Akhavan et al. (2020); Bach & Perchet (2016); Duchi et al. (2015); Gasnikov et al. (2017); Ghadimi & Lan (2013); Hu et al. (2016). In particular, Ghadimi & Lan (2013) consider the following widely-used stochastic two-point feedback

$$\widetilde{\nabla} f(x) = \frac{u}{\delta}\big(F(x+\delta u,\xi) - F(x,\xi)\big) \tag{3}$$

and show that ZO with this stochastic two-point feedback has the same convergence rate as ZO with the two-point feedback scheme in (2) for deterministic problems (P). Similarly, Duchi et al. (2015) further analyzed the oracle in (3) in a mirror descent framework and showed a similar convergence speed. Stochastic one-point and two-point feedback schemes with improved convergence rates have also been studied in Gasnikov et al. (2017). However, these stochastic two-point feedback schemes assume that the data sample ξ is controllable, i.e., one can fix the data sample ξ and evaluate the function value at two distinct points x and $x+\delta u$. This assumption is unrealistic in many applications. For example, in reinforcement learning, controlling the sample ξ requires applying the same sequence of noises to the dynamical system and reward function. Hence, two-point feedback schemes with fixed data samples can be impractical. To address this challenge, Akhavan et al. (2020); Bach &

[3] In convex setting, the accuracy is meaured by $f(x) - f(x^*) \le \epsilon$, where $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$, while in the non-convex setting, it is measured by $\|\nabla f(x)\|^2 \le \epsilon$ when the objective function is smooth. When the objective function is non-smooth, we enforce two optimality measures, $|f(x) - f_\delta(x)| \le \epsilon_f$ and $\|\nabla f_\delta(x)\|^2 \le \epsilon$ together, where function $f_\delta(x)$ is a smoothed function defined as $f_\delta(x) := \mathbb{E}_{u\sim\mathcal{N}(0,1)}[f(x+\delta u)]$. (UN) means the oracle considers uncontrollable data samples. The notations $C^{0,0}$ and $C^{1,1}$ represent the function classes that are either Lipschitz, or have Lipschitz gradient. The detailed definition of these notations can be found in Definition 1.

Perchet (2016); Hu et al. (2016) proposed a more practical noisy two-point feedback method that replaces the fixed sample ξ in (3) with two independent samples ξ, ξ'. Its convergence rate was shown to match that of the stochastic one-point feedback $\widetilde{\nabla} f(x) = \frac{u}{\delta} F(x + \delta u, \xi)$. Still though, this two-point feedback method with independent data samples produces gradient estimates with lower variance compared to the conventional one-point feedback method. Therefore, an additional fundamental question we seek to answer in this paper is:

- *(Q2): Can we develop a stochastic one-point feedback that achieves the same practical performance as that of the noisy two-point feedback?*

Contributions: In this paper, we provide positive answers to these open questions by introducing a new one-point residual feedback scheme and theoretically analyzing the convergence of zeroth-order optimization using this feedback scheme. Specifically, our contributions are as follows. We propose a new one-point feedback scheme which requires a single function evaluation at each iteration. This feedback scheme estimates the gradient using the residual between two consecutive feedback points and we refer to it as residual feedback. We show that our residual feedback induces a smaller estimation variance than the one-point feedback (1) considered in Flaxman et al. (2005); Gasnikov et al. (2017). Specifically, in deterministic optimization where the objective function is Lipschitz-continuous, we show that ZO with our residual feedback achieves the same convergence rate as existing ZO with two-point feedback schemes. To the best of our knowledge, this is the first one-point feedback scheme with provably comparable performance to two-point feedback schemes in ZO. Moreover, when the objective function has an additional smoothness structure, we further establish an improved convergence rate of ZO with residual feedback. In the stochastic case where only noisy function values are available, we show that the convergence rate of ZO with residual feedback matches the state-of-the-art result of ZO with two-point feedback under uncontrollable data samples. Hence, our residual feedback bridges the theoretical gap between ZO with one-point feedback and ZO with two-point feedback. A summary of the complexity results for the proposed residual-feedback scheme can be found in Table 1.

Applications in Learning and Control: The proposed one-point residual-feedback oracle has important applications in a variety of learning and control problems where the gradients are unavailable or difficult to compute. For example, it can be used to reduce the number of black-box function evaluations, compared to the conventional one-point oracle, in optimal charging problems for electrical vechicles Li et al. (2021), extreme seeking problems for ABS control for automotive brakes Nešić (2009); Poveda & Li (2021). In addition, residual feedback can reduce the computational cost of ZO methods for distributed reinforcement learning problems, while maintaining a similar convergence rate as that achieved by two-point methods Zhang & Zavlanos (2020). This is because residual feedback, being a one-point method, requires only a single policy evaluation (generally an expensive calculation) at each iteration to estimate the policy gradient. Moreover, residual feedback can significantly improve the convergence speed of ZO algorithms for non-stationary reinforcement learning problems, as shown in Zhang et al. (2020a). Note that two-point methods can not be used for non-stationary reinforcement learning problems because they require two different policy evaluations in the same environment, which is not possible when the environment is non-stationary and changes after each policy evaluation. Compared to these works, here we focus on the iteration complexity of ZO methods with one-point residual feedback for static optimization problems, under different assumptions on the objective functions and their evaluation. This analysis, that is summarized in Table 1, lays the theoretical foundations of residual feedback and justifies its use for the more challenging learning and control problems discussed above.

2 Preliminaries

In this section, we present definitions and preliminary results needed throughout our analysis. Following Bach & Perchet (2016); Nesterov & Spokoiny (2017), we introduce the following classes of Lipschitz and smooth functions.

Definition 1 (Lipschitz functions) *The class of Lipschtiz-continuous functions $C^{0,0}$ satisfy: for any $f \in C^{0,0}$, $|f(x) - f(y)| \leq L_0\|x - y\|$, $\forall x, y \in \mathbb{R}^d$, for some Lipschitz parameter $L_0 > 0$. The class of smooth functions $C^{1,1}$ satisfy: for any $f \in C^{1,1}$, $\|\nabla f(x) - \nabla f(y)\| \leq L_1\|x - y\|$, $\forall x, y \in \mathbb{R}^d$, for some Lipschitz parameter $L_1 > 0$.*

In ZO, the objective is to estimate the first-order gradient of a function using zeroth-order oracles. Necessarily, we need to perturb the function around the current point along all the directions uniformly in order to estimate the gradient. This motivates us to consider the Gaussian-smoothed version of the function f as introduced in Nesterov & Spokoiny (2017), $f_\delta(x) := \mathbb{E}_{u\sim\mathcal{N}(0,1)}[f(x+\delta u)]$, where the coordinates of the vector u are i.i.d standard Gaussian random variables. The following bounds on the approximation error of the function $f_\delta(x)$ have been developed in Nesterov & Spokoiny (2017).

Lemma 2 (Gaussian approximation) *Consider a function f and its Gaussian-smoothed version f_δ. It holds that*

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

$$\text{and } \|\nabla f_\delta(x) - \nabla f(x)\| \leq \delta L_1 (d+3)^{3/2}, \text{ if } f \in C^{1,1}.$$

Moreover, the smoothed function $f_\delta(x)$ has the following nice geometrical property as proved in Nesterov & Spokoiny (2017).

Lemma 3 *If function $f \in C^{0,0}$ is L_0-Lipschitz, then its Gaussian-smoothed version f_δ belongs to $C^{1,1}$ with Lipschitz constant $L_1 = \sqrt{d}\delta^{-1}L_0$.*

We also introduce the following notions of convexity.

Definition 4 (Convexity) *A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called convex if for all $x, y \in \mathbb{R}^d$, $f(x) \geq f(y) + \langle x - y, \nabla f(y)\rangle$.*

3 Deterministic ZO with Residual Feedback

In this section, we consider the problem (P), where the objective function evaluation is fully deterministic. To solve this problem, we propose a zeroth-order estimate of the gradient based on the following *one-point residual feedback* scheme

$$\widetilde{g}(x_t) := \frac{u_t}{\delta}\big(f(x_t + \delta u_t) - f(x_{t-1} + \delta u_{t-1})\big), \text{ [4]} \qquad (4)$$

where u_{t-1} and u_t are independent random vectors sampled from the standard multivariate Gaussian distribution. To elaborate, the gradient estimate in (4) evaluates the function value at one perturbed point $x_t + \delta u_t$ at each iteration t and the other function value evaluation $f(x_{t-1} + \delta u_{t-1})$ is inherited from the previous iteration. Therefore, it is a one-point feedback scheme based on the residual between two consecutive feedback points, and we name it *one-point residual feedback*. Next, we show that this estimator is an unbiased gradient estimate of the smoothed function $f_\delta(x)$ at x_t.

Lemma 5 *We have $\mathbb{E}\big[\tilde{g}(x_t)\big] = \nabla f_\delta(x_t)$ for all $x_t \in \mathbb{R}^d$.*

PROOF. The proof is straightforward because u_t is independent from u_{t-1} and has zero mean. □

Since $\tilde{g}(x_t)$ is an unbiased estimate of $\nabla f_\delta(x_t)$, we can use it in Stochastic Gradient Descent (SGD) as follows

$$x_{t+1} = x_t - \eta\tilde{g}(x_t), \qquad (5)$$

where η is the stepsize. To analyze the convergence of the above ZO algorithm with residual feedback, we need to bound the variance of the gradient estimate under proper choices of the exploration parameter δ in (4) and the stepsize η. In the following result, we present the bounds on the second moment of the gradient estimate $\mathbb{E}[\|\tilde{g}(x_t)\|^2]$, which will be used in our analysis later.

Lemma 6 *Consider a function $f \in C^{0,0}$ with Lipschitz constant L_0. Then, under the SGD update rule in (5), the second moment of the residual feedback satisfies*

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \frac{2dL_0^2\eta^2}{\delta^2}\mathbb{E}[\|\tilde{g}(x_{t-1})\|^2] + 8L_0^2(d+4)^2.$$

Furthermore, if $f(x)$ also belongs to $C^{1,1}$ with constant L_1, then the second moment of the residual feedback satisfies

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \frac{2dL_0^2\eta^2}{\delta^2}\mathbb{E}[\|\tilde{g}(x_{t-1})\|^2] + 8(d+4)^2\|\nabla f(x_{t-1})\|^2 + 4L_1^2(d+6)^3\delta^2. \qquad (6)$$

The proof of above Lemma 6 can be found in Appendix A. Lemma 6 shows that the second moment of the residual feedback $\mathbb{E}[\|\tilde{g}(x_t)\|^2]$ can be bounded by a perturbed contraction under the SGD update rule. This perturbation term is crucial to establish the iteration complexity of ZO with our residual feedback. In particular, with the traditional one-point feedback, the perturbation term is in the order of $O(\delta^{-2})$ and significantly degrades the convergence speed Hu et al. (2016). In comparison, our residual feedback induces a much smaller perturbation term. Specifically, when $f \in C^{0,0}$, the perturbation is the order of $O(L_0^2d^2)$ that is independent of δ, and when $f \in C^{1,1}$, the perturbation is in the order of $O(d^2\|\nabla f(x_{t-1})\|^2 + L_1^2d^3\delta^2)$. Therefore, ZO with our residual feedback can achieve a better iteration complexity than that of ZO with the traditional one-point feedback.

3.1 Convergence Analysis

We first consider the case where the objective function f is nonconvex. When f is differentiable, we say a solution x is ϵ-accurate if $\mathbb{E}[\|\nabla f(x)\|^2] \leq \epsilon$. However, when f is nonsmooth, the gradient of the original objective function $\nabla f(x)$ does not exist. On the other hand, the smoothed objective function $f_\delta(x)$ is differentiable. Therefore, we find an ϵ-accurate solution of the smoothed problem such that $\mathbb{E}[\|\nabla f_\delta(x)\|^2] \leq \epsilon$. In the meantime, we require f_δ to be ϵ_f-close to the original objective function f, which requires $\delta \leq \frac{\epsilon_f}{L_0\sqrt{d}}$ according to Lemma 2. Similar optimality conditions have also been considered in Nesterov & Spokoiny (2017). Under this setup, the convergence rate of ZO with residual feedback is presented below. For simplicity, all the complexity results in this paper are presented in $\mathcal{O}$ notations. The proofs and the explicit form of the constant terms can be found in the supplementary material.

[4] At time $t = 0$, we can query the objective function f at $x_0 + \delta u_0$ and update x_0 using the conventional one-point oracle (1). Then, starting from time $t = 1$, we can update using estimator (4).

Theorem 7 *Assume that $f \in C^{0,0}$ with Lipschitz constant L_0 and that f is also bounded below by f^*. Moreover, assume that SGD in (5) with residual feedback is run for $T > 1/\epsilon_f$ iterations and that $\tilde{x}$ is selected from the T iterates uniformly at random. Let also $\eta = \frac{\sqrt{\epsilon_f}}{2dL_0^2\sqrt{T}}$ and $\delta = \frac{\epsilon_f}{L_0 d^{\frac{1}{2}}}$. Then, we have that $\mathbb{E}\big[\|\nabla f_\delta(\tilde{x})\|^2\big] = \mathcal{O}(d^2\epsilon_f^{-0.5}T^{-0.5})$.*

The proof can be found in Appendix B. Based on the above convergence rate result, the required iteration complexity to achieve a point x that satisfies $|f(x) - f_\delta(x)| \leq \epsilon_f$ as well as $\mathbb{E}[\|\nabla f(\tilde{x})\|^2] \leq \epsilon$ is of the order $\mathcal{O}(\frac{d^4}{\epsilon_f\epsilon^2})$. This complexity result is close to the complexity result $\mathcal{O}(\frac{d^3}{\epsilon_f\epsilon^2})$ of ZO with two-point feedback in Nesterov & Spokoiny (2017). When $f(x) \in C^{1,1}$ is a smooth function, we obtain the following convergence rate result for ZO with residual feedback.

Theorem 8 *Assume that $f(x) \in C^{0,0}$ with Lipschitz constant L_0 and that $f(x) \in C^{1,1}$ with Lipschitz constant L_1. Moreover, assume that SGD in (5) with residual feedback is run for T iterations and that $\tilde{x}$ is selected from the T iterates uniformly at random. Let also $\eta = \frac{1}{\widetilde{L}(d+4)^2T^{\frac{1}{3}}}$, and $\delta = \frac{1}{\sqrt{d}T^{\frac{1}{3}}}$, where $\widetilde{L} = \max(2L_0, 32L_1)$. Then, we have that $\mathbb{E}\big[\|\nabla f(\tilde{x})\|^2\big] = \mathcal{O}(d^2T^{-\frac{2}{3}})$.*

The proof can be found in Appendix C. In particular, to achieve a point x that satisfies $\mathbb{E}\big[\|\nabla f(\tilde{x})\|^2\big] \leq \epsilon$, the required iteration complexity is of the order $\mathcal{O}(d^3\epsilon^{-\frac{3}{2}})$. To the best of our knowledge, the best complexity result for ZO with two-point feedback is of the order $\mathcal{O}(d\epsilon^{-1})$, which is established in Nesterov & Spokoiny (2017). Next, we consider the case where the objective function f is convex. In this case, the optimality of a solution x is measured via the loss gap $f(x) - f(x^*)$, where x^* is the global optimum of f.

Theorem 9 *Assume that $f(x) \in C^{0,0}$ is convex with Lipschitz constant L_0. Moreover, assume that SGD in (5) with residual feedback is run for T iterations and define the running average $\bar{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_t$. Let also $\eta = \frac{1}{2dL_0\sqrt{T}}$ and $\delta = \frac{1}{\sqrt{T}}$. Then, we have that $f(\bar{x}) - f(x^*) = \mathcal{O}(dT^{-0.5})$.*

Moreover, assume that additionally $f(x) \in C^{1,1}$ with Lipschitz constant L_1, and let $\eta = \frac{1}{2\tilde{L}(d+4)^2T^{\frac{1}{3}}}$ and $\delta = \frac{\sqrt{d}}{T^{\frac{1}{3}}}$, where $\tilde{L} = \max\{L_0, 16L_1\}$. Then, we have that $f(\bar{x}) - f(x^*) = \mathcal{O}(d^2T^{-\frac{2}{3}})$.

The proof can be found in Appendix D. To elaborate, to achieve a solution x that satisfies $f(\bar{x}) - f(x^*) \leq \epsilon$, the required iteration complexity is of the order $\mathcal{O}(d^2\epsilon^{-2})$ when $f \in C^{0,0}$. Such a complexity result significantly improves the complexity $\mathcal{O}(d^2\epsilon^{-4})$ of ZO with the traditional one-point feedback and is slightly worse than the best complexity $\mathcal{O}(d\epsilon^{-2})$ of ZO with two-point feedback. On the other hand, when $f(x) \in C^{1,1}$, the required iteration complexity of ZO with residual feedback further reduces to $\mathcal{O}(d^3\epsilon^{-1.5})$, which is better than the complexity $\mathcal{O}(d\epsilon^{-3})$ of ZO with the traditional one-point feedback whenever $\epsilon < d^{-4/3}$.

4 Online ZO with Stochastic Residual Feedback

In this section, we study the Problem (Q) where the objective function takes the form $f(x) := \mathbb{E}[F(x,\xi)]$ and only noisy samples of the function value $F(x,\xi)$ are available. Specifically, we propose the following stochastic residual feedback

$$\tilde{g}(x_t) := \frac{u_t}{\delta}\big(F(x_t + \delta u_t, \xi_t) - F(x_{t-1} + \delta u_{t-1}, \xi_{t-1})\big), \tag{7}$$

where ξ_{t-1} and ξ_t are independent random samples that are sampled in iterations $t-1$ and t, respectively. We note that our stochastic residual feedback is more practical than most existing two-point feedback schemes, which require the data samples to be controllable, i.e., one can query the function value at two different variables using the same data sample. This assumption is unrealistic in applications where the environment is dynamic. For example, in reinforcement learning Malik et al. (2018), these data samples can correspond to random initial states, noises added to the dynamical system, and reward functions. Therefore, controlling the data samples requires to hard reset the system to the exact same initial state and apply the same sequence of noises, which is impossible when the data is collected from a real-world system. Our stochastic residual feedback scheme in (7) does not suffer from the same issue since it does not restrict the data sampling procedure. Instead, it simply takes the residual between two consecutive stochastic feedback points. In particular, it is straightforward to show that (7) is an unbiased gradient estimate of the objective function $f_\delta(x)$. Next, we present some assumptions that are used in our analysis later.

Assumption 10 *(Bounded Variance) We assume that for any $x \in \mathbb{R}^d$ there exists $\sigma > 0$ such that*

$$\mathbb{E}\big[\big(F(x,\xi) - f(x)\big)^2\big] \leq \sigma^2.$$

Assumption 10 implies that $\mathbb{E}\big[\big(F(x,\xi_1) - F(x,\xi_2)\big)^2\big] \leq 4\sigma^2$. Furthermore, we make the following smoothness assumption in the stochastic setting.

Assumption 11 *Let function $F(x,\xi) \in C^{0,0}$ with Lipschitz constant $L_0(\xi)$. We assume that $L_0(\xi) \leq L_0$ for all $\xi \in \Xi$. In addition, let the function $F(x,\xi) \in C^{1,1}$ with*

Lipschitz constant $L_1(\xi)$. We assume that $L_1(\xi) \leq L_1$ for all $\xi \in \Xi$.

The following lemma provides an upper bound of $\mathbb{E}[\|\tilde{g}(x_t)\|^2]$ in this stochastic setting.

Lemma 12 *Let Assumptions 10 and 11 hold and assume $F(x,\xi) \in C^{0,0}$ with Lipschitz constant $L_0(\xi)$. We have that*

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \frac{4L_0^2 d\eta^2}{\delta^2}\mathbb{E}[\|\tilde{g}(x_{t-1})\|^2]+16L_0^2(d+4)^2+\frac{8\sigma^2 d}{\delta^2}.$$

The proof can be found in Appendix E. If we assume that $F(x,\xi) \in C^{1,1}$, the upper bound on the above second moment can be further improved (see supplementary material for the details). However, this improvement does not yield a better iteration complexity due to the uncontrollable samples ξ_t and ξ_{t-1}. More specifically, the uncontrollable samples lead to an additional term $\frac{8\sigma^2 d}{\delta^2}$ in the above second moment bound. According to the analysis in Hu et al. (2016), such a term can significantly degrade the iteration complexity.

4.1 Convergence Analysis

Next, we analyze the iteration complexity of ZO with stochastic residual feedback for both non-convex and convex problems.

Theorem 13 *Let Assumptions 10 and 11 hold and assume also that $F(x,\xi) \in C^{0,0}$. Moreover, assume that SGD in (5) with residual feedback is run for $T > 1/(d\epsilon_f)$ iterations and that $\tilde{x}$ is selected from the T iterates uniformly at random. Let also $\eta = \frac{\epsilon_f^{1.5}}{2\sqrt{2}L_0^2 d^{1.5}\sqrt{T}}$ and $\delta = \frac{\epsilon_f}{L_0\sqrt{d}}$. Then, we have that $\mathbb{E}\big[\|\nabla f_\delta(\tilde{x})\|^2\big] = \mathcal{O}(d^{1.5}\epsilon_f^{-1.5}T^{-0.5})$.*

Furthermore, assume that additionally $F(x,\xi) \in C^{1,1}$, and that SGD in (5) with residual feedback is run for $T > 2$ iterations. Let also $\eta = \frac{1}{2L_0 d^{\frac{4}{3}}T^{\frac{2}{3}}}$ and $\delta = \frac{1}{d^{\frac{5}{6}}T^{\frac{1}{6}}}$. Then, the output $\tilde{x}$ that is sampled uniformly from the T iterates satisfies $\mathbb{E}\big[\|\nabla f(\tilde{x})\|^2\big] = \mathcal{O}(d^{\frac{4}{3}}T^{-\frac{1}{3}})$.

The proof can be found in Appendix F. Based on the above results, when $F(x,\xi)$ is non-smooth, to achieve the $\epsilon-$stationary point $\mathbb{E}\big[\|\nabla f_\delta(\tilde{x})\|^2\big] \leq \epsilon$ and $|f(x) - f_\delta(x)| \leq \epsilon_f$, $\mathcal{O}(\frac{d^3}{\epsilon_f^3\epsilon^2})$ iterations are needed. In addition, if the function $F(x,\xi)$ also satisfies $F(x,\xi) \in C^{1,1}$, then $\mathcal{O}(\frac{d^4}{\epsilon^3})$ iterations are needed to find the $\epsilon-$stationary point of the original function $f(x)$. Next, we provide the iteration complexity results when the Problem (Q) is convex.

Theorem 14 *Let Assumptions 10 and 11 hold and assume that the function $F(x,\xi) \in C^{0,0}$ is also convex. Moreover, assume that SGD in (5) with residual feedback is run for T iterations and define the running average $\bar{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_t$. Let also $\eta = \frac{1}{2\sqrt{2}L_0\sqrt{d}T^{\frac{3}{4}}}$ and $\delta = \frac{1}{T^{\frac{1}{4}}}$. Then, we have that $f(\bar{x}) - f(x^*) = \mathcal{O}(\sqrt{d}T^{-\frac{1}{4}})$. Moreover, assume that additionally $F(x,\xi) \in C^{1,1}$, and let $\eta = \frac{1}{2\sqrt{2}L_0 d^{\frac{2}{3}}T^{\frac{2}{3}}}$ and $\delta = \frac{1}{d^{\frac{1}{6}}T^{\frac{1}{6}}}$. Then, we have that $f(\bar{x}) - f(x^*) = \mathcal{O}(d^{\frac{2}{3}}T^{-\frac{1}{3}})$.*

The proof can be found in Appendix G. According to Theorem 14, $\mathcal{O}(\frac{d^2}{\epsilon^4})$ iterations are needed to achieve $f(\bar{x}) - f(x^*) \leq \epsilon$ with a nonsmooth objective function. On the other hand, if $f(x) \in C^{1,1}$, the iteration complexity is improved to $\mathcal{O}(\frac{d^2}{\epsilon^3})$.

5 ZO with Mini-batch Stochastic Residual Feedback

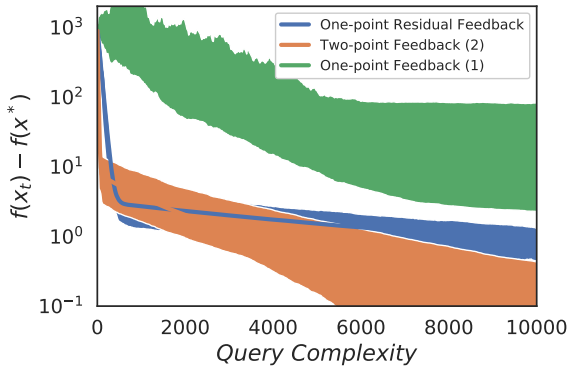
When applying zeroth-order oracles to practical applications, instead of directly using the oracle (7), a mini-batch scheme can be implemented to further reduce the variance of the gradient estimate, as discussed in Fazel et al. (2018). To be more specific, consider the gradient estimate with batch size b:

$$\tilde{g}_b(x_t) = \frac{u_t}{b\delta}\big(F(x_t+\delta u_t, \xi_{1:b}) - F(x_{t-1}+\delta u_{t-1}, \xi'_{1:b})\big),$$

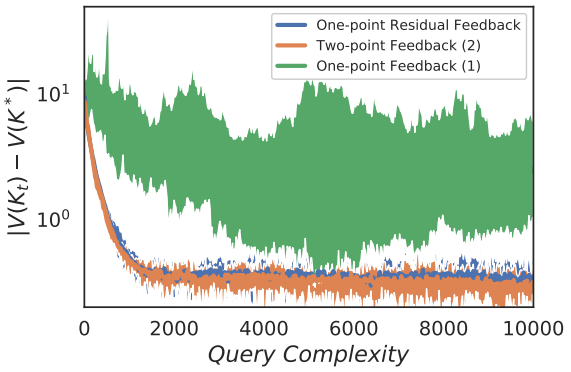
where $F(x_t+\delta u_t, \xi_{1:b}) = \sum_{j=1}^{b} F(x_t+\delta u_t, \xi_j)$. It is straightforward to see that the variance of $\tilde{g}_b(x_t)$ is b^2 times smaller than that of the oracle (7). This is particularly useful when the problem is sensitive to bad search directions. For example, in the policy optimization problem Fazel et al. (2018), when the gradient has large variance, it can drive the policy parameter to divergence and result in infinite cost. A Mini-batch scheme can reduce the variance of the policy gradient (search direction) estimate and therefore is of particular interest in this scenario. In this paper, we show that using the oracle (7) in a mini-batch scheme can achieve the same query complexity as standard SGD. Its analysis is provided in Zhang et al. (2020b).

6 Numerical Experiments

In this section, we demonstrate the effectiveness of the residual one-point feedback scheme for both deterministic and stochastic problems. In the deterministic case, we compare the performance of the proposed oracle with the original one-point feedback and two-point feedback schemes, for the quadratic programming (QP) example considered in Shamir (2013). In the stochastic case, we employ the stochastic variants of above oracles to optimize the policy parameters in a Linear Quadratic



(a)



(b)

Fig. 1. The convergence rate of applying the proposed residual one-point feedback (4) (blue), the two-point oracle (2) in Nesterov & Spokoiny (2017) (orange) and the one-point oracle (1) in Flaxman et al. (2005) (green) to two problems. In (a), the convergence of $f(x_t) - f(x^*)$ in a deterministic QP problem is presented. In (b), the convergence of the costs of policies in the stochastic LQR problem is presented.

Regulation (LQR) problem considered in Fazel et al. (2018); Malik et al. (2018). It is shown that the proposed residual one-point feedback significantly outperforms the traditional one-point feedback and its convergence rate matches that of the two-point oracles in both deterministic and stochastic cases. Furthermore, we apply our residual-feedback zeroth-order gradient estimate to solve a large-scale stochastic multi-stage decision making problem to demonstrate its performance in the high dimensional problems. All experiments are conducted using Matlab R2018b on a 2018 Macbook Pro with a 2.3 GHz Quad-Core Intel Core i5 and 8GB 2133MHz memory.

In all the experiments, we first manually select the exploration parameter δ. Then, we tune the stepsize η so that all algorithms converge at their fastest speed.

6.1 A Deterministic Scenario: QP Problem

As in Shamir (2013), consider the QP example $\min \frac{1}{2}(x-c)^T M(x-c)$, where $x, c \in \mathbb{R}^{30}$ and $M \in \mathbb{R}^{30\times 30}$ is a positive semi-definite matrix. This constitutes a convex and smooth problem. The vector c is randomly generated from a uniform distribution in $[0, 2]$. The matrix $M = PP^T$, where each entry in $P \in \mathbb{R}^{30\times 29}$ is sampled from a uniform distribution in $[0, 1]$. The initial point is the origin. For every algorithm, we manually optimize the selection of the exploration parameter δ and stepsize η and run it 100 times. Specifically, we select δ as $\delta = 0.1$, and the stepsizes for the proposed residual feedback estimator, the two-point estimator and the conventional one-point estimator are $0.05, 0.1, 0.01$, respectively. The convergence of the function value $f(x) - f(x^*)$ is presented in Figure 1(a). We observe that the proposed oracle converges as fast as the two-point oracle (2) when the iterates are far from the optimizer but achieve less accuracy in the end. Both methods find the optimal function value much faster than the one-point feedback studied in Flaxman et al. (2005); Gasnikov et al. (2017). These observations validate our theoretical results in Section 3.

6.2 A Stochastic Scenario: Policy Optimization

We use the proposed residual feedback to optimize the policy parameters in a LQR problem, as in Fazel et al. (2018); Malik et al. (2018). Specifically, consider a system whose state $x_k \in \mathbb{R}^{n_x}$ at time k is subject to the dynamical equation $x_{k+1} = Ax_k + Bu_k + w_k$, where $u_k \in \mathbb{R}^{n_u}$ is the control input at time k, $A \in \mathbb{R}^{n_x\times n_x}$ and $B \in \mathbb{R}^{n_x\times n_u}$ are dynamical matrices that are unknown, and w_k is the noise on the state transition. Moreover, consider a state feedback policy $u_k = Kx_k$, where $K \in \mathbb{R}^{n_u\times n_x}$ is the policy parameter. Policy optimization essentially aims to find the optimal policy parameter K so that the discounted accumulated cost function $V(K) := \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t (x_k^T Q x_k + u_k^T R u_k)\big]$ is minimized, where $\gamma \leq 1$ is the discount factor.

In our simulation, we select $n_x = 6$, $n_u = 6$ and $\gamma = 0.5$. Therefore, the problem has dimension $d = 36$. When implementing the policy $u_k = K_t x_k$, due to the noise w_k, evaluation of the cost of the policy K_t is noisy. We apply the one-point feedback (1) with noise Gasnikov et al. (2017), two-point feedback with uncontrolled noise Bach & Perchet (2016); Hu et al. (2016) and the residual one-point feedback (7) to solve the above policy optimization problem. To evaluate the cost $V(K_t)$ given the policy parameter K_t at iteration t, we run one episode with a finite horizon length $H = 50$. The dynamical matrices A and B are randomly generated and the noise w_k is sampled from a Gaussian distribution $\mathcal{N}(0, 0.1^2)$. We select the exploration parameter δ as $\delta = 0.1$, and the stepsizes for the proposed residual feedback estimator, the two-point estimator and the conventional one-point estimator are $2 \times 10^{-3}, 2.5 \times 10^{-3}, 1.5 \times 10^{-4}$, respectively. We run each algorithm 10 times. At each trial, all the algorithms start from the same initial guess of the policy parameter K_0, which is generated by perturbing the optimal policy parameter K^* with a random matrix,

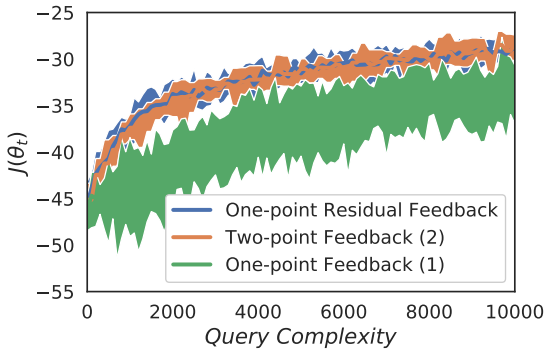
as in Malik et al. (2018). Each entry in this random perturbation matrix is sampled from a uniform distribution in $[0, 0.2]$. The performance of all the algorithms over 10 trials is measured in terms of $|V(K_t) - V(K^*)|$ and is presented in Figure 1(b). We observe that the residual one-point feedback (7) converges much faster than the one-point oracle in Gasnikov et al. (2017) and has comparable query complexity to the two-point feedback under uncontrolled noises considered in Bach & Perchet (2016); Hu et al. (2016). This corroborates our theoretical analysis in Section 4.

6.3 *Zeroth-Order Policy Optimization for a Large-Scale Multi-Stage Decision Making Problem*

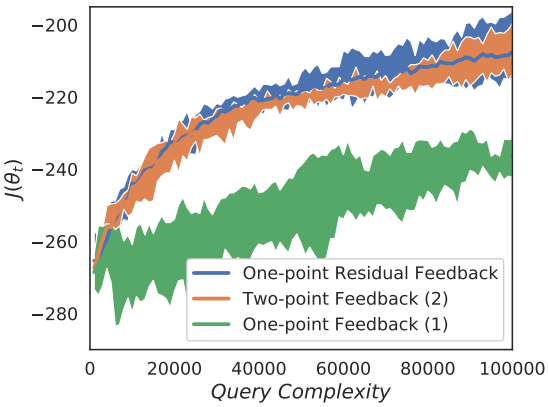
In this section, we consider a large-scale multi-stage resource allocation problem. Specifically, we consider 16 agents that are located on a 4×4 grid. At agent i, resources are stored in the amount of $m_i(k)$ and there is also a demand for resources in the amount of $d_i(k)$ at instant k. In the meantime, agent i also decides what fraction of resources $a_{ij}(k) \in [0, 1]$ it sends to its neighbors $j \in \mathcal{N}_i$ on the grid. The local amount of resources and demands at 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 $d_i(k) = A_i \sin(\omega_i k + \phi_i) + w_{i,k}$, where the amplitude A_i is sampled uniformly from $[1, 2]$, $\omega_i = 2\pi$, ϕ_i is uniformly sampled from $[0, \pi]$, and $w_{i,k}$ is the noise in the demand sampled from the normal distribution $\mathcal{N}(0, A_i^2/100)$. At time k, agent i receives a local reward $r_i(k)$, such that $r_i(k) = 0$ when $m_i(k) \geq 0$ and $r_i(k) = -m_i(k)^2$ when $m_i(k) < 0$. Let agent i makes its decisions according to a parameterized policy function $\pi_{i,\theta_i}(o_i) : \mathcal{O}_i \to [0, 1]^{|\mathcal{N}_i|}$, where θ_i is the parameter of the policy function π_i, $o_i \in \mathcal{O}_i$ denotes agent i's observation, and $|\mathcal{N}_i|$ represents the number of agent i's neighbors on the grid.

Our goal is to train a policy that can be executed in a fully distributed way based on agents' local information. Specifically, during the execution of policy functions $\{\pi_{i,\theta_i}(o_i)\}$, we let each agent only observe its local amount of resource $m_i(k)$ and demand and $d_i(k)$, i.e., $o_i(k) = [m_i(k), d_i(k)]^T$. In addition, the policy function $\pi_{i,\theta_i}(o_i)$ is parameterized as the following: $a_{ij} = \exp(z_{ij})/\sum_j \exp(z_{ij})$, where $z_{ij} = \sum_{p=1}^{9} \psi_p(o_i)\theta_{ij}(p)$ and $\theta_i = [\dots, \theta_{ij}, \dots]^T$. 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. Specifically, c_p are selected as vectors lying in the two-dimensional grid $(-0.5, 0, 0.5)^2$. The goal for the agents is to find an optimal policy $\pi^* = \{\pi_{i,\theta_i}(o_i)\}$ so that the global accumulated reward

$$J(\theta) = \sum_{i=1}^{16} \sum_{k=0}^{K} \gamma^k r_i(k) \tag{8}$$



(a)



(b)

Fig. 2. The convergence rate of applying the proposed residual one-point feedback (4) (blue), the two-point oracle (2) in Nesterov & Spokoiny (2017) (orange) and the one-point oracle (1) in Flaxman et al. (2005) (green) to the large-scale stochastic multi-stage resource allocation problem and the multi-robot cooperative navigation problem. The vertical axis represents the total rewards and the horizontal axis represents the number of episodes the agents take to evaluate their policy parameter iterates during the policy optimization procedure.

is maximized, where $\theta = [\dots, \theta_i, \dots]$ is the global policy parameter, K is the horizon of the problem, and γ is the discount factor. Effectively, the agents need to make decisions on 64 actions, and each action is decided by 9 parameters. Therefore, the problem dimension is $d = 576$. To implement zeroth-order policy gradient estimators (1) and (7) to find the optimal policy, at iteration t, we let all agents implement the policy with parameter $\theta_t + \delta u_t$, collect rewards $\{r_i(k)\}$ at time instants $k = 0, 1, \dots, K$ and compute the noisy policy value according to (8). Then, the zeroth-order policy gradient is estimated using (1) or (7). On the contrary, when the two-point zeroth-order policy gradient estimator (2) is used, at each iteration k, all agents need to evaluate two policies $\theta_t \pm \delta u_t$ to update the policy parameter once. In Figure 2(a), we present the performance of using zeroth-order policy gradients (1), (2) and (7)

to solve this large-scale multi-stage resource allocation problem, where the discount factor is set as $\gamma = 0.75$ and the length of horizon $K = 30$. We select the exploration parameter δ as $\delta = 0.1$, and the stepsizes for the proposed residual feedback estimator, the two-point estimator and the conventional one-point estimator are $1\times10^{-4}, 1\times10^{-4}, 5\times10^{-5}$, respectively. Each algorithm is run for 10 trials. We observe that policy optimization with the proposed residual-feedback gradient estimate (7) improves the optimal policy parameters with the same learning rate as the two-point zeroth-order gradient estimator (2), where the learning rate is measured by the number of episodes the agents take to evaluate the policy parameter iterates. In the meantime, both estimators perform much better than the one-point policy gradient estimate (1) considered in Fazel et al. (2018); Malik et al. (2018).

6.4 Zeroth-Order Policy Optimization for a Multi-Robot Cooperative Navigation Problem

In this section, we demonstrate the effectiveness of the proposed one-point residual-feedback gradient estimator using the benchmark multi-agent particle environment Lowe et al. (2017). Specifically, we consider the two-agent two-landmark cooperative nagivation task, where the agents navigate to the landmarks in the environment without colliding into each other. At each time step, agent i observes a vector $o_i \in \mathbb{R}^{12}$ consisting of all agents' states, i.e., their positions and velocities, and the two landmarks' positions. Then, agent i selects a 5 dimensional action vector based on its observation o_i, to drive itself around the world . The dynamics of the agents' states are governed by the physical engine used in the particle environment. At each time, the team of agents receive a team reward $r(k) = -\sum_{l=1,2} \min_{i=1,2} \|pos_i - pos_l\|$, where l denotes the landmark index, pos_i and pos_l represent the position vectors of agent i and landmark l. In addition, if the agents collide at time step k, the team receives -1 as a penalty. In each episode, there are 25 time steps.

We let each agent learn a policy function $\pi_{i,\theta_i}(o_i*)$ that is designed as a ReLU neural network with one hidden layer, where θ_i denotes the weights. The hidden layer consists of 32 neurons. Therefore, each neural network policy function has $(12+1)\times32+(32+1)\times5 = 581$ parameters to learn. The dimension of the problem is $d = 1162$. Since a ReLU activation function is used, the policy optimization problem is non-smooth. We implement the proposed residual-feedback policy gradient estimator, as well as the conventional one-point estimator (1) and the two-point estimator (2), for 5 trials. Specifically, we select the exploration parameter δ as $\delta = 0.1$, and the stepsizes for the proposed residual feedback estimator, the two-point estimator and the conventional one-point estimator are $5\times10^{-6}, 1\times10^{-5}, 1\times10^{-6}$, respectively. The learning rates for these algorithms are manually tuned to achieve their best performance respectively. The results are presented in Figure 2(b). In this non-smooth setting, we observe that policy optimization with the proposed residual-feedback gradient still has comparable performance to that of the two-point policy gradient estimator (2) and both estimators perform much better than the one-point policy gradient estimate (1), similar to the smooth case in Section 6.3.

7 Conclusion

In this paper, we proposed a residual one-point feedback oracle for zeroth-order optimization, which estimates the gradient of the objective function using a single query of the function value at each iteration. When the function evaluation is noiseless, we showed that ZO using the proposed oracle can achieve the same iteration complexity as ZO using two-point oracles when the function is non-smooth. When the function is smooth, this complexity of ZO can be further improved. This is the first time that a one-point zeroth-order oracle is shown to match the performance of two-point oracles in ZO. In addition, we considered a more realistic scenario where the function evaluation is corrupted by noise. We showed that the convergence rate of ZO using the proposed oracle matches the best known results using one-point feedback or two-point feedback with uncontrollable data samples. We provided numerical experiments that showed that the proposed oracle outperforms the one-point oracle and is as effective as two-point feedback methods.

References

Agarwal, A., Dekel, O., & Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT* (pp. 28–40). Citeseer.

Akhavan, A., Pontil, M., & Tsybakov, A. (2020). Exploiting higher order smoothness in derivative-free optimization and continuous bandits. *Advances in Neural Information Processing Systems*, *33*, 9017–9027.

Bach, F., & Perchet, V. (2016). Highly-smooth zero-th order online optimization. In *Conference on Learning Theory* (pp. 257–283).

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

Dekel, O., Eldan, R., & Koren, T. (2015). Bandit smooth convex optimization: Improving the bias-variance tradeoff. In *Advances in Neural Information Processing Systems* (pp. 2926–2934).

Duchi, J. C., Jordan, M. I., Wainwright, M. J., & Wibisono, A. (2015). Optimal rates for zero-order

convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, *61*, 2788–2806.

Fazel, M., Ge, R., Kakade, S., & Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*. volume 80.

Flaxman, A. D., Kalai, A. T., & McMahan, H. B. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 385–394). Society for Industrial and Applied Mathematics.

Gasnikov, A. V., Krymova, E. A., Lagunovskaya, A. A., Usmanova, I. N., & Fedorenko, F. A. (2017). Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case. *Automation and remote control*, *78*, 224–234.

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

Hu, X., Prashanth, L., György, A., & Szepesvári, C. (2016). (bandit) convex optimization with biased noisy gradient oracles. In *Artificial Intelligence and Statistics* (pp. 819–828).

Larson, J., Menickelly, M., & Wild, S. M. (2019). Derivative-free optimization methods. *Acta Numerica*, *28*, 287–404.

Li, Z., Dong, Z., Liang, Z., & Ding, Z. (2021). Surrogate-based distributed optimisation for expensive black-box functions. *Automatica*, *125*, 109407.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, .

Luo, X., Zhang, Y., & Zavlanos, M. M. (2020). Socially-aware robot planning via bandit human feedback. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)* (pp. 216–225). IEEE.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P. L., & Wainwright, M. J. (2018). Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *arXiv preprint arXiv:1812.08305*, .

Nešić, D. (2009). Extremum seeking control: Convergence analysis. *European Journal of Control*, *15*, 331–347.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course* volume 87. Springer Science & Business Media.

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

Poveda, J. I., & Li, N. (2021). Robust hybrid zero-order optimization algorithms with acceleration via averaging in time. *Automatica*, *123*, 109361.

Saha, A., & Tewari, A. (2011). Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (pp. 636–642).

Shamir, O. (2013). On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory* (pp. 3–24).

Shamir, O. (2017). An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, *18*, 1–11.

Zhang, Y., & Zavlanos, M. M. (2020). Cooperative multi-agent reinforcement learning with partial observations. *arXiv preprint arXiv:2006.10822*, .

Zhang, Y., Zhou, Y., Ji, K., & Zavlanos, M. M. (2020a). Boosting one-point derivative-free online optimization via residual feedback. *arXiv preprint arXiv:2010.07378*, .

Zhang, Y., Zhou, Y., Ji, K., & Zavlanos, M. M. (2020b). Improving the convergence rate of one-point zeroth-order optimization using residual feedback. *arXiv preprint arXiv:2006.10820*, .

Appendix

A Proof of Lemma 6

First, we show the bound when $f(x) \in C^{0,0}$. Recalling the expression of $\tilde{g}(x_t)$ in (4), we have that

$$\begin{aligned}\mathbb{E}[\|\tilde{g}(x_t)\|^2] &= \mathbb{E}[\frac{1}{\delta^2}\big(f(x_t+\delta u_t) - f(x_{t-1}+\delta u_{t-1})\big)^2\|u_t\|^2] \\ &\leq \frac{2}{\delta^2}\mathbb{E}[\big(f(x_t+\delta u_t) - f(x_{t-1}+\delta u_t)\big)^2\|u_t\|^2] \\ &+ \frac{2}{\delta^2}\mathbb{E}[\big(f(x_{t-1}+\delta u_t) - f(x_{t-1}+\delta u_{t-1})\big)^2\|u_t\|^2].\end{aligned}$$

Since function $f \in C^{0,0}$ with Lipschitz constant L_0, we obtain that

$$\begin{aligned}\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq &\frac{2L_0^2}{\delta^2}\mathbb{E}[\|x_t - x_{t-1}\|^2\|u_t\|^2] \\ &+ 2L_0^2\mathbb{E}[\|u_t - u_{t-1}\|^2\|u_t\|^2]. \qquad \text{(A.1)}\end{aligned}$$

Since u_t is independently sampled from $x_t - x_{t-1}$, we have that $\mathbb{E}[\|x_t - x_{t-1}\|^2\|u_t\|^2] = \mathbb{E}[\|x_t - x_{t-1}\|^2]\mathbb{E}[\|u_t\|^2]$. Since u_t is subject to standard multivariate normal distribution, $\mathbb{E}[\|u_t\|^2] = d$. Furthermore, using Lemma 1 in Nesterov & Spokoiny (2017), we get that $\mathbb{E}[\|u_t - u_{t-1}\|^2\|u_t\|^2] \leq 2\mathbb{E}[(\|u_t\|^2 + \|u_{t-1}\|^2)\|u_t\|^2] = 2\mathbb{E}[(\|u_t\|^4] + 2\mathbb{E}[\|u_{t-1}\|^2\|u_t\|^2] \leq 4(d+4)^2$. Plugging these bounds into inequality (A.1), we have that

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \frac{2dL_0^2}{\delta^2}\mathbb{E}[\|x_t - x_{t-1}\|^2] + 8L_0^2(d+4)^2.$$

Since $x_t = x_{t-1} - \eta\tilde{g}(x_{t-1})$, we get that

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \frac{2dL_0^2\eta^2}{\delta^2}\mathbb{E}[\|\tilde{g}(x_{t-1})\|^2] + 8L_0^2(d+4)^2.$$

Next, we show the bound when we have the additional smoothness condition $f(x) \in C^{1,1}$ with constant L_1. Given the gradient estimate in (4), we have that

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \mathbb{E}[\frac{(f(x_t+\delta u_t) - f(x_{t-1}+\delta u_{t-1}))^2}{\delta^2}\|u_t\|^2]. \tag{A.2}$$

Next, we bound the term $(f(x_t+\delta u_t)-f(x_{t-1}+\delta u_{t-1}))^2$. Adding and subtracting $f(x_{t-1}+\delta u_t)$ inside the square, and applying the inequality $(a+b)^2 \leq 2a^2+2b^2$, we can obtain

$$\begin{aligned}&(f(x_t+\delta u_t) - f(x_{t-1}+\delta u_{t-1}))^2\\ &\leq 2(f(x_t+\delta u_t) - f(x_{t-1}+\delta u_t))^2\\ &\quad + 2(f(x_{t-1}+\delta u_t) - f(x_{t-1}+\delta u_{t-1}))^2. \end{aligned}\tag{A.3}$$

Since the function $f(x)$ is also Lipschitz continuous with constant L_0, we get that

$$\begin{aligned}(f(x_t+\delta u_t) - f(x_{t-1}+\delta u_t))^2 &\leq L_0^2\|x_t - x_{t-1}\|^2\\ &= L_0^2\eta^2\|\tilde{g}(x_{t-1})\|^2. \end{aligned}\tag{A.4}$$

Next, we bound the term $(f(x_{t-1}+\delta u_t) - f(x_{t-1}+\delta u_{t-1}))^2$. Adding and subtracting $f(x_{t-1})$, $\langle\nabla f(x_{t-1}), \delta u_t\rangle$ and $\langle\nabla f(x_{t-1}), \delta u_{t-1}\rangle$ inside the square term, we have that

$$\begin{aligned}&(f(x_{t-1}+\delta u_t) - f(x_{t-1}+\delta u_{t-1}))^2\\ &\leq 2\langle\nabla f(x_{t-1}), \delta(u_t - u_{t-1})\rangle^2\\ &+ 4(f(x_{t-1}+\delta u_t) - f(x_{t-1}) - \langle\nabla f(x_{t-1}), \delta u_t\rangle)^2\\ &+ 4(f(x_{t-1}+\delta u_{t-1}) - f(x_{t-1}) - \langle\nabla f(x_{t-1}), \delta u_{t-1}\rangle)^2. \end{aligned}\tag{A.5}$$

Since $f(x) \in C^{1,1}$ with constant L_1, we get that $|f(x_{t-1}+\delta u_t) - f(x_{t-1}) - \langle\nabla f(x_{t-1}), \delta u_t\rangle| \leq \frac{1}{2}L_1\delta^2\|u_t\|^2$, according to (6) in Nesterov & Spokoiny (2017). And similarly, we also have $|f(x_{t-1}+\delta u_{t-1}) - f(x_{t-1}) - \langle\nabla f(x_{t-1}), \delta u_{t-1}\rangle| \leq \frac{1}{2}L_1\delta^2\|u_{t-1}\|^2$. Substituting these inequalities into (A.5), we obtain that

$$\begin{aligned}(f(x_{t-1}+\delta u_t) - f(x_{t-1}+\delta u_{t-1}))^2 &\leq 2\langle\nabla f(x_{t-1}),\\ \delta(u_t - u_{t-1})\rangle^2 + L_1^2\delta^4\|u_t\|^4 + L_1^2\delta^4\|u_{t-1}\|^4.\end{aligned}\tag{A.6}$$

Moreover, substituting the inequalities (A.4) and (A.6) in the upper bound in (A.3), we get that

$$\begin{aligned}&(f(x_t+\delta u_t) - f(x_{t-1}+\delta u_{t-1}))^2\\ &\leq 2L_0^2\eta^2\|\tilde{g}(x_{t-1})\|^2 + 4\langle\nabla f(x_{t-1}), \delta(u_t - u_{t-1})\rangle^2\\ &\quad + 2L_1^2\delta^4\|u_t\|^4 + 2L_1^2\delta^4\|u_{t-1}\|^4 \end{aligned}\tag{A.7}$$

Using the bound (A.7) in inequality (A.2), and applying the bounds $\mathbb{E}[\|u_t\|^6] \leq (d+6)^3$ and $\mathbb{E}[\|u_{t-1}\|^4\|u_t\|^2] \leq (d+6)^3$, we have that

$$\begin{aligned}&\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \frac{2dL_0^2\eta^2}{\delta^2}\mathbb{E}[\|\tilde{g}(x_{t-1})\|^2]\\ &+ 4\mathbb{E}[\langle\nabla f(x_{t-1}), u_t - u_{t-1}\rangle^2\|u_t\|^2] + 4L_1^2(d+6)^3\delta^2. \end{aligned}\tag{A.8}$$

Since $\langle\nabla f(x_{t-1}), u_t - u_{t-1}\rangle^2 \leq 2\langle\nabla f(x_{t-1}), u_t\rangle^2 + 2\langle\nabla f(x_{t-1}), u_{t-1}\rangle^2$, we get that

$$\begin{aligned}&\mathbb{E}[\langle\nabla f(x_{t-1}), u_t - u_{t-1}\rangle^2\|u_t\|^2] \leq 2\mathbb{E}[\langle\nabla f(x_{t-1}),\\ &\quad u_t\rangle^2\|u_t\|^2] + 2\mathbb{E}[\langle\nabla f(x_{t-1}), u_{t-1}\rangle^2\|u_t\|^2]. \end{aligned}\tag{A.9}$$

For the term $\mathbb{E}[\langle\nabla f(x_{t-1}), u_{t-1}\rangle^2\|u_t\|^2]$, we have that $\mathbb{E}[\langle\nabla f(x_{t-1}), u_{t-1}\rangle^2\|u_t\|^2] \leq \mathbb{E}[\|\nabla f(x_{t-1})\|^2\|u_{t-1}\|^2\|u_t\|^2] \leq d^2\mathbb{E}[\|\nabla f(x_{t-1})\|^2]$. For the term $\mathbb{E}[\langle\nabla f(x_{t-1}), u_t\rangle^2\|u_t\|^2]$, according to Theorem 3 in Nesterov & Spokoiny (2017), we have a stronger bound $\mathbb{E}[\langle\nabla f(x_{t-1}), u_t\rangle^2\|u_t\|^2] \leq (d+4)\mathbb{E}[\|\nabla f(x_{t-1})\|^2]$. Substituting these bounds into (A.9), and because $d^2+d+4 \leq (d+4)^2$, we have that

$$\begin{aligned}&\mathbb{E}[\langle\nabla f(x_{t-1}), u_t - u_{t-1}\rangle^2\|u_t\|^2]\\ &\leq 2(d+4)^2\mathbb{E}[\|\nabla f(x_{t-1})\|^2]. \end{aligned}\tag{A.10}$$

Substituting the bound (A.10) into inequality (A.8), we complete the proof.

B Proof of Theorem 7

Since we have that $f(x) \in C^{0,0}$, according to Lemma 2, the function $f_\delta(x)$ has $L_1(f_\delta)$-Lipschitz continuous gradient where $L_1(f_\delta) = \frac{\sqrt{d}}{\delta}L_0$. Furthermore, according to Lemma 1.2.3 in Nesterov (2013), we can get the following inequality

$$\begin{aligned}f_\delta(x_{t+1}) &\leq f_\delta(x_t) + \langle\nabla f_\delta(x_t), x_{t+1} - x_t\rangle\\ &\quad + \frac{L_1(f_\delta)}{2}\|x_{t+1} - x_t\|^2\\ &= f_\delta(x_t) - \eta\langle\nabla f_\delta(x_t), \tilde{g}(x_t)\rangle + \frac{L_1(f_\delta)\eta^2}{2}\|\tilde{g}(x_t)\|^2\\ &= f_\delta(x_t) - \eta\langle\nabla f_\delta(x_t), \Delta_t\rangle - \eta\|\nabla f_\delta(x_t)\|^2\\ &\quad + \frac{L_1(f_\delta)\eta^2}{2}\|\tilde{g}(x_t)\|^2, \end{aligned}\tag{B.1}$$

where $\Delta_t = \tilde{g}(x_t) - \nabla f_\delta(x_t)$. According to Lemma 5, we can get that $\mathbb{E}_{u_t}[\tilde{g}(x_t)] = \nabla f_\delta(x_t)$. Therefore, taking expectation over u_t on both sides of inequality (B.1) and rearranging terms, we have that

$$\begin{aligned}\eta\mathbb{E}[\|\nabla f_\delta(x_t)\|^2] &\leq \mathbb{E}[f_\delta(x_t)] - \mathbb{E}[f_\delta(x_{t+1})]\\ &\quad + \frac{L_1(f_\delta)\eta^2}{2}\mathbb{E}[\|\tilde{g}(x_t)\|^2]. \end{aligned}\tag{B.2}$$

Telescoping above inequalities from $t=0$ to $T-1$ and dividing both sides by η, we obtain that

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(x_t)\|^2] \leq \frac{\mathbb{E}[f_\delta(x_0)]-\mathbb{E}[f_\delta(x_T)]}{\eta} + \frac{L_1(f_\delta)\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \frac{\mathbb{E}[f_\delta(x_0)]-f_\delta^*}{\eta}+\frac{L_1(f_\delta)\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2], \quad \text{(B.3)}$$

where f_δ^* is the lower bound of the smoothed function $f_\delta(x)$. f_δ^* must exist because we assume the orignal function $f(x)$ is lower bounded and the smoothed function has a bounded distance from $f(x)$ due to Lemma 2.

Recall the contraction result of the second moment $\mathbb{E}[\|\tilde{g}(x_t)\|^2]$ in Lemma 6 when $f(x)\in C^{0,0}$. Denote the contraction rate $\frac{2dL_0^2\eta^2}{\delta^2}$ as α and the constant perturbation term $M=8L_0^2(d+4)^2$. Then, we get that

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \alpha^t\mathbb{E}[\|\tilde{g}(x_0)\|^2]+\frac{1-\alpha^t}{1-\alpha}M. \quad \text{(B.4)}$$

Summing the above inequality over time, we obtain

$$\sum_{t=0}^{T-1}\|\tilde{g}(x_t)\|^2 \leq \frac{1-\alpha^T}{1-\alpha}\mathbb{E}[\|\tilde{g}(x_0)\|^2]+\sum_{t=0}^{T-1}\big(\frac{1-\alpha^t}{1-\alpha}M\big) \leq \frac{1}{1-\alpha}\mathbb{E}[\|\tilde{g}(x_0)\|^2]+\frac{1}{1-\alpha}MT. \quad \text{(B.5)}$$

Plugging the bound in (B.5) into inequality (B.3), and since $L_1(f_\delta)=\frac{\sqrt{d}}{\delta}L_0$, we have that

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(x_t)\|^2] \leq \frac{\mathbb{E}[f_\delta(x_0)]-f_\delta^*}{\eta}+\frac{d^{\frac{1}{2}}L_0\eta}{2\delta}\big(\frac{1}{1-\alpha}\mathbb{E}[\|\tilde{g}(x_0)\|^2]+\frac{1}{1-\alpha}8L_0^2(d+4)^2T\big). \quad \text{(B.6)}$$

To fullfill the requirement that $|f(x)-f_\delta(x)|\leq\epsilon_f$, we set the exporation parameter $\delta=\frac{\epsilon_f}{d^{\frac{1}{2}}L_0}$. In addition, let the stepsize be $\eta=\frac{\sqrt{\epsilon_f}}{2dL_0^2T^{\frac{1}{2}}}$. We have that $\alpha=\frac{1}{2T\epsilon_f}\leq\frac{1}{2}$ and $\frac{1}{1-\alpha}\leq 2$, when $T\geq\frac{1}{\epsilon_f}$. Plugging the choices of η and δ into inequality (B.6), we obtain that

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(x_t)\|^2] \leq 2L_0^2\big(\mathbb{E}[f_\delta(x_t)]-f_\delta^*\big)\frac{d}{\sqrt{\epsilon_f}}\sqrt{T}+\frac{1}{2\sqrt{\epsilon_f T}}\mathbb{E}[\|\tilde{g}(x_0)\|^2]+4L_0^2\frac{(d+4)^2}{\sqrt{\epsilon_f}}\sqrt{T}.$$

Dividing both sides of above inequality by T, we complete the proof.

C Proof of Theorem 8

Following the same process in the beginning of the proof of Theorem 7, we can get

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(x_t)\|^2] \leq \frac{\mathbb{E}[f_\delta(x_0)]-f_\delta^*}{\eta}+\frac{L_1\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2]. \quad \text{(C.1)}$$

Since $\frac{1}{2}\mathbb{E}[\|\nabla f(x_t)\|^2]\leq\mathbb{E}[\|\nabla f_\delta(x_t)\|^2]+\mathbb{E}[\|\nabla f(x_t)-\nabla f_\delta(x_t)\|^2]$, and according to the bound (C.1) and Lemma 2, we have that

$$\frac{1}{2}\sum_{t=0}^{T-1}\|\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{\mathbb{E}[f_\delta(x_0)]-f_\delta^*}{\eta}+\frac{L_1\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2]+L_1^2(d+3)^3\delta^2T. \quad \text{(C.2)}$$

In addition, similar to the process to derive the bound in (B.5), according to Lemma 6, when $f(x)\in C^{1,1}$, we can get that

$$\sum_{t=0}^{T-1}\|\tilde{g}(x_t)\|^2 \leq \frac{1}{1-\alpha}\mathbb{E}[\|\tilde{g}(x_0)\|^2]+\frac{8}{1-\alpha}(d+4)^2\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2+\frac{4}{1-\alpha}L_1^2(d+6)^3\delta^2T. \quad \text{(C.3)}$$

Plugging the bound (C.3) into (C.2), we have that

$$\frac{1}{2}\sum_{t=0}^{T-1}\|\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{\mathbb{E}[f_\delta(x_0)]-f_\delta^*}{\eta}+\frac{L_1\eta}{2}\big(\frac{1}{1-\alpha}\mathbb{E}[\|\tilde{g}(x_0)\|^2]+\frac{4}{1-\alpha}L_1^2(d+6)^3\delta^2T+\frac{8}{1-\alpha}(d+4)^2\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(x_t)\|^2]\big)+L_1^2(d+3)^3\delta^2T. \quad \text{(C.4)}$$

Recalling that $\tilde{L}=\max\{32L_1,2L_0\}$, let $\eta=\frac{1}{\tilde{L}(d+4)^2T^{\frac{1}{3}}}$ and $\delta=\frac{1}{\sqrt{d}T^{\frac{1}{3}}}$, and we have that $\alpha=2dL_0^2\frac{\eta^2}{\delta^2}\leq\frac{1}{2}$. In addition, the coefficient before the term $\|\nabla f(x_t)\|^2$ in the upper bound above $\frac{L_1\eta}{2}\frac{8}{1-\alpha}(d+4)^2\leq\frac{1}{4}$. Therefore,

we obtain that

$$\frac{1}{4}\sum_{t=0}^{T-1}\|\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \tilde{L}(\mathbb{E}[f_\delta(x_0)] - f_\delta^*)(d+4)^2 T^{\frac{1}{3}} + \frac{1}{32(d+4)^2 T^{\frac{1}{3}}}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{L_1^2}{8}\frac{(d+6)^3}{(d+4)^2 d} + L_1^2\frac{(d+3)^3}{d}T^{\frac{1}{3}}.$$

Dividing both sides of above inequality by T, we complete the proof.

D Proof of Theorem 9

First, according to iteration (5), we have that

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - \eta\widetilde{g}(x_t) - x^*\|^2 = \|x_t - x^*\|^2 - 2\eta\langle\widetilde{g}(x_t), x_t - x^*\rangle + \eta^2\|\widetilde{g}(x_t)\|^2.$$

Taking expectation on both sides, and since $\mathbb{E}[\tilde{g}(x_t)] = \nabla f_\delta(x_t)$, we obtain that

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq \mathbb{E}[\|x_t - x^*\|^2] - 2\eta\langle\nabla f_\delta(x_t), x_t - x^*\rangle + \eta^2\mathbb{E}[\|\tilde{g}(x_t)\|^2]. \tag{D.1}$$

Due to the convexity, we have that $\langle\nabla f_\delta(x_t), x_t - x^*\rangle \geq f_\delta(x_t) - f_\delta(x^*)$. Plugging this inequality into (D.1), we have that

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq \mathbb{E}[\|x_t - x^*\|^2] - 2\eta(f_\delta(x_t) - f_\delta(x^*)) + \eta^2\mathbb{E}[\|\tilde{g}(x_t)\|^2]. \tag{D.2}$$

When $f(x) \in C^{0,0}$, using Lemma (2), we can replace $f_\delta(x)$ with $f(x)$ in above inequality and get

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq \mathbb{E}[\|x_t - x^*\|^2] - 2\eta(f(x_t) - f(x^*)) + \eta^2\mathbb{E}[\|\tilde{g}(x_t)\|^2] + 4L_0\sqrt{d}\delta\eta.$$

Rearranging the terms and telescoping from $t = 0$ to $T-1$, we obtain that

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)] - Tf(x^*) \leq \frac{1}{2\eta}(\|x_0 - x^*\|^2 - \mathbb{E}[\|x_T - x^*\|^2]) + \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2] + 2L_0\sqrt{d}\delta T$$
$$\leq \frac{1}{2\eta}\|x_0 - x^*\|^2 + \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2] + 2L_0\sqrt{d}\delta T$$

Since function $f(x) \in C^{0,0}$, we can plug the bound (B.5) into the above inequality and get that

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)] - Tf(x^*) \leq \frac{1}{2\eta}\|x_0 - x^*\|^2 + \frac{\eta}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{4\eta}{1-\alpha}L_0^2(d+4)^2 T + 2L_0\sqrt{d}\delta T.$$

Let $\eta = \frac{1}{2dL_0\sqrt{T}}$ and $\delta = \frac{1}{\sqrt{T}}$. We have that $\alpha = 2dL_0^2\frac{\eta^2}{\delta^2} = \frac{1}{2d} \leq \frac{1}{2}$. Therefore, $\frac{1}{1-\alpha} \leq 2$. Applying this bound and the choice of η and δ into above inequality, we have that

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)] - Tf(x^*) \leq L_0\|x_0 - x^*\|^2 d\sqrt{T} + \frac{1}{2dL_0\sqrt{T}}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + 4L_0\frac{(d+4)^2}{d}\sqrt{T} + 2L_0\sqrt{d}\sqrt{T}.$$

Recalling that $f(\bar{x}) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t)$ due to convexity and dividing both sides of above inequality by T, the proof of the nonsmooth case is complete.

When function $f(x) \in C^{1,1}$, it is straightforward to see that we also have the inequality (D.2). In addition, according to Lemma 2, we can replace $f_\delta(x)$ with $f(x)$ in above inequality and get

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq \mathbb{E}[\|x_t - x^*\|^2] - 2\eta(f(x_t) - f(x^*)) + \eta^2\mathbb{E}[\|\tilde{g}(x_t)\|^2] + 4L_1 d\delta^2\eta. \tag{D.3}$$

Similarly to the above analysis, we telescope the above inequality from $t = 0$ to $T-1$, apply the bound on $\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2]$ in (C.3) and obtain that

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)] - Tf(x^*) \leq \frac{1}{2\eta}\|x_0 - x^*\|^2 + \frac{\eta}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{2\eta}{1-\alpha}L_1^2(d+6)^3\delta^2 T + \frac{4\eta}{1-\alpha}(d+4)^2\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(x_t)\|^2] + 2L_1 d\delta^2 T.$$

Since $f(x) \in C^{1,1}$ is convex, we have that $\|\nabla f(x_t)\|^2 \leq 2L_1(f(x_t) - f(x^*))$ according to (2.1.7) in Nesterov (2013). Applying this bound into the above inequality,

we get that

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)]-Tf(x^*)\leq\frac{1}{2\eta}\|x_0-x^*\|^2 + \frac{\eta}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}(x_0)\|^2]+\frac{2\eta}{1-\alpha}L_1^2(d+6)^3\delta^2T + \frac{8\eta}{1-\alpha}L_1(d+4)^2\big(\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)]-Tf(x^*)\big)+2L_1d\delta^2T.$$

Let $\eta=\frac{1}{2\tilde{L}(d+4)^2T^{\frac{1}{3}}}$ and $\delta=\frac{\sqrt{d}}{T^{\frac{1}{3}}}$ where $\tilde{L}=\max\{L_0,16L_1\}$. Then, we have that $\alpha=2dL_0^2\frac{\eta^2}{\delta^2}\leq\frac{1}{2(d+4)^4}\leq\frac{1}{2}$. In addition, we have that $\frac{8\eta}{1-\alpha}L_1(d+4)^2\leq\frac{1}{2T^{\frac{1}{3}}}\leq\frac{1}{2}$. Applying these two bounds into above inequality and rearranging terms, we have that

$$\frac{1}{2}\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)]-Tf(x^*)\leq\tilde{L}\|x_0-x^*\|^2(d+4)^2T^{\frac{1}{3}} + \frac{1}{2\tilde{L}(d+4)^2T^{\frac{1}{3}}}\mathbb{E}[\|\tilde{g}(x_0)\|^2]+\frac{L_1}{8}\frac{(d+6)^3d}{(d+4)^2}+2L_1d^2T^{\frac{1}{3}}.$$

Recalling that $f(\bar{x})\leq\frac{1}{T}\sum_{t=0}^{T-1}f(x_t)$ due to convexity and dividing both sides of above inequality by T, the proof of the smooth case is complete.

E Proof of Lemma 12

The analysis is similar to the proof in Section A. First, consider the case when $F(x,\xi)\in C^{0,0}$ with $L_0(\xi)$. According to (7), we have that

$$\begin{aligned}&\mathbb{E}[\|\tilde{g}(x_t)\|^2]\\&=\mathbb{E}[\frac{1}{\delta^2}\big(F(x_t+\delta u_t,\xi_t)-F(x_{t-1}+\delta u_{t-1},\xi_{t-1})\big)^2\|u_t\|^2]\\&\leq\frac{2}{\delta^2}\mathbb{E}[\big(F(x_t+\delta u_t,\xi_t)-F(x_{t-1}+\delta u_{t-1},\xi_t)\big)^2\|u_t\|^2]\\&+\frac{2}{\delta^2}\mathbb{E}[\big(F(x_{t-1}+\delta u_{t-1},\xi_t)-F(x_{t-1}+\delta u_{t-1},\xi_{t-1})\big)^2\|u_t\|^2].\end{aligned}$$

Using the bound in Assumption 10, we get that $\frac{2}{\delta^2}\mathbb{E}[\big(F(x_{t-1}+\delta u_{t-1},\xi_t)-F(x_{t-1}+\delta u_{t-1},\xi_{t-1})\big)^2\|u_t\|^2]\leq\frac{8d\sigma^2}{\delta^2}$. In addition, adding and subtracting $F(x_{t-1}+\delta u_t,\xi_t)$ in $\big(F(x_t+\delta u_t,\xi_t)-F(x_{t-1}+\delta u_{t-1},\xi_t)\big)^2$ in above inequality, we obtain that

$$\begin{aligned}&\mathbb{E}[\|\tilde{g}(x_t)\|^2]\leq\frac{8d\sigma^2}{\delta^2}+\\&\frac{4}{\delta^2}\mathbb{E}[\big(F(x_t+\delta u_t,\xi_t)-F(x_{t-1}+\delta u_t,\xi_t)\big)^2\|u_t\|^2]\\&+\frac{4}{\delta^2}\mathbb{E}[\big(F(x_{t-1}+\delta u_t,\xi_t)-F(x_{t-1}+\delta u_{t-1},\xi_t)\big)^2\|u_t\|^2]\end{aligned}$$

Using Assumption 11, we can bound the last two items on the right hand side of above inequality following the same procedure after inequality (A.1) and get that

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2]\leq\frac{4dL_0^2\eta^2}{\delta^2}\mathbb{E}[\|\tilde{g}(x_{t-1})\|^2]+16L_0^2(d+4)^2+\frac{8d\sigma^2}{\delta^2}.$$

The proof is complete.

F Proof of Theorem 13

When function $F(x)\in C^{0,0}$ with $L_0(\xi)$, using Assumption 11 and following the same procedure in Section B, we have that

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

where $L_1(f_\delta)=\frac{\sqrt{d}}{\delta}L_0$. In addition, according to Lemma 12, we get that

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2]\leq\frac{1}{1-\alpha}\mathbb{E}[\|\tilde{g}(x_0)\|^2]+\frac{16L_0^2}{1-\alpha}(d+4)^2T+\frac{8\sigma^2}{1-\alpha}\frac{d}{\delta^2}T,\quad\text{(F.2)}$$

where $\alpha=\frac{4dL_0^2\eta^2}{\delta^2}$. Plugging (F.2) into the bound in (F.1), we obtain that

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(x_t)\|^2]\leq\frac{\mathbb{E}[f_\delta(x_0)]-f_\delta^*}{\eta}+\frac{4\sigma^2L_0}{1-\alpha}d^{1.5}\frac{\eta}{\delta^3}T+\frac{\sqrt{d}L_0}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}(x_0)\|^2]\frac{\eta}{\delta}+\frac{8L_0^3\sqrt{d}}{1-\alpha}(d+4)^2\frac{\eta}{\delta}T.\quad\text{(F.3)}$$

Similar to Section B, to fullfill the requirement that $|f(x)-f_\delta(x)|\leq\epsilon_f$, we set the exporation parameter $\delta=\frac{\epsilon_f}{d^{\frac{1}{2}}L_0}$. In addition, let the stepsize be $\eta=\frac{\epsilon_f^{1.5}}{2\sqrt{2}L_0^2d^{1.5}T^{\frac{1}{2}}}$. Then, we have that $\alpha=\frac{4dL_0^2\eta^2}{\delta^2}=\frac{\epsilon_f}{2dT}\leq\frac{1}{2}$ when $T\geq\frac{1}{d\epsilon_f}$. Therefore, we have that $\frac{1}{1-\alpha}\leq2$. Applying this bound and the choices of η and δ into the bound (F.3),

we obtain that

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(x_t)\|^2] \le 2\sqrt{2}L_0^2(\mathbb{E}[f_\delta(x_0)] - f_\delta^*)\frac{d^{1.5}\sqrt{T}}{\epsilon_f^{1.5}} + \frac{L_0\epsilon_f^{0.5}}{2\sqrt{2dT}}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + 4\sqrt{2}L_0^2\frac{(d+4)^2}{\sqrt{d}}\sqrt{\epsilon_f T} + 2\sqrt{2}\sigma^2L_0^2\frac{d^{1.5}\sqrt{T}}{\epsilon_f^{1.5}}.$$

Dividing both sides by T, the proof for the nonsmooth case is complete.

When function $F(x,\xi) \in C^{1,1}$ with $L_1(\xi)$, according to Assumption 11, we also have that $f_\delta(x), f(x) \in C^{1,1}$ with constant L_1. Similarly to the proof in Section C, we get that

$$\frac{1}{2}\sum_{t=0}^{T-1}\|\mathbb{E}[\|\nabla f(x)\|^2]\| \le \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} + \frac{L_1\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2] + L_1^2(d+3)^3\delta^2 T. \tag{F.4}$$

Plugging inequality (F.2) into the above upper bound, we obtain that

$$\frac{1}{2}\sum_{t=0}^{T-1}\|\mathbb{E}[\|\nabla f(x)\|^2]\| \le \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} + \frac{L_1\eta}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{8L_0^2L_1}{1-\alpha}(d+4)^2\eta T + \frac{4L_1\sigma^2}{1-\alpha}\frac{d\eta}{\delta^2}T + L_1^2(d+3)^3\delta^2 T. \tag{F.5}$$

Let $\eta = \frac{1}{2\sqrt{2}L_0 d^{\frac{4}{3}}T^{\frac{2}{3}}}$ and $\delta = \frac{1}{d^{\frac{5}{6}}T^{\frac{1}{6}}}$. Then, $\alpha = \frac{4dL_0^2\eta^2}{\delta^2} = \frac{1}{2T} \le \frac{1}{2}$ and $\frac{1}{1-\alpha} \le 2$. Plugging these results into the above inequality, we get that

$$\frac{1}{2}\sum_{t=0}^{T-1}\|\mathbb{E}[\|\nabla f(x)\|^2]\| \le 2\sqrt{2}L_0(\mathbb{E}[f_\delta(x_0)] - f_\delta^*)d^{\frac{4}{3}}T^{\frac{2}{3}} + \frac{L_1}{2\sqrt{2}L_0 d^{\frac{4}{3}}T^{\frac{2}{3}}}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + 4\sqrt{2}L_0L_1\frac{(d+4)^2}{d^{\frac{4}{3}}}T^{\frac{1}{3}} + \frac{2\sqrt{2}L_1\sigma^2}{L_0 d^{\frac{1}{3}}}T^{\frac{1}{3}} + L_1^2\frac{(d+3)^3}{d^{\frac{5}{3}}}T^{\frac{2}{3}}. \tag{F.6}$$

Dividing both sides by T, the proof for the smooth case is complete.

G Proof of Theorem 14

When the function $f(x) \in C^{0,0}$ with constant $L_0(\xi)$ is convex, we can follow the same procedure as in Section D and get that

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)] - Tf(x^*) \le \frac{1}{2\eta}\|x_0 - x^*\|^2 + \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2] + 2L_0\sqrt{d}\delta T.$$

Plugging the bound (F.2) into above inequality, we have that

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)] - Tf(x^*) \le \frac{1}{2\eta}\|x_0 - x^*\|^2 + \frac{\eta}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{8L_0^2}{1-\alpha}(d+4)^2\eta T + \frac{4\sigma^2}{1-\alpha}\frac{d\eta}{\delta^2}T + 2L_0\sqrt{d}\delta T. \tag{G.1}$$

Let $\eta = \frac{1}{2\sqrt{2}L_0\sqrt{d}T^{\frac{3}{4}}}$ and $\delta = \frac{1}{T^{\frac{1}{4}}}$. Then, we have that $\alpha = \frac{4dL_0^2\eta^2}{\delta^2} = \frac{1}{2T} \le \frac{1}{2}$. Plugging these results into the above inequality, we get that

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)] - Tf(x^*) \le \sqrt{2}L_0\|x_0 - x^*\|^2\sqrt{d}T^{\frac{3}{4}} + \frac{1}{2\sqrt{2}L_0\sqrt{d}T^{\frac{3}{4}}}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + 4\sqrt{2}L_0\frac{(d+4)^2}{\sqrt{d}}T^{\frac{1}{4}} + \frac{2\sqrt{2}\sigma^2}{L_0}\sqrt{d}T^{\frac{3}{4}} + 2L_0\sqrt{d}T^{\frac{3}{4}}. \tag{G.2}$$

Dividing both sides by T, the proof for the nonsmooth case is complete.

When the function $f(x) \in C^{1,1}$ with constant $L_1(\xi)$, we can also get the inequality (D.3) in Section D. Telescoping this inequality from $t = 0$ to $T-1$ and rearranging terms, we obtain

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)] - Tf(x^*) \le \frac{1}{2\eta}\|x_0 - x^*\|^2 + \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{g}(x_t)\|^2] + 2L_1d\delta^2 T. \tag{G.3}$$

Plugging the bound (F.2) into above inequality, we have that

$$\sum_{t=0}^{T-1}\mathbb{E}[f(x_t)] - Tf(x^*) \le \frac{1}{2\eta}\|x_0 - x^*\|^2 + 2L_1d\delta^2 T + \frac{\eta}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{8L_0^2}{1-\alpha}(d+4)^2\eta T + \frac{4\sigma^2}{1-\alpha}\frac{d\eta}{\delta^2}T.$$

Let $\eta = \frac{1}{2\sqrt{2}L_0 d^{\frac{2}{3}} T^{\frac{2}{3}}}$ and $\delta = \frac{1}{d^{\frac{1}{6}} T^{\frac{1}{6}}}$. Then, we have that $\alpha = \frac{4dL_0^2\eta^2}{\delta^2} = \frac{1}{2T} \leq \frac{1}{2}$. Plugging these parameters into above inequality, we get that

$$\sum_{t=0}^{T-1} \mathbb{E}[f(x_t)] - Tf(x^*) \leq \sqrt{2}L_0\|x_0 - x^*\|^2 d^{\frac{2}{3}} T^{\frac{2}{3}} + \frac{1}{2\sqrt{2}L_0 d^{\frac{2}{3}} T^{\frac{2}{3}}} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + 4\sqrt{2}L_0 \frac{(d+4)^2}{d^{\frac{2}{3}}} T^{\frac{1}{3}} + \frac{2\sqrt{2}\sigma^2}{L_0} d^{\frac{2}{3}} T^{\frac{2}{3}} + 2L_1 d^{\frac{2}{3}} T^{\frac{2}{3}}.$$

Dividing both sides by T, the proof for the smooth case is complete.