

A New Zeroth-Order Oracle for Distributed and  
Non-Stationary Learning

by

Yan Zhang

Department of Mechanical Engineering and Material Science  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_

Michael M. Zavlanos, Advisor

\_\_\_\_\_

Leila Bridgeman

\_\_\_\_\_

Vahid Tarokh

\_\_\_\_\_

Ronald Parr

\_\_\_\_\_

Dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy  
in the Department of Mechanical Engineering and Material Science  
in the Graduate School of  
Duke University

2021

ABSTRACT

A NEW ZEROth-ORDER ORACLE FOR DISTRIBUTED  
AND NON-STATIONARY LEARNING

by

Yan Zhang

Department of Mechanical Engineering and Material Science  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_

Michael M. Zavlanos, Advisor

\_\_\_\_\_

Leila Bridgeman

\_\_\_\_\_

Vahid Tarokh

\_\_\_\_\_

Ronald Parr

\_\_\_\_\_

An abstract of a dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy  
in the Department of Mechanical Engineering and Material Science  
in the Graduate School of  
Duke University

2021

Copyright © 2021 by Yan Zhang  
All rights reserved

# Abstract

Zeroth-Order (ZO) methods have been widely used to solve black-box or simulation-based optimization problems. These problems arise in many important modern day applications including generating adversarial attacks on machine learning systems or learning to control systems with complicated physical structure or humans in the loop. In these problems, the objective function to optimize does not have an explicit mathematical form and therefore its gradient cannot be obtained. As a result, traditional gradient-based optimization methods cannot be applied to solve these problems. Instead, ZO methods approximate the gradient by using the objective function values. Many existing ZO methods adopt two-point feedback to approximate the unknown gradient since this feedback scheme has low estimation variance and results in fast convergence speed. Specifically, two-point ZO methods estimate the gradient at the current iterate of the algorithm by querying the objective function value twice at two distinct neighboring points around the current iterate. However, two-point feedback is not possible or difficult to implement when the objective function is time-varying, or when multiple agents collaboratively optimize a global objective function that depends on all agents' decisions, because the value of the objective function can be queried only once at a single decision point. In this case, one-point ZO methods can be used which are known though to produce gradient estimates with large variance that slows down the convergence.

In this dissertation, we propose a novel one-point ZO method based on residual feedback. Specifically, the residual feedback scheme estimates the gradient using the residual between the values of the objective function at two consecutive iterates of the algorithm. 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, we show that ZO with one-point residual feedback achieves the same convergence rate as that of the two-point scheme with uncontrollable data samples.

Next, we apply the proposed one-point residual-feedback gradient estimator to solve online optimization problems, where the objective function varies over time. In this case, since each objective function can only be evaluated once at a single decision point, existing two-point ZO methods are not feasible and only one-point ZO methods can be used. We develop regret bounds for ZO with the proposed one-point residual feedback scheme for both convex and 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. Additionally, our regret bounds rely on weaker assumptions than those used in conventional one-point feedback methods.

The proposed residual-feedback scheme is next extended to solve distributed policy optimization problems that arise in multi-agent reinforcement learning (MARL). Existing MARL algorithms often assume that every agent can observe the states and actions of all the other agents in the network. This can be impractical in large-scale problems, where sharing the state and action information with multi-hop neighbors may incur significant communication overhead. The advantage of the proposed zeroth-order policy optimization method is that it allows the agents to compute

the local policy gradients needed to update their local policy functions using local estimates of the global accumulated rewards that depend on partial state and action information only and can be obtained using consensus. Specifically, one-point residual-feedback significantly reduces the variance of the local policy gradient estimates compared to using conventional one-point feedback, improving, in this way, the learning performance. We show that the proposed distributed zeroth-order policy optimization method with constant stepsize converges to a neighborhood of the global optimal policy that depends on the number of consensus steps used to calculate the local estimates of the global accumulated rewards.

Finally, to address situations where the agents do not have access to a global clock that they can use to synchronize their updates, we propose an asynchronous zeroth-order distributed optimization method that relies on the proposed one-point residual feedback gradient estimator. We show that this estimator is unbiased under asynchronous updating, and theoretically analyze its convergence.

We demonstrate the effectiveness of all proposed algorithms via extensive numerical experiments.

All contents presented in this dissertation have appeared in the papers<sup>1-4</sup>.

## Acknowledgements

First and foremost, I would like to thank my advisor Dr. Michael M. Zavlanos for all his guidance and patience during my research. Besides the countless and priceless advices that Michael has provided me with that has nurtured me into a skilled researcher who can think critically, synthesize new knowledge, and solve difficult engineering problems, he has also served as a role model that relies on the power of mutual understanding and encouragement to help young students get through their most difficult times when they first set foot into their career. I can not express my gratitude to Michael enough using words.

I would also like to thank all my Ph.D. committee members, Dr. Leila Bridgeman, Dr. Vahid Tarokh and Dr. Ronald Parr. Their feedback has been of utmost importance in directing my research to the correct direction. The works presented here would not be possible without their tremendous support.

I had the privilege of working with my wonderful colleagues, Charles Freundlich, Soomin Lee, Yiannis Kantaros, Reza Khodayi-mehr, Luke Calkins, Davood Hajinezhad, Xusheng Luo, Kavinayan Sivakumar, Yi Shen, Panagiotis Vlantis and so many excellent M.Sc. and undergraduate students in our lab. I also had many exciting discussions with Robert Ravier and Yi Zhou, who have helped me make a lot of progress during my research.

I'm hesitating to write down thank you to my dearest parents because words on paper make my gratitude to them sound lighter than it should be. They have supported me with unbelievable amount of love and tolerance long ago before I started to pursue my Ph.D., and their support is still here, warm and fresh during the more recent years of my life, even though it comes from thousands of miles away. Hugs from overseas.

To my dearest parents and friends.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Contents and Background . . . . .	3
1.2.1 A New One-Point Residual Feedback Zeroth-Order Oracle . . . . .	4
1.2.2 Zeroth-Order Online Learning using Residual Feedback . . . . .	6
1.2.3 Zeroth-Order Distributed MARL using Residual Feedback under Partial Observations . . . . .	8
1.2.4 Asynchronous Distributed Optimization using Residual Feedback . . . . .	9
1.3 Contributions . . . . .	10
<b>2 A New One-Point Residual Feedback Zeroth-Order Oracle</b>	<b>13</b>
2.1 Preliminaries . . . . .	13
2.2 Deterministic ZO with Residual Feedback . . . . .	14
2.2.1 Convergence Analysis . . . . .	16
2.3 Stochastic ZO with Residual Feedback . . . . .	18
2.3.1 Convergence Analysis . . . . .	19
2.4 Numerical Experiments . . . . .	20
2.4.1 A Deterministic Scenario: QP Problem . . . . .	20

2.4.2	A Stochastic Scenario: Policy Optimization . . . . .	21
<b>3</b>	<b>Zeroth-Order Online Learning using Residual Feedback</b>	<b>23</b>
3.1	Preliminaries and Problem Formulation . . . . .	23
3.2	ZO with Residual Feedback for Convex Online Optimization . . . . .	28
3.3	ZO with Residual Feedback for Non-Convex Online Optimization . . . . .	31
3.4	ZO with Residual Feedback for Stochastic Online Optimization . . . . .	33
3.5	Numerical Experiments . . . . .	35
3.5.1	Nonstationary LQR Control . . . . .	35
3.5.2	Nonstationary Resource Allocation . . . . .	37
<b>4</b>	<b>Zeroth-Order Distributed MARL using Residual Feedback under Partial Observations</b>	<b>39</b>
4.1	Preliminaries and Problem Formulation . . . . .	39
4.2	Algorithm Design and Theoretical Analysis . . . . .	42
4.2.1	Distributed Residual-Feedback Zeroth-Order Policy Optimization . . . . .	44
4.2.2	Distributed Residual-Feedback Zeroth-Order Policy Optimization with Value Tracking . . . . .	46
4.3	Numerical Experiments . . . . .	48
<b>5</b>	<b>Asynchronous Distributed Optimization using Residual Feedback</b>	<b>54</b>
5.1	Preliminaries and Problem Formulation . . . . .	54
5.2	Algorithm Design and Theoretical Analysis . . . . .	56
5.3	Numerical Experiments . . . . .	65
<b>6</b>	<b>Conclusions</b>	<b>69</b>
6.1	Future Research Directions . . . . .	70

<b>A</b>	<b>Proofs for Chapter 2</b>	<b>73</b>
A.1	Proof of Lemma 2.6 . . . . .	73
A.2	Proof of Theorem 2.7 . . . . .	75
A.3	Proof of Theorem 2.8 . . . . .	77
A.4	Proof of Theorem 2.9 . . . . .	78
A.5	Proof of Lemma 2.12 . . . . .	81
A.6	Proof of Theorem 2.13 . . . . .	82
A.7	Proof of Theorem 2.14 . . . . .	83
A.8	Zeroth-Order Policy Optimization for A Large-Scale Multi-Stage Decision Making Problem . . . . .	85
<b>B</b>	<b>Proofs for Chapter 3</b>	<b>88</b>
B.1	Implementation Details of the Numerical Experiments . . . . .	88
B.2	Proof of Lemma 3.2 and Lemma 3.3 . . . . .	89
B.3	Proof of Lemma 3.7 . . . . .	90
B.4	Proof of Theorem 3.9 . . . . .	91
B.5	Proof of Theorem 3.11 . . . . .	93
B.6	Proof of Theorem 3.14 . . . . .	94
B.7	Proof of Theorem 3.15 . . . . .	95
B.8	Analysis for Projected SGD with Residual-Feedback Oracle . . . . .	96
B.9	Proof of Lemma 3.18 . . . . .	103
B.10	Residual-Feedback Convex Optimization with Unit Sphere Sampling .	103
B.11	Proof of the Second Moment Bound (3.6) . . . . .	108
<b>C</b>	<b>Proofs for Chapter 4</b>	<b>109</b>
C.1	Proof of Lemma 4.6 . . . . .	109

C.2 Proof of Theorem 4.7 . . . . .	110
C.3 Proof of Lemma 4.8 . . . . .	115
C.4 Proof of Lemma 4.9 . . . . .	116
C.5 Proof of Theorem 4.10 . . . . .	117
<b>D Proofs for Chapter 5</b>	<b>121</b>
<b>Biography</b>	<b>128</b>

# List of Figures

2.1	The convergence rate of applying three oracles in the two problems. . . . .	21
3.1	Comparative results of ZO with three different oracles for online policy optimization in nonstationary LQR. . . . .	36
3.2	Comparative results of ZO with three different oracles for the nonstationary resource allocation problem. . . . .	37
4.1	Distributed zeroth-order policy optimization with the proposed residual-feedback estimator and the one-point estimator. . . . .	49
4.2	Distributed zeroth-order policy optimization with value tracking versus without value tracking. . . . .	50
4.3	Comparative results for Algorithm 1 with different algorithm hyperparameters. . . . .	52
5.1	Convergence results of the distributed feature learning problem with the proposed asynchronous zeroth-order oracles. . . . .	68
A.1	The convergence rate of applying three distributed zeroth-order oracles to the large-scale stochastic multi-stage resource allocation problem. . . . .	87

# List of Tables

1.1	Iteration Complexity of Zeroth-order Methods with One-point, Two-point and Proposed Residual Feedback Schemes . . . . .	6
-----	---	---

# Chapter 1

## Introduction

### 1.1 Motivation

Zeroth-Order (ZO) optimization methods have been widely applied to solve machine learning and control problems, where the gradient of the objective function cannot be computed, either because the target system is a black box<sup>5;6</sup>, or the mathematical formulation of the objective function is too complicated to write down<sup>7;8</sup>. Such problem arises in many practical applications. For example, in the human-in-the-loop trajectory planning problem<sup>6</sup>, the objective is to find the trajectory to navigate the robot through a crowd of people so that the robot create the least disturbance to human's works. However, human's preference for the robot trajectory is unknown *a priori*. e.g., their positions and how far they would like the robot to be away from them are black box to the robot. The only information the robot can obtain is the number of complaints received from the people who are affected by the robot when it follows a certain trajectory. It is desirable to use such information to find the robot trajectory that results in the fewest number of complaints. Another example is learning to control the heating, ventilation and air conditioning (HVAC) system in a building<sup>8</sup>. The goal is to find the best control policy for the switches of the HVAC system so that the temperature in the building is maintained in a comfortable zone. However, the effect of turning on and off the switches on the air temperature is governed by partial differential equations with complicated boundary conditions given by the structure of the building, which is impossible to model using mathematical formulations, and therefore traditional optimal control methods cannot be applied. On the contrary, it is easy to obtain the total duration when the temperature of the building is out of the desirable range given a fixed control policy. And we would like to use such information to find the optimal control policy.

A common feature for the above problems is that only the values of the objective

function at certain decision points can be used to find the optimal solution. Such problems fall into the research of derivative-free optimization<sup>9</sup>. ZO optimization methods are among one of the most popular derivative-free optimization approaches. Specifically, ZO methods randomly select one or multiple points around the current iterate, and use the values of the objective function at these points to estimate the gradient at the current iterate. Then, the ZO gradient estimate is used in the same way as in the gradient-based optimization methods. Other derivative-free methods, e.g., ellipsoid approach<sup>10</sup> and model-based approach<sup>11</sup>, also exist. However, the analysis of these approaches are usually more involved. In addition, the theoretical guarantee of these methods have poor dependency on the problem dimension. Furthermore, these methods are not easy to be extended to more challenging optimization problems, e.g., non-stationary optimization or distributed optimization settings. On the other hand, ZO methods approximate the gradient-based algorithms. Therefore, similar to the works where gradient-based optimization methods are extended to solve the non-stationary and distributed optimization problems, ZO methods can also be applied to these settings in a similar way. For these reasons, we focus on studying ZO optimization methods in this dissertation.

Existing ZO methods can be divided into two categories, namely, ZO with one-point feedback<sup>12</sup> and ZO with two-point feedback<sup>13</sup>. Specifically, ZO with one-point feedback estimates the gradient using the value at a single decision point at each iteration, while ZO with two-point feedback uses the values at two distinct decision points. It is well known that ZO with two-point feedback finds the optimal solution much faster than ZO with one-point feedback<sup>9</sup>, because the ZO gradient estimates with two-point feedback enjoy much lower variance than those with one-point feedback. Therefore, in the last few years, most of the research community have focused on analyzing the two-point ZO methods and designing new algorithms based on the two-point ZO estimators to solve more challenging settings. However, it remains unknown for years whether it is possible to design a one-point ZO method, which only requires the value of the objective function at a single decision point, to achieve comparative performance to the two-point ZO methods<sup>9;14</sup>.

Finding such competitive one-point ZO method is not only of theoretical interest. This is because the existing two-point ZO methods can be infeasible or difficult to implement in many practical problems. For example, in the online optimization problems where the objective function varies over time, it is impossible to obtain the values of the same objective function at two distinct decision points as required by the two-point ZO methods, if the evaluation of each decision point takes some time. Another example is that when multiple agents collaborate to find the optimal solution to a global objective function that depends on all their actions. In this case, each evaluation of the objective function requires all agents to synchronize to perturb their current decision. Therefore, distributed optimization methods based on two-point feedback require synchronization twice per update, while the ones based on one-point feedback only require synchronization once per update. Furthermore, when the agents conduct local updates in an asynchronous way, two-point ZO methods become infeasible. This is because when a local agent conducts two consecutive evaluations of the objective function at two decision points, the other agents in the network may have updated their local decision variables, making the local agent's objective function vary over these two evaluations. Therefore, it is desirable to implement a one-point ZO method in above settings, if such one-point ZO method can be shown to have comparable performance to that of the two-point ZO methods.

## 1.2 Contents and Background

In this section, we discuss the contents in each chapter and the existing literature that are relevant to each chapter.

## 1.2.1 A New One-Point Residual Feedback Zeroth-Order Oracle

In Chapter 2, our goal is to solve the following generic optimization problem

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

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 algorithms can be divided into two categories, namely, ZO with one-point feedback and ZO with two-point feedback.<sup>12</sup> 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  $\tilde{\nabla} f(x)$  takes the form<sup>1</sup>

$$(\text{One-point feedback}): \tilde{\nabla} f(x) = \frac{u}{\delta} f(x + \delta u), \quad (1.1)$$

where  $\delta$  is an exploration parameter and  $u \in \mathbb{R}^d$  is sampled from the standard normal distribution element-wise. In particular,<sup>12</sup> 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  $\mathcal{O}(T^{-\frac{1}{4}})$ , which is much slower than that of gradient descent algorithms used to solve problem (P). Assuming smoothness and relying on self-concordant regularization,<sup>16;17</sup> 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,<sup>13;15;18</sup> introduced the following two-point gradient estimators

$$\begin{aligned} (\text{Two-point feedback}): \tilde{\nabla} f(x) &= \frac{u}{\delta} f(x + \delta u) - f(x), \\ \text{or } \frac{u}{2\delta} &(f(x + \delta u) - f(x - \delta u)), \end{aligned} \quad (1.2)$$

---

<sup>1</sup>In<sup>12</sup>, the estimator is  $\tilde{\nabla} f(x) = \frac{du}{\delta} f(x + \delta u)$  where  $x \in \mathbb{R}^d$  and  $u$  is uniformly sampled from a unit sphere in  $\mathbb{R}^d$ . In this section, we follow<sup>15</sup> and sample  $u$  from the standard normal distribution.

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.

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

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

where only a noisy function evaluation  $F(x, \xi)$  with a random data sample  $\xi$  is available. ZO algorithms have also been developed to solve the above problem (Q), e.g.,<sup>7;14;19–21</sup>. In particular,<sup>7</sup> consider the following widely-used stochastic two-point feedback

$$\tilde{\nabla} f(x) = \frac{u}{\delta} (F(x + \delta u, \xi) - F(x, \xi)) \quad (1.3)$$

and show that ZO with this stochastic two-point feedback has the same convergence rate as ZO with the two-point feedback scheme in (1.2) for deterministic problems (P). Similarly,<sup>19</sup> further analyzed the oracle in (1.3) 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<sup>21</sup>. However, these stochastic two-point feedback schemes assume that the data sample  $\xi$  is controllable, i.e., one can fix the data sample  $\xi$  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  $\xi$  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,<sup>14;20</sup> proposed a more practical noisy two-point feedback method that replaces the fixed sample  $\xi$  in (1.3) with two independent samples  $\xi, \xi'$ . Its convergence rate was shown to match that of the stochastic one-point feedback  $\tilde{\nabla} f(x) = \frac{u}{\delta} F(x + \delta u, \xi)$ . Still though, this two-point feedback method with independent data samples produces gradient estimates with lower variance compared to the conventional one-point feedback

**Table 1.1:** Iteration Complexity of Zeroth-order Methods with One-point, Two-point and Proposed Residual Feedback Schemes

Complexity <sup>3</sup>		Convex $C^{0,0}$	Convex $C^{1,1}$	Nonconvex $C^{0,0}$	Nonconvex $C^{1,1}$
One-point	21	$d^2\epsilon^{-4}$	$d\epsilon^{-3}$	–	–
Two-point	19	$d\log(d)\epsilon^{-2}$	$d\epsilon^{-2}$	–	–
	18	$d\epsilon^{-2}$	–	–	–
	15	$d^2\epsilon^{-2}$	$d\epsilon^{-1}$	$d^3\epsilon_f^{-1}\epsilon^{-2}$	$d\epsilon^{-1}$
	20	–	$d^2\epsilon^{-3}$ (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	$d^2\epsilon^{-4}$	$d^2\epsilon^{-3}$	$d^3\epsilon_f^{-3}\epsilon^{-2}$	$d^4\epsilon^{-3}$

method.

In this chapter, we close the gap between the theoretical guarantee between the two-point ZO methods and the one-point ZO methods. Specifically, we propose a one-point residual feedback ZO gradient estimator, and we show that the performance of the proposed one point residual feedback ZO oracle achieves comparable performance to the two-point ZO methods, through both theoretical analysis and numerical experiments. We summarize the theoretical guarantee of ZO methods with the proposed residual-feedback and compare it to ZO methods with both two-point and one-point feedback schemes in Table 1.1.

## 1.2.2 Zeroth-Order Online Learning using Residual Feedback

In Chapter 3, we study the time-varying optimization problem using the proposed one-point residual feedback ZO oracle. 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  $x_t$

<sup>3</sup>In convex setting, the accuracy is measured by  $f(x) - f(x^*) \leq \epsilon$ , while in the non-convex setting, it is measured by  $\|\nabla f(x)\|^2 \leq \epsilon$  when the objective function is smooth. When the objective function is non-smooth, we enforce two optimality measures,  $|f(x) - f_\delta(x)| \leq \epsilon_f$  and  $\|\nabla f_\delta(x)\|^2 \leq \epsilon$  together. (UN) means the oracle considers uncontrollable data samples.

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 can be measured using notions of regret, that generally 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.

As we have discussed above, ZO methods with two-point feedback like<sup>13;19</sup> can not be used for non-stationary online optimization problems that arise frequently in online learning. The reason is that in these non-stationary online optimization problems, the objective function being queried is time-varying, and hence it can only be evaluated at a single decision variable at a given time instant. ZO with one-point feedback<sup>12;21</sup>, on the other hand, can be used in the non-stationary setting. However, they produce 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 can not be used for practical non-stationary optimization problems.

Besides ZO methods, other derivative-free optimization algorithms have also been applied to solve online optimization problems. The works in<sup>22</sup> employs the exploration and exploitation bandit learning framework and the proposed analysis is restricted to a special class of non-convex objective functions. Finally,<sup>11;23;24</sup> study online bandit algorithms using ellipsoid or model-based methods. In particular, these methods induce heavy computation per step and achieve regret bounds that have bad dependence on the problem dimension.

In this chapter, we apply the proposed residual-feedback ZO oracle to solve the online optimization problems. We provide regret analysis for the proposed algorithm. We show that it outperforms the conventional one-point ZO method in such online setting through both theoretical analysis and numerical experiments.

### 1.2.3 Zeroth-Order Distributed MARL using Residual Feedback under Partial Observations

In Chapter 4, we extend the proposed one-point residual-feedback ZO estimator to solve multi-agent reinforcement learning problems under partial observations. The goal is to enable a team of agents to collaboratively determine the global optimal policy that maximizes the sum of their local accumulated rewards. To do so, the agents typically need to communicate with each other in order to obtain information about the global state and action of the team. This is because their states and rewards are generally affected by the actions of their other peers. However, sharing such information can be undesirable, due to significant communication overhead or privacy concerns. Therefore, there is a great need for MARL algorithms that rely only on partial observations of the global state and action information.

A major challenge in developing cooperative MARL methods under partial observations is that the environment, as it is perceived by every individual agent, is non-stationary since it changes as a result of changes in the policies of the other agents<sup>25</sup>. In<sup>25-27</sup>, this challenge is addressed using a centralized Critic function that can mitigate the effect of non-stationarity in the learning process. Then, the trained policies can be executed in a decentralized way. In<sup>28</sup>, a distributed offline experience replay technique is developed to enable fully decentralized training, which requires that all agents receive a global reward at each timestep. However, when the global reward is defined as the sum of local agent rewards, as in cooperative MARL, this global reward can not be easily available to the local agents in practice. Cooperative MARL methods that maximize the sum of local rewards are considered in<sup>29-31</sup>. These works develop fully decentralized Actor-Critic methods where the agents maintain local estimates of the global value or policy functions, that depend on the states and actions of all other agents and update those estimates until they reach consensus. Then, these local estimates of the global value or policy functions are used to compute the policy gradient estimates needed for optimization. Since these policy gradient estimates require knowledge of the global state and action information, such Actor-Critic methods

can not be used for cooperative MARL with partial state and action information.

All aforementioned MARL algorithms are gradient-based algorithms. On the other hand, zeroth-order policy optimization has been considered in<sup>32;33</sup> for a special case of single-agent RL problems, namely, Linear Quadratic Regulation (LQR) problems. These results were extended in<sup>8</sup> for distributed LQR problems. All these works use the one-point zeroth-order policy gradient estimator proposed in<sup>12;15</sup>, which is known to have large variance that slows down learning.

In this chapter, we propose a distributed policy optimization algorithms by decentralizing the proposed residual-feedback ZO oracle. We theoretically analyze the convergence of the proposed decentralized ZO policy optimization algorithm in MARL problems with partial observations. And we show that it achieves better performance than the distributed policy optimization algorithms based on conventional one-point ZO methods<sup>8</sup> using numerical experiments.

## 1.2.4 Asynchronous Distributed Optimization using Residual Feedback

In Chapter 5, we extend the proposed residual-feedback ZO oracle to the asynchronous setting, where a group of agents collaboratively minimize a common cost function that depends on their joint decisions and at each time step, a single agent is randomly activated to query the objective function value at one decision point and update its local decision variable.

We note that existing two-point ZO oracle (1.2) cannot be directly extended to the asynchronous setting. This is because the perturbation  $u$  is according to the full decision vector  $x$ . When the vector  $x$  concatenates all agents' local decisions  $\{x_i\}$ , perturbing  $x$  with random vector  $u$  must happen simultaneously, which is infeasible when the agents do not have access to a global clock. If the objective function gradient is known, asynchronous distributed optimization methods with a common objective function have been studied in<sup>34-36</sup>. However, these works can not be directly extended to solve the black-box

optimization problems considered here.

Perhaps the most related works in this setting is<sup>37</sup>. Specifically, the authors in proposed an asynchronous zeroth-order optimization algorithm that relies on the two-point gradient estimator (1.2). Unlike the method proposed here,<sup>37</sup> assumes that when a single agent queries the values of the objective function at decision points  $x_k + \mu u_k$  and  $x_k$  (or  $x_k - \mu u_k$ ), the other agents in the network cannot update their local decision variables even if they are activated. This assumption limits the number of updates the agents can make during a fixed period of time and affects the performance of the asynchronous system.

In this chapter, we study an asynchronous version of the proposed residual-feedback ZO oracle and prove its convergence. We also demonstrate the proposed asynchronous residual-feedback ZO oracle outperforms the two-point methods under asynchronous setting.

### 1.3 Contributions

In this section, we summarize the contributions of the proposed algorithms in all the chapters separately.

In Chapter 2, we propose a new one-point feedback scheme which requires a single function evaluation at each iteration, the residual-feedback scheme, which estimates the gradient using the residual between two consecutive feedback points. We show that our residual feedback induces a smaller estimation variance than the one-point feedback (1.1) considered in<sup>12;21</sup>. 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 in static optimization problems.

In Chapter 3, we propose a novel one-point gradient estimator for zeroth-order online optimization and develop new regret bounds to study its performance. Compared to solving the online optimization problem with conventional one-point gradient estimator in <sup>12;21</sup>, our proposed method obtains tighter regret bounds both for convex and non-convex 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. 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 non-convex 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.

In Chapter 4, we propose a new distributed zeroth-order policy optimization method for general cooperative MARL problems based on the proposed residual-feedback ZO gradient estimator. Compared to the one-point policy gradient estimators in <sup>8;32;33</sup>, our proposed residual-feedback policy gradient estimator reduces the variance of the policy gradient estimates and, therefore, improves the learning performance. Compared to the centralized estimator studied in Chapters 2 and 3 that produces unbiased gradient estimates, the proposed distributed policy gradient estimator is biased due to possible consensus errors in distributedly estimating the sum of local accumulated rewards needed for the estimation of the policy gradients. We show that the proposed zeroth-order policy optimization method with constant stepsize converges to a neighborhood of the global optimal policy whose size depends on the number of consensus steps needed to control the bias in the policy gradient estimates. Moreover, we propose a value tracking method to reduce the number of consensus steps needed to achieve a desired user-specified solution accuracy.

In Chapter 5, we propose an asynchronous zeroth-order distributed optimization algorithm based on an extension of the centralized residual-feedback gradient estimator studied in Chapters 2 and 3, so that it can handle fully asynchronous queries and updates. Specifically, we show that the proposed zeroth-order gradient estimator provides an unbiased estimate of the gradient with respect to each agent's local decision. Also, we provide bounds on the second moment of this estimator, the first of their kind for any asynchronous zeroth-order gradient of this type, which we then use to show convergence of the proposed method.

# Chapter 2

## A New One-Point Residual Feedback Zeroth-Order Oracle

In this chapter, we propose a novel one-point ZO gradient estimator based on the residual feedback. We will show that although the proposed estimator only evaluates the objective function once at a single decision point per iteration as conventional one-point ZO method, it achieves comparable performance of that of the existing two-point ZO methods. We will study this estimator both theoretically and numerically. The proposed one-point ZO oracle based on residual feedback is not only of theoretical importance, it is also more flexible to implement and enables the ZO optimization scheme to be applied to solve more challenging optimization and learning problems in practice, as we will demonstrate in the next few chapters. The content in this chapter can also be found in the paper<sup>1</sup>.

### 2.1 Preliminaries

In this section, we present definitions and preliminary results needed throughout our analysis. Following<sup>15;20</sup>, we introduce the following classes of Lipschitz and smooth functions.

**Definition 2.1** (Lipschitz functions). *The class of Lipschitz-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<sup>15</sup>,  $f_\delta(x) := \mathbb{E}_{u \sim \mathcal{N}(0,1)}[f(x +$

$\delta u$ ], 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<sup>15</sup>.

**Lemma 2.2** (Gaussian approximation). *Consider a function  $f$  and its Gaussian-smoothed version  $f_\delta$ . 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<sup>15</sup>.

**Lemma 2.3.** *If function  $f \in C^{0,0}$  is  $L_0$ -Lipschitz, then its Gaussian-smoothed version  $f_\delta$  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 2.4** (Convexity). *A continuously differentiable function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is called convex if for all  $x, y \in \mathbb{R}^d$ ,  $f(x) \geq f(y) + \langle x - y, \nabla f(y) \rangle$ .*

## 2.2 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

$$\tilde{g}(x_t) := \frac{u_t}{\delta} (f(x_t + \delta u_t) - f(x_{t-1} + \delta u_{t-1})), \quad (2.1)$$

where  $u_{t-1}$  and  $u_t$  are independent random vectors sampled from the standard multivariate Gaussian distribution. To elaborate, the gradient estimate in (2.1) 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 2.5.** *We have  $\mathbb{E}[\tilde{g}(x_t)] = \nabla f_\delta(x_t)$  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.  $\square$

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

$$x_{t+1} = x_t - \eta \tilde{g}(x_t), \quad (2.2)$$

where  $\eta$  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  $\delta$  in (2.1) and the stepsize  $\eta$ . 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 2.6.** *Consider a function  $f \in C^{0,0}$  with Lipschitz constant  $L_0$ . Then, under the SGD update rule in (2.2), 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*

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \frac{2dL_0^2\eta^2}{\delta^2} \mathbb{E}[\|\tilde{g}(x_{t-1})\|^2] + 8(d+4)^2 \|\nabla f(x_{t-1})\|^2 + 4L_1^2(d+6)^3 \delta^2. \quad (2.3)$$

The proof of above Lemma 2.6 can be found in Appendix A.1. Lemma 2.6 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<sup>14</sup>. In comparison, our residual feedback induces a much smaller perturbation term. Specifically, when  $f \in C^{0,0}$ , the perturbation is the order of  $O(L_0^2 d^2)$  that is independent of  $\delta$ , and when  $f \in C^{1,1}$ , the perturbation is in the order of  $O(d^2 \|\nabla f(x_{t-1})\|^2 + L_1^2 d^3 \delta^2)$ . Therefore, ZO with our residual feedback can achieve a better iteration complexity than that of ZO with the traditional one-point feedback.

### 2.2.1 Convergence Analysis

We first consider the case where the objective function  $f$  is nonconvex. When  $f$  is differentiable, we say a solution  $x$  is  $\epsilon$ -accurate if  $\mathbb{E}[\|\nabla f(x)\|^2] \leq \epsilon$ . However, when  $f$  is nonsmooth, we follow the convention adopted in<sup>15</sup> and define a solution  $x$  to be  $\epsilon$ -accurate if  $\mathbb{E}[\|\nabla f_\delta(x)\|^2] \leq \epsilon$  holds for the Gaussian-smoothed function. In addition, we also require  $f_\delta$  to be  $\epsilon_f$ -close to the original  $f$ , which requires  $\delta \leq \frac{\epsilon_f}{L_0 \sqrt{d}}$  according to Lemma 2.2. Under this setup, the convergence rate of ZO with residual feedback is presented below. For simplicity, all the complexity results in this section are presented in  $\mathcal{O}$  notations. The proofs and the explicit form of the constant terms can be found in the supplementary material.

**Theorem 2.7.** *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 (2.2) 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  $\mathbb{E}[\|\nabla f_\delta(\tilde{x})\|^2] = \mathcal{O}(d^2 \epsilon_f^{-0.5} T^{-0.5})$ .*

The proof can be found in Appendix A.2. 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<sup>15</sup>. When  $f(x) \in C^{1,1}$  is a smooth function, we obtain the following convergence rate result for ZO with residual feedback.

**Theorem 2.8.** *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 (2.2) with residual feedback is run for  $T$  iterations and that  $\tilde{x}$  is selected from the  $T$  iterates uniformly at random. Let also  $\eta = \frac{1}{\tilde{L}(d+4)^2 T^{\frac{1}{3}}}$ , and  $\delta = \frac{1}{\sqrt{dT^{\frac{1}{3}}}}$ , where  $\tilde{L} = \max(2L_0, 32L_1)$ . Then, we have that  $\mathbb{E}[\|\nabla f(\tilde{x})\|^2] = \mathcal{O}(d^2 T^{-\frac{2}{3}})$ .*

The proof can be found in Appendix A.3. In particular, to achieve a point  $x$  that satisfies  $\mathbb{E}[\|\nabla f(\tilde{x})\|^2] \leq \epsilon$ , 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<sup>15</sup>. 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 2.9.** *Assume that  $f(x) \in C^{0,0}$  is convex with Lipschitz constant  $L_0$ . Moreover, assume that SGD in (2.2) 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  $\eta = \frac{1}{2\tilde{L}(d+4)^2 T^{\frac{1}{3}}}$  and  $\delta = \frac{\sqrt{d}}{T^{\frac{1}{3}}}$ , where  $\tilde{L} = \max\{L_0, 16L_1\}$ . Then, we have that  $f(\bar{x}) - f(x^*) = \mathcal{O}(d^2 T^{-\frac{2}{3}})$ .

The proof can be found in Appendix A.4. 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  $\epsilon < d^{-4/3}$ .

## 2.3 Stochastic ZO with 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

$$\tilde{g}(x_t) := \frac{u_t}{\delta} (F(x_t + \delta u_t, \xi_t) - F(x_{t-1} + \delta u_{t-1}, \xi_{t-1})), \quad (2.4)$$

where  $\xi_{t-1}$  and  $\xi_t$  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<sup>33</sup>, 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 (2.4) 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 (2.4) 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 2.10.** (*Bounded Variance*) We assume that for any  $x \in \mathbb{R}^d$  there exists  $\sigma > 0$  such that

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

Assumption 2.10 implies that  $\mathbb{E}[(f(x, \xi_1) - f(x, \xi_2))^2] \leq 4\sigma^2$ . Furthermore, we make the following smoothness assumption in the stochastic setting.

**Assumption 2.11.** 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 2.12.** *Let Assumptions 2.10 and 2.11 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 A.5. 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  $\xi_t$  and  $\xi_{t-1}$ . 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<sup>14</sup>, such a term can significantly degrade the iteration complexity.

### 2.3.1 Convergence Analysis

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

**Theorem 2.13.** *Let Assumptions 2.10 and 2.11 hold and assume also that  $F(x, \xi) \in C^{0,0}$ . Moreover, assume that SGD in (2.2) 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  $\mathbb{E}[\|\nabla f_\delta(\tilde{x})\|^2] = \mathcal{O}(d^{1.5}\epsilon_f^{-1.5}T^{-0.5})$ .*

*Furthermore, assume that additionally  $F(x, \xi) \in C^{1,1}$ , and that SGD in (2.2) 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  $\mathbb{E}[\|\nabla f(\tilde{x})\|^2] = \mathcal{O}(d^{\frac{4}{3}} T^{-\frac{1}{3}})$ .*

The proof can be found in Appendix A.6. Based on the above results, when  $F(x, \xi)$  is non-smooth, to achieve the  $\epsilon$ -stationary point  $\mathbb{E}[\|\nabla f_\delta(\tilde{x})\|^2] \leq \epsilon$  and  $|f(x) - f_\delta(x)| \leq \epsilon_f$ ,  $\mathcal{O}(\frac{d^3}{\epsilon_f^2 \epsilon^2})$  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 2.14.** *Let Assumptions 2.10 and 2.11 hold and assume that the function  $F(x, \xi) \in C^{0,0}$  is also convex. Moreover, assume that SGD in (2.2) 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  $\eta = \frac{1}{2\sqrt{2}L_0\sqrt{dT}^{\frac{3}{4}}}$  and  $\delta = \frac{1}{T^{\frac{1}{4}}}$ . Then, we have that  $f(\bar{x}) - f(x^*) = \mathcal{O}(\sqrt{dT}^{-\frac{1}{4}})$ . Moreover, assume that additionally  $F(x, \xi) \in C^{1,1}$ , and let  $\eta = \frac{1}{2\sqrt{2}L_0d^{\frac{2}{3}}T^{\frac{2}{3}}}$  and  $\delta = \frac{1}{d^{\frac{1}{6}}T^{\frac{1}{6}}}$ . Then, we have that  $f(\bar{x}) - f(x^*) = \mathcal{O}(d^{\frac{2}{3}}T^{-\frac{1}{3}})$ .*

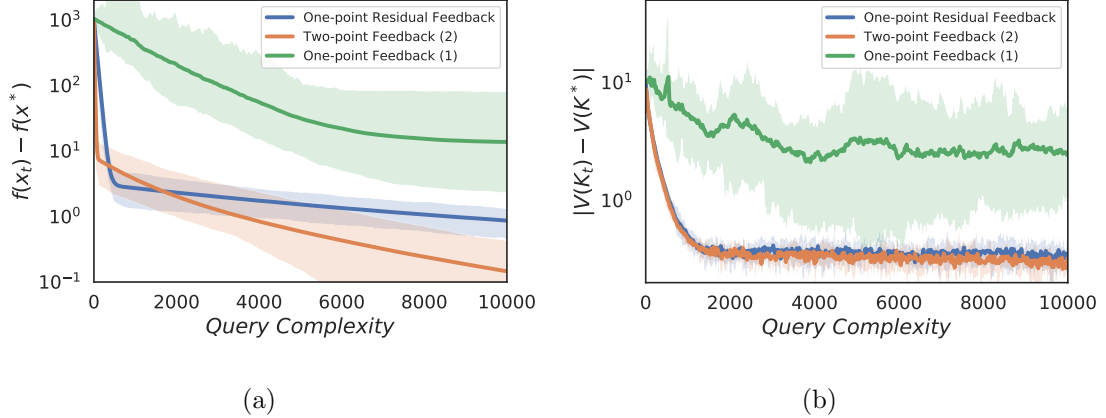
The proof can be found in Appendix A.7. According to Theorem 2.14,  $\mathcal{O}(\frac{d^2}{\epsilon^4})$  iterations are needed to achieve  $f(\bar{x}) - f(x^*) \leq \epsilon$  with a nonsmooth objective function. On the other hand, if  $f(x) \in C^{1,1}$ , the iteration complexity is improved to  $\mathcal{O}(\frac{d^2}{\epsilon^3})$ .

## 2.4 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<sup>38</sup>. In the stochastic case, we employ the stochastic variants of above oracles to optimize the policy parameters in a Linear Quadratic Regulation (LQR) problem considered in<sup>32;33</sup>. 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. 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.

### 2.4.1 A Deterministic Scenario: QP Problem

As in<sup>38</sup>, 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



**Figure 2.1:** The convergence rate of applying the proposed residual one-point feedback (2.1) (blue), the two-point oracle (1.2) in<sup>15</sup> (orange) and the one-point oracle (1.1) in<sup>12</sup> (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.

initial point is the origin. For every algorithm, we manually optimize the selection of the exploration parameter  $\delta$  and stepsize  $\eta$  and run it 100 times. The convergence of the function value  $f(x) - f(x^*)$  is presented in Figure 2.1(a). We observe that the proposed oracle converges as fast as the two-point oracle (1.2) 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<sup>12;21</sup>. These observations validate our theoretical results in Section 2.2.

## 2.4.2 A Stochastic Scenario: Policy Optimization

We use the proposed residual feedback to optimize the policy parameters in a LQR problem, as in<sup>32;33</sup>. 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  $V(K) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (x_k^T Q x_k + u_k^T R u_k)]$  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 (1.1) with noise<sup>21</sup>, two-point feedback with uncontrolled noise<sup>14;20</sup> and the residual one-point feedback (2.4) 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 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<sup>33</sup>. 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 2.1(b). We observe that the residual one-point feedback (2.4) converges much faster than the one-point oracle in<sup>21</sup> and has comparable query complexity to the two-point feedback under uncontrolled noises considered in<sup>14;20</sup>. This corroborates our theoretical analysis in Section 2.3. Furthermore, we apply our residual-feedback zeroth-order gradient estimate to solve a large-scale stochastic multi-stage decision making problem with problem dimension  $d = 576$ . The implementation details and results of the simulation can be found in Appendix A.8, where it can be seen that our residual feedback achieves similar improvement of the convergence rate over the conventional one-point feedback scheme.

# Chapter 3

## Zeroth-Order Online Learning using Residual Feedback

In this chapter, we extend the residual-feedback ZO gradient estimator proposed in Chapter 2 to the non-stationary online optimization problems. Such problems arise in many real world problems. For example, when we want to optimize according to human’s preference model, it is natural to assume that human may change its opinion when his/her contextual information, e.g., location, job, age, changes. Another example is when we learn the optimal policy to interact with an environment which consists of other agents. When the other agents change their strategy, it also changes the environment we face with. Therefore, studying ZO methods to solve non-stationary learning problems is of practical importance. We will show that when applying the proposed residual-feedback ZO estimator to optimize a time-varying objective function, it leads to a regret depending on how fast the objective function changes. And we demonstrate the proposed ZO method outperforms the conventional one-point ZO methods when tracking the trajectory of the time-varying optimizers using extensive numerical experiments. The contents in this chapter are also presented in the paper<sup>2</sup>.

### 3.1 Preliminaries and Problem Formulation

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 that we propose to solve online optimization problems with unknown gradient information. Consider the following online bandit optimization problem

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

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 (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  $\mathcal{X}_\delta := \{z : z = x + \delta u, \text{ for any } x \in \mathcal{X} \text{ and } u \in \text{US}^d\}$ , where  $\text{US}^d$  represents the unit sphere in space  $\mathbb{R}^d$ .

**Definition 3.1** (Lipschitz functions). *The class of Lipschitz-continuous functions  $C^{0,0}$  satisfies: for any  $f_t \in C^{0,0}$ ,  $|f_t(x) - f_t(y)| \leq L_0 \|x - y\|$ ,  $\forall x, y \in \mathcal{X}_\delta$ , where  $L_0 > 0$  is the Lipschitz parameter over set  $\mathcal{X}_\delta$ . The class of smooth functions  $C^{1,1}$  satisfies: for any  $f_t \in C^{1,1}$ ,  $\|\nabla f_t(x) - \nabla f_t(y)\| \leq L_1 \|x - y\|$ ,  $\forall x, y \in \mathcal{X}_\delta$ , 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  $f_t$  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  $f_t$  defined as  $f_{\delta,t}(x) := \mathbb{E}_{u \sim \text{UB}^d}[f_t(x + \delta u)]$ , where the coordinates of the vector  $u$  are uniformly sampled from a unit ball  $\text{UB}^d$  in space  $\mathbb{R}^d$ . Specifically, we have the following results bounding the approximation errors of the function  $f_{\delta,t}(x)$ .

**Lemma 3.2.** *Consider a function  $f_t$  and its smoothed version  $f_{\delta,t}$ . It holds that for all  $t$*

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

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

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

**Lemma 3.3.** *If  $f_t \in C^{0,0}$  is  $L_0$ -Lipschitz, then  $f_{\delta,t} \in C^{1,1}$  with Lipschitz constant  $L_1 = d\delta^{-1}L_0$  for all  $t$ .*

The proofs of above lemmas are included in Appendix B.2 in the supplementary material for completeness.

**Definition 3.4.** *(Objective functions) We call the sequence of objective functions  $\{f_0, f_1, \dots, f_t\}$  naturally non-stationary 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 non-stationary if the selection of  $f_t$  depends also on the agent's current decision  $x_t + \delta u_t$ .*

In this section we consider both natural and adversarial objective function sequences, as defined in Definition 3.4. Natural non-stationary 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 multi-agent 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 non-stationary environment. In such problems where the system evolves from  $f_t$  to  $f_{t+1}$ , two point feedback (1.2) can not 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 (1.1) in<sup>21</sup>. However, the gradient estimates produced by the one-point feedback method in (1.1) 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 section 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

$$\text{(Residual feedback):} \quad \tilde{g}_t(x_t) := \frac{d}{\delta} (f_t(x_t + \delta u_t) - f_{t-1}(x_{t-1} + \delta u_{t-1})) u_t, \quad (3.1)$$

where  $u_{t-1}, u_t \sim \mathbb{US}^d$  are independent random vectors. To elaborate, the proposed residual feedback estimator in (3.1) 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 3.5.** *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  $u_t$  has zero mean and is independent from  $u_{t-1}, x_{t-1}$ . □

**Remark 3.6.** *We note that the existing two-point estimators cannot be easily modified to be used for non-stationary optimization. The difficulty is in ensuring that the returned gradient estimates are unbiased as in the case of residual feedback in Lemma 2.4. To see this, consider the simple modification of the online two-point gradient estimator (7) proposed in<sup>20</sup>*

$$\tilde{g}_t(x_t) = \frac{d}{2\delta} (f_t(x_t + \delta u_t) - f_{t-1}(x_t - \delta u_t)) u_t.$$

Then, it is easy to see that this modified two-point gradient estimator is biased since  $\mathbb{E}[\tilde{g}_t(x_t)] \neq \nabla f_{\delta,t}(x_t)$ . Specifically, let  $\tilde{g}_t(x_t) = \frac{d}{2\delta} (f_t(x_t + \delta u_t) - f_{t-1}(x_t - \delta u_t)) u_t = \frac{d}{2\delta} (f_t(x_t + \delta u_t) - f_t(x_t - \delta u_t) + \epsilon_t) u_t$ , where  $\epsilon_t = f_t(x_t - \delta u_t) - f_{t-1}(x_t - \delta u_t)$ . Although  $\mathbb{E}[\frac{d}{2\delta} (f_t(x_t + \delta u_t) - f_t(x_t - \delta u_t)) u_t] = \nabla f_{\delta,t}(x_t)$ , we have that  $\mathbb{E}[\frac{d}{2\delta} \epsilon_t u_t] \neq 0$  since  $\epsilon_t$  is correlated with  $u_t$ . Therefore, for this modified estimator we have that  $\mathbb{E}[\tilde{g}_t(x_t)] = \mathbb{E}[\frac{d}{2\delta} (f_t(x_t + \delta u_t) - f_t(x_t - \delta u_t)) u_t] + \mathbb{E}[\frac{d}{2\delta} \epsilon_t u_t] \neq \nabla f_{\delta,t}(x_t)$ . Note that the original two-point estimator proposed in<sup>20</sup> is unbiased, because the function  $f_t$  is queried at two points,  $x_t + \delta u_t$  and  $x_t - \delta u_t$ , and the noise  $\epsilon_t$  in this case is simply the evaluation noise that is zero mean for any  $u_t$ .

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

feedback:

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta \tilde{g}_t(x_t)), \quad (3.2)$$

where  $\eta$  is the learning rate and  $\Pi_{\mathcal{X}}$  is the projection operator onto the set  $\mathcal{X}$ . The update (3.2) can be implemented assuming that the objective function can be queried at points outside the feasible set  $\mathcal{X}$ , similar to the methods considered in<sup>19–21</sup>. Note that it is possible to modify the update (3.2) so that the iterates are guaranteed to be within the feasible set  $\mathcal{X}$ . This modification and related analysis can be found in Section B.10 in the supplementary material. The requirement that the objective function is evaluated at feasible points in derivative-free optimization algorithms has also been considered in<sup>11;39</sup>. Specifically,<sup>11</sup> develop the so called ellipsoid method, which requires computation of an ellipsoid containing the optimizer at each time step. The following result bounds the second moment of the gradient estimate generated by using residual feedback.

**Lemma 3.7** (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 (3.2), the second moment of the residual feedback (3.1) satisfies:*

$$\mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \leq \frac{4d^2 L_0^2 \eta^2}{\delta^2} \mathbb{E}[\|\tilde{g}_{t-1}(x_{t-1})\|^2] + D_t, \quad (3.3)$$

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

The proof of above lemma can be found in Appendix B.3 in the supplementary material. 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  $\alpha = 4d^2 L_0^2 \eta^2 \delta^{-2} < 1$ . 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 (3.2).

## 3.2 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  $\mathcal{X}$  can be either compact or the whole domain. In particular, we are interested in analyzing the static regret of algorithm (3.2) defined as

$$R_T := \mathbb{E} \left[ \sum_{t=0}^{T-1} f_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x) \right]. \quad (3.4)$$

First, we make the following assumption on the non-stationarity of the online learning problem.

**Assumption 3.8** (Bounded variation). *There exists  $V_f > 0$  such that for all  $t$  and all  $x \in \mathcal{X}_\delta$ ,*

$$|f_t(x) - f_{t-1}(x)| \leq V_f. \quad (3.5)$$

Assumption 3.8 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.,  $|f_t(x)| \leq B, \forall t, x$ , which is used in the analysis of ZO with conventional one-point feedback in<sup>12;21</sup>. In particular, under Assumption 3.8, the perturbation term in Lemma 3.7 can be bounded as  $D_t \leq 16L_0^2 d^2 + 2d^2 V_f^2 \delta^{-2}$ . Then, by telescoping the contraction inequality, we obtain the following bound for the second moment of the residual-feedback gradient estimate

$$\mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \leq \max \left\{ \mathbb{E}[\|\tilde{g}_0(x_0)\|^2], \frac{1}{1-\alpha} \left( 16L_0^2 d^2 + \frac{2d^2}{\delta^2} V_f^2 \right) \right\}. \quad (3.6)$$

The detailed proof can be found in Appendix B.11 in the supplementary material. In practice,  $\delta$  needs to be sufficiently small so that the smoothed function  $f_{\delta,t}$  is close to the original function  $f_t$  according to Lemma 3.2. In this case, the above bound on the second

moment of the residual-feedback gradient estimates is dominated by  $\mathcal{O}(d^2\delta^{-2}V_f^2)$ , which is much smaller than the bound on the second moment of the conventional one-point gradient estimates  $\mathcal{O}(d^2\delta^{-2}B^2)$ , where  $B$  is the uniform bound on  $|f_t|$  over time. For example, consider the time-varying objective functions,  $f_0(x) = 1/2x^2$  and  $f_t(x) = f_{t-1}(x) + n_t$ , where  $n_t$  is Gaussian noise with zero mean at time  $t$ . Then, it can be verified that Assumption 3.8 holds with a finite  $V_f$  whereas the second moment of  $f_t(x)$  is unbounded over time. As a result, the variance of the residual feedback gradient estimates can be significantly smaller than that of the conventional one-point feedback gradient estimates.

For any sequence of objective functions, natural or adversarial, as defined in Definition 3.4, the following result characterizes the regret of ZO with residual feedback when the objective function  $f_t$  is convex and Lipschitz.

**Theorem 3.9** (Regret for Convex Lipschitz  $f_t$ ). *Let Assumption 3.8 hold. Assume that  $f_t \in C^{0,0}$  is convex with Lipschitz constant  $L_0$  over set  $\mathcal{X}_\delta$  for all  $t$  and  $\|x_0 - x^*\| \leq R$ . Run ZO with residual feedback for  $T > R^2$  iterations with  $\eta = R^{\frac{3}{2}}(2\sqrt{2d}L_0T^{\frac{3}{4}})^{-1}$  and  $\delta = \sqrt{dRT}^{-\frac{1}{4}}$ . Then, we have that*

$$\begin{aligned} R_T \leq & \sqrt{2}L_0\sqrt{dRT}^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]R^{\frac{3}{2}}}{2\sqrt{2d}L_0T^{\frac{3}{4}}} + 2L_0\sqrt{dRT}^{\frac{3}{4}} \\ & + 4\sqrt{2}d^{\frac{3}{2}}L_0R^{\frac{3}{2}}T^{\frac{1}{4}} + \sqrt{dR/2}V_f^2L_0^{-1}T^{\frac{3}{4}}. \end{aligned} \quad (3.7)$$

Asymptotically,  $R_T = \mathcal{O}((L_0 + L_0^{-1}V_f^2)\sqrt{dRT}^{\frac{3}{4}})$ .

The proof can be found in Appendix B.4 in the supplementary material. To the best of our knowledge, the best known regret for ZO with conventional one-point feedback is of the order  $\mathcal{O}(\sqrt{dL_0RB}T^{\frac{3}{4}})^{21}$ . Therefore, our regret bound is tighter if the function variation satisfies  $V_f^2 \leq \mathcal{O}(B^{\frac{1}{2}}L_0^{\frac{3}{2}})$ . Essentially, using the proposed residual feedback gradient estimator, the regret of 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. Note that knowledge of the neighborhood  $R$  in Theorem 3.9 allows to select the step-size  $\eta$  and the parameter  $\delta$  so that a better regret rate can be achieved that depends on  $R$ . However, knowledge of  $R$  is not required and ZO with residual feedback converges from any

initial point  $x_0$ . When the parameter  $R$  is unknown, we can choose  $\eta = (2\sqrt{2}L_0\sqrt{dT}^{\frac{3}{4}})^{-1}$  and  $\delta = \sqrt{dT}^{-\frac{1}{4}}$  and obtain the regret bound  $R_T \leq \mathcal{O}(L_0R^2\sqrt{dT}^{\frac{3}{4}} + L_0^{-1}\sqrt{d}V_f^2T^{\frac{3}{4}})$ . The proof can also be found in Appendix B.4.

**Remark 3.10.** *We note that the complexity bound in Theorem 3.2 generally depends on the values of the Lipschitz parameters  $L_0$ ,  $L_1$  and the constant  $V_f^2$ . Specifically, choose  $\eta = R^{\frac{3}{2}}(2\sqrt{2}L_0\sqrt{dT}^{\frac{3}{4}})^{-1}$  and  $\delta = \sqrt{dRL_0^{-q}T^{-\frac{1}{4}}}$  with  $q > 0$  as a tuning parameter, and we obtain that  $R_T = \mathcal{O}((L_0 + L_0^{1-q} + L_0^{2q-1}V_f^2)\sqrt{dRT}^{\frac{3}{4}})$  when  $T \geq L_0^{2q}R^2$ . If  $L_0 < 1$ , we can choose  $q = 1$  to achieve the bound  $R_T = \mathcal{O}((L_0 + L_0V_f^2)\sqrt{dRT}^{\frac{3}{4}})$ . On the other hand, if  $L_0 \geq 1$ , we can choose  $q = 0$  to achieve the bound  $R_T = \mathcal{O}((L_0 + L_0^{-1}V_f^2)\sqrt{dRT}^{\frac{3}{4}})$ . We note that the dependence of the bounds in Theorems 3.11, 3.14 and 3.15 on  $L_0, L_1$  can also be optimized in a similar way by properly choosing  $\delta$ .*

Next, we present the regret of ZO with residual feedback when the objective function  $f_t$  is convex and smooth. As before, the sequence of objective functions can be either natural or adversarial, as per Definition 3.4.

**Theorem 3.11** (Regret for Convex Smooth  $f_t$ ). *Let Assumption 3.8 hold. Assume that  $f_t(x) \in C^{0,0} \cap C^{1,1}$  is convex with Lipschitz constant  $L_0$  and smoothness constant  $L_1$  over set  $\mathcal{X}_\delta$  for all  $t$ , and assume that  $\|x_0 - x^*\| \leq R$ . Run ZO with residual feedback for  $T > R^2$  iterations with  $\eta = R^{\frac{4}{3}}(2\sqrt{2}L_0d^{\frac{2}{3}}T^{\frac{2}{3}})^{-1}$  and  $\delta = d^{\frac{1}{3}}R^{\frac{1}{3}}T^{-\frac{1}{6}}$ . Then, we have that*

$$\begin{aligned} R_T \leq & \sqrt{2}L_0d^{\frac{2}{3}}R^{\frac{2}{3}}T^{\frac{2}{3}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]R^{\frac{4}{3}}}{2\sqrt{2}L_0d^{\frac{2}{3}}T^{\frac{2}{3}}} + 2L_1d^{\frac{2}{3}}R^{\frac{2}{3}}T^{\frac{2}{3}} \\ & + (\sqrt{2}L_0)^{-1}d^{\frac{2}{3}}R^{\frac{2}{3}}V_f^2T^{\frac{2}{3}} + 4\sqrt{2}L_0d^{\frac{4}{3}}R^{\frac{4}{3}}T^{\frac{1}{3}} \end{aligned}$$

*Asymptotically, we have that  $R_T = \mathcal{O}((L_0 + L_1 + L_0^{-1}V_f^2)(dRT)^{\frac{2}{3}})$ .*

The proof can be found in Appendix B.5 in the supplementary material. To the best of our knowledge, the best known regret for ZO with conventional one-point feedback for convex and smooth problems is of the order  $\mathcal{O}(L_1^{\frac{1}{3}}(dRBT)^{\frac{2}{3}})^{21}$ . Therefore, our regret bound is tighter if the function variation satisfies  $V_f^2 \leq \mathcal{O}(B^{\frac{2}{3}}L_1^{\frac{1}{3}}L_0)$ . Our numerical experiments in Section 3.5 show that ZO with residual feedback always outperforms ZO with conventional one-point feedback in practice.

### 3.3 ZO with Residual Feedback for Non-Convex 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  $\{f_t\}_{t=0,\dots,T-1}$  are non-convex. To the best of our knowledge, this is the first time that a one-point zeroth-order method is studied for non-convex online optimization. Throughout this section, we make the following assumption on the objective functions.

**Assumption 3.12.** *There exist  $W_T, \widetilde{W}_T > 0$  such that the conditions below hold for all  $t$ .*

1.  $\sum_{t=1}^T |f_{\delta,t}(x_t) - f_{\delta,t-1}(x_t)| \leq W_T$ , for all  $x_t \in \mathbb{R}^d$ .
2.  $\sum_{t=1}^T |f_t(x_t) - f_{t-1}(x_t)|^2 \leq \widetilde{W}_T$ , for all  $x_t \in \mathbb{R}^d$ .

The above two conditions in Assumption 3.12 measure the accumulated first-order and second-order function variations. Such variations are called the regularity measures in online non-stationary learning problems and are also assumed in<sup>40;41</sup>.

**Remark 3.13.** *The analysis in this section requires that the sequence of objective functions satisfy Assumption 3.12 and the Lipschitz conditions  $C^{0,0}$  or  $C^{0,0} \cap C^{1,1}$  over the whole domain. However, these assumptions are only needed for theoretical soundness. In practice, the estimator proposed in this section can still be used if these conditions are satisfied locally along the trajectory of the iterates of the online algorithm. In fact, in Section 2.4, we show that the proposed estimator can be used to solve practical learning problems with objective functions that are not necessarily globally Lipschitz.*

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 non-smooth optimization problems in<sup>15</sup>. In addition, similar to<sup>15</sup>, 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 3.2. 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 3.4.

**Theorem 3.14** (Nonconvex Lipschitz  $f_t$ ). *Let Assumptions 3.12 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  $\eta = \epsilon_f^{\frac{3}{2}}(2\sqrt{2}L_0^2d^{\frac{3}{2}}T^{\frac{1}{2}})^{-1}$  and  $\delta = \epsilon_f L_0^{-1}$ . Then, we have that*

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

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 B.6. Theorem 3.14 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 3.12 holds, then we have  $\widetilde{W}_T \leq \mathcal{O}(TV_f^2)$ , 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 non-convex and smooth and study 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 3.4.

**Theorem 3.15** (Nonconvex smooth  $f_t$ ). *Let Assumptions 3.12 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  $\eta = (2\sqrt{2}L_0d^{\frac{4}{3}}T^{\frac{1}{2}})^{-1}$  and  $\delta = (d^{\frac{1}{3}}T^{\frac{1}{4}})^{-1}$ . Then,*

$$\begin{aligned} R_g^T &\leq 4\sqrt{2}L_0(\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T)d^{\frac{4}{3}}T^{\frac{1}{2}} + 8\sqrt{2}L_1L_0d^{\frac{4}{3}}T^{\frac{1}{2}} \\ &\quad + \frac{\sqrt{2}L_1}{L_0}d^{\frac{4}{3}}\widetilde{W}_T + \frac{L_1\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{\sqrt{2}L_0d^{\frac{4}{3}}T^{\frac{1}{2}}} + 2L_1^2d^{\frac{4}{3}}T^{\frac{1}{2}}. \end{aligned}$$

Asymptotically,  $R_g^T = \mathcal{O}(d^{\frac{4}{3}}L_0W_T T^{\frac{1}{2}} + d^{\frac{4}{3}}L_1L_0^{-1}\widetilde{W}_T)$ .

The proof can be found in Appendix B.7. Theorem 3.15 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  $W_T, \widetilde{W}_T$  are weaker than those in the case of nonsmooth problems, as they do not rely on the small parameter  $\epsilon_f$ .

**Remark 3.16.** *The proposed residual-feedback gradient estimator can also be used with projected stochastic gradient descent algorithms when the online nonconvex learning problems have set constraints. The analysis in such a setting is included in Appendix B.8 in the supplementary material.*

## 3.4 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

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

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

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

where  $\xi_{t-1}$  and  $\xi_t$  are independent random samples that are sampled at consecutive iterations  $t-1$  and  $t$ , respectively. 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 (3.4) 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 3.17.** (*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 3.8 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 Appendix B.9.

**Lemma 3.18.** *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 3.7, and the only difference is the perturbation term. Consequently, ZO with stochastic residual feedback achieves almost the same regret bounds as those in Theorems 3.9 and 3.11, and one simply needs to replace  $V_f$  by  $V_{f,\xi}$ . For non-convex problems, we adopt the following assumption that generalizes Assumption 3.12.

**Assumption 3.19.** *There exists  $W_T, \widetilde{W}_{T,\xi} > 0$  such that the following two conditions hold for all  $t$ .*

1.  $\sum_{t=1}^T (f_{\delta,t}(x_t) - f_{\delta,t-1}(x_t)) \leq W_T$  for all  $x_t \in \mathbb{R}^d$ .
2.  $\sum_{t=1}^T \mathbb{E}[|F_t(x_{t-1}; \xi_t) - F_{t-1}(x_{t-1}; \xi_{t-1})|^2] \leq \widetilde{W}_{T,\xi}$  for all  $x_{t-1} \in \mathbb{R}^d$ , 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 3.14 and 3.15, we can obtain similar regret bounds for ZO with stochastic residual feedback (simply replace  $W_T, \widetilde{W}_T$  in Theorems 3.14 and 3.15 by  $W_T, \widetilde{W}_{T,\xi}$ , respectively).

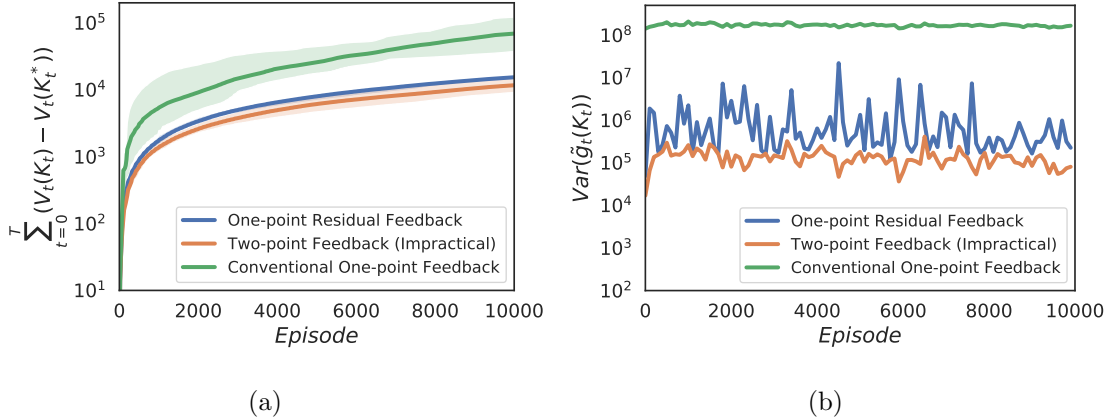
## 3.5 Numerical Experiments

In this section, we compare the performance of ZO with one-point, two-point and residual feedback in solving two non-stationary reinforcement learning problems, i.e., LQR control and resource allocation, in which either the reward or transition functions are varying over episodes. The details of the numerical experiment implementation can be found in Appendix B.1.

### 3.5.1 Nonstationary LQR Control

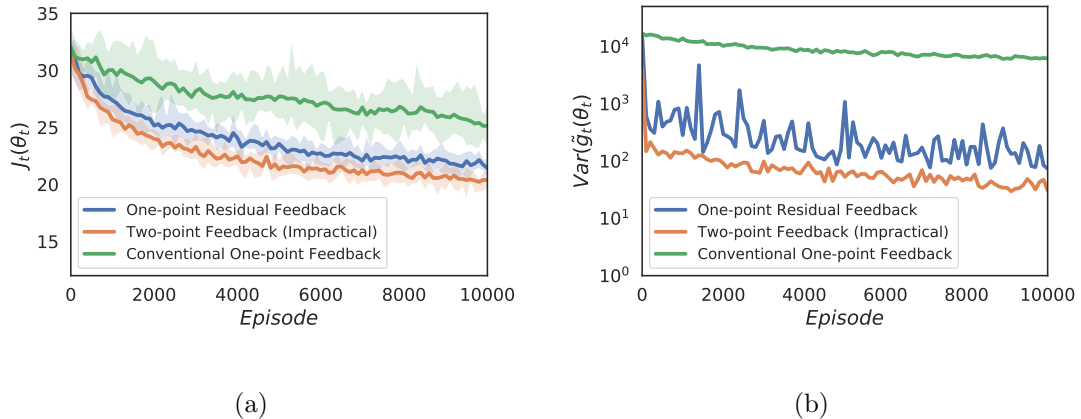
We consider an LQR problem with noisy system dynamics. The static version of this problem is considered in<sup>32;33</sup>. Specifically, consider a system whose state  $x_k \in \mathbb{R}^{n_x}$  at step  $k$  is subject to a transition function  $x_{k+1} = A_t x_k + B_t u_k + w_k$ , where  $u_k \in \mathbb{R}^{n_u}$  is the action at step  $k$ , and  $A_t \in \mathbb{R}^{n_x \times n_x}$  and  $B_t \in \mathbb{R}^{n_x \times n_u}$  are dynamical matrices in episode  $t$ . These matrices are unknown and changing over episodes. The vector  $w_k$  is the noise on the state transition. Specifically, the entries of the dynamical matrices  $A_0$  and  $B_0$  at episode 0 are randomly generated from a Gaussian distribution  $\mathcal{N}(0, 0.1^2)$ . Then, we generate the time-varying dynamical matrices as  $A_{t+1} = A_t + 0.01M_t$  and  $B_{t+1} = B_t + 0.01N_t$ , where  $M_t$  and  $N_t$  are random matrices whose entries are uniformly sampled from  $[0,1]$ . Moreover, consider a state feedback policy  $u_k = K_t x_k$ , where  $K_t \in \mathbb{R}^{n_u \times n_x}$  is the policy parameter that is fixed during episode  $t$ . We assume that there exists an optimal policy  $K_t^*$  so that the discounted accumulated cost function  $V_t(K) := \mathbb{E}[\sum_{k=0}^{H-1} \gamma^k (x_k^T Q x_k + u_k^T R u_k)]$  at episode  $t$  is minimized, where  $\gamma \leq 1$  is the discount factor and  $H$  is the horizon. The goal is to track the time-varying optimal policy parameter  $K_t^*$  so that  $V_t(K_t) - V_t(K_t^*)$  is small in every episode.

We apply the conventional one-point method in<sup>21</sup> and the proposed residual-feedback method (3.4) to solve the above non-stationary LQR problem. The performance of the two-point method in<sup>20</sup> is also presented to serve as a benchmark, although it is not possible to implement in practice for non-stationary problems. This is because the two-point



**Figure 3.1:** Comparative results of ZO with the proposed one-point residual feedback (3.1) (blue), the two-point oracle in<sup>20</sup> (orange) and the conventional one-point oracle in<sup>21</sup> (green) for online policy optimization in nonstationary LQR. Figure 1(a) presents the regrets  $\sum_{t=0}^T |V(K_t) - V(K^*)|$  achieved using the three different oracles and Figure 1(b) presents the variance of the gradient estimates 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.

method in<sup>20</sup> requires to evaluate value function  $V_t$  for two different policy functions at two consecutive episodes. However, evaluating the value function  $V_t$  for a given policy during episode  $t$  requires to collect samples by executing this policy. Then, during the subsequent episode  $t + 1$ , since the problem is non-stationary, the dynamic matrices change to  $A_{t+1}, B_{t+1}$  and so does the value function  $V_{t+1}$ . Therefore, it is not possible to evaluate the same value function  $V_t$  at two different episodes and, as a result, the two-point method in<sup>20</sup> is not applicable here. Each algorithm is run for 10 trials, and the stepsizes are optimized for each algorithm separately. The accumulated regrets  $\sum_{t=0}^{T-1} |V(K_t) - V(K^*)|$  of the three algorithms are presented in Figure 3.1(a). We observe that ZO with residual feedback achieves a much lower regret than the conventional one-point method and has a comparable performance to that of the two-point method. Moreover, we present in Figure 3.1(b) the estimated variance of the gradient estimates returned by these three oracles



**Figure 3.2:** Comparative results of ZO with the proposed one-point residual feedback (3.1) (blue), the two-point oracle in<sup>20</sup> (orange) and the conventional one-point oracle in<sup>21</sup> (green) for the non-stationary resource allocation problem. Figure 2(a) presents the varying cost  $J_t(\theta_t)$  achieved using three different oracles and Figure 2(b) presents the 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.

at the policy iterates over episodes. It can be seen that the variance of the gradient estimates returned by our proposed residual-feedback is close to that of the gradient estimates returned by the two-point feedback and is much smaller than that of the gradient estimates returned by the conventional one-point feedback. This observation validates our theoretical characterization of the second moment of the residual feedback gradient estimates.

### 3.5.2 Nonstationary Resource Allocation

We consider a multi-stage 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  $i$  makes its decisions according to a parameterized policy function  $\pi_{i,t}(o_i; \theta_{i,t}) : \mathcal{O}_i \rightarrow [0, 1]^{|M_i|}$ , 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 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.

In Figure 3.2(a), we present the cost  $J_t(\theta_t)$  achieved during each episode after 10 trials of ZO with residual-feedback, one-point, and two-point feedback which, as before, is impossible to use in practice for this non-stationary 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 non-stationary environment. In particular, ZO with both residual and two-point feedback performs much better than ZO with conventional one-point feedback. Figure 3.2(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.

## Chapter 4

# Zeroth-Order Distributed MARL using Residual Feedback under Partial Observations

In this chapter, we decentralize the proposed residual-feedback ZO gradient estimator to solve a distributed learning problem. Specifically, we consider the scenario where the evaluation of the objective function cannot be conducted by each local agent on its own. Instead, the global objective function is the summation of each agent’s local objective, and therefore, the agents need to evaluate the global objective function through collaboration. Due to limited communication capability, the global objective function cannot be exactly evaluated at each local agent, and a systematic error exists in their local evaluations. We will study how such objective function evaluation errors affect the convergence of the decentralized ZO method with residual-feedback oracle and how to alleviate the effects of such errors. Our analysis in this chapter will be conducted considering a MARL problem with partial observations. When each agent only observes partial states in the system, it cannot compute the policy gradients because the environment becomes non-Markovian. This naturally leads us to find solutions via ZO methods. To the best of our knowledge, our results presented here are the first works that provide theoretical convergence guarantee on finding approximately optimal policies for MARL problems with partial observations. The content in this chapter has been included in the paper<sup>3</sup>.

### 4.1 Preliminaries and Problem Formulation

Consider a multi-agent system consisting of  $N$  agents. The agent dynamics are governed by a Markov Decision Process (MDP) defined by a tuple  $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$ , where  $s_t = [s_{1,t}, s_{2,t}, \dots, s_{N,t}] \in \mathcal{S}$  and  $a_t = [a_{1,t}, a_{2,t}, \dots, a_{N,t}] \in \mathcal{A}$  denote the joint state and ac-

tion spaces of the  $N$  agents at time instant  $t$ . The reward vector  $r = [r_{1,t}, r_{2,t}, \dots, r_{N,t}] \in \mathcal{R}$  denotes the local rewards received by each agent at time  $t$ . The local reward  $r_{i,t}(s_t, a_t, w_t)$  is affected by the joint state and action of all the agents in the network, and is also subject to noise  $w_t$ . The transition function  $P(s_t, a_t, s_{t+1}) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1] \in \mathcal{P}$  denotes the probability of transitioning to state  $s_{t+1}$  when the agents take action  $a_t$  at state  $s_t$ . Let  $o_{i,t} \in \mathcal{O}_i$  represent the local observation received at agent  $i$  at time  $t$ , which contains partial entries of the joint state and action vectors,  $s_t$  and  $a_t$ . Agent  $i$  selects its action  $a_{i,t}$  based on the observation  $o_{i,t}$  using its local policy function  $\pi_i : \mathcal{O}_i \rightarrow \mathcal{A}_i$ . Let  $\pi$  denote the joint policy function which consists of local policy functions  $\pi_i$ . Then, the accumulated discounted reward received by agent  $i$  is defined as  $Q_i^\pi(s, a) = \mathbb{E}[\sum_{t=0}^T \gamma^t r_{i,t} | s_0 = s, a_0 = a]$  or  $V_i^\pi(s) = \mathbb{E}[\sum_{t=0}^T \gamma^t r_{i,t} | s_0 = s]$ , when the agents start from the state-action pair  $(s, a)$  or state  $s$ , follow the policy  $\pi$ , and apply a discount factor  $\gamma \leq 1$  to their future rewards <sup>1</sup>.

Our goal in this section is to find an optimal joint policy function  $\pi^*$  that solves the problem  $\max_{\pi} \frac{1}{N} \sum_{i=1}^N J_i(\pi)$ , where  $J_i(\pi) = \mathbb{E}_{(s_0, a_0) \sim \rho_0} [Q_i^\pi(s_0, a_0)]$  and  $\rho_0$  is a distribution that the initial state-action pair is sampled from. To do so, we assume that the local policy function  $\pi_i$  is parameterized as  $\pi_i(\theta_{i,k})$ , where  $\theta_{i,k} \in \mathbb{R}^{d_i}$  is the local policy parameter during episode  $k$ . Stacking these local policy parameters into the global policy parameter vector  $\theta \in \mathbb{R}^d$ , we can rewrite the problem we consider in this section as

$$\max_{\theta} J(\theta) := \frac{1}{N} \sum_{i=1}^N J_i(\theta). \quad (4.1)$$

Problem (4.1) can be solved using distributed Actor-Critic methods as in<sup>29;30</sup>. These methods require that all agents maintain local estimates of the global value function or the global policy function and that these local estimates are parameterized in the same way and depend on the global states and actions of all other agents. Therefore, they cannot be used for MARL with partial state and action information. Instead, in this section, we propose a new distributed zeroth-order policy optimization method that relies on the stochastic

---

<sup>1</sup>Although we consider a task with accumulated discounted rewards in this section, our proposed methods can be easily adapted to task considering averaged rewards.

gradient ascent update

$$\theta_{k+1} = \theta_k + \nabla J(\theta_k) \quad (4.2)$$

to determine optimal policy parameters  $\theta_k$  that solve Problem (4.1). The key idea that makes it possible to use partial state and action information in the update (4.2) is the use zeroth-order estimators  $\tilde{\nabla}J(\theta_k)$  of the true policy gradient  $\nabla J(\theta_k)$ . In this section, we adopt the one-point residual-feedback policy gradient estimator

$$\tilde{\nabla}J(\theta_k) = \frac{J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})}{\delta} u_k \quad (4.3)$$

As discussed in Chapter 2, same as the estimator (1.1), the estimator (4.3) only requires one policy evaluation at each iteration, but can use the history of policy evaluations to effectively reduce the variance of the current policy gradient estimate and, therefore, improve the learning rate.

According to<sup>1;15</sup>, both estimators (1.1) and (4.3) provide unbiased gradient estimates of a smoothed function  $J_\delta(\theta)$  at  $\theta_k$ , where  $J_\delta(\theta)$  is defined as  $J_\delta(\theta) := \mathbb{E}_u [J(\theta + \delta u)]$  and  $u$  is subject to a standard multivariate normal distribution. Therefore, updating the policy parameter  $\theta_k$  as in (4.2) using the gradient estimates (1.1) or (4.3) will in fact converge to a stationary point of the smoothed function  $J_\delta(\theta)$  rather than a stationary point of the value function  $J(\theta)$  that may be nonsmooth. To ensure that the stationary point found by this process is meaningful for the original MARL problem, we need to define appropriate optimality conditions that additionally ensure that  $J_\delta(\theta)$  and  $J(\theta)$  are close to each other. Specifically, we consider the following optimality criterion

$$\|\nabla J_\delta(\theta)\|^2 \leq \epsilon, \quad \text{and} \quad |J_\delta(\theta) - J(\theta)| \leq \epsilon_J, \quad (4.4)$$

which suggests that  $\theta$  is an  $\epsilon$ -stationary point of the smoothed value function  $J_\delta(\theta)$ , and that the smoothed value function  $J_\delta(\theta)$  is  $\epsilon_J$ -close to the true value function  $J(\theta)$ . To bound the distance between the smoothed function  $J_\delta(\theta)$  and the original value function  $J(\theta)$  in (4.4), we need the following assumption on the value function  $J(\theta)$ .

**Assumption 4.1.** *The function  $J(\theta)$  is Lipschitz with constant  $L_0$ , that is,  $|J(\theta_1) - J(\theta_2)| \leq L_0 \|\theta_1 - \theta_2\|$ , for all  $\theta_1, \theta_2 \in \mathbb{R}^d$ .*

Given Assumption 4.1, the following result for the smoothed value function  $J_\delta(\theta)$  holds.

**Lemma 4.2.** *(Gaussian Approximation<sup>15</sup>) Given Assumption 4.1, the smoothed function  $J_\delta(\theta)$  satisfies  $|J_\delta(\theta) - J(\theta)| \leq \delta L_0 \sqrt{d}$ , for all  $\theta \in \mathbb{R}^d$ .*

According to Lemma 4.2, to control the approximation accuracy of the smoothed function  $J_\delta(\theta)$ , the parameter  $\delta$  needs to be selected appropriately. The choice of this parameter will be discussed in Section 4.2. Note that although the random exploration direction  $u_k \sim \mathcal{N}(0, I_d)$  needed to evaluate the estimator (4.3) can be sampled in a fully decentralized way, the global value  $J(\theta_k + \delta u_k) = \frac{1}{N} \sum_{i=1}^N J_i(\theta_k + \delta u_k, \xi_k)$  is not accessible by the local agents. In the next section, we design a new algorithm that relies on partial state and action information only to produce a fully decentralized implementation of the estimator (4.3).

## 4.2 Algorithm Design and Theoretical Analysis

In this section, we propose a fully distributed zeroth-order policy optimization algorithm for MARL that employs the residual-feedback zeroth-order policy gradient estimator (4.3). Specifically, we first introduce a consensus step so that the global value  $J(\theta_k + \delta u_k, \xi_k)$  in the estimator (4.3) can be computed locally. Given a finite number of consensus iterations, the local estimates of  $J(\theta_k + \delta u_k, \xi_k)$  will be inexact and, therefore, the local policy gradient estimates will be biased. To control this bias, we then introduce a value tracking technique that reduces the bias at the current episode using the local estimates of  $J(\theta_k + \delta u_k, \xi_k)$  from previous episodes. Finally, we provide convergence results showing that the proposed distributed zeroth-order policy optimization method with constant stepsize converges to a neighborhood of the global optimal policy that is controlled by the number of consensus steps during each episode. Proofs of all theoretical results that follow can be found in the supplementary materials.

---

**Algorithm 1:** Distributed Residual-Feedback Zeroth-Order Policy Optimization
 

---

**Input:** Exploration parameter  $\delta$ , stepsize  $\alpha$ , consensus matrix  $W$ , number of consensus steps  $N_c$ , initial policy parameter  $\theta_0$ , discount ratio  $\gamma$ , maximum number of time steps run per episode  $t_{\max}$ , number of episodes  $K$ , and the logic variable **DoTracking**.

- 1 Set  $\mu_i^{-1}(N_c) = 0$  for all  $i = 1, 2, \dots, N$  ;
- 2 **for** *episode*  $k = 0, 1, 2, \dots, K$  **do**
- 3     For agents  $i = 1, 2, \dots, N$ , let agent  $i$  sample a random exploration direction  $u_{i,k}$  from the standard multivariate normal distribution ;
- 4     Let all agents implement their perturbed policy  $\pi_i(\theta_{i,k} + \delta u_{i,k})$  for  $t_{\max}$  time steps and construct unbiased estimates of their local accumulated rewards  $\{J_i(\theta_k + \delta u_k, \xi_k)\}$  ;
- 5     For all agent  $i = 1, 2, \dots, N$ ,
- 6     **if** **DoTracking** = **False** **or**  $k == 0$  **then**
- 7         | set  $\mu_i^k(0) = J_i(\theta_k + \delta u_k, \xi_k)$  ;
- 8     **else**
- 9         | set  $\mu_i^k(0) = \mu_i^{k-1}(N_c) + J_i(\theta_k + \delta u_k, \xi_k) - J_i(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})$  ;
- 10    **end**
- 11    **for**  $m = 0, 2, \dots, N_c - 1$  **do**
- 12         For agents  $i = 1, 2, \dots, N$ , let agent  $i$  send  $\mu_i^k(m)$  to its direct neighbors  $j \in \mathcal{N}_i$  and conduct local averaging by computing  $\mu_i^k(m+1) = \sum_{j \in \mathcal{N}_i} W_{ij} \mu_j^k(m)$  ;
- 13    **end**
- 14    For agents  $i = 1, 2, \dots, N$ , let agent  $i$  update its current policy parameter  $\theta_{i,k}$  by
 
$$\theta_{i,k+1} = \theta_{i,k} + \alpha \frac{\mu_i^k(N_c) - \mu_i^{k-1}(N_c)}{\delta} u_{i,k}. \quad (4.5)$$
- 15 **end**

---

Our proposed algorithm is summarized in Algorithm 1. In what follows, we also assume that the  $N$  agents form a communication graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V} = \{1, 2, \dots, N\}$  is the index set of agents and  $\mathcal{E}$  represents the set of edges. The edge  $(i, j) \in \mathcal{E}$  if agents  $i$  and  $j \in \mathcal{N}$  can directly send information to each other. Moreover, we define by  $W \in \mathbb{R}^{N \times N}$  a weight matrix associated with the graph  $\mathcal{G}$  such that the entry  $W_{ij} > 0$  when  $(i, j) \in \mathcal{E}$  and  $W_{ij} = 0$  otherwise. Note that the communication graph  $\mathcal{G}$  is independent of the coupling between agents in their state transition function  $\mathcal{P}(s_t, a_t, s_{t+1})$  and reward functions  $\{r_i(s_t, a_t)\}$  defined in Section 4.1.

## 4.2.1 Distributed Residual-Feedback Zeroth-Order Policy Optimization

In this section, we describe and analyze our proposed residual-feedback zeroth-order policy optimization algorithm in the absence of value tracking, i.e., when `DoTracking = False` in Algorithm 1 (line 6). Specifically, at the beginning of episode  $k$ , the agents randomly perturb their current policy parameters  $\theta_k$  using a random exploration direction  $u_k$  and conduct on-policy local policy evaluation to obtain an unbiased estimate of the local accumulated rewards  $\{J_i(\theta_k + \delta u_k)\}_{i=1,2,\dots,N}$  (lines 3-4). To conduct local policy evaluations, existing MARL methods<sup>29;30</sup> usually assume that the global state-action pairs  $(s_t, a_t)$  are available to all local agents. Under this assumption, it is possible to update the local Critic functions  $Q_i^\pi(s_t, a_t)$  in<sup>29;30</sup> to reduce the variance of policy evaluations and, therefore, the variance of the policy gradient estimates<sup>42</sup>. However, when the agents only have access to local observations  $o_{i,t}$  which contain partial entries of  $(s_t, a_t)$ , these methods cannot be used. Therefore, in this section, evaluate the local policies as  $J_i(\theta_k + \delta u_k, \xi_k) = r_i(1) + \gamma r_i(2) + \gamma^2 r_i(3) + \dots$ , same as in REINFORCE<sup>43</sup>. This policy evaluation method can be implemented in a fully decentralized way but is subject to large variance, which increases the variance of the zeroth-order policy gradient estimates and degrades the convergence speed of the algorithm. The residual-feedback policy gradient estimator (4.3) can effectively reduce this variance as we will discuss later. In what follows, we make the following assumption on the local policy value estimator.

**Assumption 4.3.** *For all agents, the local policy evaluation is unbiased and subject to bounded variance. That is,  $\mathbb{E}_\xi[J_i(\theta, \xi)] = J_i(\theta)$ , and  $\mathbb{E}[(J_i(\theta, \xi) - J_i(\theta))^2] \leq \sigma^2$  for  $i = 1, 2, \dots, N$ .*

After all agents compute unbiased local policy values  $J_i(\theta_{i,k}, \xi_k, \xi_k)$ , they conduct  $N_c$  rounds of local averaging on their local policy values  $\{J_i(\theta_k + \delta u_k, \xi_k)\}_{i=1,2,\dots,N}$  (lines 7, 11-13). As a result, they obtain inexact estimates  $\mu_i^k(N_c + 1)$  of the global accumulated rewards  $J(\theta_k + \delta u_k, \xi_k)$ . To bound this estimation error, we need the following two assumptions.

**Assumption 4.4.** *The undirected communication graph  $\mathcal{G}$  is connected and fixed for all episodes. In addition, the associated weight matrix  $W$  is doubly stochastic. That is,  $W\mathbf{1} = \mathbf{1}$  and  $W^T\mathbf{1} = \mathbf{1}$ .*

**Assumption 4.5.** *The local values  $J_i(\theta, \xi)$  are upper bounded by  $J_u$  and lower bounded by  $J_l$  for all  $i = 1, 2, \dots, N$  and all policy parameters  $\theta$ .*

Assumption 4.5 can be easily satisfied when the rewards are bounded and the discount factor  $\gamma < 1$ . When the rewards are unbounded, Assumption 4.5 holds when all policy iterates are stabilizing policies<sup>32</sup>. Let  $\mu^k(m) = [\mu_1^k(m), \dots, \mu_N^k(m)]^T$ . Then, we can show the following lemma.

**Lemma 4.6.** *Given Assumptions 4.4 and 4.5, we have that  $\|\mu^k(N_c) - J(\theta_k + \delta u_k, \xi_k)\mathbf{1}\| \leq \rho_W^{N_c} \sqrt{N}(J_u - J_l)$ , where  $\rho_W = \|W - \frac{1}{N}\mathbf{1}\mathbf{1}^T\| < 1$ .*

Lemma 4.6 shows that the bias in the local estimate  $\mu^k(N_c)$  can be controlled by choosing a large enough  $N_c$ , when the local policy values are upper and lower bounded by  $J_u$  and  $J_l$ . Using the estimates  $\mu^k(N_c)$ , the agents can then construct the decentralized policy gradients (4.5) and update their policy parameters  $\theta_{i,k}$  (line 14). This completes episode  $k$ . The decentralized residual-feedback estimator (4.5) can reduce the variance of the policy gradient estimates, since the value estimate of the last policy iterate  $\mu_i^{k-1}(N_c)$  can provide a baseline to compare  $\mu_i^k(N_c)$  to. Effectively, the value estimate of the last policy iterate has an analogous variance reduction effect to the state value  $V^\pi(s)$  that is used as a baseline for the action value  $Q^\pi(s, a)$  in Actor-Critic methods<sup>42</sup>.<sup>3</sup> Next, we show how to select  $N_c$  so that the optimality criterion (4.4) is satisfied.

**Theorem 4.7. (Learning Rate of Algorithm 1 without Value Tracking)** *Let Assumptions 4.1, 4.3, 4.4 and 4.5 hold and define  $\delta = \frac{\epsilon_J}{\sqrt{d}L_0}$ ,  $\alpha = \frac{\epsilon_J^{1.5}}{4d^{1.5}L_0^2\sqrt{K}}$ , and  $N_c \geq$*

---

<sup>3</sup>Note that the fact that the value of past policy iterates can also be used as a baseline to reduce the variance of policy gradient estimates may be of interest in its own right in the development of policy gradient methods.

$\log(\frac{\sqrt{\epsilon}\epsilon_J}{\sqrt{2}d^{1.5}L_0(J_u-J_l)})/\log(\rho_W)$ . Then, running Algorithm 1 with `DoTracking = False`, we have that  $\frac{1}{K} \sum_{k=1}^{K-1} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] \leq \mathcal{O}(d^{1.5}\epsilon_J^{-1.5}K^{-0.5}) + \frac{\epsilon}{2}$ .

As shown in Theorem 4.7, Algorithm 1 converges to a neighborhood of the global optimal policy whose size at steady-state is equal to  $\frac{\epsilon}{2}$  and can be controlled by choosing the number of consensus steps  $N_c$ . Moreover, the number of consensus steps  $N_c$  depends not only on the user-specified accuracy level  $\epsilon$  and  $\epsilon_J$ , but also on the range of the policy bounds  $J_u$  and  $J_l$ . This is because the consensus iteration at each episode is independent of those at previous episodes. Therefore, to select  $N_c$  to control the estimation bias  $|\mu_i^k(N_c) - J(\theta_k + \delta u_k, \xi_k)|$ , we need to select the term  $J_u - J_l$  in the definition of  $N_c$  in Theorem 4.7 as the difference between the initial estimates  $\mu_i^k(0) = J_i(\theta_k + \delta u_k, \xi_k) \in [J_l, J_u]$  that have the maximum and minimum value.

## 4.2.2 Distributed Residual-Feedback Zeroth-Order Policy Optimization with Value Tracking

As discussed before, the estimation bias  $|\mu_i^k(N_c) - J(\theta_k + \delta u_k, \xi_k)|$  at episode  $k$  can be reduced using local policy estimates from previous episodes. Specifically, rather than resetting  $\mu_i^k(0) = J_i(\theta_k + \delta u_k, \xi_k)$  in line 7 of Algorithm 1, we update it using the estimate  $\mu_i^{k-1}(N_c)$  from the last episode as  $\mu_i^k(0) = \mu_i^{k-1}(N_c) + J_i(\theta_k + \delta u_k, \xi_k) - J_i(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})$ . Then, we run  $N_c$  consensus iterations on  $\mu_i^k(0)$  as before. Let  $\bar{\mu}^k(m) = \frac{1}{N} \sum_{i=1}^N \mu_i^k(m)$ . The following lemma shows that the value tracking updates preserve the global information  $J(\theta_k + \delta u_k, \xi_k)$ .

**Lemma 4.8.** *Let Assumption 4.4 hold. Then, running Algorithm 1 with `DoTracking = True`, we have that  $\bar{\mu}^k(m) = J(\theta_k + \delta u_k, \xi_k) = \frac{1}{N} \sum_{i=1}^N J_i(\theta_{i,k}, \xi_k)$ , for all  $m = 1, 2, \dots, N_c$  and all  $k$ .*

Lemma 4.8 implies that the local estimation bias  $|\mu_i^k(N_c) - J(\theta_k + \delta u_k, \xi_k)|$  is equal to the consensus error  $|\mu_i^k(N_c) - \bar{\mu}^k(N_c)|$ . Using value tracking, the bias at episode  $k$  can be

controlled by the consensus steps of past episodes. This is formally shown in the following lemma.

**Lemma 4.9.** *Let Assumptions 4.1, 4.3, 4.4 hold and define  $E_\mu^k = \|\mu_k(N_c) - \bar{\mu}_k(N_c)\|$ . Then, running Algorithm 1 with `DoTracking = True`, we have that*

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2] \leq & \left( 2\mathbb{E}[(E_\mu^{k-1})^2] + 32dL_0^2 \frac{\alpha^2}{\delta^2} \mathbb{E}[(E_\mu^{k-1})^2 \|u_{k-1}\|^2] + 32d^2 L_0^2 \frac{\alpha^2}{\delta^2} \mathbb{E}[(E_\mu^{k-2})^2] \right) \rho_W^{2N_c} \\ & + 16NL_0^2 \alpha^2 \mathbb{E}[\|\tilde{\nabla} J(\theta_{k-1})\|^2] \rho_W^{2N_c} + 32NdL_0^2 \delta^2 \rho_W^{2N_c} + 16N\sigma^2 \rho_W^{2N_c}. \end{aligned} \quad (4.6)$$

Compared to Lemma 4.6, the proposed value tracking technique makes it possible to bound the consensus error at episode  $k$  with the consensus errors from episodes  $k-1$  and  $k-2$ . Furthermore, at each episode, this error is perturbed by the second order momentum of the policy gradient estimate (4.3) which can be controlled by choosing a small stepsize  $\alpha$  and a large number  $N_c$ . To see the benefit of this result, when the consensus errors from episodes  $k-1$  and  $k-2$  are small, value tracking needs fewer consensus iterations to achieve small consensus error at episode  $k$ . This is in contrast to the case without value tracking, where  $N_c$  is selected regardless of previous consensus errors. The following result shows convergence of Algorithm 1 using value tracking.

**Theorem 4.10. (Learning Rate of Algorithm 1 with Value Tracking)** *Let Assumptions 4.1, 4.3, 4.4 hold and define  $\delta = \frac{\epsilon_J}{\sqrt{dL_0}}$ ,  $\alpha = \frac{\epsilon_J^{1.5}}{4d^{1.5}L_0^2\sqrt{K}}$ , and*

$$N_c \geq \max\left(\log(\rho_W)^{-1} \log\left(\frac{1}{2\sqrt{2}}\right), \log(\rho_W)^{-1} \log\left(\sqrt{\frac{\epsilon}{2G^2\epsilon_J + 64(d+4)^2 dL_0^2 + 32d^3 L_0^2 \sigma^2 / \epsilon_J^2}}\right)\right),$$

where  $G^2 = \max\left(\mathbb{E}[\|\tilde{\nabla} J(\theta_0)\|^2], \frac{2\epsilon_J\epsilon}{dK} + 32L_0^2(d+4)^2 + 16d^2 L_0^2 \frac{\sigma^2}{\epsilon_J^2}\right)$ . Then, running Algorithm 1 with `DoTracking = True`, we have that

$$\frac{1}{K} \sum_{k=1}^{K-1} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] \leq \mathcal{O}(d^{1.5}\epsilon_J^{-1.5}K^{-0.5}) + \frac{\epsilon}{2}.$$

In Theorem 4.10, the constant  $G^2$  represents the uniform bound on  $\mathbb{E}[\|\tilde{\nabla} J(\theta_k)\|^2]$  for all  $k = 1, 2, \dots, K$ . Moreover, from the bounds on  $N_c$  in Theorems 4.7 and 4.10, when  $\epsilon$  and  $\epsilon_J$  are close to 0, we obtain that  $N_c \sim \mathcal{O}(\log(\frac{\sqrt{\epsilon}\epsilon_J}{d^{1.5}\sigma}) / \log(\rho_W))$  in Theorem 4.10 and

$N_c \sim \mathcal{O}(\log(\frac{\sqrt{\epsilon}\epsilon_J}{d^{1.5}(J_u - J_l)})/\log(\rho_W))$  in Theorem 1. This suggests that the choice of  $N_c$  in Theorem (4.10) depends on the variance of function evaluation  $\sigma^2$ , while in Theorem 4.7 the choice of  $N_c$  depends on the range of value functions  $[J_l, J_u]$ . In practice, the variance of function evaluation  $\sigma^2$  can be much smaller than the range of its value  $[J_l, J_u]$ . Therefore, Algorithm 1 with value tracking requires fewer consensus steps per episode than without value tracking.

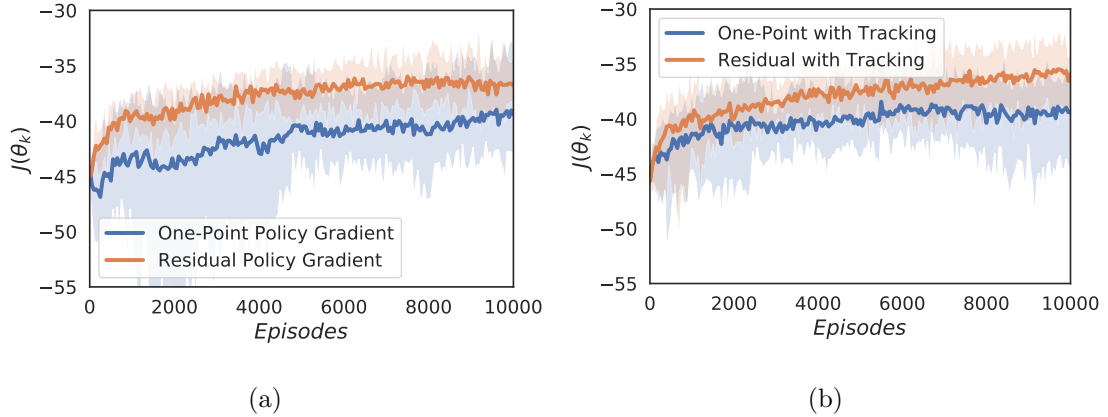
**Remark 4.11.** *The proposed value-tracking technique can also be combined with the existing distributed one-point policy gradient estimator<sup>8</sup> to reduce the variance of its gradient estimates. To see this, note that the global value function  $J(\theta_k + \delta u_k, \xi_k)$  used in the one-point estimator (1.1) can be replaced by the local estimate of the value  $J(\theta_k + \delta u_k, \xi_k)$ , i.e.,  $\mu_i^k(N_c)$ . Then, we obtain the following distributed one-point policy gradient estimator with value tracking:*

$$\begin{aligned} \tilde{\nabla}_{\theta_{i,k}} J(\theta_k) &\approx \frac{\mu_i^k(N_c)}{\delta} u_{i,k} \\ &= \frac{\sum_{j \in \mathcal{N}_i} [W^{N_c}]_{ij} (\mu_j^{k-1}(N_c) + J_j(\theta_k + \delta u_k, \xi_k) - J_j(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1}))}{\delta} u_{i,k}. \end{aligned} \quad (4.7)$$

We observe that the estimator (4.7) has the same structure as the distributed residual-feedback policy gradient estimator without value tracking (4.5) except for an additional noise term  $\frac{\sum_{j \in \mathcal{N}_i} [W^{N_c}]_{ij} \mu_j^{k-1}(N_c)}{\delta} u_{i,k}$ . Therefore, the variance of the estimator (4.7) is reduced through a similar mechanism as that of the distributed residual-feedback policy gradient estimator without value tracking. As a result, the learning performance is improved compared to that of the existing distributed one-point policy gradient estimator<sup>8</sup>, as we will demonstrate in the next section.

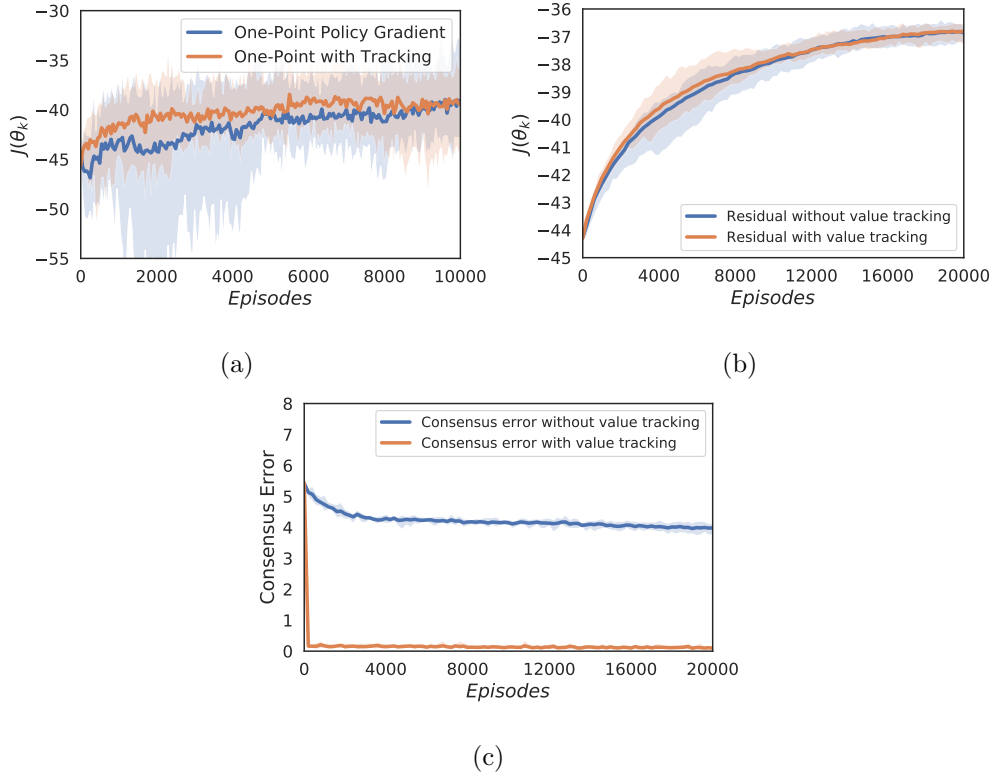
### 4.3 Numerical Experiments

In this section, we illustrate our proposed MARL algorithm on stochastic multi-agent multi-stage decision making problems. Specifically, we conduct an ablation study to demonstrate the benefits of applying the decentralized residual-feedback zeroth-order policy gradient



**Figure 4.1:** Distributed zeroth-order policy optimization with the proposed residual-feedback estimator (4.3) (orange) versus the one-point estimator (1.1) (blue). In each case, Algorithm 1 is run 10 times. (a): Comparative results without value tracking. (b): Comparative results with value tracking.

estimate (4.5) and the value tracking technique separately. We consider 16 agents that are located on a  $4 \times 4$  grid. Agent  $i$  stores resources and receives a local demand in the amount of  $m_i(t)$  and  $d_i(t)$  at time  $t$ . In the meantime, agent  $i$  also shares resources with its neighbors in the grid in the amount of  $[\dots, a_{ij}(t), \dots]_{j \in \mathcal{N}_i}$ , where  $a_{ij}(t) \in [0, 1]$  denotes the fraction of resources agent  $i$  sends to its neighbor  $j$  at time  $t$ . The local resources and demands at agent  $i$  are defined as  $m_i(t+1) = m_i(t) - \sum_{j \in \mathcal{N}_i} a_{ij}(t)m_i(t) + \sum_{j \in \mathcal{N}_i} a_{ji}(t)m_j(t) - d_i(t)$  and  $d_i(t) = A_i \sin(\omega_i t + \phi_i) + w_{i,t}$ , where  $w_{i,t}$  is the noise in the demand. At time  $t$ , agent  $i$  receives a local reward  $r_i(t)$ , such that  $r_i(t) = 0$  when  $m_i(t) \geq 0$  and  $r_i(t) = -m_i(t)^2$  when  $m_i(t) < 0$ . We consider a partial observation scenario, where agent  $i$  can only observe its local resources and demands, that is,  $o_i(t) = [m_i(t), d_i(t)]^T$ . Agent  $i$  determines its actions  $\{a_{ij}(t)\}$  using its local policy function  $\pi_i(o_i(t))$ . Specifically, we have that  $a_{ij} = \exp(z_{ij}) / \sum_j \exp(z_{ij})$ , where  $z_{ij} = \sum_{p=1}^9 \|o_i - c_p\|^2 \theta_{ij}(p)$  and  $c_p$  is the  $p$ -th feature parameter. We consider episodes of length  $T = 30$ , and select the discount factor as  $\gamma = 0.75$ . The communication graph is assumed to be a chain graph. Moreover, we select the number of consensus steps  $N_c = 1$ , and show that Algorithm 1 can achieve policy improvement even in this challenging



**Figure 4.2:** Distributed zeroth-order policy optimization with value tracking (orange) versus without value tracking (blue). In each case, Algorithm 1 is run 10 times. (a): Comparative results for the one-point estimator (1.1). (b): Comparative results for the proposed residual-feedback estimator (4.3). (c) Maximum absolute consensus errors  $\max_i |\mu_i^k(N_c) - J(\theta_k + \delta u_k, \xi_k)|$  over episodes.

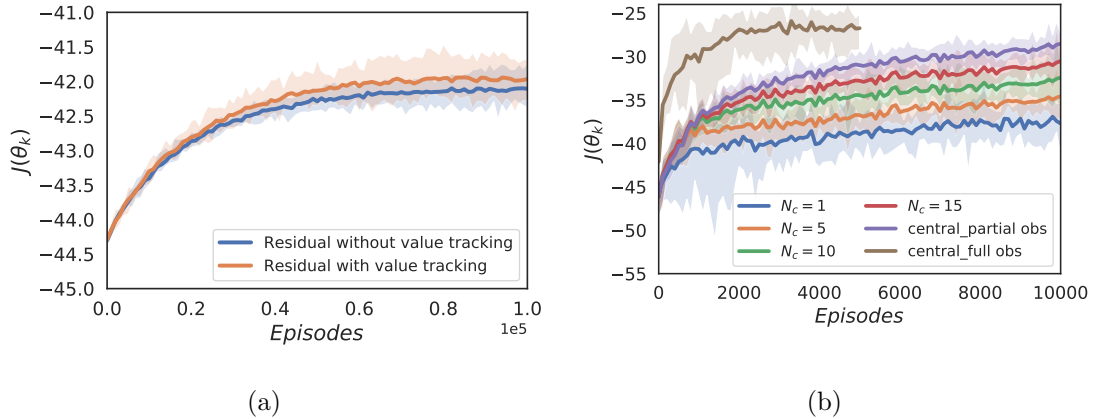
scenario. The stepsizes are selected so that the convergence speed is optimized.

First, we compare the performance of Algorithm 1 using the decentralized policy gradients (4.3) and (1.1), without value tracking. The learning progress is presented in Figure 4.1(a). We observe that the decentralized residual-feedback policy gradient estimator has less variance than the existing one-point policy gradient estimator and, therefore, improves faster and finds a better policy in the end of the learning. The same effect is observed in Figure 4.1(b), when both estimators are implemented with value tracking. This suggests that the residual-feedback zeroth-order policy gradient estimator is superior to the one-

point policy gradient estimator for decentralized policy optimization problems.

Next, we demonstrate the merit of using value tracking. Specifically, we first run Algorithm 1 with the decentralized one-point policy gradient estimator (1.1), with and without value tracking. The difference in the performance is shown in Figure 4.2(a). We observe that using value tracking results in less variance and also achieves better policies. This is because the decentralized one-point policy gradient estimator with value tracking (4.7) has the same variance reduction effect on the policy gradient estimates as the residual feedback estimator 4.3, as we have discussed in Remark 4.11. Figure 4.2(b) shows the results of using the residual-feedback policy gradient estimator (4.3) with and without value tracking. We observe that Algorithm 1 with value tracking performs slightly better in the mean than without value tracking. This is because value tracking can track the value of the global objective function better, as shown in Figure 4.2(c) where the maximum consensus error  $\max_i |\mu_i^k(N_c) - J(\theta_k + \delta u_k, \xi_k)|$  at each episode  $k$  is presented. The improvement achieved by value tracking is not very significant in Figure 4.2(b) because the underlying communication graph respects the coupling relationship among agents. Specifically, we say that the communication graph respects the coupling relationship between agents if the action of every agent  $i$  that can directly communicate with an agent  $j$  also directly affects the reward and transition function of that agent  $j$ . In this case, the rewards received from an agent’s local neighbors can approximate this agent’s contribution to the global reward well even without tracking the information from other distant neighbors in the graph.

To further demonstrate the advantage of combining the decentralized residual-feedback gradient estimator with value tracking, we consider a challenging scenario where the communication graph does not respect the coupling relationship among agents as described above. The performance of Algorithm 1 using the residual-feedback estimator (4.3) with and without value tracking is presented in Figure 4.3(a). In this case, using rewards from the local neighbors does not approximate well the local agent’s contribution to the global reward. Therefore, value tracking can help obtain a better estimate of the global reward information. As a result, the decentralized residual-feedback policy gradient estimator with



**Figure 4.3:** (a) Algorithm 1 with value tracking (orange) versus without value tracking (blue) under the communication graph which does not respect the coupling relationship among agents. (b) Comparative results for Algorithm 1 under different number of consensus steps  $N_c$  and the centralized algorithms with partial and full observations. In each case, Algorithm 1 is run 10 times.

value tracking outperforms the one without value tracking, as shown in Figure 4.3(a). We note that, while the numerical results presented above show that the performance of Algorithm 1 is affected by the structure of the communication graph among the agents, in this section we focus on the sampling complexity of Algorithm 1 provided the communication graph is connected. A further investigation of the relationship between the communication topology and the learning rate of Algorithm 1 is an interesting problem that is left for our future research.

Finally, we demonstrate the effect of the number of consensus steps  $N_c$  on the performance of Algorithm 1 by comparing to a centralized algorithm that uses the gradient estimator (4.3) with both full and partial observations. Specifically, in the centralized algorithm, the value of the global objective function  $J(\theta_k + \delta u_k, \xi_k)$  is directly provided to each local agent at each episode, and the local agents' policy functions receive all agents' states as inputs when full observations are assumed and only receive the neighboring agents' states as inputs when partial observations are assumed. As shown in Figure 4.3(b), as the

number of consensus steps  $N_c$  increases, the performance of Algorithm 1 approaches that of the centralized algorithm with partial observations. And the performance of the centralized algorithm with partial observations slightly underperforms that of the centralized algorithm with full observations. This is because policy functions learned using partial observations constitute a subset of those that can be learned using full observations.

# Chapter 5

## Asynchronous Distributed Optimization using Residual Feedback

In this chapter, we study the proposed residual-feedback oracle in another setting of distributed learning problems. Different from the one studied in Chapter 4, each agent has access to the global objective function values. However, they do not have access to a global clock, so they perturb their local decisions, obtain a value feedback and update their local decision variables in an asynchronous fashion. This leads to an asynchronous ZO gradient estimator that is in a completely different form from the ZO estimators we studied in the last few chapters. In practice, synchronizing all agents to conduct ZO gradient estimator is expensive in a large-scale network. Therefore, studying the asynchronous ZO gradient estimator is of great interest. Unlike the existing two-point methods, we will demonstrate the asynchronous ZO estimator based on the proposed residual-feedback scheme is still unbiased. We study its convergence under such asynchronous setting and demonstrate its effectiveness using a distributed learning example. The contents in this chapter are also presented in the paper<sup>4</sup>.

### 5.1 Preliminaries and Problem Formulation

Consider a multi-agent system consisting of  $N$  agents that collaboratively solve the unconstrained optimization problem

$$\min_x f(x), \tag{5.1}$$

where the cost function  $f$  is non-convex and smooth,  $x := (x^1, \dots, x^N) \in \mathbb{R}^n$  is the joint decision vector, and  $x^i \in \mathbb{R}^{n_i}$  is the local decision vector of agent  $i \in \{1, \dots, N\}$ . We first make the following assumptions on the cost function  $f$ .

**Assumption 5.1.** *The cost function  $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$  is bounded below by  $f^*$ . It is  $L_0$ -Lipschitz and  $L_1$ -smooth, i.e.,*

$$|f(x) - f(y)| \leq L_0 \|x - y\|, \text{ and } \|\nabla f(x) - \nabla f(y)\| \leq L_1 \|x - y\|,$$

for all  $x, y \in \mathbb{R}^n$ .

As shown in<sup>15</sup>,  $L_1$ -smoothness is equivalent to the condition;

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{1}{2} L_1 \|x - y\|^2, \quad (5.2)$$

for all  $x, y \in \mathbb{R}^n$ . For each agent  $i$ , define the local smoothing function:

$$f_{\mu_i}(x) = \frac{1}{\kappa_i} \int f(x + \mu_i u_i) e^{-\frac{1}{2} \|u_i\|^2} du_i^i, \quad (5.3)$$

where  $\kappa_i = \int e^{-\frac{1}{2} \|u_i\|^2} du_i^i$ . The random sampling vector  $u_i = \{u_i^1, \dots, u_i^N\} \in \mathbb{R}^n$  is a vector of all zeros except for the entry  $u_i^i$  that is sampled from  $\mathcal{N}(0, I_{n_i})$ .<sup>1</sup> Note that  $f_{\mu_i}$  preserves all the Lipschitz conditions of  $f$  as proved in<sup>15</sup>. Specifically, we have the following lemma.

**Lemma 5.2.** *Under Assumption 5.1, we have that, for all agents  $i$ ,  $f_{\mu_i}(x) : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L_0$ -Lipschitz and  $L_1$ -smooth.*

As a result, (5.2) also holds for  $f_{\mu_i}$ , which allows us to bound the approximation errors of  $f_{\mu_i}$  and  $\nabla_i f_{\mu_i}$  with respect to (w.r.t.) to  $f$  and  $\nabla_i f$ , as shown in Lemma 2 below that is adopted from<sup>15</sup>. In Lemma 2 and the following analysis, we denote by  $\nabla f(x) \in \mathbb{R}^n$  the gradient of  $f(x)$ . Moreover, we define by  $\nabla_i f(x) \in \mathbb{R}^n$  the projection of  $\nabla f(x)$  onto the index  $i$  by setting the entries of  $\nabla f(x)$  not equal to  $i$  to be 0.

**Lemma 5.3.** *Under Assumption 5.1, the cost function  $f$  and its corresponding smoothed function  $f_{\mu_i}$  satisfy  $\forall i \in \{1, \dots, N\}$*

$$|f_{\mu_i}(x) - f(x)| \leq \frac{\mu_i^2}{2} L_1 n_i, \quad (5.4)$$

$$\|\nabla_i f_{\mu_i}(x) - \nabla_i f(x)\| \leq \frac{\mu_i}{2} L_1 (n_i + 3)^{3/2}. \quad (5.5)$$

---

<sup>1</sup>In the following analysis, we drop the agent index  $i$  in  $u_i$  for simplicity.

Next, we define the model that the agents use to asynchronously update their decision variables.

**Definition 5.4** (Asynchrony Model). *At each time step, one agent is independently and randomly selected according to a fixed distribution  $P = [p_1, \dots, p_N]$ . The selected agent  $i$  can query the value of the cost function<sup>2</sup> once and update its local decision variable, while the decisions of the other agents  $\{x^j\}_{j \neq i}$  are fixed. The agents only communicate with a central entity that has access to the global cost function but not with each other.*

**Remark 5.5.** *Definition 5.4 can be satisfied when all agents make queries and update their decisions according to their local clock without any central coordination. Specifically, let the time interval between each agent's consecutive queries be called a waiting time. Then, if each local waiting time is random and subject to an exponential distribution, according to Chapter 2.1 in<sup>44</sup>, Definition 5.4 will be satisfied.*

## 5.2 Algorithm Design and Theoretical Analysis

In this section, we present the proposed asynchronous zeroth-order distributed optimization algorithm and analyze its convergence rate. To do so, we first propose an asynchronous zeroth-order gradient estimator based on the centralized residual feedback estimator (2.1).

For every agent  $i$ , at time step  $k$ , this estimator takes the form

$$G_{\mu_i}(x_k) = \frac{f(x_k + \mu_i u_k) - f(x_{k-M} + \mu_i u_{k-M})}{\mu_i} u_k, \quad (5.6)$$

where  $k-M$  is a random index denoting the iteration when agent  $i$  conducted its most recent update. This index takes values on a global time scale. The random sampling vector  $u_k$  is as defined in (5.3). Note that  $G_{\mu_i}$  is different from the centralized zeroth-order gradient estimator (2.1), where  $u_k$  is a perturbation along the full decision vector  $x$ . Here,  $G_{\mu_i}$  estimates the gradient by perturbing the function  $f$  along a random direction restricted

---

<sup>2</sup>Here, we assume that each agent receives noiseless feedback  $f(x)$ . The proposed method can be extended to noisy feedback with bounded variance.

to agent  $i$ 's block of the full decision vector  $x$  and uses the previous query to reduce the variance. Indeed,  $G_{\mu_i}$  provides an unbiased gradient estimate of the corresponding smoothed function  $f_{\mu_i}$  restricted to agent  $i$ 's block, as shown in the following lemma.

**Lemma 5.6.** *For each agent  $i$ , we have that*

$$\mathbb{E}[G_{\mu_i}(x_k)] = \nabla_i f_{\mu_i}(x_k).$$

*Proof.* Taking the expectation of both sides of (5.6), we obtain that

$$\begin{aligned} \mathbb{E}[G_{\mu_i}(x_k)] &= \mathbb{E}\left[\frac{f(x_k + \mu_i u_k) - f(x_{k-M} + \mu_i u_{k-M})}{\mu_i} u_k\right] \\ &= \mathbb{E}\left[\frac{f(x + \mu_i u_k)}{\mu_i} u_k\right] = \nabla_i f_{\mu_i}(x_k), \end{aligned}$$

where the second equality follows from the fact that  $x_{k-M}$  and  $u_{k-M}$  are independent from  $u_k$ . The last equality follows from the definitions of  $f_{\mu_i}$  and  $\nabla_i f_{\mu_i}$ .  $\square$

**Remark 5.7.** *Note that both  $G_{\mu_i}(x_k)$  and  $\nabla_i f_{\mu_i}$  are vectors in  $\mathbb{R}^n$  with entries equal to zero at blocks other than  $i$ .*

Using the local gradient estimate  $G_{\mu_i}(x_k)$ , we can define the update rule for every agent  $i$  as

$$x_{k+1} = x_k - \alpha_i G_{\mu_i}(x_k), \tag{5.7}$$

where  $\alpha_i$  is the step size. The proposed asynchronous zeroth-order distributed optimization algorithm with residual feedback is described in Algorithm 2<sup>3</sup>.

Without loss of generality, we assume that decision variables  $x^i \in \mathbb{R}^{\bar{n}}$  of all agents  $i$  have the same dimensions, and that the step sizes and smoothing parameters of all agents are also the same, i.e.,  $\alpha_i = \alpha$  and  $\mu_i = \mu$ . To analyze the convergence of Algorithm 2, we need to

---

<sup>3</sup>Note that, we can also extend (1.2) for asynchronous problems and design the algorithm thereof. The lemmas and theorems proved in this section can be easily adapted to this case as well. In Section 5.3, we will compare these two gradient estimators empirically. However, the extension of (1.2) is non-trivial. It can be verified that Lemma 5.6 does not hold for the extension of (1.2). We leave it as future work.

---

**Algorithm 2:** Asynchronous Zeroth-Order Residual Feedback
 

---

**Input:** sampling rate  $p_i$  with  $\sum_{i=1}^N p_i = 1$ , decision variable  $x_0^i$ , smoothing parameter  $\mu_i$  and step size  $\alpha_i$  for all agents  $i$ . Set the iteration counter  $t = 0$  and let  $T$  be the maximum number of iterations.

```

1 for  $t \leq T$  do
2   | sample an index  $i_t$  according to  $\mathbb{P}(i_t = i) = p_i$ ;
3   | sample  $u^i \sim \mathcal{N}(0, I_{n_i})$ ;
4   | query the function value  $f(x + \mu_i u)$ ;
5   | compute  $G_{\mu_i}$  according to (5.6);
6   | update local decision  $x^i \leftarrow x^i - \alpha_i G_{\mu_i}(x^i)$ ;
7 end

```

---

bound the second moment of the proposed gradient estimator  $G_{\mu_i}(x_k)$ . However, under the asynchronous framework considered in this section, there can be a random number of agents updating their local decision variables between the two queries made by agent  $i$  at time steps  $k$  and  $k - M$  in (5.6). These updates by other agents introduce additional variance into the estimator (5.6) compared to the variance of the centralized estimator analyzed in<sup>1</sup>. Next, we analyze the effect of the asynchronous updates on the second moment of the zeroth-order gradient estimator. To the best of our knowledge, this is the first time that a bound on the second moment of a zeroth-order gradient estimator is provided for asynchronous problems. An additional contribution of this work, is that the proof technique presented below can be extended to obtain similar results for the two-point gradient estimator (1.2).

**Lemma 5.8.** *Let Assumptions 5.1 hold under the Asynchrony Model, and define by  $\mathbb{E}[\|G_{\bar{\mu}}(x_k)\|^2] := \mathbb{E}_{i_k}[\mathbb{E}_{u_{[k]}, i_{[k-1]}}[\|G_{\mu_i}(x_k)\|^2 | i_k = i]]$ , where  $u_{[k]} = (u_1, \dots, u_k)$  and  $i_{[k]} = (i_1, \dots, i_k)$ . Then, running the asynchronous Algorithm 2, we have that  $\mathbb{E}[\|G_{\bar{\mu}}(x_k)\|^2]$  satisfies*

$$\begin{aligned} \mathbb{E}[\|G_{\bar{\mu}}(x_k)\|^2] &\leq \frac{2\bar{n}L_0^2\alpha^2k}{\mu^2} \sum_{m=0}^{k-1} (1 - p_{\min})^m \mathbb{E}[\|G_{\bar{\mu}}(x_{k-m-1})\|^2] \\ &\quad + 4L_0^2((4 + \bar{n})^2 + \bar{n}^2), \end{aligned} \tag{5.8}$$

where  $p_{\min} = \min_i p_i$ , and the expectations are taken w.r.t. the sequence of random exploration directions  $\{u_k\}$  and the sequence of random indices of activated agents  $\{i_k\}$ .

*Proof.* Suppose that at time step  $k$ , agent  $i$  is selected. Taking the second moment of  $G_{\mu_i}$

and using equation (5.6), we we have<sup>4</sup>

$$\mathbb{E}_{u_{[k]}, i_{[k-1]}} \left[ \|G_{\mu_i}(x_k)\|^2 | i_k = i \right] \leq \mathbb{E} \left[ \frac{(f(x_k + \mu_i u_k) - f(x_{k-M} + \mu_i u_{k-M}))^2 \|u_k\|^2}{\mu_i^2} \right] \quad (5.9)$$

Notice that

$$\begin{aligned} (f(x_k + \mu_i u_k) - f(x_{k-M} + \mu_i u_{k-M}))^2 &\leq 2 \underbrace{(f(x_k + \mu_i u_k) - f(x_{k-M} + \mu_i u_k))^2}_a \\ &\quad + 2 \underbrace{(f(x_{k-M} + \mu_i u_k) - f(x_{k-M} + \mu_i u_{k-M}))^2}_b. \end{aligned} \quad (5.10)$$

Substituting (5.10) into (5.9), we obtain that

$$\begin{aligned} \mathbb{E}_{u_{[k]}, i_{[k-1]}} \left[ \|G_{\mu_i}(x_k)\|^2 | i_k = i \right] &\leq \mathbb{E} \left[ \frac{2a + 2b}{\mu_i^2} \|u_k\|^2 \right] \\ &\leq \mathbb{E} \left[ \frac{2L_0^2 \|x_k - x_{k-M}\|^2}{\mu_i^2} \|u_k\|^2 \right] + \mathbb{E} \left[ \frac{2b}{\mu_i^2} \|u_k\|^2 \right] \\ &\leq \frac{2L_0^2 \bar{n}}{\mu_i^2} \mathbb{E} \left[ \|x_k - x_{k-M}\|^2 \right] + \mathbb{E} \left[ \frac{2b}{\mu_i^2} \|u_k\|^2 \right], \end{aligned} \quad (5.11)$$

where the second inequality holds due to Lipschitzness of  $f$  and the last inequality holds since  $x_k - x_{k-M}$  is independent from  $u_k$  and  $\mathbb{E}[\|u_k\|^2] = \bar{n}$ . We first bound the second term in the right-hand-side of (5.11). Specifically, we have that

$$\begin{aligned} \mathbb{E} \left[ \frac{2b}{\mu_i^2} \|u_k\|^2 \right] &\leq \mathbb{E} \left[ 2L_0^2 \|u_k - u_{k-M}\|^2 \|u_k\|^2 \right] \\ &\leq \mathbb{E} \left[ 4L_0^2 \left( \|u_k\|^2 + \|u_{k-M}\|^2 \right) \|u_k\|^2 \right] \\ &\leq \mathbb{E} \left[ 4L_0^2 \|u_k\|^4 \right] + \mathbb{E} \left[ 4L_0^2 \|u_{k-M}\|^2 \right] \mathbb{E} \left[ \|u_k\|^2 \right] \\ &\leq 4L_0^2 \left( (4 + \bar{n})^2 + \bar{n}^2 \right), \end{aligned} \quad (5.12)$$

where the first inequality holds due to Lipschitzness of  $f$ , the third inequality holds since  $u_{k-M}$  is independent from  $u_k$  and the last inequality follows from Lemma 1 in<sup>15</sup>.

Next, we bound the first term in the right-hand-side of (5.11) containing the second moment of  $x_k - x_{k-M}$ . Given that agent  $i$  updates at time step  $k$ , we can partition the

---

<sup>4</sup>To simplify the notation, when it is clear from the context, we drop the subscript of the expectation and the conditional event, e.g.,  $\mathbb{E}[\|G_{\mu_i}(x_k)\|^2] := \mathbb{E}_{u_{[k]}, i_{[k-1]}}[\|G_{\mu_i}(x_k)\|^2 | i_k = i]$ .

sequence of all past updates into  $k$  events  $A_m^i = \{M = m\}$ , where  $A_m^i$  represents all sequences of updates such that the most recent update by agent  $i$  is at global time step  $k - m$ . In particular,  $A_k^i$  indicates that agent  $i$  has not been updated before and  $k$  is the first time that this agent gets updated. It is easy to see that the sets  $\{A_m^i\}_{m=1}^k$  are disjoint and contain all possible sequences of updates by the team of agents. Using the definition of these events, we can rewrite the conditional expectation of  $\|x_k - x_{k-M}\|^2$  as

$$\begin{aligned} Y_i &:= \mathbb{E}_{u_{[k]}, i_{[k-1]}} \left[ \|x_k - x_{k-M}\|^2 \mid i_k = i \right] \\ &= \sum_{m=1}^k \mathbb{E}_{u_{[k]}} \left[ \|x_k - x_{k-m}\|^2 \mid A_m^i, i_k = i \right] \mathbb{P}(A_m^i). \end{aligned} \quad (5.13)$$

Equation (5.13) can be rewritten as

$$\begin{aligned} Y_i &= \sum_{m=1}^k \mathbb{E} \left[ \|x_k - x_{k-m}\|^2 \mid A_m^i \right] \mathbb{P}(A_m^i) \\ &= \sum_{m=1}^k \mathbb{E} \left[ \left\| \sum_{l=0}^{m-1} (x_{k-l} - x_{k-l-1}) \right\|^2 \mid A_m^i \right] \mathbb{P}(A_m^i) \\ &\leq \sum_{m=1}^k \mathbb{E} \left[ m \sum_{l=0}^{m-1} \|(x_{k-l} - x_{k-l-1})\|^2 \mid A_m^i \right] \mathbb{P}(A_m^i), \end{aligned}$$

where the last inequality holds due to the fact that  $\left(\sum_{m=1}^k a_m\right)^2 \leq k \sum_{m=1}^k a_m^2$ . We first collect all the terms containing  $\|x_k - x_{k-1}\|^2$ , which we denote by  $\{Y_i : \|x_k - x_{k-1}\|^2\}$ .

Then, we have that

$$\begin{aligned} \{Y_i : \|x_k - x_{k-1}\|^2\} &= \sum_{m=1}^k m \mathbb{E} \left[ \|x_k - x_{k-1}\|^2 \mid A_m^i \right] \mathbb{P}(A_m^i) \\ &\leq k \sum_{m=1}^k \mathbb{E} \left[ \|x_k - x_{k-1}\|^2 \mid A_m^i \right] \mathbb{P}(A_m^i) \\ &= k \mathbb{E} \left[ \|x_k - x_{k-1}\|^2 \right], \end{aligned} \quad (5.14)$$

where the last equation holds by the definition of conditional expectation. Next we collect all the terms containing  $\|x_{k-s} - x_{k-s-1}\|^2$  for all  $s \in \{1, \dots, k-1\}$ . Specifically, we have

that

$$\begin{aligned} \{Y_i : \|x_{k-s} - x_{k-s-1}\|^2\} &= \sum_{m=s+1}^k m \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 | A_m^i \right] \mathbb{P}(A_m^i) \\ &\leq k \sum_{m=s+1}^k \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 | A_m^i \right] \mathbb{P}(A_m^i). \end{aligned} \quad (5.15)$$

We claim that the right hand side of (5.15) satisfies the following equation

$$k \sum_{m=s+1}^k \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 | A_m^i \right] \mathbb{P}(A_m^i) = k \mathbb{P}(A_{1:s}^{ic}) \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 \right], \quad (5.16)$$

where  $A_{1:s}^{ic} = \cup_{m=s+1}^k A_m^i$ . To see this, we first observe that

$$\begin{aligned} \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 \right] &= \sum_{m=1}^k \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 | A_m^i \right] \mathbb{P}(A_m^i) \\ &= \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 | A_{1:s}^i \right] \mathbb{P}(A_{1:s}^i) + \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 | A_{1:s}^{ic} \right] \mathbb{P}(A_{1:s}^{ic}), \end{aligned} \quad (5.17)$$

where  $A_{1:s}^i = \cup_{m=1}^s A_m^i$  and  $A_{1:s}^{ic}$  is the complement of  $A_{1:s}^i$ . The second equality follows from the property of conditional expectation of disjoint events. Note that  $\mathbb{E}[\|x_{k-s} - x_{k-s-1}\|^2 | A_{1:s}^i]$  and  $\mathbb{E}[\|x_{k-s} - x_{k-s-1}\|^2 | A_{1:s}^{ic}]$  are equal. Specifically, event  $A_{1:s}^i$  and event  $A_{1:s}^{ic}$  only differ after time step  $k-s$ , where  $A_{1:s}^i$  contains all sequences of updates where  $i$  update after  $k-s$  and  $A_{1:s}^{ic}$  contains all sequences of updates where  $i$  does not update after  $k-s$ . Since both events  $A_{1:s}^i$  and  $A_{1:s}^{ic}$  do not affect the agents' updates before time step  $k-s$ , we have that

$$\mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 \right] = \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 | A_{1:s}^{ic} \right].$$

Combining the above equality with (5.15), we have that

$$\begin{aligned} k \sum_{m=s+1}^k \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 | A_m^i \right] \mathbb{P}(A_m^i) &= k \mathbb{P}(A_{1:s}^{ic}) \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 | A_{1:s}^{ic} \right] \\ &= k \mathbb{P}(A_{1:s}^{ic}) \mathbb{E} \left[ \|x_{k-s} - x_{k-s-1}\|^2 \right], \end{aligned}$$

which completes the proof of (5.16). Then, we can bound  $Y_i$  in (5.13) by combining (5.14) and (5.16) and have that

$$\mathbb{E} \left[ \|x_k - x_{k-M}\|^2 \right] \leq k \mathbb{E} \left[ \|x_k - x_{k-1}\|^2 \right] + k \sum_{m=1}^{k-1} \mathbb{P}(A_{1:m}^{ic}) \mathbb{E} \left[ \|x_{k-m} - x_{k-m-1}\|^2 \right]. \quad (5.18)$$

By definition, we have  $\mathbb{P}(A_m^i) = p_i(1 - p_i)^{m-1}$  for  $1 \leq m < k$ , and  $\mathbb{P}(A_k^i) = (1 - p_i)^k$  for  $m = k$ , where  $p_i$  is the probability of agent  $i$  being sampled at each time step. As a result,  $\mathbb{P}(A_{1:m}^{ic}) = (1 - p_i)^m$ . Substituting these probabilities into (5.18), we have that

$$\mathbb{E} \left[ \|x_k - x_{k-M}\|^2 \right] \leq k \sum_{m=0}^{k-1} (1 - p_{\min})^m \mathbb{E} \left[ \|x_{k-m} - x_{k-m-1}\|^2 \right], \quad (5.19)$$

where  $p_{\min} = \min_i p_i$ . Substituting (5.19) and (5.12) into (5.11), we get that

$$\begin{aligned} \mathbb{E} \left[ \|G_{\mu_i}(x_k)\|^2 \right] &\leq \frac{2L_0^2 \bar{n} k}{\mu_i^2} \sum_{m=0}^{k-1} (1 - p_{\min})^m \mathbb{E} \left[ \|x_{k-m} - x_{k-m-1}\|^2 \right] \\ &\quad + 4L_0^2 \left( (4 + \bar{n})^2 + \bar{n}^2 \right). \end{aligned} \quad (5.20)$$

Recall that all expectations from the beginning of the proof are taken conditioned on the event  $\{i_k = i\}$ . Now, taking the expectation w.r.t.  $i_k$  on both sides of (5.20), and substituting the step size  $\alpha$  and smoothing parameter  $\mu$  into (5.20), we get that

$$\mathbb{E} \left[ \|G_{\bar{\mu}}(x_k)\|^2 \right] \leq \frac{2\bar{n}L_0^2 \alpha^2 k}{\mu^2} \sum_{m=0}^{k-1} (1 - p_{\min})^m \mathbb{E} \left[ \|G_{\bar{\mu}}(x_{k-m-1})\|^2 \right] + 4L_0^2 \left( (4 + \bar{n})^2 + \bar{n}^2 \right),$$

where  $\mathbb{E}[\|x_{k-m} - x_{k-m-1}\|^2] = \alpha^2 \mathbb{E}[\|G_{\bar{\mu}}(x_{k-m-1})\|^2]$  according to the update rule (5.7) and the definition of  $\mathbb{E}[\|G_{\bar{\mu}}\|^2]$  as in Lemma 5.8. The proof is complete.  $\square$

Lemma 5.8 indicates that the second moment of the zeroth-order gradient estimate at time step  $k$  is related to the second moments of all previous gradient estimates. Specifically, the effect of the second moment of the past gradient estimates on the current estimate diminishes geometrically over time. Next, using Lemma 5.8, we present a bound on the accumulated second moments of the residual-feedback gradient estimates from  $k = 0$  to  $T - 1$ , which we will later use to prove our main theorem.

**Lemma 5.9.** *Let Assumptions 5.1 hold under the Asynchrony Model. Then, running the asynchronous updates Algorithm 2, we have that*

$$\begin{aligned} \sum_{k=0}^{T-1} \mathbb{E} \left[ \|G_{\bar{\mu}}(x_k)\|^2 \right] &\leq \frac{1 - \beta}{1 - (\gamma + \beta)} \mathbb{E} \left[ \|G_{\bar{\mu}}(x_0)\|^2 \right] \\ &\quad + (T - 1) \frac{1 - \beta}{1 - (\gamma + \beta)} M - \frac{\gamma}{(1 - (\gamma + \beta))^2} M, \end{aligned}$$

where  $\gamma = \frac{2\bar{n}L_0^2\alpha^2(T-1)}{\mu^2}$ ,  $\beta = 1 - p_{\min}$ ,  $M = 4L_0^2((4 + \bar{n})^2 + \bar{n}^2)$  and provided with  $0 < \gamma + \beta < 1$ .

The proof follows from Lemma D.1 in the Appendix.

**Theorem 5.10.** *Let Assumptions 5.1 hold under the Asynchrony Model. Moreover, run the asynchronous algorithm Algorithm 2 for  $T$  iterations and let  $\tilde{x}$  be uniformly randomly selected from  $T$  iterations. Then, selecting the step size  $\alpha = \frac{\sqrt{p_{\min}}}{T^{\frac{3}{2}}}$  and the smoothing parameter  $\mu = \frac{2L_0\sqrt{\bar{n}}}{T^{\frac{1}{6}}}$ , we have  $\mathbb{E} \left[ \|\nabla f(\tilde{x})\|^2 \right] \leq \mathcal{O}(\bar{n}^3 T^{-\frac{1}{3}})$ .*

*Proof.* Substituting  $x_{k+1}$  and  $x_k$  in the version of (5.2) for the smoothed function  $f_{\mu_i}$ , we obtain that

$$\begin{aligned} f_{\mu_i}(x_{k+1}) &\leq f_{\mu_i}(x_k) + \langle \nabla f_{\mu_i}(x_k), x_{k+1} - x_k \rangle + \frac{L_1}{2} \|x_{k+1} - x_k\|^2 \\ &= f_{\mu_i}(x_k) - \alpha \langle \nabla_i f_{\mu_i}(x_k), \Delta_{i,k} \rangle - \alpha \|\nabla_i f_{\mu_i}(x_k)\|^2 + \frac{L_1\alpha^2}{2} \|G_{\mu_i}(x_k)\|^2, \end{aligned} \quad (5.21)$$

where  $\Delta_{i,k} := G_{\mu_i}(x_k) - \nabla_i f_{\mu_i}(x_k)$ . The first equality follows by (5.7) and the fact that  $\langle \nabla f_{\mu_i}(x_k), x_{k+1} - x_k \rangle = \langle \nabla_i f_{\mu_i}(x_k), x_{k+1} - x_k \rangle$ , which holds since  $x_{k+1}$  and  $x_k$  only differ at block  $i$ . Taking expectation w.r.t.  $u_{[k]}$  and  $i_{[k-1]}$  on both sides of (5.21) conditioned on the event  $\{i_k = i\}$ , we get that

$$\mathbb{E} \left[ \|\nabla_i f_{\mu_i}(x_k)\|^2 \right] \leq \frac{\mathbb{E}[f_{\mu_i}(x_k)] - \mathbb{E}[f_{\mu_i}(x_{k+1})]}{\alpha} + \frac{L_1\alpha}{2} \mathbb{E} \left[ \|G_{\mu_i}(x_k)\|^2 \right], \quad (5.22)$$

where the inner-product term  $\langle \nabla_i f_{\mu_i}(x_k), \Delta_{i,k} \rangle$  disappears since  $\mathbb{E}[\Delta_{i,k}] = 0$ . According to Lemma 5.3 and using the fact that  $(a + b)^2 \leq 2a^2 + 2b^2$ , we have

$$\|\nabla_i f(x)\|^2 \leq 2 \|\nabla_i f_{\mu_i}(x)\|^2 + \mu_i^2 L_1^2 (n_i + 3)^3. \quad (5.23)$$

Combining (5.22) and (5.23) and, we obtain that

$$\begin{aligned} \frac{1}{2} \mathbb{E} \left[ \|\nabla_i f(x_k)\|^2 \right] &\leq \frac{\mathbb{E}[f_{\mu_i}(x_k)] - \mathbb{E}[f_{\mu_i}(x_{k+1})]}{\alpha} \\ &\quad + \frac{L_1\alpha}{2} \mathbb{E} \left[ \|G_{\mu_i}(x_k)\|^2 \right] + \frac{1}{2} \mu^2 L_1^2 (\bar{n} + 3)^3, \end{aligned} \quad (5.24)$$

where the last term follows by substituting the common smoothing parameter  $\mu$  and agents' dimension  $\bar{n}$ . Taking expectation on both sides of (5.24) w.r.t.  $i_k$ , we have that

$$\begin{aligned} \frac{1}{2} \mathbb{E}_{i_k} \left[ \mathbb{E}_{u_{[k]}, i_{[k-1]}} \left[ \|\nabla_i f(x_k)\|^2 \mid i_k = i \right] \right] &\leq \frac{\mathbb{E}[f_{\bar{\mu}}(x_k)] - \mathbb{E}[f_{\bar{\mu}}(x_{k+1})]}{\alpha} \\ &+ \frac{L_1 \alpha}{2} \mathbb{E} \left[ \|G_{\bar{\mu}}(x_k)\|^2 \right] + \frac{1}{2} \mu^2 L_1^2 (\bar{n} + 3)^3, \end{aligned} \quad (5.25)$$

where  $\mathbb{E}[f_{\bar{\mu}}(x_k)] := \mathbb{E}_{i_k} [\mathbb{E}_{u_{[k]}, i_{[k-1]}} [f_{\mu_{i_k}}(x_k) \mid i_k = i]]$  and  $\mathbb{E}[f_{\bar{\mu}}(x_{k+1})] := \mathbb{E}_{i_k} [\mathbb{E}_{u_{[k]}, i_{[k-1]}} [f_{\mu_{i_k}}(x_{k+1}) \mid i_k = i]]$ .  $\mathbb{E}[\|G_{\bar{\mu}}(x_k)\|^2]$  follows from the definition in Lemma 5.8. Next, we show that the left hand side of (5.25) satisfies

$$\mathbb{E}_{i_k} [\mathbb{E}_{u_{[k]}, i_{[k-1]}} [\|\nabla_i f(x_k)\|^2 \mid i_k = i]] \geq p_{\min} \mathbb{E}_{u_{[k]}, i_{[k]}} \|\nabla f(x_k)\|^2. \quad (5.26)$$

To see this, by definitions of the projected gradient as in Section 5.1, since  $\nabla_i f(x_k)$  is only nonzero at block  $i$ , we have  $\|\nabla f(x_k)\|^2 = \sum_{i=1}^N \|\nabla_i f(x_k)\|^2$ . Therefore, we can further get that

$$\mathbb{E}_{u_{[k]}, i_{[k]}} \|\nabla f(x_k)\|^2 = \sum_{i=1}^N \mathbb{E}_{u_{[k]}, i_{[k]}} \|\nabla_i f(x_k)\|^2 = \sum_{i=1}^N \mathbb{E}_{u_{[k]}, i_{[k-1]}} [\|\nabla_i f(x_k)\|^2 \mid i_k = i], \quad (5.27)$$

where the second equality holds since  $x_k$  is independent from  $i_k$ . Therefore, according to (31), to show the inequality (30), it is sufficient to show  $\mathbb{E}_{i_k} [\mathbb{E}_{u_{[k]}, i_{[k-1]}} [\|\nabla_i f(x_k)\|^2 \mid i_k = i]] \geq p_{\min} \sum_{i=1}^N \mathbb{E}_{u_{[k]}, i_{[k-1]}} [\|\nabla_i f(x_k)\|^2 \mid i_k = i]$ . This is simple to prove because  $\mathbb{E}_{i_k} [\mathbb{E}_{u_{[k]}, i_{[k-1]}} [\|\nabla_i f(x_k)\|^2 \mid i_k = i]] = \sum_i p_i \mathbb{E}_{u_{[k]}, i_{[k-1]}} [\|\nabla_i f(x_k)\|^2 \mid i_k = i]$ . Therefore, inequality (30) is true. Substituting (5.26) into (5.25) and then summing (5.25) from  $k = 0$  to  $T - 1$ , we have that

$$\begin{aligned} \frac{p_{\min}}{2} \sum_{k=0}^{T-1} \mathbb{E} \left[ \|\nabla f(x_k)\|^2 \right] &\leq \frac{\mathbb{E}[f_{\bar{\mu}}(x_0)] - \mathbb{E}[f_{\bar{\mu}}(x_T)]}{\alpha} \\ &+ \sum_{k=0}^{T-1} \frac{L_1 \alpha}{2} \mathbb{E} \left[ \|G_{\bar{\mu}}(x_k)\|^2 \right] + \frac{\mu^2}{2} L_1^2 (\bar{n} + 3)^3 T. \end{aligned} \quad (5.28)$$

Applying Lemma 5.9 to (5.28), we get that

$$\begin{aligned} \frac{p_{\min}}{2} \sum_{k=0}^{T-1} \mathbb{E} \left[ \|\nabla f(x_k)\|^2 \right] &\leq \frac{\mathbb{E}[f_{\bar{\mu}}(x_0)] - \mathbb{E}[f_{\bar{\mu}}(x_T)]}{\alpha} + \frac{L_1 \alpha}{2} (T-1) \frac{1-\beta}{1-(\gamma+\beta)} M \\ &+ \frac{L_1 \alpha}{2} \frac{1-\beta}{1-(\gamma+\beta)} \mathbb{E} \left[ \|G_{\bar{\mu}}(x_0)\|^2 \right] + \frac{\mu^2}{2} L_1^2 (\bar{n} + 3)^3 T \\ &- \frac{L_1 \alpha}{2} \frac{\gamma}{(1-(\gamma+\beta))^2} M, \end{aligned} \quad (5.29)$$

where  $\gamma, \beta$  and  $M$  are as defined in Lemma 5.9. Selecting  $\mu = \frac{2L_0}{T^{\frac{1}{6}}}$  and  $\alpha = \frac{\sqrt{p_{\min}}}{\sqrt{\bar{n}}T^{\frac{2}{3}}}$ , we have  $\gamma \leq \frac{p_{\min}}{2}$  and  $1 - (\gamma + \beta) \geq \frac{p_{\min}}{2}$ . Substituting these values into (5.29) and omitting the negative term, we obtain that

$$\begin{aligned} \frac{p_{\min}}{2} \sum_{k=0}^{T-1} \mathbb{E} \left[ \|\nabla f(x_k)\|^2 \right] &\leq \frac{\mathbb{E}[f_{\bar{\mu}}(x_0)] - f_{\bar{\mu}}^*}{\sqrt{p_{\min}}} \sqrt{\bar{n}}T^{\frac{2}{3}} + L_1 \frac{\sqrt{p_{\min}}}{\sqrt{\bar{n}}T^{\frac{2}{3}}} \mathbb{E} \left[ \|G_{\bar{\mu}}(x_0)\|^2 \right] \\ &\quad + L_1 \frac{\sqrt{p_{\min}}}{\sqrt{\bar{n}}} T^{\frac{1}{3}} M + 2L_0^2 \mu^2 L_1^2 (\bar{n} + 3)^3 T^{\frac{2}{3}}, \end{aligned} \quad (5.30)$$

where  $f_{\bar{\mu}}^*$  is a lower bound on  $\mathbb{E}[f_{\bar{\mu}}(x)]$ . The existence of such lower bound is due to (5.4), the definition of  $\mathbb{E}[f_{\bar{\mu}}(x)]$  and Assumption 5.1. The result in Theorem 5.10 follows by dividing both sides of the above inequality by  $T$ .  $\square$

The convergence rate shown above has the same order as that of applying the residual-feedback gradient estimator (2.1) to optimize a stochastic objective function as shown in<sup>1</sup>. This is because of this asynchronous scenario, the updates conducted by the other agents between two queries of a given agent introduce noise in the function evaluations from the perspective of this given agent. Furthermore, the bound on the non-stationarity of the solution in (5.30) increases as  $p_{\min}$  becomes smaller. In practice, it means that the convergence of Algorithm 2 slows down if one of the agents is activated less frequently than others.

## 5.3 Numerical Experiments

In this section, we demonstrate the effectiveness of the proposed asynchronous distributed zeroth-order optimization algorithm on a distributed feature learning example common in Internet of Things (IoT) applications. All the experiments are conducted using Python 3.8.5 on a 2017 iMac with 4.2GHz Quad-Core Intel Core i7 and 32GB 2400MHz DDR4.

Specifically, we consider the biomarker learning example described in<sup>45</sup>, where a network of health monitoring edge devices collect heterogeneous raw input data  $\{D_{i,j}\}_{i=1:N}$ , e.g., different types of biosignals. Then, each device encodes its local raw data  $D_{i,j}$  into a biomarker  $d_{i,j}$  via a feature extraction function  $\phi(D_{i,j}; x_i)$ , e.g., a neural network with

weights  $x_i$ , and sends it to a third-party entity that uses the collected biomarkers as predictors to learn a disease diagnosis for user  $j$ . The goal of the edge device  $i$  is to learn a better feature extraction function  $\phi(\cdot; x_i)$  to help the third-party entity to make better predictions. In practice, the prediction process at the third-party entity can be complicated and hard to model using an explicit function. Moreover, it may need to remain confidential. As a result, the edge device cannot obtain gradient information from this third-party entity. In the meantime, it is unreasonable to expect that the edge devices can update their feature extraction models synchronously or that they know the other edge devices' feature extraction models and parameters. Therefore, this problem presents an ideal case for the asynchronous distributed zeroth-order method proposed in this chapter.

For simulation purposes, in this section, we assume that the third-party entity uses the logistic regression model

$$P(y_j; d_j) = 1/(1 + \exp(-y_j W^T d_j)), \quad (5.31)$$

where  $j$  represents the data point index,  $y_j = \{1, -1\}$  and  $d_j$  denote the label and predictors for data point  $j$ , and  $W_T$  is a fixed classifier parameter. Specifically, let  $d_j = [d_{1,j}, \dots, d_{N,j}]^T$  represent the concatenated biomarker vector. The agents aim to collaboratively minimize the following loss function

$$f(\{x_i\}_{i=1:N}) = -\frac{1}{J} \sum_{j=1}^J \log(P(y_j; d_j(\cdot; \{x_i\}_{i=1:N}))), \quad (5.32)$$

where  $J$  is the total number of data points.

Next, we apply the proposed Algorithm 2 to this distributed feature learning problem and compare its performance to an asynchronous extension of the centralized two-point gradient estimate (1.2) defined by

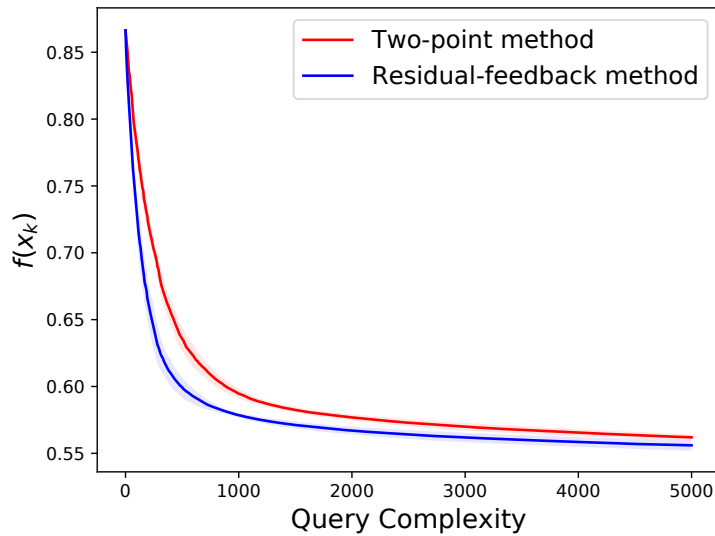
$$G_{\mu_i}(x_k) = \frac{f(x_k + \mu_i u_k) - f(x_{k-M})}{\mu_i} u_k, \quad (5.33)$$

where  $x_{k-M}$  is the most recent decision point agent  $i$  queries at iteration  $k - M$ .

Note that the convergence of the stochastic gradient descent update (5.7) using the asynchronous two-point estimator (5.33) has not been studied yet. We compare our proposed algorithm to the one with (5.33) simply to demonstrate the efficacy of our proposed

approach. Specifically, we consider a network of 5 agents who collaboratively deal with  $J = 20$  data samples. The feature extraction model  $\phi(\cdot; x_i)$  at agent  $i$  is a single layer neural network with the input  $D_{i,j} \in \mathbb{R}^{10}$  and a single output  $d_{i,j} \in \mathbb{R}$ . The activation function of the neural network is the sigmoid function. The weight of the neural network at agent  $i$  is denoted as  $x_i$ , which is initialized by sampling from a standard Gaussian distribution. We apply both the asynchronous residual-feedback gradient estimator (5.6) and the asynchronous two-point gradient estimator (5.33) to solve this problem. Specifically, for both gradient estimators, we run 10 trials. In addition, the smoothing parameter is  $\mu = 0.1$ , and the stepsizes  $\alpha$  for gradient estimators (5.6) and (5.33) are selected as 0.5 and 0.5, respectively, so that they both achieve their fastest convergence speed during 10 trials of experiments. At each iteration, each agent has equal probability to be activated.

The comparative performance results of using the two zeroth-order gradient estimators (5.6) and (5.33) are presented in Figure 5.1. We observe that during 10 trials, asynchronous learning with the residual-feedback gradient estimator (5.6) converges faster than asynchronous learning with the two-point gradient estimator (5.33). This is because the asynchronous residual-feedback gradient estimator (5.6) is subject to almost the same level of variance as the two-point gradient estimator (5.33), but can make twice the number of updates compared to the two-point gradient estimator (5.33) for the same number of queries. Note that we compare the two algorithms in terms of the number of queries rather than the number of updates, because the number of queries corresponds to the length of the global wall time required to run the algorithm.



**Figure 5.1:** Convergence results of the distributed feature learning problem. The red curve is obtained by applying the asynchronous two-point gradient estimator (5.33) and the blue curve is by the asynchronous residual-feedback estimator (5.6). The y axis denotes the value of the loss function (5.32) and the x axis represents the number of queries made in total by the team of agents. The shaded area around each curve represents the standard deviation of the function values over 10 trials.

# Chapter 6

## Conclusions

In this dissertation, we proposed a novel one-point ZO gradient estimator based on residual-feedback scheme, and studied the performance of optimization algorithms with the proposed ZO oracle under different settings of learning and control problems. Next, we summarize our contributions under these settings separately.

In Chapter 2, 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.

In Chapter 3, we applied the residual-feedback ZO gradient estimator to solve online optimization problems. 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 non-convex 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 non-convex online problems. Numerical experiments on two non-stationary reinforcement learning problems were conducted and the proposed

residual-feedback estimator was shown to significantly outperform the conventional one-point method.

In Chapter 4, we proposed a new distributed zeroth-order policy optimization method for MARL problems, through decentralizing the proposed residual-feedback ZO gradient estimator. Compared to existing MARL algorithms that require all the agents' states and actions to be accessible by every local agent, our algorithm can be applied even when each agent only observes partial states and actions. Specifically, we developed a new distributed residual-feedback zeroth-order estimator of the policy gradient and analyzed the effect of bias in the local policy gradient estimates on the convergence of the proposed MARL algorithm. Furthermore, we introduced a value tracking technique to reduce the number of consensus steps needed at each episode to control the bias in the estimation of the policy gradient. Finally, we provided numerical experiments on a stochastic multi-agent multi-stage decision making problem that demonstrated the effectiveness of both the decentralized residual-feedback policy gradient estimator and the value tracking technique.

In Chapter 5, we proposed an asynchronous residual-feedback gradient estimator for distributed zeroth-order optimization. More importantly, only the local decision vector is needed for estimating the gradient and no communication among agents is required. We showed that the convergence rate of the proposed method matches the results for centralized residual-feedback methods when the function evaluation has noise. To the best of our knowledge, this is the first time the convergence of a fully asynchronous ZO gradient estimator is studied. Numerical experiments on a distributed logistic regression problem are presented to show the effectiveness of the proposed method.

## 6.1 Future Research Directions

The convergence of ZO methods can be much slower when the problem is of high dimension. To deal with such challenge, existing literatures usually assume that the underlying true gradient of the original problem is sparse, which usually cannot be satisfied in practice.

Therefore, how to apply ZO methods, e.g., the method with the proposed residual-feedback ZO oracle in this dissertation, to high dimensional problem is still an open problem. Two future directions can be explored to deal with this issue.

First, instead of running ZO methods directly on the original parameter space which is of high dimension, we can optimize some intermediate output given the original parameter. One example is to optimize the weights of neural networks in machine learning problems. Rather than directly perturbing the neural network’s weights, which can usually be of tens of thousands dimensions, and updating these weights with ZO gradient estimates, we can perturb the output of the neural networks, which is usually of much lower dimension, get the gradient with respect to the neural network’s output, and obtain the gradient with respect to the neural network’s weights through the chain rule. It will be interesting to study the theoretical convergence rate of such algorithms and compare its dependency on the problem dimension to that of the existing ZO optimization scheme.

Second, rather than directly assuming that the underlying problem enjoys sparsity property as assumed by existing works, it is interesting to learn a transformed space from the original parameter space, so that such sparsity assumption is satisfied by learning such transformation. One possible approach to learn such transformed parameter space is to use the loss function with  $L_1$  regularization on the gradient vectors at the transformed space. It is interesting to test this learning to optimize idea both empirically and theoretically.

Besides improving the dependency of ZO methods on problem dimensions, another important question to answer is how to guarantee safety when implementing ZO methods. Almost all existing ZO methods apply omni-directional perturbation schemes, e.g., uniform sphere sampling or Gaussian sampling, to estimate the gradient of the unknown objective functions. Such scheme assume the perturbed decision variable in ZO gradient estimator must be successfully and safely implemented so that the value of the objective function can be returned. However, in practice, the perturbed decision variable may violate the safety routine and be denied. In this scenario, the feedback the algorithm received is a binary signal, whether the perturbed decision is safe, and the objective function value is only

returned to the algorithm if the perturbed decision is safe. Otherwise, the algorithm receives signal 0. Whether ZO methods with such safety constraints of function evaluation can still converge, and if converges, how much bias is in the solution, are still open questions.

# Appendix A

## Proofs for Chapter 2

### A.1 Proof of Lemma 2.6

First, we show the bound when  $f(x) \in C^{0,0}$ . Recalling the expression of  $\tilde{g}(x_t)$  in (2.1), we have that

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

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

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

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<sup>15</sup>, 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 (2.1), we have that

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

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

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

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 <sup>15</sup>. 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

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

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

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

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

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

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<sup>15</sup>, 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

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

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

## A.2 Proof of Theorem 2.7

Since we have that  $f(x) \in C^{0,0}$ , according to Lemma 2.2, 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<sup>46</sup>, we can get the following inequality

$$\begin{aligned}
f_\delta(x_{t+1}) &\leq f_\delta(x_t) + \langle \nabla f_\delta(x_t), x_{t+1} - x_t \rangle \\
&\quad + \frac{L_1(f_\delta)}{2} \|x_{t+1} - x_t\|^2 \\
&= f_\delta(x_t) - \eta \langle \nabla f_\delta(x_t), \tilde{g}(x_t) \rangle + \frac{L_1(f_\delta)\eta^2}{2} \|\tilde{g}(x_t)\|^2 \\
&= f_\delta(x_t) - \eta \langle \nabla f_\delta(x_t), \Delta_t \rangle - \eta \|\nabla f_\delta(x_t)\|^2 \\
&\quad + \frac{L_1(f_\delta)\eta^2}{2} \|\tilde{g}(x_t)\|^2,
\end{aligned} \tag{A.11}$$

where  $\Delta_t = \tilde{g}(x_t) - \nabla f_\delta(x_t)$ . According to Lemma 2.5, 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 (A.11) and rearranging terms, we have that

$$\begin{aligned}
\eta \mathbb{E}[\|\nabla f_\delta(x_t)\|^2] &\leq \mathbb{E}[f_\delta(x_t)] - \mathbb{E}[f_\delta(x_{t+1})] \\
&\quad + \frac{L_1(f_\delta)\eta^2}{2} \mathbb{E}[\|\tilde{g}(x_t)\|^2].
\end{aligned} \tag{A.12}$$

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

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(x_t)\|^2] &\leq \frac{\mathbb{E}[f_\delta(x_0)] - \mathbb{E}[f_\delta(x_T)]}{\eta} \\
&\quad + \frac{L_1(f_\delta)\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}(x_t)\|^2] \\
&\leq \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} + \frac{L_1(f_\delta)\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}(x_t)\|^2],
\end{aligned} \tag{A.13}$$

where  $f_\delta^*$  is the lower bound of the smoothed function  $f_\delta(x)$ .  $f_\delta^*$  must exist because we assume the original function  $f(x)$  is lower bounded and the smoothed function has a bounded distance from  $f(x)$  due to Lemma 2.2.

Recall the contraction result of the second moment  $\mathbb{E}[\|\tilde{g}(x_t)\|^2]$  in Lemma 2.6 when  $f(x) \in C^{0,0}$ . Denote the contraction rate  $\frac{2dL_0^2\eta^2}{\delta^2}$  as  $\alpha$  and the constant perturbation term  $M = 8L_0^2(d+4)^2$ . Then, we get that

$$\mathbb{E}[\|\tilde{g}(x_t)\|^2] \leq \alpha^t \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{1 - \alpha^t}{1 - \alpha} M. \tag{A.14}$$

Summing the above inequality over time, we obtain

$$\begin{aligned} \sum_{t=0}^{T-1} \|\tilde{g}(x_t)\|^2 &\leq \frac{1-\alpha^T}{1-\alpha} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \sum_{t=0}^{T-1} \left(\frac{1-\alpha^t}{1-\alpha} M\right) \\ &\leq \frac{1}{1-\alpha} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{1}{1-\alpha} MT. \end{aligned} \quad (\text{A.15})$$

Plugging the bound in (A.15) into inequality (A.13), and since  $L_1(f_\delta) = \frac{\sqrt{d}}{\delta} L_0$ , we have that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(x_t)\|^2] &\leq \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} + \frac{d^{\frac{1}{2}} L_0 \eta}{\delta} \\ &\quad \left(\frac{1}{1-\alpha} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{1}{1-\alpha} 8L_0^2(d+4)^2 T\right). \end{aligned} \quad (\text{A.16})$$

To fulfill the requirement that  $|f(x) - f_\delta(x)| \leq \epsilon_f$ , we set the exploration parameter  $\delta = \frac{\epsilon_f}{d^{\frac{1}{2}} L_0}$ . In addition, let the stepsize be  $\eta = \frac{\sqrt{\epsilon_f}}{2dL_0^2 T^{\frac{1}{2}}}$ . 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  $\eta$  and  $\delta$  into inequality (A.16), we obtain that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(x_t)\|^2] &\leq 2L_0^2 (\mathbb{E}[f_\delta(x_0)] - f_\delta^*) \frac{d}{\sqrt{\epsilon_f}} \sqrt{T} \\ &\quad + \mathbb{E}[\|\tilde{g}(x_0)\|^2] + 8L_0^2 \frac{(d+4)^2}{\sqrt{\epsilon_f}} \sqrt{T}. \end{aligned}$$

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

### A.3 Proof of Theorem 2.8

Following the same process in the beginning of the proof of Theorem 2.7, we can get

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(x_t)\|^2] \leq \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} + \frac{L_1 \eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}(x_t)\|^2]. \quad (\text{A.17})$$

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 (A.17) and Lemma 2.2, we have that

$$\begin{aligned} \frac{1}{2} \sum_{t=0}^{T-1} \|\mathbb{E}[\|\nabla f(x_t)\|^2]\| &\leq \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} \\ &\quad + \frac{L_1 \eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}(x_t)\|^2] + L_1^2 (d+3)^3 \delta^2 T. \end{aligned} \quad (\text{A.18})$$

In addition, similar to the process to derive the bound in (A.15), according to Lemma 2.6, when  $f(x) \in C^{1,1}$ , we can get that

$$\begin{aligned} \sum_{t=0}^{T-1} \|\tilde{g}(x_t)\|^2 &\leq \frac{1}{1-\alpha} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{8}{1-\alpha} (d+4)^2 \\ &\quad \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 + \frac{4}{1-\alpha} L_1^2 (d+6)^3 \delta^2 T. \end{aligned} \quad (\text{A.19})$$

Plugging the bound (A.19) into (A.18), we have that

$$\begin{aligned} \frac{1}{2} \sum_{t=0}^{T-1} \|\mathbb{E}[\|\nabla f(x_t)\|^2]\| &\leq \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} \\ &\quad + \frac{L_1 \eta}{2} \left( \frac{1}{1-\alpha} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{4}{1-\alpha} L_1^2 (d+6)^3 \delta^2 T \right) \\ &\quad + \frac{8}{1-\alpha} (d+4)^2 \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \\ &\quad + L_1^2 (d+3)^3 \delta^2 T. \end{aligned} \quad (\text{A.20})$$

Recalling that  $\tilde{L} = \max\{32L_1, 2L_0\}$ , let  $\eta = \frac{1}{\tilde{L}(d+4)^2 T^{\frac{1}{3}}}$  and  $\delta = \frac{1}{\sqrt{dT^{\frac{1}{3}}}}$ , 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

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

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

## A.4 Proof of Theorem 2.9

First, according to iteration (2.2), we have that

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

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

$$\begin{aligned} \mathbb{E}[\|x_{t+1} - x^*\|^2] &\leq \mathbb{E}[\|x_t - x^*\|^2] - 2\eta \langle \nabla f_\delta(x_t), x_t - x^* \rangle \\ &\quad + \eta^2 \mathbb{E}[\|\tilde{g}(x_t)\|^2]. \end{aligned} \tag{A.21}$$

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 (A.21), we have that

$$\begin{aligned} \mathbb{E}[\|x_{t+1} - x^*\|^2] &\leq \mathbb{E}[\|x_t - x^*\|^2] - 2\eta(f_\delta(x_t) - f_\delta(x^*)) \\ &\quad + \eta^2 \mathbb{E}[\|\tilde{g}(x_t)\|^2]. \end{aligned} \tag{A.22}$$

When  $f(x) \in C^{0,0}$ , using Lemma (2.2), we can replace  $f_\delta(x)$  with  $f(x)$  in above inequality and get

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

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

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

Since function  $f(x) \in C^{0,0}$ , we can plug the bound (A.15) into the above inequality and get that

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

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  $\eta$  and  $\delta$  into above inequality, we have that

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

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 (A.22). In addition, according to Lemma 2.2, we can replace  $f_\delta(x)$  with  $f(x)$  in above inequality and get

$$\begin{aligned} \mathbb{E}[\|x_{t+1} - x^*\|^2] &\leq \mathbb{E}[\|x_t - x^*\|^2] - 2\eta(f(x_t) - f(x^*)) \\ &+ \eta^2 \mathbb{E}[\|\tilde{g}(x_t)\|^2] + 4L_1 d \delta^2 \eta. \end{aligned} \tag{A.23}$$

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 (A.19) and obtain that

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

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<sup>46</sup>. Applying this bound into the above inequality, we get that

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

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

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

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.

## A.5 Proof of Lemma 2.12

The analysis is similar to the proof in Section A.1. First, consider the case when  $F(x, \xi) \in C^{0,0}$  with  $L_0(\xi)$ . According to (2.4), we have that

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

Using the bound in Assumption 2.10, we get that  $\frac{2}{\delta^2} \mathbb{E}[(F(x_{t-1} + \delta u_{t-1}, \xi_t) - F(x_{t-1} + \delta u_{t-1}, \xi_{t-1}))^2 \|u_t\|^2] \leq \frac{8d\sigma^2}{\delta^2}$ . In addition, adding and subtracting  $F(x_{t-1} + \delta u_t, \xi_t)$  in  $(F(x_t + \delta u_t, \xi_t) - F(x_{t-1} + \delta u_{t-1}, \xi_t))^2$  in above inequality, we obtain that

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

Using Assumption 2.11, 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.

## A.6 Proof of Theorem 2.13

When function  $F(x) \in C^{0,0}$  with  $L_0(\xi)$ , using Assumption 2.11 and following the same procedure in Section A.2, we have that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(x_t)\|^2] &\leq \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} \\ &\quad + \frac{L_1(f_\delta)\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}(x_t)\|^2], \end{aligned} \quad (\text{A.24})$$

where  $L_1(f_\delta) = \frac{\sqrt{d}}{\delta} L_0$ . In addition, according to Lemma 2.12, we get that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}(x_t)\|^2] &\leq \frac{1}{1-\alpha} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{16L_0^2}{1-\alpha} (d+4)^2 T \\ &\quad + \frac{8\sigma^2}{1-\alpha} \frac{d}{\delta^2} T, \end{aligned} \quad (\text{A.25})$$

where  $\alpha = \frac{4dL_0^2\eta^2}{\delta^2}$ . Plugging (A.25) into the bound in (A.24), we obtain that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(x_t)\|^2] &\leq \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} + \frac{4\sigma^2 L_0}{1-\alpha} d^{1.5} \frac{\eta}{\delta^3} T \\ &\quad + \frac{\sqrt{d}L_0}{2(1-\alpha)} \mathbb{E}[\|\tilde{g}(x_0)\|^2] \frac{\eta}{\delta} + \frac{8L_0^3\sqrt{d}}{1-\alpha} (d+4)^2 \frac{\eta}{\delta} T. \end{aligned} \quad (\text{A.26})$$

Similar to Section A.2, to fulfill the requirement that  $|f(x) - f_\delta(x)| \leq \epsilon_f$ , we set the exploration parameter  $\delta = \frac{\epsilon_f}{d^{1/2} L_0}$ . In addition, let the stepsize be  $\eta = \frac{\epsilon_f^{1.5}}{2\sqrt{2}L_0^2 d^{1.5} T^{1/2}}$ . 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  $\frac{1}{1-\alpha} \leq 2$ .

Applying this bound and the choices of  $\eta$  and  $\delta$  into the bound (A.26), we obtain that

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

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 2.11, we also have that  $f_\delta(x), f(x) \in C^{1,1}$  with constant  $L_1$ . Similarly to the proof in Section A.3, we get that

$$\begin{aligned} \frac{1}{2} \sum_{t=0}^{T-1} \|\mathbb{E}[\|\nabla f(x)\|^2]\| &\leq \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} \\ &+ \frac{L_1 \eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}(x_t)\|^2] + L_1^2 (d+3)^3 \delta^2 T. \end{aligned} \quad (\text{A.27})$$

Plugging inequality (A.25) into the above upper bound, we obtain that

$$\begin{aligned} \frac{1}{2} \sum_{t=0}^{T-1} \|\mathbb{E}[\|\nabla f(x)\|^2]\| &\leq \frac{\mathbb{E}[f_\delta(x_0)] - f_\delta^*}{\eta} \\ &+ \frac{L_1 \eta}{2(1-\alpha)} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{8L_0^2 L_1}{1-\alpha} (d+4)^2 \eta T \\ &+ \frac{4L_1 \sigma^2}{1-\alpha} \frac{d\eta}{\delta^2} T + L_1^2 (d+3)^3 \delta^2 T. \end{aligned} \quad (\text{A.28})$$

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,  $\alpha = \frac{4dL_0^2 \eta^2}{\delta^2} = \frac{1}{2T} \leq \frac{1}{2}$  and  $\frac{1}{1-\alpha} \leq 2$ . Plugging these results into the above inequality, we get that

$$\begin{aligned} \frac{1}{2} \sum_{t=0}^{T-1} \|\mathbb{E}[\|\nabla f(x)\|^2]\| &\leq 2\sqrt{2}L_0 (\mathbb{E}[f_\delta(x_0)] - f_\delta^*) d^{\frac{4}{3}} T^{\frac{2}{3}} \\ &+ \frac{L_1}{2\sqrt{2}L_0 d^{\frac{4}{3}} T^{\frac{2}{3}}} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + 4\sqrt{2}L_0 L_1 \frac{(d+4)^2}{d^{\frac{4}{3}}} T^{\frac{1}{3}} \\ &+ \frac{2\sqrt{2}L_1 \sigma^2}{L_0 d^{\frac{1}{3}}} T^{\frac{1}{3}} + L_1^2 \frac{(d+3)^3}{d^{\frac{5}{3}}} T^{\frac{2}{3}}. \end{aligned} \quad (\text{A.29})$$

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

## A.7 Proof of Theorem 2.14

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 A.4 and get that

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

Plugging the bound (A.25) into above inequality, we have that

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[f(x_t)] - Tf(x^*) &\leq \frac{1}{2\eta} \|x_0 - x^*\|^2 \\
&+ \frac{\eta}{2(1-\alpha)} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{8L_0^2}{1-\alpha} (d+4)^2 \eta T \\
&+ \frac{4\sigma^2}{1-\alpha} \frac{d\eta}{\delta^2} T + 2L_0 \sqrt{d} \delta T.
\end{aligned} \tag{A.30}$$

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

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[f(x_t)] - Tf(x^*) &\leq \sqrt{2}L_0 \|x_0 - x^*\|^2 \sqrt{dT}^{\frac{3}{4}} \\
&+ \frac{1}{2\sqrt{2}L_0\sqrt{dT}^{\frac{3}{4}}} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + 4\sqrt{2}L_0 \frac{(d+4)^2}{\sqrt{d}} T^{\frac{1}{4}} \\
&+ \frac{2\sqrt{2}\sigma^2}{L_0} \sqrt{dT}^{\frac{3}{4}} + 2L_0 \sqrt{dT}^{\frac{3}{4}}.
\end{aligned} \tag{A.31}$$

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 (A.23) in Section A.4. Telescoping this inequality from  $t = 0$  to  $T - 1$  and rearranging terms, we obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[f(x_t)] - Tf(x^*) &\leq \frac{1}{2\eta} \|x_0 - x^*\|^2 \\
&+ \frac{\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}(x_t)\|^2] + 2L_1 d \delta^2 T.
\end{aligned} \tag{A.32}$$

Plugging the bound (A.25) into above inequality, we have that

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[f(x_t)] - Tf(x^*) &\leq \frac{1}{2\eta} \|x_0 - x^*\|^2 + 2L_1 d \delta^2 T \\
&+ \frac{\eta}{2(1-\alpha)} \mathbb{E}[\|\tilde{g}(x_0)\|^2] + \frac{8L_0^2}{1-\alpha} (d+4)^2 \eta T + \frac{4\sigma^2}{1-\alpha} \frac{d\eta}{\delta^2} T.
\end{aligned}$$

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

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

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

## A.8 Zeroth-Order Policy Optimization for A Large-Scale Multi-Stage Decision Making Problem

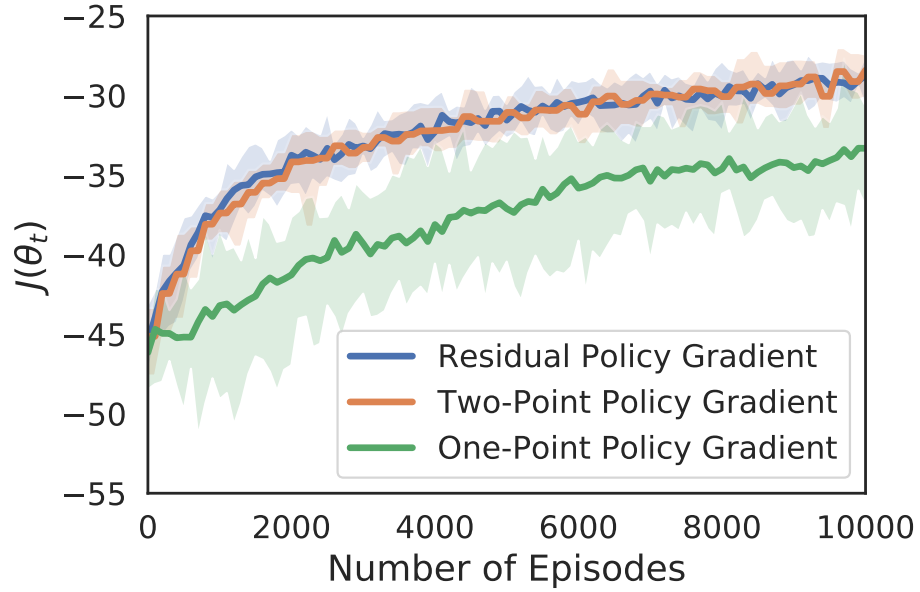
In this section, we consider a large-scale multi-stage resource allocation problem. Specifically, we consider 16 agents that are located on a  $4 \times 4$  grid. At agent  $i$ , resources are stored in the amount of  $m_i(k)$  and there is also a demand for resources in the amount of  $d_i(k)$  at instant  $k$ . 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  $w_{i,k}$  is the noise in the demand. 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  $\pi_{i,\theta_i}(o_i) : \mathcal{O}_i \rightarrow [0, 1]^{|\mathcal{N}_i|}$ , where  $\theta_i$  is the parameter of the policy function  $\pi_i$ ,  $o_i \in \mathcal{O}_i$  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. The goal for the agents is to find an optimal policy  $\pi^* = \{\pi_{i, \theta_i}(o_i)\}$  so that the global accumulated reward

$$J(\theta) = \sum_{i=1}^{16} \sum_{k=0}^K \gamma^k r_i(k) \quad (\text{A.33})$$

is maximized, where  $\theta = [\dots, \theta_i, \dots]$  is the global policy parameter,  $K$  is the horizon of the problem, and  $\gamma$  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 (1.1) and (2.4) 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 (A.33). Then, the zeroth-order policy gradient is estimated using (1.1) or (2.4). On the contrary, when the two-point zeroth-order policy gradient estimator (1.2) 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 A.1, we present the performance of using zeroth-order policy gradients (1.1), (1.2) and (2.4) 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  $K = 30$ . Each algorithm is run for 10 trials. We observe that policy optimization with the proposed residual-feedback gradient estimate (2.4) improves the optimal policy parameters with the same learning rate as the two-point zeroth-order gradient estimator (1.2), 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 (1.1) considered in<sup>32;33</sup>.



**Figure A.1:** The convergence rate of applying the proposed residual one-point feedback (2.1) (blue), the two-point oracle (1.2) in<sup>15</sup> (orange) and the one-point oracle (1.1) in<sup>12</sup> (green) to the large-scale stochastic multi-stage resource allocation 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.

# Appendix B

## Proofs for Chapter 3

### B.1 Implementation Details of the Numerical Experiments

All experiments are conducted using Matlab R2019a on Ubuntu 18.04 with the AMD Ryzen 2700X 8-core processor and 16GB 2133MHz memory.

For the non-stationary LQR experiments, we select  $n_x = 6$ ,  $n_u = 6$  and  $\gamma = 0.5$ . The dynamical matrices  $A_0$  and  $B_0$  at episode 0 are randomly generated from a Gaussian distribution  $\mathcal{N}(0, 0.1^2)$ . Then, we generate the time-varying dynamical matrices according to  $A_{t+1} = A_t + 0.01M_t$  and  $B_{t+1} = B_t + 0.01N_t$ , where  $M_t$  and  $N_t$  are random matrices whose entries are uniformly sampled from  $[0,1]$ . To evaluate the cost function  $V_t(K_t)$  given the policy parameter  $K_t$  at episode  $t$ , we roll out a trajectory of length  $H = 50$  using the policy parameter  $K_t$  and sum up the collected rewards.

For the non-stationary resource allocation experiments, 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 9 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  $[-1, 1]$ .

## B.2 Proof of Lemma 3.2 and Lemma 3.3

*Proof.* Recall that  $f_\delta(x) = \mathbb{E}_{v \in \mathbb{B}}[f(x + \delta v)]$ . Then, we have that

$$\begin{aligned} |f_\delta(x) - f(x)| &= |\mathbb{E}_{v \in \mathbb{B}}[f(x + \delta v) - f(x)]| \\ &\leq \mathbb{E}_{v \in \mathbb{B}}[|f(x + \delta v) - f(x)|] \\ &\leq \mathbb{E}_{v \in \mathbb{B}}[L_0 \|\delta v\|]. \end{aligned} \quad (\text{B.1})$$

Furthermore, since  $v \in \mathbb{B}$ , we have that  $\|\delta v\| \leq \delta$ . Combining this inequality with (B.1), we have that  $|f_\delta(x) - f(x)| \leq \mathbb{E}_{v \in \mathbb{B}}[\delta L_0] = L_0 \delta$ . When the function  $f \in C^{1,1}$  with Lipschitz constant  $L_1$ , we have that

$$\langle \nabla f(x), \delta v \rangle - \frac{L_1}{2} \|\delta v\|^2 \leq f(x + \delta v) - f(x) \leq \langle \nabla f(x), \delta v \rangle + \frac{L_1}{2} \|\delta v\|^2, \quad (\text{B.2})$$

for all  $v \in \mathbb{B}$ . Taking the expectation of (B.2) over  $v$  sampled uniformly from the unit ball  $\mathbb{B}$  and recalling that  $v$  is sampled independently from  $x$  and has zero mean, we get that

$$-L_1 \delta^2 \leq -\frac{L_1}{2} \mathbb{E}_{v \in \mathbb{B}}[\|\delta v\|^2] \leq \mathbb{E}_{v \in \mathbb{B}}[f(x + \delta v) - f(x)] \leq \frac{L_1}{2} \mathbb{E}_{v \in \mathbb{B}}[\|\delta v\|^2] \leq L_1 \delta^2. \quad (\text{B.3})$$

In addition, because  $|f_\delta(x) - f(x)| = |\mathbb{E}_{v \in \mathbb{B}}[f(x + \delta v) - f(x)]|$ , we obtain that  $|f_\delta(x) - f(x)| \leq L_1 \delta^2$ . The proof is complete.

Next, we show that  $\|\nabla_\delta f(x) - \nabla f(x)\| \leq \delta L_1 d$  when  $f \in C^{1,1}$ . First, recall that  $\nabla_\delta f(x) = \mathbb{E}_u[\frac{d}{\delta}(f(x + \delta u) - f(x))u]$  and  $\nabla f(x) = \mathbb{E}_u[d\langle \nabla f(x), u \rangle u]$ , we have that

$$\|\nabla_\delta f(x) - \nabla f(x)\| = \|\mathbb{E}_u[\frac{d}{\delta}(f(x + \delta u) - f(x))u] - \mathbb{E}_u[d\langle \nabla f(x), u \rangle u]\| \quad (\text{B.4})$$

$$\leq \frac{d}{\delta} \mathbb{E}_u[\|(f(x + \delta u) - f(x) - \langle \nabla f(x), \delta u \rangle)u\|], \quad (\text{B.5})$$

where the second inequality is due to Jensen's inequality. Since vector  $u$  is randomly sampled from the unit sphere, we get that

$$\|\nabla_\delta f(x) - \nabla f(x)\| \leq \frac{d}{\delta} \mathbb{E}_u[\|(f(x + \delta u) - f(x) - \langle \nabla f(x), \delta u \rangle)\|]. \quad (\text{B.6})$$

Furthermore, since  $f \in C^{1,1}$ , we have that  $|f(x + \delta u) - f(x) - \langle \nabla f(x), \delta u \rangle| \leq L_1 \|\delta u\|^2 \leq L_1 \delta^2$ . Combining this bound with inequality (B.6), we complete the proof.

Finally, we show Lemma 2.3. Recall that  $\nabla_\delta f(x) = \mathbb{E}_u[\frac{d}{\delta}f(x + \delta u)u]$ . Therefore, for any  $x_1, x_2 \in \mathcal{X}$ , we have that

$$\begin{aligned} \|\nabla_\delta f(x_1) - \nabla_\delta f(x_2)\| &\leq \|\mathbb{E}_u[\frac{d}{\delta}f(x_1 + \delta u)u] - \mathbb{E}_u[\frac{d}{\delta}f(x_2 + \delta u)u]\| \\ &\leq \frac{d}{\delta}\mathbb{E}_u[\|(f(x_1 + \delta u) - f(x_2 + \delta u))u\|] \leq \frac{d}{\delta}\mathbb{E}_u[\|f(x_1 + \delta u) - f(x_2 + \delta u)\|], \end{aligned} \quad (\text{B.7})$$

where the second inequality is due to Jensen's inequality. Since  $f \in C^{0,0}$  with Lipschitz constant  $L_0$ , we have that  $\|f(x_1 + \delta u) - f(x_2 + \delta u)\| \leq L_0\|x_1 - x_2\|$ . Therefore, we get that

$$\|\nabla_\delta f(x_1) - \nabla_\delta f(x_2)\| \leq \frac{dL_0}{\delta}\|x_1 - x_2\|. \quad (\text{B.8})$$

The proof is complete.  $\square$

### B.3 Proof of Lemma 3.7

*Proof.* By definition of the residual feedback (3.1), we have that

$$\begin{aligned} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] &= \mathbb{E}[\frac{d^2}{\delta^2}(f_t(x_t + \delta u_t) - f_{t-1}(x_{t-1} + \delta u_{t-1}))^2\|u_t\|^2] \\ &\leq \frac{2d^2}{\delta^2}\mathbb{E}[(f_t(x_t + \delta u_t) - f_t(x_{t-1} + \delta u_{t-1}))^2\|u_t\|^2] \\ &\quad + \frac{2d^2}{\delta^2}\mathbb{E}[(f_t(x_{t-1} + \delta u_{t-1}) - f_{t-1}(x_{t-1} + \delta u_{t-1}))^2\|u_t\|^2] \quad (\text{B.9}) \\ &\leq \frac{2d^2}{\delta^2}\mathbb{E}[(f_t(x_t + \delta u_t) - f_t(x_{t-1} + \delta u_{t-1}))^2] \\ &\quad + \frac{2d^2}{\delta^2}\mathbb{E}[(f_t(x_{t-1} + \delta u_{t-1}) - f_{t-1}(x_{t-1} + \delta u_{t-1}))^2], \end{aligned}$$

where the last inequality is because  $u_t \in \mathbb{U}^d$ . Moreover, adding and subtracting  $f_t(x_{t-1} + \delta u_t)$  to the term  $(f_t(x_t + \delta u_t) - f_t(x_{t-1} + \delta u_{t-1}))^2$  in the inequality (B.9), we obtain

$$\begin{aligned} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] &\leq \frac{4d^2}{\delta^2}\mathbb{E}[(f_t(x_t + \delta u_t) - f_t(x_{t-1} + \delta u_t))^2] \\ &\quad + \frac{4d^2}{\delta^2}\mathbb{E}[(f_t(x_{t-1} + \delta u_t) - f_t(x_{t-1} + \delta u_{t-1}))^2] \quad (\text{B.10}) \\ &\quad + \frac{2d^2}{\delta^2}\mathbb{E}[(f_t(x_{t-1} + \delta u_{t-1}) - f_{t-1}(x_{t-1} + \delta u_{t-1}))^2]. \end{aligned}$$

Since  $f_t \in C^{0,0}$  is Lipschitz with constant  $L_0$ , we further obtain that

$$\begin{aligned} \mathbb{E}[\|\tilde{g}(x_t)\|^2] &\leq \frac{4d^2L_0^2}{\delta^2}\mathbb{E}[\|x_t - x_{t-1}\|^2] + 4d^2L_0^2\mathbb{E}[\|u_t - u_{t-1}\|^2] \\ &\quad + \frac{2d^2}{\delta^2}\mathbb{E}[(f_t(x_{t-1} + \delta u_{t-1}) - f_{t-1}(x_{t-1} + \delta u_{t-1}))^2]. \end{aligned} \quad (\text{B.11})$$

Since  $u_t \in \text{US}^d$ , we get that  $\mathbb{E}[\|u_t - u_{t-1}\|^2] \leq 4$ . Substituting this bound into inequality (B.11), we obtain that

$$\begin{aligned} \mathbb{E}[\|\tilde{g}(x_t)\|^2] &\leq \frac{4d^2L_0^2}{\delta^2}\mathbb{E}[\|x_t - x_{t-1}\|^2] + 16d^2L_0^2 \\ &\quad + \frac{2d^2}{\delta^2}\mathbb{E}[(f_t(x_{t-1} + \delta u_{t-1}) - f_{t-1}(x_{t-1} + \delta u_{t-1}))^2]. \end{aligned} \quad (\text{B.12})$$

Since  $x_t = \Pi_{\mathcal{X}}[x_{t-1} - \eta\tilde{g}(x_{t-1})]$ , we get that  $\|x_t - x_{t-1}\| = \|\Pi_{\mathcal{X}}[x_{t-1} - \eta\tilde{g}(x_{t-1})] - \Pi_{\mathcal{X}}[x_{t-1}]\| \leq \eta\|\tilde{g}(x_{t-1})\|$  due to the nonexpansiveness of the projection operator onto a convex set. Therefore, we have that

$$\begin{aligned} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] &\leq \frac{4d^2L_0^2\eta^2}{\delta^2}\mathbb{E}[\|\tilde{g}_{t-1}(x_{t-1})\|^2] + 16d^2L_0^2 \\ &\quad + \frac{2d^2}{\delta^2}\mathbb{E}[(f_t(x_{t-1} + \delta u_{t-1}) - f_{t-1}(x_{t-1} + \delta u_{t-1}))^2]. \end{aligned} \quad (\text{B.13})$$

The proof is complete.  $\square$

## B.4 Proof of Theorem 3.9

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

$$f_{\delta,t}(x_t) - f_{\delta,t}(x) \leq \langle \nabla f_{\delta,t}(x_t), x_t - x \rangle, \text{ for all } x \in \mathcal{X}, \quad (\text{B.14})$$

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

$$\mathbb{E}[f_{\delta,t}(x_t) - f_{\delta,t}(x)] \leq \mathbb{E}[\langle \tilde{g}_t(x_t), x_t - x \rangle]. \quad (\text{B.15})$$

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

$$\begin{aligned} \|x_{t+1} - x\|^2 &= \|\Pi_{\mathcal{X}}[x_t - \eta\tilde{g}(x_t)] - \Pi_{\mathcal{X}}[x]\|^2 \\ &\leq \|x_t - \eta\tilde{g}(x_t) - x\|^2 \\ &= \|x_t - x\|^2 - 2\eta\langle \tilde{g}_t(x_t), x_t - x \rangle + \eta^2\|\tilde{g}_t(x_t)\|^2. \end{aligned} \quad (\text{B.16})$$

Rearranging the above inequality yields that

$$\langle \tilde{g}_t(x_t), x_t - x \rangle \leq \frac{1}{2\eta} (\|x_t - x\|^2 - \|x_{t+1} - x\|^2) + \frac{\eta}{2} \|\tilde{g}_t(x_t)\|^2. \quad (\text{B.17})$$

Taking expectation on both sides of the above inequality over  $u_t$ , using inequality (B.15), and telescoping the resulting bound from  $t = 0$  to  $T$ , we obtain that

$$\mathbb{E} \left[ \sum_{t=0}^T f_{\delta,t}(x_t) - \sum_{t=0}^T f_{\delta,t}(x) \right] \leq \frac{1}{2\eta} \|x_0 - x\|^2 + \frac{\eta}{2} \mathbb{E} \left[ \sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2 \right]. \quad (\text{B.18})$$

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

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x) \right] &= \mathbb{E} \left[ \sum_{t=0}^T f_{\delta,t}(x_t) - \sum_{t=0}^T f_{\delta,t}(x) \right] \\ &\quad + \mathbb{E} \left[ \sum_{t=0}^T (f_t(x_t) - f_{\delta,t}(x_t)) - \sum_{t=0}^T (f_t(x) - f_{\delta,t}(x)) \right] \\ &\leq \frac{1}{2\eta} \|x_0 - x\|^2 + \frac{\eta}{2} \mathbb{E} \left[ \sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2 \right] + 2L_0\delta T. \end{aligned} \quad (\text{B.19})$$

On the other hand, telescoping the second moment bound in (3.3) 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 3.8, we obtain that

$$\mathbb{E} \left[ \sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2 \right] \leq \frac{1}{1-\alpha} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \frac{16}{1-\alpha} d^2 L_0^2 T + \frac{2d^2 V_f^2}{1-\alpha} \frac{1}{\delta^2} T, \quad (\text{B.20})$$

where  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2}$ . Substituting the above bound into (B.19) yields that

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x) \right] &\leq \frac{1}{2\eta} \|x_0 - x\|^2 + \frac{\eta}{2(1-\alpha)} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \frac{8}{1-\alpha} L_0^2 d^2 \eta T \\ &\quad + 2L_0\delta T + \frac{d^2 V_f^2}{1-\alpha} \frac{\eta}{\delta^2} T. \end{aligned} \quad (\text{B.21})$$

Since 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  $\eta = \frac{R^{\frac{3}{2}}}{2\sqrt{2d}L_0T^{\frac{3}{4}}}$  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

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x^*) \right] &\leq \sqrt{2d}L_0\sqrt{RT}^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] R^{\frac{3}{2}}}{2\sqrt{2d}L_0T^{\frac{3}{4}}} \\ &\quad + 4\sqrt{2}d^{\frac{3}{2}}L_0R^{\frac{3}{2}}T^{\frac{1}{4}} + 2L_0\sqrt{dRT}^{\frac{3}{4}} + \frac{\sqrt{dRV_f^2}}{\sqrt{2}L_0} T^{\frac{3}{4}}. \end{aligned} \quad (\text{B.22})$$

When  $R$  is unknown, let  $\eta = \frac{1}{2\sqrt{2d}L_0T^{\frac{3}{4}}}$  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

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x^*)\right] &\leq \sqrt{2d}L_0R^2T^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{2\sqrt{2d}L_0T^{\frac{3}{4}}} + 4\sqrt{2}d^{\frac{3}{2}}L_0T^{\frac{1}{4}} \\ &\quad + 2\sqrt{d}L_0T^{\frac{3}{4}} + \frac{\sqrt{d}V_f^2}{\sqrt{2}L_0}T^{\frac{3}{4}}. \end{aligned} \quad (\text{B.23})$$

On the other hand, we can let  $\eta = \frac{R^{\frac{3}{2}}}{2\sqrt{2d}L_0T^{\frac{3}{4}}}$  and  $\delta = \frac{\sqrt{dR}}{L_0^qT^{\frac{1}{4}}}$ , where  $q \in \mathbb{R}$  is a user-specific parameter. With this choice of parameters, we get  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2} = \frac{L_0^{2q}R^2}{2T} \leq \frac{1}{2}$  when  $T \geq L_0^{2q}R^2$  and, as a result, we obtain that

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x^*)\right] &\leq \sqrt{2d}L_0\sqrt{RT}^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]R^{\frac{3}{2}}}{2\sqrt{2d}L_0T^{\frac{3}{4}}} + 4\sqrt{2}d^{\frac{3}{2}}L_0R^{\frac{3}{2}}T^{\frac{1}{4}} \\ &\quad + 2L_0^{1-q}\sqrt{dRT}^{\frac{3}{4}} + \sqrt{2}^{-1}\sqrt{dRL_0^{2q-1}}V_f^2T^{\frac{3}{4}}. \end{aligned} \quad (\text{B.24})$$

## B.5 Proof of Theorem 3.11

Since  $f_t(x) \in C^{1,1}$ , we know that  $|f_{\delta,t}(x) - f_t(x)| \leq \delta^2L_1$ . Following the same proof logic as that for proving (B.19), we obtain that

$$\mathbb{E}\left[\sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x)\right] \leq \frac{1}{2\eta}\|x_0 - x\|^2 + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2\right] + 2L_1\delta^2T. \quad (\text{B.25})$$

Substituting the bound in (B.20) into the above inequality, we obtain that

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x)\right] &\leq \frac{1}{2\eta}\|x_0 - x\|^2 + \frac{\eta}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \frac{8}{1-\alpha}L_0^2d^2\eta T \\ &\quad + 2L_1\delta^2T + \frac{d^2V_f^2}{1-\alpha}\frac{\eta}{\delta^2}T. \end{aligned} \quad (\text{B.26})$$

Since above 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  $\eta = \frac{R^{\frac{4}{3}}}{2\sqrt{2}L_0d^{\frac{2}{3}}T^{\frac{2}{3}}}$  and  $\delta = \frac{d^{\frac{1}{3}}R^{\frac{1}{3}}}{T^{\frac{1}{6}}}$  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

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^T f_t(x_t) - \sum_{t=0}^T f_t(x)\right] &\leq \sqrt{2}L_0d^{\frac{2}{3}}R^{\frac{2}{3}}T^{\frac{2}{3}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]R^{\frac{4}{3}}}{2\sqrt{2}L_0d^{\frac{2}{3}}T^{\frac{2}{3}}} + 4\sqrt{2}L_0d^{\frac{4}{3}}R^{\frac{4}{3}}T^{\frac{1}{3}} \\ &\quad + 2L_1d^{\frac{2}{3}}R^{\frac{2}{3}}T^{\frac{2}{3}} + (\sqrt{2}L_0)^{-1}d^{\frac{2}{3}}R^{\frac{2}{3}}V_f^2T^{\frac{2}{3}}. \end{aligned} \quad (\text{B.27})$$

The proof is complete.

## B.6 Proof of Theorem 3.14

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 Lemma 1.2.3 in<sup>46</sup>, we have the following inequality

$$\begin{aligned} f_{\delta,t}(x_{t+1}) &\leq f_{\delta,t}(x_t) + \langle \nabla f_{\delta,t}(x_t), x_{t+1} - x_t \rangle + \frac{L_{1,\delta}}{2} \|x_{t+1} - x_t\|^2 \\ &= f_{\delta,t}(x_t) - \eta \langle \nabla f_{\delta,t}(x_t), \tilde{g}_t(x_t) \rangle + \frac{L_{1,\delta}\eta^2}{2} \|\tilde{g}_t(x_t)\|^2 \\ &= f_{\delta,t}(x_t) - \eta \langle \nabla f_{\delta,t}(x_t), \Delta_t \rangle - \eta \|\nabla f_{\delta,t}(x_t)\|^2 + \frac{L_{1,\delta}\eta^2}{2} \|\tilde{g}_t(x_t)\|^2, \end{aligned} \quad (\text{B.28})$$

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 (B.28) and rearranging terms, we obtain that

$$\begin{aligned} \eta \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] &\leq \mathbb{E}[f_{\delta,t}(x_t)] - \mathbb{E}[f_{\delta,t}(x_{t+1})] + \frac{L_{1,\delta}\eta^2}{2} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \\ &\leq \mathbb{E}[f_{\delta,t}(x_t)] - \mathbb{E}[f_{\delta,t+1}(x_{t+1})] + \frac{L_{1,\delta}\eta^2}{2} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] + \mathbb{E}[f_{\delta,t+1}(x_{t+1})] - \mathbb{E}[f_{\delta,t}(x_{t+1})], \end{aligned} \quad (\text{B.29})$$

where the expectation is conditional on  $x_t$ . Then, we can further condition both sides of (B.29) on  $x_0$  without changing the sign of inequality, and then apply the tower rule of conditional expectation to make the expectation in (B.29) become full expectation. Telescoping the above inequality over  $t = 0, \dots, T-1$  and dividing both sides by  $\eta$ , according to Assumption 3.12, we obtain that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] &\leq \frac{\mathbb{E}[f_{\delta,0}(x_0)] - \mathbb{E}[f_{\delta,T}(x_T)]}{\eta} + \frac{L_{1,\delta}\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] + \frac{W_T}{\eta} \\ &\leq \frac{\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^*}{\eta} + \frac{L_{1,\delta}\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] + \frac{W_T}{\eta}, \end{aligned} \quad (\text{B.30})$$

where  $f_{\delta,T}^*$  is the lower bound of the smoothed function  $f_{\delta,T}(x)$ .  $f_{\delta,T}^*$  must exist because we assume 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  $t$ .

Next, we derive the bound on  $\sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2]$ . Summing the bound in (??) 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 3.12, we obtain that

$$\mathbb{E}\left[\sum_{t=0}^T \|\tilde{g}_t(x_t)\|^2\right] \leq \frac{1}{1-\alpha}\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \frac{16}{1-\alpha}L_0^2d^2T + \frac{2d^2}{1-\alpha}\frac{\widetilde{W}_T}{\delta^2}, \quad (\text{B.31})$$

Substituting this bound into the inequality (B.30), we obtain that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] &\leq \frac{\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^*}{\eta} + \frac{W_T}{\eta} + \frac{dL_0\eta}{2\delta} \frac{1}{1-\alpha} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2] \\ &\quad + \frac{dL_0\eta}{2\delta} \frac{16}{1-\alpha} L_0^2 d^2 T + \frac{dL_0\eta}{2\delta} \frac{2d^2}{1-\alpha} \frac{\widetilde{W}_T}{\delta^2}. \end{aligned} \quad (\text{B.32})$$

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  $\eta = \frac{\epsilon_f^{1.5}}{2\sqrt{2}L_0^2d^{1.5}T^{\frac{1}{2}}}$ . 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 (B.32), we finally obtain that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] &\leq 2\sqrt{2}L_0^2(\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T) \frac{d^{1.5}}{\epsilon_f^{1.5}} T^{\frac{1}{2}} + \frac{\epsilon_f^{\frac{1}{2}} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{2\sqrt{2}dT} \\ &\quad + 4\sqrt{2}L_0^2\epsilon_f^{\frac{1}{2}} d^{1.5} T^{\frac{1}{2}} + \frac{L_0^2 d^{1.5} \widetilde{W}_T}{\sqrt{2} \epsilon_f^{1.5} T^{\frac{1}{2}}}. \end{aligned} \quad (\text{B.33})$$

The proof is complete.

## B.7 Proof of Theorem 3.15

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 3.14 but replacing  $L_{1,\delta}$  with  $L_1$ , we obtain that

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2] \leq \frac{\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^*}{\eta} + \frac{L_1\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] + \frac{W_T}{\eta}. \quad (\text{B.34})$$

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

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] &= \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t) - \nabla f_{\delta,t}(x_t) + \nabla f_{\delta,t}(x_t)\|^2] \\ &\leq 2 \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t) - \nabla f_{\delta,t}(x_t)\|^2] + 2 \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{\delta,t}(x_t)\|^2]. \end{aligned} \quad (\text{B.35})$$

Substituting the bound in (B.31) into (B.34) and using the bound in (B.35), we obtain that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] &\leq 2 \frac{\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^*}{\eta} + 2 \frac{W_T}{\eta} + \frac{L_1}{1-\alpha} \mathbb{E}[\|\tilde{g}_0(x_0)\|^2] \eta + \frac{16L_1}{1-\alpha} L_0^2 d^2 \eta T \\ &\quad + \frac{2d^2 L_1 \widetilde{W}_T}{1-\alpha} \frac{\eta}{\delta^2} + 2d^2 L_1^2 \delta^2 T, \end{aligned} \quad (\text{B.36})$$

Choose  $\eta = \frac{1}{2\sqrt{2}L_0 d^{\frac{4}{3}} T^{\frac{1}{2}}}$  and  $\delta = \frac{1}{d^{\frac{1}{3}} T^{\frac{1}{4}}}$ . Then,  $\alpha = \frac{4d^2 L_0^2 \eta^2}{\delta^2} = \frac{1}{2\sqrt{T}} \leq \frac{1}{2}$  and  $\frac{1}{1-\alpha} \leq 2$ .

Substituting these results into the above inequality, we finally obtain that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] &\leq 4\sqrt{2}L_0(\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T) d^{\frac{4}{3}} T^{\frac{1}{2}} + \frac{L_1 \mathbb{E}[\|\tilde{g}_0(x_0)\|^2]}{\sqrt{2}L_0 d^{\frac{4}{3}} T^{\frac{1}{2}}} \\ &\quad + 8\sqrt{2}L_1 L_0 d^{\frac{2}{3}} T^{\frac{1}{2}} + \frac{\sqrt{2}L_1}{L_0} d^{\frac{4}{3}} \widetilde{W}_T + 2L_1^2 d^{\frac{4}{3}} T^{\frac{1}{2}}. \end{aligned} \quad (\text{B.37})$$

The proof is complete.

## B.8 Analysis for Projected SGD with Residual-Feedback Oracle

In this section, we analyze the regret of ZO with residual feedback for online bandit problem (P) where the objective functions  $\{f_t\}_{t=0,\dots,T-1}$  are non-convex and the problem is constrained. Specifically, we consider a mini-batch gradient estimator based on the proposed residual feedback

$$\tilde{G}_{t,M}(x_t) = \frac{1}{M} \sum_{i=1}^M \tilde{g}_{t,i}(x_t), \quad (\text{B.38})$$

where  $\tilde{g}_{t,i}(x_t) = \frac{d}{\delta}(f_t(x_t + \delta u_{t,i}) - f_{t-1}(x_{t-1} + \delta u_{t-1,i}))u_{t,i}$ , and  $u_{t,i}$  and  $u_{t-1,i}$  are independent from each other. Furthermore, we consider the update

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta \tilde{G}_{t,M}(x_t)). \quad (\text{B.39})$$

Mini-batched gradient estimator is commonly adopted in stochastic static constrained non-convex optimization problems<sup>47</sup> when stochastic gradient information is available. In this section, we study the case when the stochastic gradient cannot be obtained and we use estimator (B.38) to approximate such information. Estimator (B.38) requires to evaluate the values of objective function  $f_t$  at multiple points,  $x_t + \delta u_{t,i}$ , similar to the two-point gradient estimator (1.2). However, we note that studying the regret using zeroth-order gradient estimators to solve constrained general non-convex online optimization problems is of theoretical importance on its own. To the best of our knowledge, the regret in this setting has not been studied yet even for the two-point gradient estimator (1.2). Perhaps the most related work is<sup>40</sup>, which studied the two-point estimator's regret for solving constrained quasi-convex online optimization problems.

**Lemma B.1.** *(Second moment bound for mini-batch estimator) Assume that  $f_t \in C^{0,0}$  with Lipschitz constant  $L_0$  for all time  $t$ . Then, under the ZO update rule in (B.39), the second moment of the residual feedback (B.38) satisfies:*

$$\mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2] \leq \frac{4d^2 L_0^2 \eta^2}{\delta^2} \mathbb{E}[\|\tilde{G}_{t-1,M}(x_{t-1})\|^2] + D_t, \quad (\text{B.40})$$

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

*Proof.* By definition of the residual feedback (B.38), we have that

$$\mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2] = \frac{1}{M^2} \mathbb{E}[\|\sum_{i=1}^M \tilde{g}_{t,i}(x_t)\|^2] = \mathbb{E}[\|\tilde{g}_{t,i}(x_t)\|^2], \quad (\text{B.41})$$

where the second equality is due to the fact that  $\tilde{g}_{t,i}(x_t)$  are independent among each other given  $x_{t-1}$  and  $u_{t-1,i=1:M}$ , and that  $\mathbb{E}[\|\tilde{g}_{t,i}(x_t)\|] = \mathbb{E}[\|\tilde{g}_{t,j}(x_t)\|]$  for all  $i, j$ . Next, we provide the bound for the term  $\mathbb{E}[\|\tilde{g}_{t,j}(x_t)\|^2]$ . Following the same procedure as in the proof in B.3,

we can obtain the same inequality as (B.12)

$$\begin{aligned} \mathbb{E}[\|\tilde{g}_{t,i}(x_t)\|^2] &\leq \frac{4d^2L_0^2}{\delta^2}\mathbb{E}[\|x_t - x_{t-1}\|^2] + 16d^2L_0^2 \\ &\quad + \frac{2d^2}{\delta^2}\mathbb{E}[(f_t(x_{t-1} + \delta u_{t-1,i}) - f_{t-1}(x_{t-1} + \delta u_{t-1,i}))^2]. \end{aligned} \quad (\text{B.42})$$

According to update (B.39), we have that  $\|x_t - x_{t-1}\| = \|\Pi_{\mathcal{X}}[x_{t-1} - \eta\tilde{G}_{t-1,M}(x_{t-1})] - \Pi_{\mathcal{X}}[x_{t-1}]\| \leq \eta\|\tilde{G}_{t-1,M}(x_{t-1})\|$ . Using this bound with inequality (B.42), we complete the proof.  $\square$

**Lemma B.2.** *Given the mini-batch estimator (B.38), we have that*

$$\mathbb{E}[\|\tilde{G}_{t,M}(x_t) - \nabla f_{\delta,t}(x_t)\|^2] \leq \frac{1}{M}\mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2]. \quad (\text{B.43})$$

*Proof.* By definition of the residual feedback (B.38), we have that

$$\mathbb{E}[\|\tilde{G}_{t,M}(x_t) - \nabla f_{\delta,t}(x_t)\|^2] = \frac{1}{M}\mathbb{E}[\|\tilde{g}_{t,i}(x_t) - \nabla f_{\delta,t}(x_t)\|^2]. \quad (\text{B.44})$$

Since  $\mathbb{E}[\tilde{g}_{t,i}(x_t)] = \nabla f_{\delta,t}(x_t)$ , we get that  $\mathbb{E}[\|\tilde{g}_{t,i}(x_t) - \nabla f_{\delta,t}(x_t)\|^2] \leq \mathbb{E}[\|\tilde{g}_{t,i}(x_t)\|^2]$ . This is because  $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] = \mathbb{E}[\|X\|^2] - \mathbb{E}[X]^2 \leq \mathbb{E}[\|X\|^2]$  for any random vector  $X$ .

Therefore, we have that

$$\mathbb{E}[\|\tilde{G}_{t,M}(x_t) - \nabla f_{\delta,t}(x_t)\|^2] \leq \frac{1}{M}\mathbb{E}[\|\tilde{g}_{t,i}(x_t)\|^2]. \quad (\text{B.45})$$

Using equation (B.41) and inequality (B.45), we complete the proof.  $\square$

**Theorem B.3** (Nonconvex Lipschitz  $f_t$ ). *Let Assumptions 3.12 hold. Assume that  $f_t \in C^{0,0}$  with Lipschitz constant  $L_0$  over set  $\mathcal{X}_\delta$  and that  $f_t$  is bounded below by  $f_t^*$  for all  $t$ . Run ZO with residual feedback for  $T > (d\epsilon_f)^{-1}$  iterations with  $\eta = \epsilon_f^{\frac{3}{2}}(2\sqrt{2}L_0^2d^{\frac{3}{2}}T^{\frac{1}{2}})^{-1}$  and  $\delta = \epsilon_f L_0^{-1}$ . Then, we have that*

$$\begin{aligned} R_{g,\delta}^T &\leq 4\sqrt{2}L_0^2(\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T)d^{\frac{3}{2}}\epsilon_f^{-\frac{3}{2}}T^{\frac{1}{2}} \\ &\quad + \left(\frac{8}{M} + \frac{\sqrt{\epsilon_f}}{\sqrt{2dT^{\frac{1}{2}}}}\right)(\mathbb{E}[\|\tilde{G}_{0,M}(x_0)\|^2] \\ &\quad \quad \quad + 16d^2L_0^2T + \frac{2d^2L_0^2}{\epsilon_f^2}\tilde{W}_T) \end{aligned} \quad (\text{B.46})$$

*Asymptotically, when the batch size  $M$  is of order  $\mathcal{O}(\sqrt{dT}\sqrt{\epsilon_f^{-1}})$ , we have  $R_{g,\delta}^T = \mathcal{O}(d^{\frac{3}{2}}L_0^2\epsilon_f^{-\frac{3}{2}}(W_T + \tilde{W}_T T^{-1})T^{\frac{1}{2}} + d^{\frac{3}{2}}L_0^2\epsilon_f^{\frac{1}{2}}T^{\frac{1}{2}})$ .*

*Proof.* Let  $\bar{x}_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta \nabla f_{\delta,t}(x_t))$ . Recall that  $\mathcal{G}_{\eta,\delta,t}(x_t) = \frac{1}{\eta}(x_t - \Pi_{\mathcal{X}}(x_t - \eta \nabla f_{\delta,t}(x_t))) = \frac{1}{\eta}(x_t - \bar{x}_{t+1})$  and  $\tilde{\mathcal{G}}_{\eta,\delta,t}(x_t) = \frac{1}{\eta}(x_t - \Pi_{\mathcal{X}}(x_t - \eta \tilde{G}_{t,M}(x_t))) = \frac{1}{\eta}(x_t - x_{t+1})$ . 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 Lemma 1.2.3 in<sup>46</sup>, we have the following inequality

$$\begin{aligned}
f_{\delta,t}(x_{t+1}) &\leq f_{\delta,t}(x_t) + \langle \nabla f_{\delta,t}(x_t), x_{t+1} - x_t \rangle + \frac{L_{1,\delta}}{2} \|x_{t+1} - x_t\|^2 \\
&\leq f_{\delta,t}(x_t) + \langle \nabla f_{\delta,t}(x_t), x_{t+1} - x_t \rangle + \frac{L_{1,\delta}\eta^2}{2} \|\tilde{G}_{t,M}(x_t)\|^2, \\
&\leq f_{\delta,t}(x_t) + \langle \tilde{G}_{t,M}(x_t), x_{t+1} - x_t \rangle + \langle \nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t), x_{t+1} - x_t \rangle \\
&\quad + \frac{L_{1,\delta}\eta^2}{2} \|\tilde{G}_{t,M}(x_t)\|^2, \tag{B.47}
\end{aligned}$$

where the second inequality is due to the nonexpansive property of projection onto the convex set  $\mathcal{X}$ , i.e.,  $\|x_{t+1} - x_t\| = \|\Pi_{\mathcal{X}}[x_t - \eta \tilde{G}_{t,M}(x_t)] - \Pi_{\mathcal{X}}[x_t]\| \leq \eta \|\tilde{G}_{t,M}(x_t)\|$ . Next, we show that

$$\eta \|\tilde{\mathcal{G}}_{\eta,\delta,t}(x_t)\|^2 \leq \langle \tilde{G}_{t,M}(x_t), x_t - x_{t+1} \rangle. \tag{B.48}$$

To prove inequality (B.48), it is sufficient to show that

$$\frac{1}{\eta} \langle x_t - \eta \tilde{G}_{t,M}(x_t) - x_{t+1}, x_t - x_{t+1} \rangle \leq 0.$$

The above inequality is true, because  $x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta \tilde{G}_{t,M}(x_t))$ ,  $x_t \in \mathcal{X}$  and the set  $\mathcal{X}$  is convex. Then, combining inequalities (B.28) and (B.48), we obtain that

$$\begin{aligned}
\eta \|\tilde{\mathcal{G}}_{\eta,\delta,t}(x_t)\|^2 &\leq f_{\delta,t}(x_t) - f_{\delta,t}(x_{t+1}) + \langle \nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t), x_{t+1} - x_t \rangle + \frac{L_{1,\delta}\eta^2}{2} \|\tilde{G}_{t,M}(x_t)\|^2 \\
&\leq f_{\delta,t}(x_t) - f_{\delta,t}(x_{t+1}) + \langle \nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t), x_{t+1} - \bar{x}_{t+1} \rangle \\
&\quad + \langle \nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t), \bar{x}_{t+1} - x_t \rangle + \frac{L_{1,\delta}\eta^2}{2} \|\tilde{G}_{t,M}(x_t)\|^2. \tag{B.49}
\end{aligned}$$

In addition, we have that

$$\begin{aligned}
\langle \nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t), x_{t+1} - \bar{x}_{t+1} \rangle &\leq \frac{\eta}{2} \|\nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t)\|^2 + \frac{1}{2\eta} \|x_{t+1} - \bar{x}_{t+1}\|^2 \\
&\leq \eta \|\nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t)\|^2. \tag{B.50}
\end{aligned}$$

Combining inequalities (B.49) and (B.50), and taking expectation over  $u_{t,i=1:M}$  on both sides, we have that

$$\begin{aligned} \eta \mathbb{E}[\|\tilde{\mathcal{G}}_{\eta,\delta,t}(x_t)\|^2] &\leq \mathbb{E}[f_{\delta,t}(x_t) - f_{\delta,t}(x_{t+1})] + \eta \mathbb{E}[\|\nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t)\|^2] \\ &\quad + \frac{L_{1,\delta}\eta^2}{2} \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2]. \end{aligned} \quad (\text{B.51})$$

Furthermore, we have that

$$\begin{aligned} \|\mathcal{G}_{\eta,\delta,t}(x_t)\|^2 &\leq 2\|\tilde{\mathcal{G}}_{\eta,\delta,t}(x_t)\|^2 + 2\|\tilde{\mathcal{G}}_{\eta,\delta,t}(x_t) - \mathcal{G}_{\eta,\delta,t}(x_t)\|^2 \\ &\leq 2\|\tilde{\mathcal{G}}_{\eta,\delta,t}(x_t)\|^2 + 2\|\nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t)\|^2. \end{aligned} \quad (\text{B.52})$$

Combining inequalities (B.51) and (B.52), we obtain that

$$\begin{aligned} \mathbb{E}[\|\mathcal{G}_{\eta,\delta,t}(x_t)\|^2] &\leq \frac{2}{\eta} \mathbb{E}[(f_{\delta,t}(x_t) - f_{\delta,t}(x_{t+1}))] + 4\mathbb{E}[\|\nabla f_{\delta,t}(x_t) - \tilde{G}_{t,M}(x_t)\|^2] \\ &\quad + L_{1,\delta}\eta \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2]. \end{aligned} \quad (\text{B.53})$$

According to Lemma B.2, we get that

$$\mathbb{E}[\|\mathcal{G}_{\eta,\delta,t}(x_t)\|^2] \leq \frac{2}{\eta} \mathbb{E}[(f_{\delta,t}(x_t) - f_{\delta,t}(x_{t+1}))] + \frac{4}{M} \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2] + L_{1,\delta}\eta \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2].$$

Telescoping the above inequalities from  $t = 0$  to  $T - 1$ , and recalling that  $L_{1,\delta} = d\delta^{-1}L_0$ , we have that

$$\begin{aligned} R_{g,\delta}^T &\leq \frac{2}{\eta} \mathbb{E}[f_{\delta,0}(x_0) - f_{\delta,T}^*] + \frac{2}{\eta} \sum_{t=1}^T |f_{\delta,t}(x_t) - f_{\delta,t-1}(x_t)| + \frac{4}{M} \sum_{t=0}^T \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2] \\ &\quad + dL_0 \frac{\eta}{\delta} \sum_{t=0}^T \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2] \end{aligned} \quad (\text{B.54})$$

Next, we use the results in Lemma B.1 to obtain the bound on  $\sum_{t=0}^T \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2]$ .

Specifically, telescoping the second moment bound in (B.40) over  $t = 1, 2, \dots, T$ , adding  $\mathbb{E}[\|\tilde{G}_{0,M}(x_0)\|^2]$  on both sides, and adding  $\frac{4d^2L_0^2\eta^2}{\delta^2} \mathbb{E}[\|\tilde{G}_{T,M}(x_T)\|^2]$  to the right hand side,

we obtain that

$$\begin{aligned} \sum_{t=0}^T \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2] &\leq \frac{1}{1-\alpha} \mathbb{E}[\|\tilde{G}_{0,M}(x_0)\|^2] + \frac{16}{1-\alpha} d^2 L_0^2 T \\ &\quad + \frac{2d^2}{1-\alpha} \frac{1}{\delta^2} \sum_{t=1}^T \mathbb{E}[(f_t(x_{t-1} + \delta u_{t-1,i}) - f_{t-1}(x_{t-1} + \delta u_{t-1,i}))^2], \end{aligned} \quad (\text{B.55})$$

where  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2}$ . Combining inequalities (B.54) and (B.55), and according to Assumption 3.12, we get that

$$R_{g,\delta}^T \leq \frac{2}{\eta} \mathbb{E}[f_{\delta,0}(x_0) - f_{\delta,T}^*] + \frac{2}{\eta} W_T \\ + \left(\frac{4}{M} + dL_0 \frac{\eta}{\delta}\right) \left(\frac{1}{1-\alpha} \mathbb{E}[\|\tilde{G}_{0,M}(x_0)\|^2]\right) + \frac{16}{1-\alpha} d^2 L_0^2 T + \frac{2d^2}{1-\alpha} \frac{1}{\delta^2} \widetilde{W}_T. \quad (\text{B.56})$$

To fulfill the requirement that  $|f_t(x) - f_{\delta,t}(x)| \leq \epsilon_f$ , according to Lemma 2.2, 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 (B.56), we finally obtain that

$$R_{g,\delta}^T \leq 4\sqrt{2}L_0^2 (\mathbb{E}[f_{\delta,0}(x_0) - f_{\delta,T}^*] + W_T) \frac{d^{1.5}}{\epsilon_f^{1.5}} T^{\frac{1}{2}} \\ + \left(\frac{8}{M} + \frac{\sqrt{\epsilon_f}}{\sqrt{2dT^{\frac{1}{2}}}}\right) (\mathbb{E}[\|\tilde{G}_{0,M}(x_0)\|^2]) + 16d^2 L_0^2 T + \frac{2d^2 L_0^2}{\epsilon_f^2} \widetilde{W}_T. \quad (\text{B.57})$$

The proof is complete.  $\square$

Next, we assume that the objectives  $f_t$  in (P) are non-convex and smooth. In this case, we study the gradient mapping

$$\mathcal{G}_\eta(x_t) = \frac{1}{\eta} (x_t - \Pi_{\mathcal{X}}(x_t - \eta \nabla f_t(x_t))),$$

and the corresponding regret  $R_g^T := \sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{G}_\eta(x_t)\|^2]$ . Specifically, we provide the following regret bound for ZO with residual-feedback.

**Theorem B.4** (Nonconvex smooth  $f_t$ ). *Let Assumptions 3.12 hold. Assume that  $f_t \in C^{0,0} \cap C^{1,1}$  with Lipschitz constant  $L_0$  and smoothness constant  $L_1$  over set  $\mathcal{X}_\delta$  and that  $f_t$  is bounded below by  $f_t^*$  for all  $t$ . Run ZO with residual feedback for  $T > d$  iterations with  $\eta = (2\sqrt{2}L_0 d T^{\frac{1}{2}})^{-1}$  and  $\delta = (d^{\frac{1}{4}} T^{\frac{1}{4}})^{-1}$ . Then,*

$$R_g^T \leq 8\sqrt{2}L_0 (\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T) d T^{\frac{1}{2}} \\ + 2dL_1^2 T^{\frac{1}{2}} + \left(\frac{16}{M} + \frac{\sqrt{2}L_1}{L_0 d T^{\frac{1}{2}}}\right) (\mathbb{E}[\|\tilde{G}_{0,M}(x_0)\|^2]) \\ + 16d^2 L_0^2 T + 2d^{2.5} \widetilde{W}_T T^{\frac{1}{2}}. \quad (\text{B.58})$$

Asymptotically, when the batch size  $M$  is in order of  $\mathcal{O}(dT^{\frac{1}{2}})$ , we have that  $R_g^T = \mathcal{O}(dL_0W_T T^{\frac{1}{2}} + d^{\frac{3}{2}}L_1L_0^{-1}\widetilde{W}_T)$ .

*Proof.* 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 3.14 but replacing  $L_{1,\delta}$  with  $L_1$ , we obtain that

$$R_{g,\delta}^T \leq \frac{2}{\eta} \mathbb{E}[f_{\delta,0}(x_0) - f_{\delta,T}^*] + \frac{2}{\eta} \sum_{t=1}^T |f_{\delta,t}(x_t) - f_{\delta,t-1}(x_t)| + \left(\frac{4}{M} + L_1\eta\right) \sum_{t=0}^T \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2] \quad (\text{B.59})$$

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

$$\begin{aligned} R_g^T &= \sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{G}_{\eta,t} - \mathcal{G}_{\eta,\delta,t} + \mathcal{G}_{\eta,\delta,t}\|^2] \\ &\leq 2R_{g,\delta}^T + 2 \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_t(x_t) - \nabla f_{\delta,t}(x_t)\|^2] \leq 2R_{g,\delta}^T + 2d^2L_1^2\delta^2T \end{aligned} \quad (\text{B.60})$$

Combining inequalities (B.34) and (B.35), and according to Assumption 3.12, we obtain that

$$R_g^T \leq \frac{4}{\eta} \mathbb{E}[f_{\delta,0}(x_0) - f_{\delta,T}^*] + \frac{4}{\eta} W_T + \left(\frac{8}{M} + 2L_1\eta\right) \sum_{t=0}^T \mathbb{E}[\|\tilde{G}_{t,M}(x_t)\|^2] + 2d^2L_1^2\delta^2T \quad (\text{B.61})$$

Furthermore, using the bound in (B.55), we have that

$$\begin{aligned} R_g^T &\leq \frac{4}{\eta} \mathbb{E}[f_{\delta,0}(x_0) - f_{\delta,T}^*] + \frac{4}{\eta} W_T + 2d^2L_1^2\delta^2T \\ &\quad + \left(\frac{8}{M} + 2L_1\eta\right) \left(\frac{1}{1-\alpha} \mathbb{E}[\|\tilde{G}_{0,M}(x_0)\|^2] + \frac{16}{1-\alpha} d^2L_0^2T + \frac{2d^2}{1-\alpha} \frac{1}{\delta^2} \widetilde{W}_T\right). \end{aligned} \quad (\text{B.62})$$

Choose  $\eta = \frac{1}{2\sqrt{2}L_0dT^{\frac{1}{2}}}$  and  $\delta = \frac{1}{d^{\frac{1}{4}}T^{\frac{1}{4}}}$ . Then,  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2} = \frac{\sqrt{d}}{2\sqrt{T}} \leq \frac{1}{2}$  and  $\frac{1}{1-\alpha} \leq 2$  when  $T > d$ . Substituting these results into the above inequality, we finally obtain that

$$\begin{aligned} R_g^T &\leq 8\sqrt{2}L_0(\mathbb{E}[f_{\delta,0}(x_0)] - f_{\delta,T}^* + W_T)dT^{\frac{1}{2}} + 2dL_1^2T^{\frac{1}{2}} \\ &\quad + \left(\frac{16}{M} + \frac{\sqrt{2}L_1}{L_0dT^{\frac{1}{2}}}\right) (\mathbb{E}[\|\tilde{G}_{0,M}(x_0)\|^2] + 16d^2L_0^2T + 2d^{2.5}\widetilde{W}_T T^{\frac{1}{2}}) \end{aligned} \quad (\text{B.63})$$

The proof is complete.  $\square$

## B.9 Proof of Lemma 3.18

Consider the case when  $F_t(x, \xi) \in C^{0,0}$  with  $L_0(\xi)$ . According to (2.4), we have that

$$\begin{aligned} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] &= \mathbb{E}\left[\frac{d^2}{\delta^2} (F_t(x_t + \delta u_t, \xi_t) - F_{t-1}(x_{t-1} + \delta u_{t-1}, \xi_{t-1}))^2 \|u_t\|^2\right] \\ &\leq \frac{2d^2}{\delta^2} \mathbb{E}[(F_t(x_t + \delta u_t, \xi_t) - F_t(x_{t-1} + \delta u_{t-1}, \xi_t))^2 \|u_t\|^2] \\ &\quad + \frac{2d^2}{\delta^2} \mathbb{E}[(F_t(x_{t-1} + \delta u_{t-1}, \xi_t) - F_{t-1}(x_{t-1} + \delta u_{t-1}, \xi_{t-1}))^2 \|u_t\|^2]. \end{aligned} \tag{B.64}$$

Using the bound in Assumption 2.10, we get that  $\frac{2d^2}{\delta^2} \mathbb{E}[(F_t(x_{t-1} + \delta u_{t-1}, \xi_t) - F_{t-1}(x_{t-1} + \delta u_{t-1}, \xi_{t-1}))^2 \|u_t\|^2] \leq \frac{2d^2}{\delta^2} V_{f,\xi}^2$ . In addition, adding and subtracting  $F_t(x_{t-1} + \delta u_t, \xi_t)$  in  $(F_t(x_t + \delta u_t, \xi_t) - F_t(x_{t-1} + \delta u_{t-1}, \xi_t))^2$  in above inequality, we obtain that

$$\begin{aligned} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] &\leq \frac{4d^2}{\delta^2} \mathbb{E}[(F_t(x_t + \delta u_t, \xi_t) - F_t(x_{t-1} + \delta u_t, \xi_t))^2 \|u_t\|^2] \\ &\quad + \frac{4d^2}{\delta^2} \mathbb{E}[(F_t(x_{t-1} + \delta u_t, \xi_t) - F_t(x_{t-1} + \delta u_{t-1}, \xi_t))^2 \|u_t\|^2] \\ &\quad + \frac{2d^2}{\delta^2} \mathbb{E}[(F_t(x_{t-1} + \delta u_{t-1}, \xi_t) - F_{t-1}(x_{t-1} + \delta u_{t-1}, \xi_{t-1}))^2]. \end{aligned} \tag{B.65}$$

By Lipschitz continuity of  $F_t(\cdot; \xi_t)$ , we can bound the first two items on the right hand side of above inequality following the same procedure after inequality (B.11) and get that

$$\begin{aligned} \mathbb{E}[\|\tilde{g}_t(x_t)\|^2] &\leq \frac{4d^2 L_0^2 \eta^2}{\delta^2} \mathbb{E}[\|\tilde{g}_t(x_{t-1})\|^2] + 16L_0^2 d^2 \\ &\quad + \frac{2d^2}{\delta^2} \mathbb{E}[(F_t(x_{t-1} + \delta u_{t-1}, \xi_t) - F_{t-1}(x_{t-1} + \delta u_{t-1}, \xi_{t-1}))^2]. \end{aligned}$$

The proof is complete.

## B.10 Residual-Feedback Convex Optimization with Unit Sphere Sampling

Consider the online bandit optimization problem (P) with convex objective functions and a compact constraint set  $\mathcal{X}$ . In this section, we assume that the objective function  $f(x)$

cannot be queried outside the constraint set  $\mathcal{X}$ . To ensure that the iterates are confined within the constraint set  $\mathcal{X}$ , we consider the update

$$x_{t+1} = \Pi_{(1-\xi)\mathcal{X}}(x_t - \eta\tilde{g}_t(x_t)), \quad (\text{B.66})$$

where the set  $(1-\xi)\mathcal{X} := \{(1-\xi)x : \forall x \in \mathcal{X}\}$  is a shrunk version of the original constraint set  $\mathcal{X}$ . The goal is to select a parameter  $\xi$  so that for every  $x_\xi \in (1-\xi)\mathcal{X}$ ,  $x_\xi + \delta u \in \mathcal{X}$  for every  $u \in \mathbb{S}$ . To achieve this, we first make the following assumption that is inspired by<sup>12;48</sup>.

**Assumption B.5.** *There exist constants  $r$  and  $\bar{r}$  such that  $r\mathbb{B} \subset \mathcal{X} \subset \bar{r}\mathbb{B}$ .*

Then, we have the following lemma.

**Lemma B.6.** *If the parameter  $\xi$  satisfies  $1 \geq \xi \geq \frac{\delta}{r}$ , then for every iterate  $x_t$  obtained using (B.66), we have that  $x_t + \delta u_t \in \mathcal{X}$  for all  $u_t \in \mathbb{S}$ .*

*Proof.* When  $1 \geq \xi \geq \frac{\delta}{r}$ , we get that  $\|\delta u\| \leq \xi r$ . Therefore, there exists  $x' \in r\mathbb{B} \subset \mathcal{X}$  such that the vector  $\delta u = \xi x'$ . Since  $x_t \in (1-\xi)\mathcal{X}$ , there exists  $x \in \mathcal{X}$  such that  $x_t = (1-\xi)x$ , and there exists  $x' \in \mathcal{X}$  such that  $\delta u = \xi x'$ . As a result, we have that  $x_t + \delta u = (1-\xi)x + \xi x' \in \mathcal{X}$ . This is because set  $\mathcal{X}$  is convex.  $\square$

Next, we study the regret  $R_T := \mathbb{E}\left[\sum_{t=0}^{T-1} f_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x)\right]$  achieved by executing the online update (B.66). We do so in the following two steps. First, in Lemma B.7, we provide an upper bound on the difference between the optimal solution that lies in the set  $(1-\xi)\mathcal{X}$  and the one that lies in the set  $\mathcal{X}$ , i.e.,  $\min_{x \in (1-\xi)\mathcal{X}} \sum_{t=0}^{T-1} f_t(x) - \min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x)$ ; Then, in Theorem B.8, we bound the regret defined by the expected difference between the function values achieved by running the update (B.66) and the term  $\min_{x \in (1-\xi)\mathcal{X}} \sum_{t=0}^{T-1} f_t(x)$ , i.e.,  $\mathbb{E}\left[\sum_{t=0}^{T-1} f_t(x_t) - \min_{x \in (1-\xi)\mathcal{X}} \sum_{t=0}^{T-1} f_t(x)\right]$ . Adding the two bounds above, we can complete the proof.

In the following lemma we provide a bound on  $\min_{x \in (1-\xi)\mathcal{X}} \sum_{t=0}^{T-1} f_t(x) - \min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x)$ .

**Lemma B.7.** *If the function  $f_t$  is convex and  $f_t \in C^{0,0}$  with Lipschitz constant  $L_0$  for all time  $t$ , we have that*

$$\sum_{t=0}^{T-1} f_t(x_\xi^*) - \sum_{t=0}^{T-1} f_t(x^*) \leq \bar{r}L_0\xi T, \quad (\text{B.67})$$

where  $x_\xi^* = \arg \min_{x \in (1-\xi)\mathcal{X}} \sum_{t=0}^{T-1} f_t(x)$  and  $x^* = \arg \min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x)$ .

*Proof.* Since  $x^* \in \mathcal{X}$ , we have that  $(1-\xi)x^* \in (1-\xi)\mathcal{X}$ . Moreover, since  $x_\xi^*$  is the minimizer in the set  $(1-\xi)\mathcal{X}$ , we get that

$$\sum_{t=0}^{T-1} f_t(x_\xi^*) \leq \sum_{t=0}^{T-1} f_t((1-\xi)x^*). \quad (\text{B.68})$$

Also, since  $f_t$  is convex and  $(1-\xi)x^* = (1-\xi)x^* + \xi 0$ , we have that

$$\begin{aligned} f_t((1-\xi)x^*) &\leq (1-\xi)f_t(x^*) + \xi f_t(0) \\ &\leq (1-\xi)f_t(x^*) + \xi f_t(x^*) - \xi f_t(x^*) + \xi f_t(0) \\ &\leq f_t(x^*) + \xi L_0 \|x^*\| \leq f_t(x^*) + \bar{r}L_0\xi, \end{aligned} \quad (\text{B.69})$$

where the last inequality is due to the fact that  $x^* \in \mathcal{X} \subset \bar{r}\mathbb{B}$ . Summing the inequality (B.69) over time, we obtain that

$$\sum_{t=0}^{T-1} f_t((1-\xi)x^*) - \sum_{t=0}^{T-1} f_t(x^*) \leq \bar{r}L_0\xi T. \quad (\text{B.70})$$

Adding up the inequalities (B.68) and (B.70) and rearranging terms completes the proof.  $\square$

Next, we study the regret  $\mathbb{E} \left[ \sum_{t=0}^{T-1} f_t(x_t) - \min_{x \in (1-\xi)\mathcal{X}} \sum_{t=0}^{T-1} f_t(x) \right]$  following similar steps as in Section 3.2. Using above lemmas, we can obtain the main theorem for online convex optimization using (B.66).

**Theorem B.8** (Regret for Convex Lipschitz  $f_t$ ). *Let Assumption 3.8 hold. Assume that  $f_t \in C^{0,0}$  is convex with Lipschitz constant  $L_0$  for all  $t$ . Run ZO with residual feedback for*

$T > \bar{r}^2 L_0^{2q}$  iterations with  $\eta = \frac{\bar{r}^{\frac{3}{2}}}{2\sqrt{2}L_0\sqrt{dT}^{\frac{3}{4}}}$  and  $\delta = \frac{\sqrt{\bar{r}d}}{L_0^q T^{\frac{1}{4}}}$ , where  $q \in \mathbb{R}$  is a user-specified parameter. Then, we have that

$$\begin{aligned} R_T \leq & 4\sqrt{2\bar{r}d}L_0T^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]\bar{r}^{\frac{3}{2}}}{2\sqrt{2d}L_0T^{\frac{3}{4}}} + 8\sqrt{2d}^{\frac{3}{2}}L_0\bar{r}^{\frac{3}{2}}T^{\frac{1}{4}} \\ & + (2 + \frac{\bar{r}}{r})L_0^{1-q}\sqrt{d\bar{r}}T^{\frac{3}{4}} + \frac{\sqrt{2d\bar{r}}V_f^2}{L_0^{1-2q}}T^{\frac{3}{4}}. \end{aligned} \quad (\text{B.71})$$

Asymptotically, we have  $R_T = \mathcal{O}((L_0 + L_0^{1-q} + L_0^{2q-1}V_f^2)\sqrt{d\bar{r}}T^{\frac{3}{4}})$ .

*Proof.* First, we provide a bound on the regret that compares the sum of the function values obtained using (B.66) to that obtained for the optimizer  $x_\xi^*$  in the shrunk constraint set  $(1 - \xi)\mathcal{X}$ , i.e.,  $\mathbb{E}\left[\sum_{t=0}^{T-1} f_t(x_t) - \min_{x \in (1-\xi)\mathcal{X}} \sum_{t=0}^{T-1} f_t(x)\right]$ . Since  $f_{\delta,t}(x)$  is convex for all  $t$ , we conclude that

$$f_{\delta,t}(x_t) - f_{\delta,t}(x) \leq \langle \nabla f_{\delta,t}(x_t), x_t - x \rangle, \text{ for all } x \in (1 - \xi)\mathcal{X}. \quad (\text{B.72})$$

Adding and subtracting  $\tilde{g}_t(x_t)$  to  $\nabla f_{\delta,t}(x_t)$  in inequality (B.72), and taking the expectation of both sides with respect to  $u_t$ , we obtain that

$$\mathbb{E}[f_{\delta,t}(x_t) - f_{\delta,t}(x)] \leq \mathbb{E}[\langle \tilde{g}_t(x_t), x_t - x \rangle]. \quad (\text{B.73})$$

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

$$\begin{aligned} \|x_{t+1} - x\|^2 &= \|\Pi_{(1-\xi)\mathcal{X}}[x_t - \eta\tilde{g}(x_t)] - \Pi_{(1-\xi)\mathcal{X}}[x]\|^2 \\ &\leq \|x_t - \eta\tilde{g}(x_t) - x\|^2 \\ &= \|x_t - x\|^2 - 2\eta\langle \tilde{g}_t(x_t), x_t - x \rangle + \eta^2\|\tilde{g}_t(x_t)\|^2. \end{aligned} \quad (\text{B.74})$$

Rearranging the terms in inequality (B.74) yields

$$\langle \tilde{g}_t(x_t), x_t - x \rangle \leq \frac{1}{2\eta}(\|x_t - x\|^2 - \|x_{t+1} - x\|^2) + \frac{\eta}{2}\|\tilde{g}_t(x_t)\|^2. \quad (\text{B.75})$$

Taking the expectation of both sides of inequality (B.75) with respect to  $u_t$  and substituting the resulting bound into (B.73), we obtain that

$$\mathbb{E}\left[\sum_{t=0}^{T-1} f_{\delta,t}(x_t) - \sum_{t=0}^{T-1} f_{\delta,t}(x)\right] \leq \frac{1}{2\eta}\|x_0 - x\|^2 + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=0}^{T-1} \|\tilde{g}_t(x_t)\|^2\right]. \quad (\text{B.76})$$

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

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^{T-1} f_t(x_t) - \sum_{t=0}^{T-1} f_t(x)\right] &= \mathbb{E}\left[\sum_{t=0}^{T-1} f_{\delta,t}(x_t) - \sum_{t=0}^{T-1} f_{\delta,t}(x)\right] \\ &\quad + \mathbb{E}\left[\sum_{t=0}^{T-1} (f_t(x_t) - f_{\delta,t}(x_t)) - \sum_{t=0}^{T-1} (f_t(x) - f_{\delta,t}(x))\right] \\ &\leq \frac{1}{2\eta}\|x_0 - x\|^2 + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=0}^{T-1} \|\tilde{g}_t(x_t)\|^2\right] + 2L_0\delta T, \end{aligned} \quad (\text{B.77})$$

where we have made use of the bound in (B.76). Telescoping the bound in (??) over  $t = 1, 2, \dots, T-1$ , adding  $\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]$  to both sides, and adding  $\frac{4d^2L_0^2\eta^2}{\delta^2}\mathbb{E}[\|\tilde{g}_{T-1}(x_{T-1})\|^2]$  to the right hand side, we obtain that

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \|\tilde{g}_t(x_t)\|^2\right] \leq \frac{1}{1-\alpha}\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \frac{16}{1-\alpha}d^2L_0^2T + \frac{2d^2V_f^2}{1-\alpha}\frac{1}{\delta^2}T, \quad (\text{B.78})$$

where  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2}$ . Substituting the bound in (B.78) into (B.77) yields

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^{T-1} f_t(x_t) - \sum_{t=0}^{T-1} f_t(x)\right] &\leq \frac{1}{2\eta}\|x_0 - x\|^2 + \frac{\eta}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \frac{16}{1-\alpha}d^2L_0^2\eta T \\ &\quad + 2L_0\delta T + \frac{2d^2V_f^2}{1-\alpha}\frac{\eta}{\delta^2}T. \end{aligned} \quad (\text{B.79})$$

Since inequality (B.79) holds for all  $x \in (1-\xi)\mathcal{X}$ , we can replace  $x$  in (B.79) with  $x_\xi^*$ . Furthermore, using Lemma B.7, we have that

$$\sum_{t=0}^{T-1} f_t(x_\xi^*) - \sum_{t=0}^{T-1} f_t(x^*) \leq \bar{r}L_0\xi T. \quad (\text{B.80})$$

Summing inequalities (B.79) and (B.80), we obtain

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^{T-1} f_t(x_t) - \sum_{t=0}^{T-1} f_t(x^*)\right] &\leq \frac{1}{2\eta}\|x_0 - x_\xi^*\|^2 + \frac{\eta}{2(1-\alpha)}\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \frac{16}{1-\alpha}d^2L_0^2\eta T \\ &\quad + 2L_0\delta T + \frac{2d^2V_f^2}{1-\alpha}\frac{\eta}{\delta^2}T + \bar{r}L_0\xi T, \end{aligned} \quad (\text{B.81})$$

where  $\|x_0 - x_\xi^*\|^2 \leq 4\bar{r}^2$ . According to Lemma B.6, we can select  $\xi = \frac{\delta}{\bar{r}}$  to guarantee that all iterates  $x_t + \delta u_t \in \mathcal{X}$  for all  $u_t \in \mathbb{S}$ . Furthermore, let  $\eta = \frac{\bar{r}^{\frac{3}{2}}}{2\sqrt{2}L_0\sqrt{dT}^{\frac{3}{4}}}$  and  $\delta = \frac{\sqrt{\bar{r}d}}{L_0^q T^{\frac{1}{4}}}$ , where

$q \in \mathbb{R}$  is a user-specified parameter. Then,  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2} = \frac{1}{2T}\bar{r}^2L_0^{2q} \leq \frac{1}{2}$  when  $T \geq \bar{r}^2L_0^{2q}$ .

Substituting these parameter values into (B.81), we obtain that

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^{T-1} f_t(x_t) - \sum_{t=0}^{T-1} f_t(x^*)\right] &\leq 4\sqrt{2\bar{r}d}L_0T^{\frac{3}{4}} + \frac{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2]\bar{r}^{\frac{3}{2}}}{2\sqrt{2d}L_0T^{\frac{3}{4}}} + 8\sqrt{2d^{\frac{3}{2}}L_0\bar{r}^{\frac{3}{2}}T^{\frac{1}{4}}} \\ &\quad + \left(2 + \frac{\bar{r}}{r}\right)L_0^{1-q}\sqrt{d\bar{r}}T^{\frac{3}{4}} + L_0^{2q-1}\sqrt{2d\bar{r}}V_f^2T^{\frac{3}{4}}. \end{aligned} \quad (\text{B.82})$$

The proof is complete.  $\square$

## B.11 Proof of the Second Moment Bound (3.6)

Let  $\alpha = \frac{4d^2L_0^2\eta^2}{\delta^2}$ , using (3.3), we have that

$$\mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \leq \alpha^t\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \sum_{j=1}^t \alpha^{t-j}D_j, \text{ for all } t \geq 1. \quad (\text{B.83})$$

According to Assumption 3.8, we obtain that

$$\mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \leq \alpha^t\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \sum_{j=1}^t \alpha^{t-j}\left(16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2\right), \text{ for all } t \geq 1. \quad (\text{B.84})$$

Therefore, we get that

$$\mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \leq \max\left\{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2], \dots, \alpha^t\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \sum_{j=1}^t \alpha^{t-j}\left(16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2\right), \dots\right\}.$$

Next, we show that this inequality is equivalent to

$$\mathbb{E}[\|\tilde{g}_t(x_t)\|^2] \leq \max\left\{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2], \frac{1}{1-\alpha}\left(16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2\right)\right\}. \quad (\text{B.85})$$

To see this, observe that the sequence  $\left\{\mathbb{E}[\|\tilde{g}_0(x_0)\|^2], \dots, \alpha^t\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \sum_{j=1}^t \alpha^{t-j}\left(16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2\right), \dots\right\}$  is monotonic. This is because if  $\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] \geq \alpha\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + 16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2$ , then we can multiply both sides by  $\alpha$  and add  $16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2$  to both sides and get that  $\alpha\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + 16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2 \geq \alpha^2\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + \alpha\left(16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2\right) + \left(16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2\right)$ . Using mathematical induction we can show that the sequence is monotonically non-increasing. Similarly, if  $\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] \leq \alpha\mathbb{E}[\|\tilde{g}_0(x_0)\|^2] + 16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2$ , then we can show that the sequence is monotonically non-decreasing and converges to  $\frac{1}{1-\alpha}\left(16L_0^2d^2 + \frac{2d^2}{\delta^2}V_f^2\right)$ .

Therefore, the proof is complete.

# Appendix C

## Proofs for Chapter 4

### C.1 Proof of Lemma 4.6

**Lemma 3.1.** Given Assumptions 4.4 and 4.5, we have that  $\|\mu^k(N_c) - J(\theta_k + \delta u_k, \xi_k)\mathbf{1}\| \leq \rho_W^{N_c} \sqrt{N} (J_u - J_l)$ , where  $\rho_W = \|W - \frac{1}{N}\mathbf{1}\mathbf{1}^T\| < 1$ .

*Proof.* First, we show that  $\frac{1}{N}\mathbf{1}^T \mu^k(m) = J(\theta_k + \delta u_k, \xi_k)$  for all  $m = 0, 1, \dots$ . Note that the consensus step  $\mu_i^k(m+1) = \sum_{j \in \mathcal{N}_i} W_{ij} \mu_j^k(m)$  (line 12 in Algorithm 1) can be equivalently written in a compact form as  $\mu^k(m+1) = W\mu^k(m)$ . Therefore, we have that

$$\frac{1}{N}\mathbf{1}^T \mu^k(m+1) = \frac{1}{N}\mathbf{1}^T W \mu^k(m) = \frac{1}{N}\mathbf{1}^T \mu^k(m), \quad (\text{C.1})$$

where the second equality is due to Assumption 4.4, that the matrix  $W$  is doubly stochastic. Extending equality (C.1) from  $m$  to 0, we obtain that  $\frac{1}{N}\mathbf{1}^T \mu^k(m) = \frac{1}{N}\mathbf{1}^T \mu^k(0) = \frac{1}{N}\mathbf{1}^T J_i(\theta_k + \delta u_k, \xi_k) = J(\theta_k + \delta u_k, \xi_k)$ , for  $m = 0, 1, \dots$

Next, we show that  $\|\mu^k(m) - \frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(m)\| \leq \|W - \frac{1}{N}\mathbf{1}\mathbf{1}^T\|^m \|\mu^k(0) - \frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(0)\|$ , for  $m = 1, 2, \dots$ . To see this, we have that

$$\begin{aligned} \|\mu^k(m+1) - \frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(m+1)\| &= \|W\mu^k(m) - \frac{1}{N}\mathbf{1}\mathbf{1}^T W\mu^k(m)\| \\ &= \|W\mu^k(m) - \frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(m)\|, \end{aligned} \quad (\text{C.2})$$

where the second equality is due to Assumption 4.4. According to (C.2), we have that  $\|\mu^k(m+1) - \frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(m+1)\| = \|(W - \frac{1}{N}\mathbf{1}\mathbf{1}^T)\mu^k(m)\| = \|(W - \frac{1}{N}\mathbf{1}\mathbf{1}^T)(\mu^k(m) - \frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(m))\|$ .

This is because  $(W - \frac{1}{N}\mathbf{1}\mathbf{1}^T)\frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(m) = 0$ . Therefore, we get that

$$\begin{aligned} \|\mu^k(m+1) - \frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(m+1)\| &\leq \|W - \frac{1}{N}\mathbf{1}\mathbf{1}^T\| \|\mu^k(m) - \frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(m)\| \\ &\leq \dots \leq \|W - \frac{1}{N}\mathbf{1}\mathbf{1}^T\|^{m+1} \|\mu^k(0) - \frac{1}{N}\mathbf{1}\mathbf{1}^T \mu^k(0)\|. \end{aligned} \quad (\text{C.3})$$

The first inequality is due to the definition of the induced matrix norm  $\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^T\|$ . According to Assumption 4.4, we have that  $\rho_W = \|W - \frac{1}{N}\mathbf{1}\mathbf{1}^T\| < 1$ . Furthermore, recalling the fact that  $\frac{1}{N}\mathbf{1}^T \mu^k(m) = \frac{1}{N}\mathbf{1}^T J_i(\theta_k + \delta u_k, \xi_k) = J(\theta_k + \delta u_k, \xi_k)$  for  $m = 0, 1, \dots$ , we obtain that

$$\|\mu^k(N_c) - J(\theta_k + \delta u_k, \xi_k)\mathbf{1}\| \leq \rho_W^{N_c} \|\mu^k(0) - J(\theta_k + \delta u_k, \xi_k)\mathbf{1}\|. \quad (\text{C.4})$$

Since  $\mu_i^k(0) = J_i(\theta_k + \delta u_k, \xi_k) \in [J_l, J_u]$  and  $J(\theta_k + \delta u_k, \xi_k) = \frac{1}{N} \sum_{i=1}^N J_i(\theta_k + \delta u_k, \xi_k) \in [J_l, J_u]$ , we have that  $\|\mu^k(0) - J(\theta_k + \delta u_k, \xi_k)\mathbf{1}\| \leq \sqrt{N}(J_u - J_l)$ . Plugging this inequality into the bound in (C.4), we complete the proof.  $\square$

## C.2 Proof of Theorem 4.7

In the subsequent proof, we need the following lemma by<sup>15</sup>.

**Lemma C.1.** (*Lipschitz Properties of the Smoothed Function*,<sup>15</sup> Given Assumption 4.1, the smoothed function  $J_\delta(\theta)$  is differentiable and its gradient is Lipschitz, that is,  $\|\nabla J_\delta(\theta_1) - \nabla J_\delta(\theta_2)\| \leq \frac{\sqrt{d}L_0}{\delta} \|\theta_1 - \theta_2\|$ , lfor all  $\theta_1, \theta_2 \in \mathbb{R}^d$ .

Next, we present a lemma that bounds the squared norm of the gradient of the smoothed function  $\nabla J_\delta(\theta_k)$  at iterate  $\theta_k$ .

**Lemma C.2.** *Let Assumptions 4.1 and 4.3 hold. Then, for all  $k \geq 0$ , we have that*

$$\begin{aligned} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] &\leq \frac{2}{\alpha} (\mathbb{E}[J_\delta(\theta_{k+1}) - J_\delta(\theta_k)]) + 2\sqrt{d}L_0 \frac{\alpha}{\delta} \mathbb{E}[\|g_\delta(\theta_k)\|^2] + \frac{d_i}{\delta^2} \mathbb{E}[\|\mu^k - \bar{\mu}^k \mathbf{1}\|^2 \|u_k\|^2] \\ &\quad + 4\sqrt{d}d_i L_0 \frac{\alpha}{\delta^3} \mathbb{E}[\|\mu^k - \bar{\mu}^k \mathbf{1}\|^2 \|u_k\|^2] + 4d^{1.5} d_i L_0 \frac{\alpha}{\delta^3} \mathbb{E}[\|\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}\|^2], \end{aligned} \quad (\text{C.5})$$

where  $g_\delta(\theta_k) = \frac{J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})}{\delta} u_k$  and  $\bar{\mu}^k = \frac{1}{N} \mathbf{1}^T \mu^k = J(\theta_k + \delta u_k, \xi_k)$ .

*Proof.* According to Assumption 4.1 and Lemma C.1, we have that the smoothed function  $f_\delta(\theta)$  has Lipschitz gradient with the Lipschitz constant  $L_{1,\delta} = \frac{\sqrt{d}L_0}{\delta}$ . Therefore, using the inequality (6) in<sup>15</sup>, we obtain that

$$\langle \nabla J_\delta(\theta_k), \theta_{k+1} - \theta_k \rangle \leq J_\delta(\theta_{k+1}) - J_\delta(\theta_k) + \frac{L_{1,\delta}}{2} \|\theta_{k+1} - \theta_k\|^2. \quad (\text{C.6})$$

Without loss of generality, we assume that each agent's local policy function  $\pi_i$  is parameterized with  $\theta_i \in \mathbb{R}^{d_i}$  and  $d_i = \frac{d}{N}$  for all  $i$ . Then, the update (4.5) can be written in the compact form

$$\theta_{k+1} = \theta_k + \frac{\alpha}{\delta} \text{diag}([\mu_1^k(N_c) - \mu_1^{k-1}(N_c), \dots, \mu_N^k(N_c) - \mu_N^{k-1}(N_c)]) \otimes I_{d_i} u_k, \quad (\text{C.7})$$

where  $\otimes$  represents the kronecker product and  $I_{d_i}$  is an identity matrix with dimension  $d_i$ . To simplify notations, we use  $\mu_i^k$  or  $\mu^k$  to denote  $\mu_i^k(N_c)$  and  $\mu^k(N_c)$  respectively. Then, we equivalently rewrite equality (C.7) as

$$\begin{aligned} \theta_{k+1} - \theta_k &= \alpha \frac{J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})}{\delta} u_k \\ &\quad + \frac{\alpha}{\delta} \text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k - \frac{\alpha}{\delta} \text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}) \otimes I_{d_i} u_k, \end{aligned} \quad (\text{C.8})$$

where  $\bar{\mu}^k = \frac{1}{N} \mathbf{1}^T \mu^k = J(\theta_k + \delta u_k, \xi_k)$  as in the proof of Lemma 4.6. Substituting (C.8) into the bound in (C.6) and rearranging terms, we get that

$$\begin{aligned} \alpha \langle \nabla J_\delta(\theta_k), g_\delta(\theta_k) \rangle &\leq J_\delta(\theta_{k+1}) - J_\delta(\theta_k) - \frac{\alpha}{\delta} \langle \nabla J_\delta(\theta_k), \text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k \rangle \\ &\quad + \frac{L_{1,\delta}}{2} \alpha^2 \|g_\delta(\theta_k) + \frac{1}{\delta} \text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k - \frac{1}{\delta} \text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}) \otimes I_{d_i} u_k\|^2 \\ &\quad + \frac{\alpha}{\delta} \langle \nabla J_\delta(\theta_k), \text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}) \otimes I_{d_i} u_k \rangle, \end{aligned} \quad (\text{C.9})$$

where  $g_\delta(\theta_k) = \frac{J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})}{\delta} u_k$ . Dividing both sides of (C.9) by  $\alpha$  and taking the expectation of both sides with respect to  $u_k$  and  $\xi_k$  conditioned on the filtration  $\mathcal{F}_{k-1} = \sigma(u_t, \xi_t | t \leq k-1)$ , we have that

$$\begin{aligned} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] &\leq \frac{\mathbb{E}[J_\delta(\theta_{k+1}) - J_\delta(\theta_k)]}{\alpha} - \frac{1}{\delta} \mathbb{E}[\langle \nabla J_\delta(\theta_k), \text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k \rangle] \\ &\quad + \frac{L_{1,\delta}}{2} \alpha \mathbb{E}[\|g_\delta(\theta_k) + \frac{1}{\delta} \text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k - \frac{1}{\delta} \text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}) \otimes I_{d_i} u_k\|^2]. \end{aligned} \quad (\text{C.10})$$

This is because  $\mathbb{E}[g_\delta(\theta_k)] = \nabla J_\delta(\theta_k)$ ,  $\nabla J_\delta(\theta_k)$  and  $\text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1})$  are fixed when conditioned on the filtration  $\mathcal{F}_{k-1}$ , and  $\mathbb{E}[u_k] = 0$ . Next, we provide bounds on the second and third terms in the right hand side (RHS) of (C.10). Specifically, because  $-\langle \nabla J_\delta(\theta_k), \frac{1}{\delta} \text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k \rangle \leq \frac{1}{2} \|\nabla J_\delta(\theta_k)\|^2 + \frac{1}{2\delta^2} \|\text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k\|^2$ , we

have that

$$\begin{aligned}
& -\frac{1}{\delta}\mathbb{E}[\langle \nabla J_\delta(\theta_k), \text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k \rangle] \\
& \leq \frac{1}{2}\mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] + \frac{1}{2\delta^2}\mathbb{E}[\|\text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k\|^2] \\
& \leq \frac{1}{2}\mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] + \frac{d_i}{2\delta^2}\mathbb{E}[\|\mu^k - \bar{\mu}^k \mathbf{1}\|^2 \|u_k\|^2], \tag{C.11}
\end{aligned}$$

where the third inequality is due to the fact that  $\|\text{diag}(v_1)v_2\|^2 \leq \|v_1\|^2\|v_2\|^2$  for all  $v_1, v_2 \in \mathbb{R}^d$ . Furthermore, we have that

$$\begin{aligned}
& \mathbb{E}[\|g_\delta(\theta_k) + \frac{1}{\delta}\text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k - \frac{1}{\delta}\text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}) \otimes I_{d_i} u_k\|^2] \\
& \leq 2\mathbb{E}[\|g_\delta(\theta_k)\|^2] + \frac{2}{\delta^2}\mathbb{E}[\|\text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k - \text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}) \otimes I_{d_i} u_k\|^2] \\
& \leq 2\mathbb{E}[\|g_\delta(\theta_k)\|^2] + \frac{4}{\delta^2}\mathbb{E}[\|\text{diag}(\mu^k - \bar{\mu}^k \mathbf{1}) \otimes I_{d_i} u_k\|^2] + \frac{4}{\delta^2}\mathbb{E}[\|\text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}) \otimes I_{d_i} u_k\|^2] \\
& \leq 2\mathbb{E}[\|g_\delta(\theta_k)\|^2] + \frac{4d_i}{\delta^2}\mathbb{E}[\|\mu^k - \bar{\mu}^k \mathbf{1}\|^2 \|u_k\|^2] + \frac{4dd_i}{\delta^2}\mathbb{E}[\|\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}\|^2], \tag{C.12}
\end{aligned}$$

where the first two inequalities are due to the fact that  $\|v_1 + v_2\|^2 \leq 2\|v_1\|^2 + 2\|v_2\|^2$  and the third inequality is due to the fact that  $\mathbb{E}[\|\text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}) \otimes I_{d_i} u_k\|^2] \leq d_i \mathbb{E}[\|\text{diag}(\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1})\|^2 \|u_k\|^2]$ , the fact that  $\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}$  is independent of  $u_k$  and the fact that  $\mathbb{E}[\|u_k\|^2] = d$ . Substituting the bounds in (C.11) and (C.12) into the bound in (C.10) and rearranging the terms, we get that

$$\begin{aligned}
\mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] & \leq \frac{2}{\alpha}(\mathbb{E}[J_\delta(\theta_{k+1}) - J_\delta(\theta_k)]) + 2\sqrt{d}L_0\frac{\alpha}{\delta}\mathbb{E}[\|g_\delta(\theta_k)\|^2] \\
& \quad + 4\sqrt{d}d_iL_0\frac{\alpha}{\delta^3}\mathbb{E}[\|\mu^k - \bar{\mu}^k \mathbf{1}\|^2 \|u_k\|^2] + 4d^{1.5}d_iL_0\frac{\alpha}{\delta^3}\mathbb{E}[\|\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}\|^2] \\
& \quad + \frac{d_i}{\delta^2}\mathbb{E}[\|\mu^k - \bar{\mu}^k \mathbf{1}\|^2 \|u_k\|^2]. \tag{C.13}
\end{aligned}$$

The proof is complete.  $\square$

Next, we present a lemma bounding the second moment of the gradient estimate  $g_\delta(\theta_k)$ .

**Lemma C.3.** *Let Assumptions 4.1 and 4.3 hold. Then, for all  $k \geq 1$ , we have that*

$$\begin{aligned}
\mathbb{E}[\|g_\delta(\theta_k)\|^2] & \leq 8dL_0^2\frac{\alpha^2}{\delta^2}\mathbb{E}[\|g_\delta(\theta_{k-1})\|^2] + 16dd_iL_0^2\frac{\alpha^2}{\delta^4}\mathbb{E}[\|\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}\|^2 \|u_{k-1}\|^2] \\
& \quad + 16d^2d_iL_0^2\frac{\alpha^2}{\delta^4}\mathbb{E}[\|\mu^{k-2} - \bar{\mu}^{k-2} \mathbf{1}\|^2] + 16(d+4)^2L_0^2 + \frac{8d\sigma^2}{\delta^2}. \tag{C.14}
\end{aligned}$$

*Proof.* Recalling that  $g_\delta(\theta_k) = \frac{J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})}{\delta} u_k$ , we have that

$$\mathbb{E}[\|g_\delta(\theta_k)\|^2] = \frac{1}{\delta^2} \mathbb{E}[\|J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})\|^2 \|u_k\|^2]. \quad (\text{C.15})$$

In addition, we have that

$$\begin{aligned} |J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})|^2 &\leq 4(J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_k, \xi_k))^2 \\ &\quad + 4(J(\theta_{k-1} + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_k))^2 \\ &\quad + 2(J(\theta_{k-1} + \delta u_{k-1}, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1}))^2. \end{aligned} \quad (\text{C.16})$$

According to Assumption 4.1, we have that  $(J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_k, \xi_k))^2 \leq L_0^2 \|\theta_k - \theta_{k-1}\|^2$  and  $(J(\theta_{k-1} + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_k))^2 \leq L_0^2 \delta^2 \|u_k - u_{k-1}\|^2$ . Furthermore, according to Assumption 4.3, we have that  $(J(\theta_{k-1} + \delta u_{k-1}, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1}))^2 \leq 4\sigma^2$ . Applying the above bounds to the RHS of (C.16), we get that

$$|J(\theta_k + \delta u_k, \xi_k) - J(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})|^2 \leq 4L_0^2 \|\theta_k - \theta_{k-1}\|^2 + 4L_0^2 \delta^2 \|u_k - u_{k-1}\|^2 + 8\sigma^2. \quad (\text{C.17})$$

Substituting the bound (C.17) into (C.15), we have that

$$\begin{aligned} \mathbb{E}[\|g_\delta(\theta_k)\|^2] &\leq \frac{4L_0^2}{\delta^2} \mathbb{E}[\|\theta_k - \theta_{k-1}\|^2 \|u_k\|^2] + 4L_0^2 \mathbb{E}[\|u_k - u_{k-1}\|^2 \|u_k\|^2] + \frac{8\sigma^2}{\delta^2} \mathbb{E}[\|u_k\|^2] \\ &\leq \frac{4dL_0^2}{\delta^2} \mathbb{E}[\|\theta_k - \theta_{k-1}\|^2] + 16(d+4)^2 L_0^2 + \frac{8d\sigma^2}{\delta^2}. \end{aligned} \quad (\text{C.18})$$

The second inequality is due to the fact that  $\|\theta_k - \theta_{k-1}\|^2$  is independent of  $u_k$ ,  $\mathbb{E}[\|u_k - u_{k-1}\|^2 \|u_k\|^2] \leq 4(d+4)^2$  and  $\mathbb{E}[\|u_k\|^2] = d$ . Specifically, we have that  $\mathbb{E}[\|u_k - u_{k-1}\|^2 \|u_k\|^2] \leq 4(d+4)^2$  because  $\|u_k - u_{k-1}\|^2 \leq 2\|u_k\|^2 + 2\|u_{k-1}\|^2$  and that  $\mathbb{E}[\|u_k\|^4] \leq (d+4)^2$  according to<sup>15</sup>. Substituting the expression for  $\theta_k - \theta_{k-1}$  in (C.8) into (C.18) and applying the bound in (C.12), we obtain that

$$\begin{aligned} \mathbb{E}[\|g_\delta(\theta_k)\|^2] &\leq 8dL_0^2 \frac{\alpha^2}{\delta^2} \mathbb{E}[\|g_\delta(\theta_{k-1})\|^2] + 16dd_i L_0^2 \frac{\alpha^2}{\delta^4} \mathbb{E}[\|\mu^{k-1} - \bar{\mu}^{k-1} \mathbf{1}\|^2 \|u_{k-1}\|^2] \\ &\quad + 16d^2 d_i L_0^2 \frac{\alpha^2}{\delta^4} \mathbb{E}[\|\mu^{k-2} - \bar{\mu}^{k-2} \mathbf{1}\|^2] + 16(d+4)^2 L_0^2 + \frac{8d\sigma^2}{\delta^2}. \end{aligned}$$

The proof is complete.  $\square$

Now, we are ready to present the proof for Theorem 4.7.

**Theorem 3.2. (Learning Rate of Algorithm 1 without Value Tracking)** Let Assumptions 4.1, 4.3, 4.4 and 4.5 hold and define  $\delta = \frac{\epsilon_J}{\sqrt{d}L_0}$ ,  $\alpha = \frac{\epsilon_J^{1.5}}{4d^{1.5}L_0^2\sqrt{K}}$ , and  $N_c \geq \log(\frac{\sqrt{\epsilon}\epsilon_J}{\sqrt{2}d^{1.5}L_0(J_u - J_l)}) / \log(\rho_W)$ . Then, running Algorithm 1 with `DoTracking = False`, we have that  $\frac{1}{K} \sum_{k=1}^{K-1} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] \leq \mathcal{O}(d^{1.5}\epsilon_J^{-1.5}K^{-0.5}) + \frac{\epsilon}{2}$ .

*Proof.* According to Lemma 4.6, and using Assumptions 4.4 and 4.5, we select  $N_c \geq \log(\frac{\sqrt{\epsilon}\epsilon_J}{\sqrt{2}d^{1.5}L_0(J_u - J_l)}) / \log(\rho_W)$  so that  $\|\mu^k - \bar{\mu}^k \mathbf{1}\| = \|\mu^k - J(\theta_k + \delta u_k, \xi_k) \mathbf{1}\| \leq E_\mu$  regardless of  $u_k$  for all  $k \geq 0$ , where  $E_\mu$  is a small constant such that  $E_\mu^2 = \frac{\epsilon\delta^2}{2dd_i}$ . Therefore, the bound in (C.5) can be simplified as

$$\begin{aligned} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] &\leq \frac{2}{\alpha}(\mathbb{E}[J_\delta(\theta_{k+1}) - J_\delta(\theta_k)]) + 2\sqrt{d}L_0\frac{\alpha}{\delta}\mathbb{E}[\|g_\delta(\theta_k)\|^2] \\ &\quad + 8d^{1.5}d_iL_0\frac{\alpha}{\delta^3}E_\mu^2 + \frac{dd_i}{\delta^2}E_\mu^2. \end{aligned} \quad (\text{C.19})$$

Telescoping the above inequality from  $k = 0$  to  $K - 1$ , we get that

$$\begin{aligned} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] &\leq \frac{2}{\alpha}(\mathbb{E}[J_\delta(\theta_K) - J_\delta(\theta_0)]) + 2\sqrt{d}L_0\frac{\alpha}{\delta} \sum_{k=0}^{K-1} \mathbb{E}[\|g_\delta(\theta_k)\|^2] \\ &\quad + 8d^{1.5}d_iL_0\frac{\alpha}{\delta^3}E_\mu^2K + \frac{dd_i}{\delta^2}E_\mu^2K. \end{aligned} \quad (\text{C.20})$$

Next, we bound the term  $\sum_{k=0}^{K-1} \mathbb{E}[\|g_\delta(\theta_k)\|^2]$  on the RHS of (C.20). Specifically, since  $N_c$  is selected so that  $\|\mu^k - \bar{\mu}^k \mathbf{1}\| = \|\mu^k - J(\theta_k + \delta u_k, \xi_k) \mathbf{1}\| \leq E_\mu$  regardless of  $u_k$  for all  $k \geq 0$ , the bound in (C.14) can be simplified as

$$\mathbb{E}[\|g_\delta(\theta_k)\|^2] \leq 8dL_0^2\frac{\alpha^2}{\delta^2}\mathbb{E}[\|g_\delta(\theta_{k-1})\|^2] + 32d^2d_iL_0^2\frac{\alpha^2}{\delta^4}E_\mu^2 + 16(d+4)^2L_0^2 + \frac{8d\sigma^2}{\delta^2}. \quad (\text{C.21})$$

Telescoping the above inequality from  $k = 1$  to  $K - 1$ , adding  $\mathbb{E}[\|g_\delta(\theta_0)\|^2]$  on both sides, adding  $8dL_0^2\frac{\alpha^2}{\delta^2}\mathbb{E}[\|g_\delta(\theta_{K-1})\|^2]$  on the RHS, and rearranging the terms, we have that

$$\begin{aligned} \sum_{k=0}^{K-1} \mathbb{E}[\|g_\delta(\theta_k)\|^2] &\leq \frac{1}{1 - \alpha_g} \mathbb{E}[\|g_\delta(\theta_0)\|^2] + \frac{32d^2d_iL_0^2}{1 - \alpha_g} \frac{\alpha^2}{\delta^4} E_\mu^2 K \\ &\quad + \frac{16(d+4)^2L_0^2}{1 - \alpha_g} K + \frac{8d\sigma^2}{(1 - \alpha_g)\delta^2} K, \end{aligned} \quad (\text{C.22})$$

where  $\alpha_g = 8dL_0^2 \frac{\alpha^2}{\delta^2}$ . When  $\delta = \frac{\epsilon_J}{\sqrt{d}L_0}$  and  $\alpha = \frac{\epsilon_J^{1.5}}{4d^{1.5}L_0^2\sqrt{K}}$ , we have that  $\alpha_g = \frac{\epsilon_J}{2dK} \leq \frac{1}{2}$  when  $\epsilon_J \leq d$  and  $K \geq 1$ . Substituting the bound on  $\alpha_g$  into (C.22), we obtain that

$$\sum_{k=0}^{K-1} \mathbb{E}[\|g_\delta(\theta_k)\|^2] \leq 2\mathbb{E}[\|g_\delta(\theta_0)\|^2] + 64d^2 d_i L_0^2 \frac{\alpha^2}{\delta^4} E_\mu^2 K + 32(d+4)^2 L_0^2 K + \frac{16d\sigma^2}{\delta^2} K. \quad (\text{C.23})$$

Moreover, substituting the bound in (C.23) into the bound in (C.20), we get that

$$\begin{aligned} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] &\leq \frac{2}{\alpha} (\mathbb{E}[J_\delta(\theta_K) - J_\delta(\theta_0)]) + 4\sqrt{d}L_0 \frac{\alpha}{\delta} \mathbb{E}[\|g_\delta(\theta_0)\|^2] + 64(d+4)^{2.5} L_0^3 \frac{\alpha}{\delta} K \\ &+ 32d^{1.5} L_0 \sigma^2 \frac{\alpha}{\delta^3} K + 128d^{2.5} d_i L_0^3 \frac{\alpha^3}{\delta^5} E_\mu^2 K + 8d^{1.5} d_i L_0 \frac{\alpha}{\delta^3} E_\mu^2 K + \frac{dd_i}{\delta^2} E_\mu^2 K. \end{aligned} \quad (\text{C.24})$$

Recalling that  $E_\mu^2 = \frac{\epsilon\delta^2}{2dd_i}$  and substituting the selected values for  $\delta = \frac{\epsilon_J}{\sqrt{d}L_0}$  and  $\alpha = \frac{\epsilon_J^{1.5}}{4d^{1.5}L_0^2\sqrt{K}}$  into (C.24), we obtain that

$$\begin{aligned} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] &\leq \frac{8d^{1.5}L_0^2}{\epsilon_J^{1.5}} \mathbb{E}[J_\delta^* - J_\delta(\theta_0)]\sqrt{K} + \frac{\epsilon_J^{0.5}}{\sqrt{dK}} \mathbb{E}[\|g_\delta(\theta_0)\|^2] + \frac{\epsilon\epsilon_J^{1.5}}{d^{1.5}\sqrt{K}} \\ &+ 16\frac{(d+4)^2}{d} L_0^2 \epsilon_J^{0.5} \sqrt{K} + \frac{8d^{1.5}L_0^2\sigma^2}{\epsilon_J^{1.5}} \sqrt{K} + \frac{\epsilon\epsilon_J^{0.5}}{\sqrt{d}} \sqrt{K} + \frac{\epsilon}{2} K, \end{aligned} \quad (\text{C.25})$$

where  $J_\delta^* \geq J_\delta(\theta)$  for all  $\theta \in \mathbb{R}^d$ . The upper bound on  $J_\delta(\theta)$  exists due to Assumption 4.5. Dividing both sides of (C.25) by  $K$ , we achieve the bound in Theorem 4.7.  $\square$

### C.3 Proof of Lemma 4.8

**Lemma 3.3.** Let Assumption 4.4 hold. Then, running Algorithm 1 with `DoTracking = True`, we have that  $\bar{\mu}^k(m) = J(\theta_k + \delta u_k, \xi_k) = \frac{1}{N} \sum_{i=1}^N J_i(\theta_k + \delta u_k, \xi_k)$ , for all  $m = 1, 2, \dots, N_c$  and all  $k$ .

*Proof.* According to (C.1), we have that

$$\bar{\mu}^k(m) = \frac{1}{N} \mathbf{1}^T \mu^k(m) = \frac{1}{N} \mathbf{1}^T \mu^k(m-1) = \dots = \frac{1}{N} \mathbf{1}^T \mu^k(0), \quad \text{for all } m, k \geq 0. \quad (\text{C.26})$$

Next, we show that  $\bar{\mu}^k(N_c) = \bar{\mu}^k(0) = \frac{1}{N} \sum_{i=1}^N J_i(\theta_k + \delta u_k, \xi_k)$  for all  $k$ . We use mathematical induction to construct the proof. Specifically, suppose that  $\bar{\mu}^{k-1}(N_c) = \bar{\mu}^{k-1}(0) =$

$\frac{1}{N} \sum_{i=1}^N J_i(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})$  holds. Then, according to line 9 in Algorithm 1, we have that

$$\begin{aligned} \frac{1}{N} \mathbf{1}^T \mu^k(N_c) &= \frac{1}{N} \mathbf{1}^T \mu^k(0) = \frac{1}{N} \mathbf{1}^T (\mu^{k-1}(N_c) + \vec{J}(\theta_k + \delta u_k, \xi_k) - \vec{J}(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})) \\ &= \frac{1}{N} \sum_{i=1}^N J_i(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1}) + \frac{1}{N} \sum_{i=1}^N J_i(\theta_k + \delta u_k, \xi_k) - \frac{1}{N} \sum_{i=1}^N J_i(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1}) \\ &= \frac{1}{N} \sum_{i=1}^N J_i(\theta_k + \delta u_k, \xi_k), \end{aligned}$$

where  $\vec{J}(\theta_k + \delta u_k, \xi_k) = [\dots, J_i(\theta_k + \delta u_k, \xi_k), \dots]^T$ . The second equality above is due to the induction hypothesis. We have that the induction hypothesis is satisfied for  $\bar{\mu}^0(N_c)$ , according to line 7 in Algorithm 1. Therefore, we have shown that  $\frac{1}{N} \mathbf{1}^T \mu^k(0) = \frac{1}{N} \sum_{i=1}^N J_i(\theta_k + \delta u_k, \xi_k)$  for all  $k$ . And due to (C.26), we have shown that  $\bar{\mu}^k(m) = J(\theta_k + \delta u_k, \xi_k) = \frac{1}{N} \sum_{i=1}^N J_i(\theta_k + \delta u_k, \xi_k)$ , for all  $m = 1, 2, \dots, N_c$  and all  $k$ . The proof is complete.  $\square$

## C.4 Proof of Lemma 4.9

**Lemma 3.4.** Let Assumptions 4.1, 4.3, 4.4 hold and define  $E_\mu^k = \|\mu^k(N_c) - \bar{\mu}^k(N_c) \mathbf{1}\|$ .

Then, running Algorithm 1 with `DoTracking = True`, we have that

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2] &\leq \left( 2\mathbb{E}[(E_\mu^{k-1})^2] + 32dL_0^2 \frac{\alpha^2}{\delta^2} \mathbb{E}[(E_\mu^{k-1})^2 \|u_{k-1}\|^2] + 32d^2 L_0^2 \frac{\alpha^2}{\delta^2} \mathbb{E}[(E_\mu^{k-2})^2] \right) \rho_W^{2N_c} \\ &\quad + 16NL_0^2 \alpha^2 \mathbb{E}[\|\tilde{\nabla} J(\theta_{k-1})\|^2] \rho_W^{2N_c} + 32NdL_0^2 \delta^2 \rho_W^{2N_c} + 16N\sigma^2 \rho_W^{2N_c}. \end{aligned} \quad (\text{C.27})$$

*Proof.* To simplify notations, in what follows, we denote  $\mu^k(N_c)$  and  $\bar{\mu}^k(N_c)$  by  $\mu^k$  and  $\bar{\mu}^k$  respectively. According to lines 9 and 12 in Algorithm 1, we have that

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2] &= \mathbb{E}[\|(I - \frac{1}{N} \mathbf{1} \mathbf{1}^T) \mu^k\|^2] \\ &= \mathbb{E}[\|(I - \frac{1}{N} \mathbf{1} \mathbf{1}^T) W^{N_c} (\mu^{k-1} + \vec{J}(\theta_k + \delta u_k, \xi_k) - \vec{J}(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1}))\|^2] \\ &= \mathbb{E}[\|(W^{N_c} - \frac{1}{N} \mathbf{1} \mathbf{1}^T) (\mu^{k-1} + \vec{J}(\theta_k + \delta u_k, \xi_k) - \vec{J}(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1}))\|^2]. \end{aligned}$$

Since  $(W^{N_c} - \frac{1}{N}\mathbf{1}\mathbf{1}^T)\bar{\mu}^{k-1}\mathbf{1} = 0$ , we obtain that

$$\begin{aligned}\mathbb{E}[(E_\mu^k)^2] &= \mathbb{E}[\|(W^{N_c} - \frac{1}{N}\mathbf{1}\mathbf{1}^T)(\mu^{k-1} - \bar{\mu}^{k-1}\mathbf{1} + \vec{J}(\theta_k + \delta u_k, \xi_k) - \vec{J}(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1}))\|^2] \\ &\leq 2\mathbb{E}[\|W^{N_c} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\|^2(E_\mu^{k-1})^2] \\ &\quad + 2\mathbb{E}[\|W^{N_c} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\|^2\|\vec{J}(\theta_k + \delta u_k, \xi_k) - \vec{J}(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})\|^2].\end{aligned}\quad (\text{C.28})$$

Moreover, we have that  $\|W^{N_c} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\|^2 = \|(W - \frac{1}{N}\mathbf{1}\mathbf{1}^T)^{N_c}\|^2 \leq \|W - \frac{1}{N}\mathbf{1}\mathbf{1}^T\|^{2N_c} = \rho_W^{2N_c}$ .

Applying this bound in (C.28), we get that

$$\mathbb{E}[(E_\mu^k)^2] \leq 2\rho_W^{2N_c}\mathbb{E}[(E_\mu^{k-1})^2] + 2\rho_W^{2N_c}\mathbb{E}[\|\vec{J}(\theta_k + \delta u_k, \xi_k) - \vec{J}(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})\|^2].\quad (\text{C.29})$$

Following the same procedure used to derive the bound in (C.17), we get that  $|J_i(\theta_k + \delta u_k, \xi_k) - J_i(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})|^2 \leq 4L_0^2\|\theta_k - \theta_{k-1}\|^2 + 4L_0^2\delta^2\|u_k - u_{k-1}\|^2 + 8\sigma^2$ , for all  $i = 1, 2, \dots, N$ . Applying this bound to the RHS in (C.29), we obtain that

$$\begin{aligned}\mathbb{E}[(E_\mu^k)^2] &\leq 2\rho_W^{2N_c}\mathbb{E}[(E_\mu^{k-1})^2] + 8NL_0^2\rho_W^{2N_c}\mathbb{E}[\|\theta_k - \theta_{k-1}\|^2] \\ &\quad + 8NL_0^2\delta^2\rho_W^{2N_c}\mathbb{E}[\|u_k - u_{k-1}\|^2] + 16N\sigma^2\rho_W^{2N_c}.\end{aligned}\quad (\text{C.30})$$

Moreover, we have that  $\mathbb{E}[\|u_k - u_{k-1}\|^2] \leq \mathbb{E}[2\|u_k\|^2 + 2\|u_{k-1}\|^2] \leq 4d$ . Substituting the expression of  $\theta_k - \theta_{k-1}$  for (C.8) into (C.30) and applying the bound in (C.12), we obtain that

$$\begin{aligned}\mathbb{E}[(E_\mu^k)^2] &\leq 2\rho_W^{2N_c}\mathbb{E}[(E_\mu^{k-1})^2] + 16NL_0^2\rho_W^{2N_c}\alpha^2\mathbb{E}[\|g_\delta(\theta_{k-1})\|^2] + 32dNL_0^2\delta^2\rho_W^{2N_c} + 16N\sigma^2\rho_W^{2N_c} \\ &\quad + 32dL_0^2\rho_W^{2N_c}\frac{\alpha^2}{\delta^2}\mathbb{E}[(E_\mu^{k-1})^2\|u_{k-1}\|^2] + 32d^2L_0^2\rho_W^{2N_c}\frac{\alpha^2}{\delta^2}\mathbb{E}[(E_\mu^{k-2})^2].\end{aligned}$$

The proof is complete.  $\square$

## C.5 Proof of Theorem 4.10

First, we present a lemma characterizing the bound on  $\mathbb{E}[(E_\mu^k)^2\|u_k\|^2]$ .

**Lemma C.4.** *Let Assumptions 4.1, 4.3, 4.4 hold. Then, for all  $k \geq 1$ , we have that*

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2 \|u_k\|^2] &\leq 2d\rho_W^{2N_c} \mathbb{E}[(E_\mu^{k-1})^2] + 16dNL_0^2 \rho_W^{2N_c} \alpha^2 \mathbb{E}[\|g_\delta(\theta_{k-1})\|^2] \\ &\quad + 32d^2 L_0^2 \rho_W^{2N_c} \frac{\alpha^2}{\delta^2} \mathbb{E}[(E_\mu^{k-1})^2 \|u_{k-1}\|^2] + 32d^3 L_0^2 \rho_W^{2N_c} \frac{\alpha^2}{\delta^2} \mathbb{E}[(E_\mu^{k-2})^2] \\ &\quad + 32(d+4)^2 NL_0^2 \delta^2 \rho_W^{2N_c} + 16dN\sigma^2 \rho_W^{2N_c}. \end{aligned} \tag{C.31}$$

*Proof.* According to the bound on  $(E_\mu^k)^2$  derived in (C.28), we have that

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2 \|u_k\|^2] &\leq 2\rho_W^{2N_c} \mathbb{E}[(E_\mu^{k-1})^2 \|u_k\|^2] \\ &\quad + 2\rho_W^{2N_c} \mathbb{E}[\|\vec{J}(\theta_k + \delta u_k, \xi_k) - \vec{J}(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})\|^2 \|u_k\|^2]. \end{aligned} \tag{C.32}$$

Since  $E_\mu^{k-1}$  is independent of  $u_k$ , we have that  $\mathbb{E}[(E_\mu^{k-1})^2 \|u_k\|^2] = \mathbb{E}[(E_\mu^{k-1})^2] \mathbb{E}[\|u_k\|^2] = d\mathbb{E}[(E_\mu^{k-1})^2]$ . Following the same procedure used to derive the bound in (C.17), we get that  $|J_i(\theta_k + \delta u_k, \xi_k) - J_i(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})|^2 \leq 4L_0^2 \|\theta_k - \theta_{k-1}\|^2 + 4L_0^2 \delta^2 \|u_k - u_{k-1}\|^2 + 8\sigma^2$ , for all  $i = 1, 2, \dots, N$ . Applying this bound to the RHS in (C.32), we obtain that

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2 \|u_k\|^2] &\leq 2d\rho_W^{2N_c} \mathbb{E}[(E_\mu^{k-1})^2] + 8NL_0^2 \rho_W^{2N_c} \mathbb{E}[\|\theta_k - \theta_{k-1}\|^2 \|u_k\|^2] \\ &\quad + 8NL_0^2 \delta^2 \rho_W^{2N_c} \mathbb{E}[\|u_k - u_{k-1}\|^2 \|u_k\|^2] + 16dN\sigma^2 \rho_W^{2N_c}. \end{aligned} \tag{C.33}$$

Moreover, we have that  $\mathbb{E}[\|u_k - u_{k-1}\|^2 \|u_k\|^2] \leq \mathbb{E}[2\|u_k\|^4 + 2\|u_{k-1}\|^2 \|u_k\|^2] \leq 4(d+4)^2$ . Substituting the expression of  $\theta_k - \theta_{k-1}$  in (C.8) into (C.33) and applying the bound (C.12), we obtain that

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2 \|u_k\|^2] &\leq 2d\rho_W^{2N_c} \mathbb{E}[(E_\mu^{k-1})^2] + 16dNL_0^2 \rho_W^{2N_c} \alpha^2 \mathbb{E}[\|g_\delta(\theta_{k-1})\|^2] \\ &\quad + 32d^2 L_0^2 \rho_W^{2N_c} \frac{\alpha^2}{\delta^2} \mathbb{E}[(E_\mu^{k-1})^2 \|u_{k-1}\|^2] + 32d^3 L_0^2 \rho_W^{2N_c} \frac{\alpha^2}{\delta^2} \mathbb{E}[(E_\mu^{k-2})^2] \\ &\quad + 32(d+4)^2 NL_0^2 \delta^2 \rho_W^{2N_c} + 16dN\sigma^2 \rho_W^{2N_c}. \end{aligned}$$

The proof is complete.  $\square$

Now, we are ready to present the proof for Theorem 4.10.

**Theorem 3.5. (Learning Rate of Algorithm 1 with Value Tracking)** Let Assumptions 4.1, 4.3, 4.4 hold and define  $\delta = \frac{\epsilon_J}{\sqrt{d}L_0}$ ,  $\alpha = \frac{\epsilon_J^{1.5}}{4d^{1.5}L_0^2\sqrt{K}}$ , and  $N_c \geq \max(\log(\frac{1}{2\sqrt{2}})/\log(\rho_W))$ ,

$\log(\sqrt{\frac{\epsilon}{2G^2\epsilon_J+64(d+4)^2dL_0^2+32d^3L_0^2\sigma^2/\epsilon_J^2}})/\log(\rho_W))$  where  $G^2 = \max\left(\mathbb{E}[\|\tilde{\nabla}J(\theta_0)\|^2], \frac{2\epsilon_J\epsilon}{dK}+32L_0^2(d+4)^2+16d^2L_0^2\frac{\sigma^2}{\epsilon_J^2}\right)$ . Then, running Algorithm 1 with `DoTracking = True`, we have that  $\frac{1}{K}\sum_{k=1}^{K-1}\mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] \leq \mathcal{O}(d^{1.5}\epsilon_J^{-1.5}K^{-0.5}) + \frac{\epsilon}{2}$ .

*Proof.* First, we show that for all  $k \geq 0$ , we have that  $\mathbb{E}[\|g_\delta(\theta_k)\|^2] \leq G^2$ ,  $\mathbb{E}[(E_\mu^k)^2] \leq E_\mu^2$  and  $\mathbb{E}[(E_\mu^k)^2\|u_k\|^2] \leq dE_\mu^2$  when we let  $\delta = \frac{\epsilon_J}{\sqrt{dL_0}}$ ,  $\alpha = \frac{\epsilon_J^{1.5}}{4d^{1.5}L_0^2\sqrt{K}}$ , and  $N_c \geq \max(\log(\frac{1}{2\sqrt{2}})/\log(\rho_W), \log(\sqrt{\frac{\epsilon}{2G^2\epsilon_J+64(d+4)^2dL_0^2+32d^3L_0^2\sigma^2/\epsilon_J^2}})/\log(\rho_W))$ , where  $E_\mu^2 = \frac{\epsilon\delta^2}{2dd_i}$ . To prove this, we use mathematical induction. Specifically, suppose we have that  $\mathbb{E}[(E_\mu^{k-1})^2] \leq E_\mu^2$ ,  $\mathbb{E}[(E_\mu^{k-2})^2] \leq E_\mu^2$ ,  $\mathbb{E}[(E_\mu^{k-1})^2\|u_{k-1}\|^2] \leq dE_\mu^2$  and  $\mathbb{E}[\|g_\delta(\theta_{k-1})\|^2] \leq G^2$ . Then, according to Lemma C.3, we have that

$$\mathbb{E}[\|g_\delta(\theta_k)\|^2] \leq 8dL_0^2\frac{\alpha^2}{\delta^2}G^2 + 32d^2d_iL_0^2\frac{\alpha^2}{\delta^4}E_\mu^2 + 16(d+4)^2L_0^2 + \frac{8d\sigma^2}{\delta^2}. \quad (\text{C.34})$$

Substituting the selected values for  $\delta$ ,  $\alpha$  and the constant  $E_\mu^2$  in (C.34), we get that

$$\mathbb{E}[\|g_\delta(\theta_k)\|^2] \leq \frac{\epsilon_J}{2dK}G^2 + \frac{\epsilon\epsilon_J}{dK} + 16L_0^2(d+4)^2 + 8d^2L_0^2\frac{\sigma^2}{\epsilon_J^2} \leq G^2, \quad (\text{C.35})$$

where the second inequality holds because  $\frac{\epsilon_J}{2dK}G^2 \leq \frac{1}{2}G^2$  when  $\frac{\epsilon_J}{dK} \leq 1$ . In addition, we have that  $\frac{\epsilon\epsilon_J}{dK} + 16L_0^2(d+4)^2 + 8d^2L_0^2\frac{\sigma^2}{\epsilon_J^2} \leq \frac{1}{2}G^2$  due to the choice of  $G^2$  in Theorem 4.10. Furthermore, according to Lemma 4.9, we have that

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2] &\leq (2 + 64d^2L_0^2\frac{\alpha^2}{\delta^2})\rho_W^{2N_c}E_\mu^2 + 16NL_0^2G^2\alpha^2\rho_W^{2N_c} \\ &\quad + 32NdL_0^2\delta^2\rho_W^{2N_c} + 16N\sigma^2\rho_W^{2N_c}. \end{aligned} \quad (\text{C.36})$$

Substituting the selected values for  $\delta$ ,  $\alpha$  and  $E_\mu^2$  into (C.36), we get that

$$\mathbb{E}[(E_\mu^k)^2] \leq (2 + 4\frac{\epsilon_J}{K})\rho_W^{2N_c}E_\mu^2 + (\frac{NG^2\epsilon_J^3}{d^3L_0^2K} + 32N\epsilon_J^2 + 16N\sigma^2)\rho_W^{2N_c} \leq E_\mu^2. \quad (\text{C.37})$$

The second inequality is because when  $\epsilon_J/K \leq \frac{1}{2}$  and  $N_c \geq \log(\frac{1}{2\sqrt{2}})/\log(\rho_W)$ , we have that  $(2+4\frac{\epsilon_J}{K})\rho_W^{2N_c}E_\mu^2 \leq \frac{1}{2}E_\mu^2$ . In addition, when  $N_c \geq \log(\sqrt{\frac{\epsilon}{2G^2\epsilon_J+64(d+4)^2dL_0^2+32d^3L_0^2\sigma^2/\epsilon_J^2}})/\log(\rho_W)$ , we get that  $(\frac{NG^2\epsilon_J^3}{d^3L_0^2K} + 32N\epsilon_J^2 + 16N\sigma^2)\rho_W^{2N_c} \leq \frac{1}{2}E_\mu^2$ .

Next, according to Lemma C.4, we have that

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2\|u_k\|^2] &\leq (2 + 64d^2L_0^2\frac{\alpha^2}{\delta^2})\rho_W^{2N_c}dE_\mu^2 + 16dNL_0^2G^2\alpha^2\rho_W^{2N_c} \\ &\quad + 32N(d+4)^2L_0^2\delta^2\rho_W^{2N_c} + 16dN\sigma^2\rho_W^{2N_c}. \end{aligned} \quad (\text{C.38})$$

Substituting the selected values for  $\delta$ ,  $\alpha$  and  $E_\mu^2$  into (C.38), we get that

$$\begin{aligned} \mathbb{E}[(E_\mu^k)^2 \|u_k\|^2] &\leq (2 + 4\frac{\epsilon_J}{K})\rho_W^{2N_c} dE_\mu^2 + (\frac{NG^2\epsilon_J^3}{d^2L_0^2K} + 32N\frac{(d+4)^2}{d}\epsilon_J^2 + 16dN\sigma^2)\rho_W^{2N_c} \\ &\leq dE_\mu^2. \end{aligned} \quad (\text{C.39})$$

The second inequality holds for similar reasons as those used to obtain (C.37). To complete the induction argument, we simply need to verify the induction hypothesis when  $k = 1$ . It is straightforward to see that  $\mathbb{E}[\|g_\delta(\theta_0)\|^2] \leq G^2$  due to the definition of  $G^2$ . In addition, due to the initialization step  $\mu^{-1}(N_c) = 0$  in line 1 in Algorithm 1, we have that  $\mathbb{E}[(E_\mu^{-1})^2] \leq E_\mu^2$ . To satisfy the conditions  $\mathbb{E}[(E_\mu^0)^2] \leq E_\mu^2$  and  $\mathbb{E}[(E_\mu^0)^2 \|u_0\|^2] \leq dE_\mu^2$ , it is sufficient to run many enough consensus steps only at the first iteration of Algorithm 1, according to Lemma 4.6. To summarize, the induction hypothesis is satisfied at the first iteration of Algorithm 1 and we have shown that for all  $k \geq 0$ , we have that  $\mathbb{E}[\|g_\delta(\theta_k)\|^2] \leq G^2$ ,  $\mathbb{E}[(E_\mu^k)^2] \leq E_\mu^2$  and  $\mathbb{E}[(E_\mu^k)^2 \|u_k\|^2] \leq dE_\mu^2$  under the choice of parameters specified in Theorem 4.10.

Finally, using the uniform bounds  $\mathbb{E}[(E_\mu^k)^2] \leq E_\mu^2$  and  $\mathbb{E}[(E_\mu^k)^2 \|u_k\|^2] \leq dE_\mu^2$ , we can follow the same procedure as in the proof of Theorem 4.7 and obtain the following optimality bound

$$\begin{aligned} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] &\leq \frac{8d^{1.5}L_0^2}{\epsilon_J^{1.5}} \mathbb{E}[J_\delta^* - J_\delta(\theta_0)]\sqrt{K} + \frac{\epsilon_J^{0.5}}{\sqrt{dK}} \mathbb{E}[\|g_\delta(\theta_0)\|^2] + \frac{\epsilon\epsilon_J^{1.5}}{d^{1.5}\sqrt{K}} \\ &\quad + 16\frac{(d+4)^2}{d}L_0^2\epsilon_J^{0.5}\sqrt{K} + \frac{8d^{1.5}L_0^2\sigma^2}{\epsilon_J^{1.5}}\sqrt{K} + \frac{\epsilon\epsilon_J^{0.5}}{\sqrt{d}}\sqrt{K} + \frac{\epsilon}{2}K. \end{aligned} \quad (\text{C.40})$$

Dividing both sides by  $K$  completes the proof.  $\square$

# Appendix D

## Proofs for Chapter 5

**Lemma D.1.** Consider a sequence of non-negative real numbers  $\{V_k\}$  with the following relations for all  $1 \leq k \leq T-1$ , provided with  $0 < \gamma + \beta < 1$ ,

$$V_k \leq \gamma (V_{k-1} + \beta V_{k-2} + \cdots + \beta^{k-1} V_0) + M,$$

where  $M$  is a constant, then we have

$$V_k \leq \gamma(\gamma + \beta)^{k-1} V_0 + \frac{1 - \beta - \gamma(\gamma + \beta)^{k-1}}{1 - (\gamma + \beta)} M.$$

In addition,

$$\sum_{k=0}^{T-1} V_k \leq \frac{1 - \beta}{1 - (\gamma + \beta)} V_0 + (T-1) \frac{1 - \beta}{1 - (\gamma + \beta)} M - \frac{\gamma}{(1 - (\gamma + \beta))^2} M.$$

*Proof.* Fix some  $k = K$ . We have

$$\begin{aligned} V_K &\leq \gamma (V_{K-1} + \beta V_{K-2} + \cdots + \beta^{K-1} V_0) + M \\ &\leq \gamma(\gamma + \beta) (V_{K-2} + \cdots + \beta^{K-2} V_0) + \gamma M + M \end{aligned}$$

Repeat the above process, we obtain that

$$\begin{aligned} V_K &\leq \gamma(\gamma + \beta)^{K-2} (\gamma V_0 + M + \beta V_0) + \gamma \sum_{k=0}^{K-2} (\gamma + \beta)^k M + M \\ &= \gamma(\gamma + \beta)^{K-1} V_0 + \frac{1 - \beta - \gamma(\gamma + \beta)^{K-1}}{1 - (\gamma + \beta)} M, \end{aligned}$$

which completes the first part of the proof. Summing  $V_k$  from 0 to  $T-1$ , we obtain that

$$\begin{aligned} \sum_{k=0}^{T-1} V_k &= \sum_{k=1}^{T-1} \left( \gamma(\gamma + \beta)^{k-1} V_0 + \frac{1 - \beta - \gamma(\gamma + \beta)^{k-1}}{1 - (\gamma + \beta)} M \right) \\ &\quad + V_0 \\ &= \frac{1 - \beta}{1 - (\gamma + \beta)} V_0 + (T-1) \frac{1 - \beta}{1 - (\gamma + \beta)} M - \frac{\gamma}{(1 - (\gamma + \beta))^2} M, \end{aligned}$$

which complete the proof of the lemma. □

## Bibliography

- [1] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, “Improving the convergence rate of one-point zeroth-order optimization using residual feedback,” *Automatica (Provisionally Accepted)*, 2021.
- [2] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, “Boosting one-point derivative-free online optimization via residual feedback,” *arXiv preprint arXiv:2010.07378*, 2020.
- [3] Y. Zhang and M. M. Zavlanos, “Cooperative multi-agent reinforcement learning with partial observations,” *Advances in Neural Information Processing Systems (Under Review)*, 2021.
- [4] Y. Shen, Y. Zhang, S. Nivison, Z. I. Bell, and M. M. Zavlanos, “Asynchronous zeroth-order distributed optimization with residual feedback,” in *2021 IEEE 60th Conference on Decision and Control (CDC) (Under Review)*, IEEE.
- [5] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, “Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models,” in *Proceedings of the 10th ACM Workshop on Artificial Intelgence and Security*, pp. 15–26, 2017.
- [6] X. Luo, Y. Zhang, and M. M. Zavlanos, “Socially-aware robot planning via bandit human feedback,” in *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPs)*, pp. 216–225, IEEE, 2020.
- [7] S. Ghadimi and G. Lan, “Stochastic first-and zeroth-order methods for nonconvex stochastic programming,” *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.
- [8] Y. Li, Y. Tang, R. Zhang, and N. Li, “Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach,” *arXiv preprint arXiv:1912.09135*, 2019.
- [9] J. Larson, M. Menickelly, and S. M. Wild, “Derivative-free optimization methods,” *arXiv preprint arXiv:1904.11585*, 2019.
- [10] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin, “Stochastic convex optimization with bandit feedback,” *SIAM Journal on Optimization*, vol. 23, no. 1, p. 213, 2013.
- [11] S. Bubeck, Y. T. Lee, and R. Eldan, “Kernel-based methods for bandit convex optimization,” in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 72–85, ACM, 2017.
- [12] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, “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, 2005.
- [13] A. Agarwal, O. Dekel, and L. Xiao, “Optimal algorithms for online convex optimization with multi-point bandit feedback.,” in *COLT*, pp. 28–40, Citeseer, 2010.
  - [14] X. Hu, L. Prashanth, A. György, and C. Szepesvári, “(bandit) convex optimization with biased noisy gradient oracles,” in *Artificial Intelligence and Statistics*, pp. 819–828, 2016.
  - [15] Y. Nesterov and V. Spokoiny, “Random gradient-free minimization of convex functions,” *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.
  - [16] A. Saha and A. Tewari, “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, 2011.
  - [17] O. Dekel, R. Eldan, and T. Koren, “Bandit smooth convex optimization: Improving the bias-variance tradeoff,” in *Advances in Neural Information Processing Systems*, pp. 2926–2934, 2015.
  - [18] O. Shamir, “An optimal algorithm for bandit and zero-order convex optimization with two-point feedback.,” *Journal of Machine Learning Research*, vol. 18, no. 52, pp. 1–11, 2017.
  - [19] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, “Optimal rates for zero-order convex optimization: The power of two function evaluations,” *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.
  - [20] F. Bach and V. Perchet, “Highly-smooth zero-th order online optimization,” in *Conference on Learning Theory*, pp. 257–283, 2016.
  - [21] A. V. Gasnikov, E. A. Krymova, 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,” *Automation and remote control*, vol. 78, no. 2, pp. 224–234, 2017.
  - [22] L. Zhang, T. Yang, R. Jin, and Z.-H. Zhou, “Online bandit learning for a special class of non-convex losses.,” in *AAAI*, pp. 3158–3164, 2015.
  - [23] A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin, “Stochastic convex optimization with bandit feedback,” in *Advances in Neural Information Processing Systems*, pp. 1035–1043, 2011.
  - [24] E. Hazan and Y. Li, “An optimal algorithm for bandit convex optimization,” *arXiv preprint arXiv:1603.04350*, 2016.
  - [25] J. K. Gupta, M. Egorov, and M. Kochenderfer, “Cooperative multi-agent control using deep reinforcement learning,” in *International Conference on Autonomous Agents and Multiagent Systems*, pp. 66–83, Springer, 2017.

- [26] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” in *Advances in Neural Information Processing Systems*, pp. 6379–6390, 2017.
- [27] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, “Counterfactual multi-agent policy gradients,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [28] S. Omidshafiei, J. Papis, C. Amato, J. P. How, and J. Vian, “Deep decentralized multi-task multi-agent reinforcement learning under partial observability,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2681–2690, JMLR. org, 2017.
- [29] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, “Fully decentralized multi-agent reinforcement learning with networked agents,” in *International Conference on Machine Learning*, pp. 5867–5876, 2018.
- [30] Y. Zhang and M. M. Zavlanos, “Distributed off-policy actor-critic reinforcement learning with policy consensus,” in *58th IEEE Conference on Decision and Control*, (Nice, France), December 2019.
- [31] W. Suttle, Z. Yang, K. Zhang, Z. Wang, T. Basar, and J. Liu, “A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning,” *arXiv preprint arXiv:1903.06372*, 2019.
- [32] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, “Global convergence of policy gradient methods for the linear quadratic regulator,” in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018.
- [33] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. L. Bartlett, and M. J. Wainwright, “Derivative-free methods for policy optimization: Guarantees for linear quadratic systems,” *arXiv preprint arXiv:1812.08305*, 2018.
- [34] J. Tsitsiklis, D. Bertsekas, and M. Athans, “Distributed asynchronous deterministic and stochastic gradient optimization algorithms,” *IEEE transactions on automatic control*, vol. 31, no. 9, pp. 803–812, 1986.
- [35] A. Agarwal and J. C. Duchi, “Distributed delayed stochastic optimization,” in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 5451–5452, IEEE, 2012.
- [36] M. Zhong and C. G. Cassandras, “Asynchronous distributed optimization with minimal communication,” in *2008 47th IEEE Conference on Decision and Control*, pp. 363–368, IEEE, 2008.
- [37] H. Cai, Y. Lou, D. McKenzie, and W. Yin, “A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization,” *arXiv preprint arXiv:2102.10707*, 2021.

- [38] O. Shamir, “On the complexity of bandit and derivative-free stochastic convex optimization,” in *Conference on Learning Theory*, pp. 3–24, 2013.
- [39] O. Bilenne, P. Mertikopoulos, and E.-V. Belmega, “Fast optimization with zeroth-order feedback in distributed, multi-user mimo systems,” *IEEE Transactions on Signal Processing*, 2020.
- [40] A. Roy, K. Balasubramanian, S. Ghadimi, and P. Mohapatra, “Multi-point bandit algorithms for nonstationary online nonconvex optimization,” *arXiv preprint arXiv:1907.13616*, 2019.
- [41] P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou, “Dynamic regret of convex and smooth functions,” *arXiv preprint arXiv:2007.03479*, 2020.
- [42] V. R. Konda and J. N. Tsitsiklis, “Actor-critic algorithms,” in *Advances in neural information processing systems*, pp. 1008–1014, 2000.
- [43] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.
- [44] R. Durrett, *Essentials of stochastic processes*, vol. 1. Springer, 1999.
- [45] B. Bent, K. Wang, E. Grzesiak, C. Jiang, Y. Qi, Y. Jiang, P. Cho, K. Zingler, F. I. Ogbeide, A. Zhao, *et al.*, “The digital biomarker discovery pipeline: An open-source software platform for the development of digital biomarkers using mhealth and wearables data,” *Journal of Clinical and Translational Science*, pp. 1–8, 2020.
- [46] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media, 2013.
- [47] S. Ghadimi, G. Lan, and H. Zhang, “Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization,” *Mathematical Programming*, vol. 155, no. 1-2, pp. 267–305, 2016.
- [48] S. Bubeck, N. Cesa-Bianchi, *et al.*, “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.
- [49] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” in *Advances in neural information processing systems*, pp. 2951–2959, 2012.
- [50] T. Tatarenko, “Stochastic learning in potential games: Communication and payoff-based approaches,” *arXiv preprint arXiv:1804.04403*, 2018.
- [51] D. Hajinezhad, M. Hong, and A. Garcia, “Zone: Zeroth-order nonconvex multiagent optimization over networks,” *IEEE Transactions on Automatic Control*, vol. 64, no. 10, pp. 3995–4010, 2019.

- [52] D. Hajinezhad and M. M. Zavlanos, “Gradient-free multi-agent nonconvex nonsmooth optimization,” in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 4939–4944, IEEE, 2018.
- [53] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel, “A unified game-theoretic approach to multiagent reinforcement learning,” in *Advances in Neural Information Processing Systems*, pp. 4190–4203, 2017.
- [54] S. Srinivasan, M. Lanctot, V. Zambaldi, J. Pérolat, K. Tuyls, R. Munos, and M. Bowling, “Actor-critic policy optimization in partially observable multiagent environments,” in *Advances in neural information processing systems*, pp. 3422–3435, 2018.
- [55] Anonymous, “Improving the convergence rate of one-point zeroth-order optimization with residual feedback,” in *NIPS*, 2020. Submitted.
- [56] M. Benaïm, J. Hofbauer, and S. Sorin, “Stochastic approximations and differential inclusions,” *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.
- [57] T. Degris, M. White, and R. Sutton, “Off-policy actor-critic,” in *International Conference on Machine Learning*, 2012.
- [58] K. Zhang, Z. Yang, and T. Basar, “Networked multi-agent reinforcement learning in continuous spaces,” in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 2771–2776, IEEE, 2018.
- [59] P. Pennesi and I. C. Paschalidis, “A distributed actor-critic algorithm and applications to mobile sensor network coordination problems,” *IEEE Transactions on Automatic Control*, vol. 55, no. 2, pp. 492–497, 2010.
- [60] S. V. Macua, J. Chen, S. Zazo, and A. H. Sayed, “Distributed policy evaluation under multiple behavior strategies,” *IEEE Transactions on Automatic Control*, vol. 60, no. 5, pp. 1260–1274, 2015.
- [61] D. Lee, H. Yoon, and N. Hovakimyan, “Primal-dual algorithm for distributed reinforcement learning: distributed gtd,” in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 1967–1972, IEEE, 2018.
- [62] M. S. Stanković and S. S. Stanković, “Multi-agent temporal-difference learning with linear function approximation: Weak convergence under time-varying network topologies,” in *2016 American Control Conference (ACC)*, pp. 167–172, IEEE, 2016.
- [63] K. Yuan, B. Ying, J. Liu, and A. H. Sayed, “Variance-reduced stochastic learning by networked agents under random reshuffling,” *IEEE Transactions on Signal Processing*, vol. 67, no. 2, pp. 351–366, 2017.
- [64] H.-T. Wai, Z. Yang, P. Z. Wang, and M. Hong, “Multi-agent reinforcement learning via double averaging primal-dual optimization,” in *Advances in Neural Information*

- Processing Systems*, pp. 9672–9683, 2018.
- [65] R. S. Sutton, C. Szepesvári, and H. R. Maei, “A convergent o (n) algorithm for off-policy temporal-difference learning with linear function approximation,” *Advances in neural information processing systems*, vol. 21, no. 21, pp. 1609–1616, 2008.
- [66] L. Baird, “Residual algorithms: Reinforcement learning with function approximation,” in *Machine Learning Proceedings 1995*, pp. 30–37, Elsevier, 1995.
- [67] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 1995.
- [68] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” in *NIPS Deep Learning Workshop*, 2013.
- [69] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, p. 529, 2015.
- [70] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*, vol. 48. Springer, 2009.
- [71] M. Razaviyayn, M. Hong, and Z.-Q. Luo, “A unified convergence analysis of block successive minimization methods for nonsmooth optimization,” *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.
- [72] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, “A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing,” *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 57–77, 2016.
- [73] K. Ciosek and S. Whiteson, “Expected policy gradients,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [74] H. R. Maei, C. Szepesvári, S. Bhatnagar, and R. S. Sutton, “Toward off-policy learning control with function approximation,” in *International Conference on Machine Learning*, pp. 719–726, 2010.
- [75] K. Zhang, Z. Yang, and T. Başar, “Networked multi-agent reinforcement learning in continuous spaces,” in *CDC*, 2018.

## Biography

Yan Zhang received his bachelor's degree in mechanical engineering from Tsinghua University, Beijing, China in 2014, and his master's degree in mechanical engineering from Duke University, Durham, NC in 2016. Currently, he is pursuing a doctoral degree in mechanical engineering at Duke University, Durham, NC. His research interests include derivative-free optimization, distributed optimization, multi-agent reinforcement learning algorithms and transfer reinforcement learning problems. He has authored 5 journal articles, 10 conference papers and 2 preprints. He has received graduate student fellowship from the department of Mechanical Engineering and Material Science at Duke University in 2018.