Federated Learning for Reinforcement Learning and Control

Han Wang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024

ABSTRACT

Federated Learning for Reinforcement Learning and Control

Han Wang

Federated learning (FL), a novel distributed learning paradigm, has attracted significant attention in the past few years. Federated algorithms take a client/server computation model, and provide scope to train large-scale machine learning models over an edge-based distributed computing architecture. In the paradigm of FL, models are trained collaboratively under the coordination of a central server while storing data locally on the edge/clients. This thesis addresses critical challenges in FL, focusing on supervised learning, reinforcement learning (RL), control systems, and personalized system identification. By developing robust, efficient algorithms, our research enhances FL's applicability across diverse, real-world environments characterized by data heterogeneity and communication constraints.

In the first part, we introduce an algorithm for supervised FL to address the challenges posed by heterogeneous client data, ensuring stable convergence and effective learning, even with partial client participation. In the federated reinforcement learning (FRL) part, we develop algorithms that leverage similarities across heterogeneous environments to improve sample efficiency and accelerate policy learning. Our setup involves N agents interacting with environments that share the same state and action space but differ in their reward functions and state transition kernels. Through rigorous theoretical analysis, we show that information exchange via FL can expedite both policy evaluation and optimization in decentralized, multi-agent settings, enabling faster, more efficient, and robust learning.

Extending FL into control systems, we propose the `FedLQR` algorithm, which enables agents with unknown but similar dynamics to collaboratively learn stabilizing policies, addressing the unique demands of closed-loop stability in federated control. Our method overcomes numerous technical challenges, such as heterogeneity in the agents'dynamics, multiple local updates, and

stability concerns. We show that our proposed algorithm `FedLQR` produces a common policy that, at each iteration, is stabilizing for all agents. We provide bounds on the distance between the common policy and each agent's local optimal policy. Furthermore, we prove that when learning each agent's optimal policy, `FedLQR` achieves a sample complexity reduction proportional to the number of agents M in a low-heterogeneity regime, compared to the single-agent setting.

In the last part, we explore techniques for personalized system identification in FL, allowing clients to obtain customized models suited to their individual environments. We consider the problem of learning linear system models by observing multiple trajectories from systems with differing dynamics. This framework encompasses a collaborative scenario where several systems seeking to estimate their dynamics are partitioned into clusters according to system similarity. Thus, the systems within the same cluster can benefit from the observations made by the others. Considering this framework, we present an algorithm where each system alternately estimates its cluster identity and performs an estimation of its dynamics. This is then aggregated to update the model of each cluster. We show that under mild assumptions, our algorithm correctly estimates the cluster identities and achieves an ϵ-approximate solution with a sample complexity that scales inversely with the number of systems in the cluster, thus facilitating a more efficient and personalized system identification.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor, James, for his support and guidance during my Ph.D. study. Thank you for giving me the freedom to delve into diverse research topics and for making this academic adventure both enriching and enjoyable. Thank Wei Family Scholarship for supporting me during my Ph.D. study.

I would also like to extend my deepest gratitude to my co-author, Professor Aritra Mitra. Aritra, I consider myself truly fortunate to have witnessed your passion for research, your strong work ethic, and your ability to approach complex problems with remarkable clarity and insight. Your calmness and patience during our weekly discussions not only provided me with invaluable knowledge but also gave me the confidence to persevere and advance my research on Federated Learning.

I would also like to thank my coauthors Leonardo F. Toso, Chenyu Zhang, Donglin Zhan, Sihong He, Fei Miao, and Siddartha Marella as well. Leo, your strong support and collaboration have been instrumental in completing so many papers, and I am deeply grateful for your help. I really enjoyed the time spent playing board games with you and Isabela every Saturday–it provided me with immense emotional support and joy throughout my Ph.D. journey. I thank all my labmates in James's group, Leonardo F. Toso, Donglin Zhan, and Yiqian Wu. I have learned a lot from our discussions.

Outside of research, I want to thank my parents for their unconditional love and support. Thank you for sending me to the United States to pursue my studies, giving me the opportunity to complete my Ph.D. Even from thousands of miles away, your words have been a constant source of strength, inspiring me to stay motivated and keep moving forward. And my younger bother, Zhi Wang, thank you for your love and always providing me new perspectives. I love you immensely and am so proud of you!

Last but not least, I want to thank my husband, Shaoru Chen. Without you, I might not have had the patience to complete my Ph.D. Whenever I faced difficulties, you were always there to patiently

comfort me and guide me on how to move forward. I deeply regret the countless arguments we had during the first two years of my Ph.D. when I was struggling to find my research direction. Thank you for your patience and for standing by me—your love has made me a better person. I will love you forever!

Chapter 1

Introduction

1.1 Overview

Recent years have seen machine learning (ML) techniques achieve spectacular success in tackling complex problems across various domains, from image classification and speech recognition to personalized recommendation engines. However, the performance of many ML algorithms largely depends on the availability of large-scale datasets, which are often widely distributed across different organizations under the protection of privacy restrictions. In response to this need, collaborative/FL [96, 132] has emerged as a popular distributed learning paradigm that leverages decentralized data sources while maintaining privacy and reducing data transfer costs. The widespread application of FL can be primarily attributed to its capability of enabling multiple clients (e.g. mobile devices or whole organizations) to collaboratively train models while keeping the raw data on edge devices (i.e., client-side). Thus, FL embodies the seemingly disparate objectives of collaboration and privacy-preservation.

Although a lot of progress has been made in applying FL in the context of supervised learning, with examples of the most popular algorithms including `FedAvg`, `FedNova`, `FedProx` and `FedMA`, challenges persist when confronted with client heterogeneity–a fundamental challenge in FL. Specifically, variations in clients' data distributions, hardware capabilities (CPU, memory), and

power constraints (battery level) can lead to a severe degradation in performance. Additionally, the incorporation of FL into reinforcement learning (RL) and control systems is still at a nascent stage. It is still unknown how to quickly and reliably plan and control the behavior of autonomous systems/intelligent agents in new or similar environments. My research aims to tackle these issues by focusing on *pushing the boundaries of FL into the realm of supervised learning* and *exploring the challenges and potential of FL in RL and Control*. In particular, the main focus of my thesis includes:

- Developing effective, robust, and efficient algorithms for supervised FL.
- Leveraging FL to improve the sample efficiency in control and RL tasks.
- Examining the effect of heterogeneity in clients' dynamical systems and environments.
- Designing personalized learning strategies for each client, tailoring the learning process to individual needs.

Answers to such questions can not only deepen our understanding of modern FL systems but also provide valuable insights into applications across diverse domains, including healthcare, finance, and smart devices. Despite significant research progress in FL and related areas over the past few years, many challenges remain unresolved, particularly concerning the impact of data heterogeneity, communication efficiency, convergence stability, and its usage in RL and control tasks. Existing literature offers partial solutions to these issues, but key questions still need further exploration. This motivates our present work. In what follows, we provide a brief overview of each of these problems, and then delve into our specific contributions.

1.1.1 Efficient Algorithms for Supervised FL

FL addresses the challenge of decentralized data, where data across clients is often non-IID and heterogeneous [96]. The most widely used algorithm for FL is `FedAvg` [132], which mitigates communication costs by performing multiple local updates before aggregating them at a central server. Although `FedAvg` has demonstrated success in some applications, its performance on heterogeneous data remains a subject of ongoing research [66, 86, 110, 111, 236]. One significant challenge posed by heterogeneity is the drift in client updates, which slows down convergence and leads to instability.

To address the heterogeneity concern, in Chapter 3, we explore the impact of heterogeneity on FL algorithms and offer a new federated learning algorithm, `FedADMM`, for solving non-convex composite optimization problems with non-smooth regularizers. We establish the convergence of `FedADMM`, demonstrating that it achieves optimal optimization and communication complexity, under the case when not all clients are able to participate in a given communication round under a very general sampling model.

1.1.2 Federated Learning for Reinforcement Learning (FRL)

The recent progress in RL, particularly in applications like video games [138] and robotic manipulation [120], is notable. However, RL faces challenges in real-world applications. One of the major problems is poor sample complexity of RL algorithms. To address this issue and expedite the learning process, FL has emerged as a solution; by leveraging the wealth of data available from numerous agents and enabling multiple similar agents to collectively learn a shared policy without disclosing agents' raw data. Federated Reinforcement Learning (FRL) has empirically shown significant success in improving sample efficiency and learning performance

in areas such as autonomous driving [116], Internet of Things (IoT) [117], and network resource management [237]. Despite these practical achievements, the theoretical aspects of FRL remain under-explored. Therefore, understanding the theoretical foundations of FRL is essential to fully realize its real world potential.

In chapter 4, we focuses on a federated policy evaluation problem: whether agents in diverse environments can collaboratively improve the efficiency of policy evaluation than if they were to work in isolation, by leveraging similarities across environments. In chapter 4, we affirmatively addressed this question through a federated policy evaluation problem. In our formation, each agent's environment has its own reward function and state transition kernel, but have identical state and action spaces. We introduced `FedTD`(0), a federated temporal difference learning algorithm, to facilitate agents to exchange their model estimates. We have rigorously demonstrated that `FedTD`(0) can achieve an N-fold speed-up in convergence compared to independent learning, even with this very general definition of environmental heterogeneity.

The key finding from this study highlights FRL's ability to accelerate policy evaluation and reduce sample complexity. In addition, we further expanded our framework to address policy optimization problems [105, 210, 242]. In Chapter 5, we designed a provably efficient FRL algorithm which can accommodate arbitrary levels of environmental heterogeneity among agents in [210]. In summary, this body of work has substantially advanced the theoretical understanding of FRL, showcasing its potential to facilitate faster learning and more robust policy development across heterogeneous environments.

1.1.3 Federated Learning for Control

Inspired by FL's success in RL, in Chapter 6, we have also explored its application in complex control tasks through a multi-agent, model-free Linear Quadratic Regulator (LQR) problem using

Policy Gradient (PG) methods. Our study in [193] involves multiple agents with unknown but similar dynamics, collaboratively learning to minimize an average quadratic cost while maintaining data privacy. The main focus in this project is to investigate: *Can each agent learn its own optimal policy faster by leveraging the similarities among the agents' dynamics?* A distinguishing challenge in this control setting, compared to standard RL, is ensuring stability, as data heterogeneity might impede the learning of stabilizing policies, leading to poor performances in control tasks.

Addressing this question has substantial significance for applications in diverse fields, such as recommendation systems and robotic manipulation. For example, a fleet of identical robots from the same manufacturer, each learning from its own dynamics that may differ due to variations in payload or manufacturing defects, can collectively develop a versatile and advanced skill set by aggregating data from interactions with various environments and tasks.

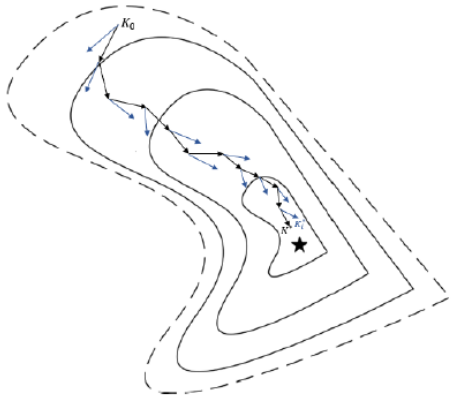


Figure 1.1: Closeness between `FedLQR`'s trajectory and PG's trajectory: (*Black*): PG with single system; (*Blue*): `FedLQR` with multiple distinct systems.

We have developed the `FedLQR` algorithm in Chapter 6 and proved that it not only converges to a policy close to each agent's optimal policy using fewer samples compared to independent learning, but also addresses the unique challenge, encountered in control problems that of providing closed-loop stability. To arrive these results, we first investigated the impact of heterogeneity across systems, revealing the presence of *bounded PG heterogeneity among similar systems*. This finding implies that the trajectory of `FedLQR` closely resembles that of a single system using PG (see Figure 1.1). Secondly, we quantitatively *characterized the heterogeneity requirement necessary for*

the existence of a universal policy capable of stabilizing all unique systems simultaneously. With this requirement, our algorithm can output policies which are simultaneously stabilizing for each distinct system. In sharp contrast to existing works, *we are the first ones to quantify the gap between the common/federated policy and locally optimal policies*. This result has profound implications for the field of meta-learning, particularly in how each agent can fine-tune its own optimal policy using this common policy.

1.1.4 Personalized System Identification

In Chapter 7, we explore techniques for achieving personalization in federated learning. Rather than using a common policy or model for all agents, we present methods for obtaining personalized solutions for each agent. We leverage the system identification problem as an illustrative example.

The system identification problem revolves around the estimation of parameters for a dynamic system based on observed data. Building upon our previous research, we introduced a FL framework tailored for efficient system identification. Our problem in Chapter 7 assumes a central server connected to M clients. Each client is observing data from a different dynamical system. Our primary objective was to investigate how multiple clients collaboratively learn dynamical models in the presence of heterogeneity. In [194], we quantitatively answered this question and demonstrated that with an increasing number of agents, each client can achieve an improved finite-time convergence rate than if a single agent uses its own data to learn the system. We also provided a theoretical analysis of *how the dissimilarity amongst those systems influences the non-asymptotic convergence rate of the proposed technique*. Based on [194], it was observed that the collaboration benefits diminish with increasing system heterogeneity. This is primarily because the approach described in [194] only provided a unified estimation for all systems, thereby limiting its applicability in scenarios involving arbitrary levels of system heterogeneity.

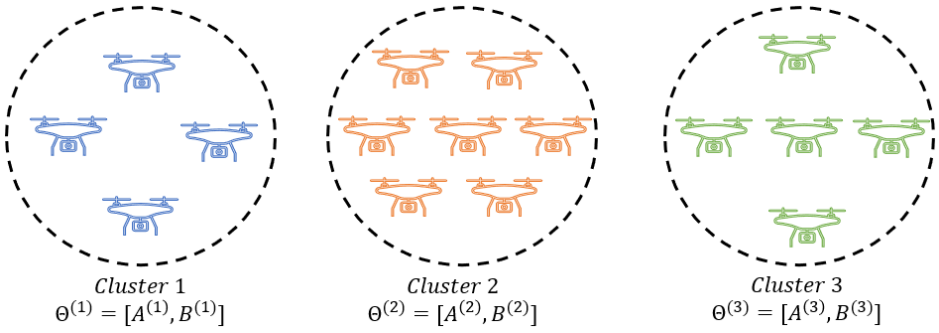


Figure 1.2: In [198], we have proposed an algorithm that iteratively identifies clusters of similar agent from data, then locally learns their dynamics.

Addressing this gap, our latest research discussed in Chapter 7 harnessed clustering techniques to attain *personalized* model estimations for each client. Our specific focus lies in exploring scenarios where there are M dynamical systems, each of which falls into one of K distinct system types, referred to as a "cluster". Unlike conventional clustering problems, our challenge lies in dealing with a data generation process in system identification that is inherently non-i.i.d. Overcoming this challenge, our algorithms can simultaneously determine the correct cluster identities for each of the M systems and achieve an approximate sample complexity that inversely scales with the number of systems in the cluster. Ultimately, *this clustered approach offers a more effective and personalized strategy for system identification*, enhancing the applicability and efficiency of learning diverse dynamical systems.

1.2 Thesis Outline

In Chapter 3, we introduce an efficient supervised FL algorithm to address challenges related to data heterogeneity and convergence stability, while achieving optimal convergence and communication efficiency, even with partial client participation. In Chapter 4 and 5, we explore several FRL problems under environmental heterogeneity. To address these problems, we develop efficient

FRL algorithms that improve sample efficiency and facilitate robust policy learning across diverse environments, supported by rigorous theoretical foundations. In Chapter 6, we extend FL to control systems, proposing the `FedLQR` algorithm to enable multiple agents with unknown but similar dynamics to collaboratively learn stabilizing policies, while quantifying the impact of system heterogeneity on performance. In Chapter 7, we propose some techniques to do the personalized system identification, enabling customized solutions for each agent in FL. Finally, Chapter 8 summarizes the contributions and future work.

1.3 Contributions

The work presented in this thesis have been published in the following conferences and journals:

- **Wang, H.**, Marella, S. and Anderson, J., 2022, December. Fedadmm: A federated primal-dual algorithm allowing partial participation. *In 2022 IEEE 61st Conference on Decision and Control (CDC) (pp. 287-294). IEEE.*

- **Wang, H.**, Mitra, A., Hassani, H., Pappas, G.J. and Anderson, J., 2024. Federated TD learning with linear function approximation under environmental heterogeneity. *Transactions on Machine Learning Research*.

- **Wang, H.**, He, S., Zhang, Z., Miao, F. and Anderson, J., 2024. Momentum for the Win: Collaborative Federated Reinforcement Learning across Heterogeneous Environments. *In 2024 International Conference on Machine Learning*.

- **Wang, H.**, Toso, L.F., Mitra, A. and Anderson, J., 2023. Model-free learning with heterogeneous dynamical systems: A federated LQR approach. *arXiv preprint arXiv:2308.11743*.

- Toso, L.F., **Wang, H.** and Anderson, J., 2023, December. Learning personalized models with clustered system identification. *In 2023 62nd IEEE Conference on Decision and Control (CDC) (pp. 7162-7169). IEEE*.

In addition to these papers, the author have also published works on:

- Zhang, C., **Wang, H.**, Mitra, A. and Anderson, J., Finite-Time Analysis of On-Policy Heterogeneous Federated Reinforcement Learning. *In The Twelfth International Conference on Learning Representations*.

- Toso, L.F., Zhan, D., Anderson, J. and **Wang, H.**, 2024. Meta-Learning Linear Quadratic Regulators: A Policy Gradient MAML Approach for the Model-free LQR. *In 2024 Learning for Dynamics and Control Conference* [202]. (**Best Paper Award**)

- Lan, G., **Wang, H.**, Anderson, J., Brinton, C. and Aggarwal, V., 2024. Improved Communication Efficiency in Federated Natural Policy Gradient via ADMM-based Gradient Updates. *Advances in Neural Information Processing Systems, 36*.

- **Wang, H.** and Anderson, J., 2022, June. Large-scale system identification using a randomized svd. *In 2022 American Control Conference (ACC) (pp. 2178-2185). IEEE* [209].

- **Wang, H.**, Toso, L.F. and Anderson, J., 2023, June. Fedsysid: A federated approach to sample-efficient system identification. *In Learning for Dynamics and Control Conference (pp. 1308-1320). PMLR*.

- Toso, L.F., **Wang, H.** and Anderson, J., 2024, July. Oracle Complexity Reduction for Model-free LQR: A Stochastic Variance-Reduced Policy Gradient Approach. *In 2024 American Control Conference (ACC) (pp. 4032-4037). IEEE* [201].

- Toso, L.F., **Wang, H.** and Anderson, J., 2024. Asynchronous Heterogeneous Linear Quadratic Regulator Design. *In 2024 63rd IEEE Conference on Decision and Control (CDC)* [200].

- **Wang, H.** and Anderson, J., 2022, May. Learning linear models using distributed iterative hessian sketching. *In Learning for Dynamics and Control Conference (pp. 427-440). PMLR*.

Chapter 2

Background

2.1 Background on Federated Learning (FL)

FL is a collaborative machine learning framework in which multiple clients collaborate in solving a machine learning problem, coordinated by a central server. Unlike traditional centralized methods, FL does not require clients to share or transfer their raw data. Instead, each client retains data locally, performing computations on-site to generate model updates, which are then sent to the central server for aggregation.

This method addresses critical challenges in data privacy, security, and accessibility, making it particularly well-suited for applications in industries such as healthcare, finance, and telecommunications where data sensitivity is paramount.

2.1.1 A Typical Federated Training Process

To better understand the operational flow of FL, we consider a typical template for federated training that includes the Federated Averaging algorithm (`FedAvg`)by [132], which is foundational to many FL variations. Pseudocode of `FedAvg` is given in Algorithm 1. This iterative process,

managed by a central server, involves the following steps:

Algorithm 1 `FedAvg`

Server executes:
  initialize x_0
for each round $t = 1, 2, \ldots, T$ **do**
    $S_t \leftarrow$ (random set of M clients)
  for each client $i \in S_t$ **in parallel do**
      $x_{t+1}^i \leftarrow \text{ClientUpdate}(i, x_t)$
  end for
    $x_{t+1} \leftarrow \frac{1}{M} \sum_{k=1}^{M} x_{t+1}^i$
end for
ClientUpdate(i, x):
for local step $j = 1, \ldots, K$ **do**
    $x \leftarrow x - \eta \nabla f(x; z)$ for $z \sim \mathcal{P}_i$
end for
  return x to server

- **Client Selection**: The server selects a subset of clients that meet certain criteria for participation. For instance, mobile devices might only connect if they are plugged in, connected to an unmetered Wi-Fi network, and idle to avoid user disruption.

- **Broadcast**: The selected clients receive the current model parameters and a training program, such as a TensorFlow graph, from the server.

- **Client Computation/Local Update**: Each client locally trains the model by executing the training program, often through stochastic gradient descent (SGD) on its local data, as implemented in `FedAvg`. This step allows each client to compute updates based on its unique data.

- **Aggregation**: The server gathers updates from the clients and combines them. For efficiency, updates from slow or inactive clients (stragglers) may be omitted once a sufficient number of updates have been received. This stage may incorporate additional techniques like secure aggregation to enhance privacy, lossy compression for communication efficiency, and differential privacy methods, such as noise addition and update clipping.

- **Model Update**: Based on the aggregated client updates, the server refines the shared global model, which is then redistributed to clients in the next round.

This process repeats until the model reaches the desired level of performance or a stopping criterion defined by the model engineer. Through this approach, FL allows for collaborative model training without compromising the privacy of individual data sources.

2.1.2 Key Challenges in FL

In typical FL tasks, the objective is to train a single global model that minimizes the empirical risk across the entire dataset, which consists of the combined data from all clients. Unlike standard distributed training methods, federated optimization algorithms must address unique challenges, as shown below:

- **Privacy Concerns**: Privacy in FL is a critical issue due to potential attacks that can expose sensitive information from local client data. Although FL enhances privacy by sharing model updates rather than raw data, it remains vulnerable to various privacy attacks, including membership inference and model inversion attacks [134, 155, 176]. Recent methods aim to strengthen privacy in FL systems using techniques such as differential privacy and secure multi-party computation. However, these privacy-preserving techniques often come at the cost

of reduced model performance or system efficiency. Balancing these trade-offs is essential for achieving optimal privacy without compromising the effectiveness of the FL solution.

- **Statistical Heterogeneity / Non-IID Data**: In FL, clients often have data distributions that differ significantly from each other (non-IID data). This heterogeneity can degrade the performance of standard FL methods. In particular, heterogeneous settings can introduce a drift in the updates of each client resulting in slow and unstable convergence.

- **System Heterogeneity:** In FL, there exist significant variations in hardware, network connectivity, and computational resources across client devices, which can range from high-performance servers to resource-constrained mobile phones. This variability poses challenges such as inconsistent device availability, computational disparity, and network variability, all of which can slow down model convergence and impact efficiency. Clients may be intermittently unavailable due to battery constraints or connectivity issues, and slower devices can create straggler problems, forcing the server to wait or drop updates, potentially losing valuable information. Therefore, it is crucial to design efficient methods to mitigate these issues and ensure robust and effective FL deployment across diverse environments.

- **Communication Cost**: Communication in FL is a major bottleneck, as resource-constrained clients may find it costly to send and receive large model updates. Reducing the number of communication rounds and the size of transmitted data packets is crucial to make FL more efficient. However, this often involves a trade-off between communication cost and model accuracy. For example, techniques designed to speed up convergence or compress the model size may slightly reduce accuracy. Therefore, it is very important to design some efficient FL algorithms which can manage this trade-off.

As discussed in [84], `FedAvg` suffers from the "client-drift" phenomenon and can not effectively

address the aforementioned challenges. In Chapter 3, we will introduce our proposed methods to enhance model accuracy, efficiency, and privacy in FL.

2.2 Background on the Linear Quadratic Regulator (LQR)

The Linear Quadratic Regulator (LQR) [6] is a foundational method in control theory for designing optimal controllers for linear systems. The goal of an LQR controller is to determine the control input that minimizes a cost function, balancing control effort with system performance. This approach is particularly popular for systems where maintaining stability while minimizing deviations and control effort is essential, such as in aerospace, robotics, and automotive applications.

In this thesis, we consider the following infinite horizon LQR problem:

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\left[\sum_{t=0}^{\infty}\left(x_t^\top Q x_t + u_t^\top R u_t\right)\right] \\ \text{subject to} \quad & x_{t+1} = Ax_t + Bu_t, \quad x_0 \sim \mathcal{D}, \end{aligned} \tag{2.1}$$

where the initial state $x_0 \sim \mathcal{D}$ is randomly distributed according to a distribution $\mathcal{D}$. The matrices $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times k}$ represent the system (or transition) matrices, while $Q \in \mathbb{R}^{d\times d}$ and $R \in \mathbb{R}^{k\times k}$ are positive definite matrices that parameterize the quadratic costs. Note that the infinite horizon LQR problem (2.1) is a non-convex problem.

Throughout this thesis, we assume that the matrices A and B are chosen such that the optimal cost remains finite—this condition is satisfied, for example, if the pair (A, B) is controllable. According to optimal control theory [6, 45], the optimal control input can be represented as a linear function of the state:

$$u_t = -K^* x_t,$$

where $K^* \in \mathbb{R}^{k\times d}$. In what follows, we consider two settings–model-based and model free

setting–to discuss how to find such K.

Model-based Setting: For the infinite-horizon LQR problem, if the system matrices A and B are known (referred to as the model-based setting) the optimal controller can be achieved by solving the Algebraic Riccati Equation (ARE):

$$P = A^T PA + Q - A^T PB(B^T PB + R)^{-1} B^T PA, \tag{2.2}$$

where P is a positive definite matrix that parameterizes the "cost-to-go" (the optimal cost from a given state onward). The optimal control gain is then given by:

$$K^* = -(B^T PB + R)^{-1} B^T PA. \tag{2.3}$$

Model-free Setting: If the system matrices A and B are unknown (referred to as the model-free setting), [187] provided guarantees on the global convergence of policy gradient methods for model-free LQR. This breakthrough has attracted considerable interest in model-free LQR methods, leading to a series of subsequent works [65, 71, 80, 82, 102, 129, 140, 152] that further analyze convergence guarantees and sample complexity within this setting. Notably, [35] provided a detailed characterization of the sample complexity for the LQR problem.

Chapter 3

An Efficient Algorithm for Supervised FL

3.1 Introduction

Federated learning (FL) [96, 132], a novel distributed learning paradigm, has attracted significant attention in the past few years. Federated algorithms take a client/server computation model, and provide scope to train large-scale machine learning models over an edge-based distributed computing architecture. In the paradigm of FL, models are trained collaboratively under the coordination of a central server while storing data locally on the edge/clients. Typically, clients (devices and entities ranging from mobile phones to hospitals, to an internet of things [83, 141]) are assumed to be heterogeneous; each client is subject to its own constraints on available computational and storage resources. By allowing data to be stored client-side, the FL paradigm has many favorable privacy properties.

In contrast to "traditional" distributed optimization, FL framework has its own unique challenges and characteristics. First, *communication* becomes problematic when the number of edge devices/clients is large, or the connection between the central server and a device is slow, e.g., when the mobile phones have limited bandwidth. Second, datasets stored in each client may be highly *heterogeneous*

in that they are sampled from different population distributions, or the amount of data belonging to each client is unbalanced. Third, *device/client heterogeneity* can severely hinder algorithm performance; differences in hardware, software, and power (connectivity) lead to varying computation speeds among clients, leading to global performance being dominated by the slowest agent. This is known as the "straggler" effect. Additionally, the server may lose control over the clients when they power down or lose connectivity. It is thus common for only a fraction of clients to participate in in each round of the training (optimization) process, and federated optimization algorithms must accommodate this *partial participation*.

A wealth of algorithms have been developed to address the aforementioned challenges. Notably, work in [132] proposed the now popular FedAvg algorithm, where each client performs multiple stochastic gradient descent (SGD) steps before sending the model to the server for aggregation. Subsequent efforts [96, 112, 161, 183, 213] provided theoretical analysis and further empirical performance evaluations. Since the proposal of FedAvg, there has been a rich body of work concentrating on developing federated optimization algorithms, such as; FedProx [166], FedSplit [149], Scaffold [84], FedLin [136], FedDyn [1], FedDR [203] and FedPD [248].

We consider a general unconstrained, composite optimization models formulated as

$$\frac{1}{n}\sum_{i=1}^{n} f_i(x) + g(x). \tag{3.1}$$

No convexity assumptions on f_i are made and g can be non-smooth. Of the previously mentioned federated algorithms, we restrict our attention to FedDR and FedPD. These algorithms are designed to alleviate the unrealistic assumptions required by FedAvg in order to realize desirable theoretical convergence rates. As described in [203], FedDR combines the nonconvex Douglas-Rachford splitting (DRS) algorithm [109] with a randomized block-coordinate strategy. FedDR provably converges when only a subset of clients participate in any given communication round. In contrast,

FedPD is a primal-dual algorithm which requires either full participation or no participation by all clients at every per round. Unlike FedDR, FedPD cannot handle optimization problems of the form of (3.1) for $g \not\equiv 0$.

The key observation of this paper is to note that the updating rules of FedPD share a similar form to those of the alternating direction method of multipliers (ADMM) [57], but specifies how the local models are updated to satisfy the flexibility need of FL. Motivated by the fact that ADMM is the dual formulation of DRS [54, 234], we provide a new algorithm called FedADMM. Specifically, our contributions are:

1. By applying FedDR to the dual formulation of problem (3.1), we propose a new algorithm called FedADMM, which allows partial participation and solves the federated composite optimization problems as in [239].

2. When $g \equiv 0$ in problem (3.1), we find that FedADMM reduces to FedPD but requires only partial participation.

3. We prove equivalence between FedDR and FedADMM and provide a one-to-one and onto mapping between the the iterates of both algorithms.

4. We provide convergence guarantees for FedADMM using the equivalence established in point 3.

Since FedADMM is the dual formulation of FedDR, it inherits all the desirable properties from FedDR. First, it can handle both *statistical* and *system* heterogeneity. Second, it allows inexact evaluation of users' proximal operators as in FedProx and FedPD. Third, by considering $g \not\equiv 0$ in (3.1), more general applications and problems with constraints can be considered [239].

Refer to [211] for all proofs in this Chapter.

3.1.1 Related Work

ADMM and DRS: DRS was first proposed in [39] in the context of providing numerical solutions to heat conduction partial differential equations. Subsequently, it found applications in the solution of convex optimization problems [119, 173] and later non-convex problems [108, 109, 195]. ADMM [19, 63] is a very popular iterative algorithm for solving composite optimization problems. The equivalence between DRS and ADMM has been subject of a lot of work [44, 56, 234, 250]. It was first established for convex problems where ADMM is equivalent to applying DRS to the dual problem [44, 56]. Recently, these ideas were extended in [195] to show equivalence in the non-convex regime. Inspired by the fact that FedDR can be viewed as a variant of nonconvex DRS applied to the FL framework, we propose a new algorithm, FedADMM and further extend the equivalence of these two algorithms to the FL paradigm.

Federated Learning: FedAvg was first proposed in [132]. However, it works well only with a homogeneous set of clients. It is difficult to analyze the convergence of FedAvg for the heterogeneous setting unless additional assumptions are made [87, 112, 114, 220]. The main reason for this is that the algorithm suffers from client-drift [252] under objective heterogeneity. To address the data and system heterogeneity, FedProx [166] was proposed by adding an extra proximal term [148] to the objective. However, this extra term might degrade the training performance so that FedProx doesn't converge to the global or local stationery points unless the step-size is carefully tuned. Another method called Scaffold [84] uses control variates (or variance reduction) to reduce client-drift at the cost of increased communication incurred by sending extra variables to the server. FedSplit [149] applied the operator splitting schemes to remedy the objective heterogeneity issues, while it only considered the convex problems and required the full participation of clients. As mentioned earlier, FedDR [203] was inspired from DRS, and allowed partial participation. From the

primal-dual optimization perspective, FedPD [248] proposed a new concept of participation, which restricted its potential application on real problems. It is also worthwhile to mention that FedDyn [1] is equivalent to FedPD [248] from [247] under the full participation setting, but it allows partial participation. Unlike [239], FedPD and FedDyn can't solve non-smooth or constrained problems. Finally, we refer readers to [83] for a comprehensive understanding of the recent advances in FL.

3.2 Preliminaries and Problem Formulation

We consider the canonical Federated learning optimization problem defined as

$$\min_{x\in\mathbb{R}^d}\left\{F(x)=f(x)+g(x)\equiv\frac{1}{n}\sum_{i=1}^{n}f_i(x)+g(x)\right\} \tag{3.2}$$

where n is the number of clients, f_i denotes the loss function associated to the i-th client. Each f_i is nonconvex and Lipschitz differentiable (see Assumptions 2.1 and 2.2 below), and g is a proper, closed, and convex function and is not necessarily smooth. For example, g could be any ℓ_p norm or an indicator function.

Assumption 1. *$F(x)$ is bounded below, i.e.,*

$$\inf_{x\in\mathbb{R}^d}F(x)>-\infty \text{ and } dom(F)\neq\emptyset.$$

Assumption 2. ***(Lipschitz differentiability)*** *Each $f_i(\cdot)$ in (3.2) has L-Lipschitz gradient, i.e.,*

$$\|\nabla f_i(x)-\nabla f_i(y)\|\leq L\|x-y\|$$

for all $i\in[n]$ and $x,y\in\mathbb{R}^d$.

The notation $[n]$ above defines the set $\{1,2,\ldots,n\}$. All the norms in the paper are ℓ_2 norm. We will frequently make use of the proximal operator [148]. Although typically defined for convex functions, we make no such assumptions.

Definition 1. ***(Proximal operator)*** *Given an L-Lipschitz (possibly nonconvex and nonsmooth) function f, then the proximal mapping $\mathbb{R}^d \to (-\infty, \infty]$ is defined as*

$$\text{prox}_{\eta f}(x) = \arg\min_y \left\{ f(y) + \frac{1}{2\eta}\|x - y\|^2 \right\}. \tag{3.3}$$

where parameter $\eta > 0$.

If f is nonconvex but L-Lipschitz, $\text{prox}_{\eta f}(x)$ is still well-defined with $0 < \eta < 1/L$.

Definition 2. ***(Conjugate function)*** *Let $f : \mathbb{R}^d \to \mathbb{R}$. The function $f^* : \mathbb{R}^d \to \mathbb{R}$ defined as*

$$f^*(y) \triangleq \sup_{x \in \text{dom } f} \left(y^T x - f(x) \right)$$

is called the conjugate function of f.

Note that the conjugate function is closed and convex even when f is not, since it is the piecewise supremum of a set of affine functions.

Definition 3. ***(ε-stationarity)*** *A vector x is said to be an ε-stationery solution to* (3.2) *if*

$$\mathbb{E}\left[\|\nabla F(x)\|^2 \right] \leq \varepsilon^2,$$

where expectation is taken with respect to all random variables in the respective algorithm.

3.3 Douglas-Rachford Algorithm

3.3.1 Douglas-Rachford Splitting

Douglas-Rachford Splitting (DRS) [39] is an iterative splitting algorithm for solving the optimization problems that can be written as

$$\text{minimize}_{x \in \mathbb{R}^d} \quad f(x) + g(x). \tag{3.4}$$

Although originally used for solving convex problems, it has been shown to work well on certain non-convex problems with additional structure. DRS solves problem (3.4) by producing a series of iterates (y_k, z_k, x_k) for $k = 1, 2, \ldots$ given by

$$\begin{cases} y_k & = \text{prox}_{\eta f}\left(x_k\right) \\ z_k & = \text{prox}_{\eta g}\left(2y_k - x_k\right) \\ x_{k+1} & = x_k + \alpha\left(z_k - y_k\right) \end{cases} \tag{3.5}$$

where α is a relaxation parameter. When $\alpha = 1$, (3.5) is the classical Douglas-Rachford splitting and when $\alpha = 2$, (3.5) is a related splitting algorithm called Peaceman-Rachford splitting [151].

If f in problem (3.4) can decomposed as $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, then (3.5) can be modified so as to run in parallel if we include a global averaging step. The resulting algorithm is given below:

$$\begin{cases} y_i^{k+1} & = y_i^k + \alpha\left(\bar{x}^k - x_i^k\right), \quad \forall i \in [n] \\ x_i^{k+1} & = \text{prox}_{\eta f_i}\left(y_i^{k+1}\right), \quad \forall i \in [n] \\ \hat{x}_i^{k+1} & = 2x_i^{k+1} - y_i^{k+1}, \quad \forall i \in [n] \\ \tilde{x}^{k+1} & = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i^{k+1}, \\ \bar{x}^{k+1} & = \text{prox}_{\eta g}\left(\tilde{x}^{k+1}\right). \end{cases} \tag{3.6}$$

A full derivation is given in [203]. Equation (3.6) is called full parallel Douglas-Rachford splitting (DRS).

3.3.2 FedDR

Implicit in the full parallel DRS (3.6), is the fact that all users are required to participate at every iteration. Instead of requiring all users $i \in [n]$ to participate as in (3.6), work in [203] proposed an inexact randomized block-coordinate DRS algorithm, called FedDR. Here, a subset $\mathcal{S}_k$ of clients is sampled from a "proper" sampling scheme $\hat{\mathcal{S}}$ (See Definition 4 below for details) at each iteration.

Each client, $i \in \mathcal{S}_k$ performs a local update (i.e., executes the first three steps in (3.6)), then sends its local model to server for aggregation. Each client $i \notin \mathcal{S}_k$ does noting. The complete FedDR algorithm is shown in Alg 2.

Algorithm 2 FL with Randomized DR (FedDR) [203]

Initialize $x^0, \eta, \alpha > 0, K,$ and tolerances $\epsilon_{i,0} \geq 0$.
Initialize the server with $\bar{x}^0 = x^0$ and $\tilde{x}^0 = x^0$
Initialize each client $i \in [n]$ with $y_i^0 = x^0, x_i^0 \approx \text{prox}_{nf_i}\left(y_i^0\right),$ and $\hat{x}_i^0 = 2x_i^0 - y_i^0$.
for $k = 0, \ldots, K$ **do**
  Randomly sample $\mathcal{S}_k \subseteq [n]$ with size S.
  $\triangleright$ User side
  for each user $i \in \mathcal{S}_k$ **do**
    receive $\bar{x}^k$ from the server.
    choose $\epsilon_{i,k+1} \geq 0$ and update
    $y_i^{k+1} = y_i^k + \alpha\left(\bar{x}^k - x_i^k\right),$
    $x_i^{k+1} \approx \text{prox}_{\eta f_i}\left(y_i^{k+1}\right),$
    $\hat{x}_i^{k+1} = 2x_i^{k+1} - y_i^{k+1}$.
    send $\Delta\hat{x}_i^k = \hat{x}_i^{k+1} - \hat{x}_i^k$ back to the server .
  end for
  $\triangleright$ Server side
  aggregation $\tilde{x}^{k+1} = \tilde{x}^k + \frac{1}{n}\sum_{i\in\mathcal{S}_k} \Delta\hat{x}_i^k$
  update $\bar{x}^{k+1} = \text{prox}_{\eta g}\left(\tilde{x}^{k+1}\right)$
end for

Convergence to an ϵ-stationary point of FedDR is guaranteed when the sampling scheme $\hat{\mathcal{S}}$ is proper and Assumption 1 and 2 hold [203].

Definition 4. *Let $p = (p_1, p_2, \cdots, p_n)$, where $p_i = \mathbb{P}(i \in \hat{\mathcal{S}})$. If $p_i > 0$ for all $i \in [n]$, we call the sampling scheme $\hat{\mathcal{S}}$ proper, i.e., every client has a nonzero probability to be selected.*

Assumption 3. *All partial participation algorithms in this paper use a proper sampling scheme.*

From the analysis in [164], this assumption includes a lot of sampling schemes such as non-overlapping uniform and doubly uniform sampling as special cases. The intuition behind proper sampling is to ensure that on average every client has a chance to be selected at every iteration.

In FedDR there are three variables that get updated: $\bar{x}^k, x_i^k$ and y_i^k. The variable $\bar{x}^k$ denotes the consensus/average variable to minimize the global model F, x_i^k denotes the local variable associated to f_i, while y_i^k measures the distance between the global variable $\bar{x}^k$ and local model x_i^k. To account for the limitations on computation resources for local users, FedDR allows the inexact calculation of the proximal step, i.e.,

$$x_i^{k+1} \approx \operatorname{prox}_{\eta f_i}\left(y_i^{k+1}\right) \iff \left\|x_i^{k+1} - \operatorname{prox}_{\eta f_i}\left(y_i^{k+1}\right)\right\| \leq \epsilon_{i,k+1}.$$

Thus $\approx$ defines an ϵ-close solution. After local clients $i \in \mathcal{S}_k$ update their model and send them back to the server, the server aggregates the updates to update the global model by executing line 16 and 17 in Algorithm 2.

3.4 From FedDR to FedADMM

Our first contribution is to derive the FedADMM algorithm from FedDR.

3.4.1 An equivalent formulation

We begin by rewriting problem (3.2) as the equivalent constrained problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^{nd},\bar{x}} & \left\{ F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i) + g(\bar{x}) \right\} \\ \text{s.t. } & \mathbb{I}_{nd} x = \mathbb{1}\bar{x} \end{aligned} \tag{3.7}$$

where $x = \left[x_1^T, x_2^T, \cdots, x_n^T\right]^T \in \mathbb{R}^{nd}$, $\mathbb{I}_d$ is the $d \times d$ identity matrix, and $\mathbb{1} = [\mathbb{I}_d \ \cdots \ \mathbb{I}_d]^T$. Here $\bar{x}$ should be interpreted as the global consensus variable.

Forming the Lagrangian of (3.7) and using the definition of the conjugate function, the dual formulation of (3.7) is

$$\max_{z \in \mathbb{R}^{nd}} \left\{ F^*(z) = -f^*(-\mathbb{I}_{nd} z) - g^*(\mathbb{1}^T z) \right\} \tag{3.8}$$

where $z = \left[z_1^T, z_2^T, \cdots, z_n^T\right]^T \in \mathbb{R}^{nd}$ is the vector of dual variables. Problem (3.8) is clearly equivalent to

$$\min_{z_1, z_2, \cdots, z_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} f_i^* \left(-z_i\right) + g^* \left(\sum_{i}^{n} z_i \right) \right\}. \tag{3.9}$$

Before proceeding to develop an algorithm for solving (3.9), we first rewrite the full parallel DRS algorithm 3.6. Changing the execution order of (3.6) and choosing $\alpha = 1$ give

$$\begin{cases} \hat{x}_i^k & = 2x_i^k - y_i^k, \quad \forall i \in [n] \\ \tilde{x}^k & = \frac{1}{n} \sum_{i=1}^{n} \hat{x}_i^k, \quad \forall i \in [n] \\ \bar{x}^k & = \mathrm{prox}_{\eta g} \left(\tilde{x}^k \right), \\ x_i^{k+1} & = \mathrm{prox}_{\eta f_i} \left(y_i^k + \bar{x}^k - x_i^k \right), \quad \forall i \in [n] \\ y_i^{k+1} & = y_i^k + \bar{x}^k - x_i^k, \quad \forall i \in [n]. \end{cases} \tag{3.10}$$

Introducing the change of variables $w_i^k = x_i^k - y_i^k$, we have the following parallel DR algorithm

$$\begin{cases} \hat{x}_i^k & = x_i^k + w_i^k, \quad \forall i \in [n] \\ \tilde{x}^k & = \frac{1}{n} \sum_{i=1}^{n} \hat{x}_i^k, \quad \forall i \in [n] \\ \bar{x}^k & = \mathrm{prox}_{\eta g} \left(\tilde{x}^k \right), \\ x_i^{k+1} & = \mathrm{prox}_{\eta f_i} \left(\bar{x}^k - w_i^k \right), \quad \forall i \in [n]. \\ w_i^{k+1} & = w_i^k + x_i^{k+1} - \bar{x}^k, \quad \forall i \in [n] \end{cases} \tag{3.11}$$

Remark 1. *Note that* (3.6),(3.10) *and* (3.11) *are essentially the same parallel algorithm under a change of execution order and variables.*

3.4.2 FedDR-II

From section 3.3, we observe that the only difference between full parallel DRS and FedDR is that FedDR only requires a subset of clients to update their variables, while full parallel DRS requires full participation. Similarly, by only considering partial participation in (3.11), we introduce the intermediate FedDR-II algorithm. We now describe each step of a single epoch of FedDR-II:

1. **Initialization:** Given an initial vector $x^0 \in \mathrm{dom}(F)$ and tolerances $\epsilon_{i,0} \geq 0$. Initialize the server with $\bar{x}^0 = x^0$. Initialize all users $i \in [n]$ with $w_i^0 = 0$ and $x_i^0 = x^0$.

2. **The k-th iteration:** $(k \geq 0)$ Sample a proper subset $\mathcal{S}_k \subseteq [n]$ so that $\mathcal{S}_k$ represents the subset of active clients.

3. **Client update (Local):** For each client $i \in \mathcal{S}_k$, update $\hat{x}_i^k = x_i^k + w_i^k$. Clients $i \notin \mathcal{S}_k$ do nothing, i.e.
$$\begin{cases} \hat{x}_i^k &= \hat{x}_i^{k-1} \\ x_i^k &= x_i^{k-1} \\ w_i^k &= w_i^{k-1} \end{cases}$$

4. **Communication:** Each user $i \in \mathcal{S}_k$ sends only $\hat{x}_i^k$ to the server.

5. **Server update:** The server aggregates $\tilde{x}^k = \frac{1}{n}\sum_{i=1}^n \hat{x}_i^k$, and then compute $\bar{x}^k = \mathrm{prox}_{\eta g}\left(\tilde{x}^k\right)$.

6. **Communication (Broadcast):** Each user $i \in \mathcal{S}_k$ receives $\bar{x}^k$ from the server.

7. **Client update (Local):** For each user $i \in \mathcal{S}_k$, given $\epsilon_{i,k+1} \geq 0$, it updates
$$\begin{cases} x_i^{k+1} \approx \mathrm{prox}_{\eta f_i}\left(\bar{x}^k - w_i^k\right) \\ w_i^{k+1} = w_i^k + x_i^{k+1} - \bar{x}^k. \end{cases}$$
Each user $i \notin \mathcal{S}_k$ does nothing, i.e.
$$\begin{cases} w_i^{k+1} &= w_i^k \\ x_i^{k+1} &= x_i^k \end{cases}$$

Remark 2. *FedDR and FedDR-II are equivalent because they are partial participation version of* (3.6) *and* (3.11) *respectively.*

3.4.3 Solving the dual problem using FedDR-II

In this subsection, we use FedDR-II to solve the dual problem (3.9), introducing a new algorithm called FedADMM. We call this algorithm FedADMM because it is derived from applying FedDR-II to the dual problem (3.9). Let us define the augmented Lagrangian functions associated to (3.7) as

$$\mathcal{L}_i(x_i, \bar{x}^k, z_i) = f_i(x_i) + g(\bar{x}^k) + \left\langle z_i^k, x_i - \bar{x}^k \right\rangle + \frac{\eta}{2} \left\| x_i - \bar{x}^k \right\|^2 \tag{3.12}$$

where η denotes penalty parameter. Finally, we define $\Delta\hat{x}_i^k = \hat{x}_i^{k+1} - \hat{x}_i^k$. With everything defined, FedADMM is shown in Algorithm 3.

When $g \equiv 0$, the server-side steps 15-16 of FedADMM reduce to the single step:

$$\bar{x}^{k+1} = \tilde{x}^{k+1} = \tilde{x}^k + \frac{1}{n} \sum_{i \in \mathcal{S}_k} \Delta\hat{x}_i^k = \frac{1}{n} \sum_{i=1}^{n} \hat{x}_i^{k+1}.$$

In this case, the updating rules of FedADMM are essentially the same as FedPD in [248]. Both compute the local model x_i^{k+1} by first minimizing (3.12), followed by updating the dual variable λ_i^{k+1}, and then aggregating $\hat{x}_i^{k+1}$ to achieve the global model $\bar{x}^{k+1}$. However, FedADMM allows for partial participation (only chooses a subset of clients to update) while FedPD requires all clients to update at each communication rounds, making it less practical and applicable in real world scenarios.

Note that FedADMM can handle the case where $g \not\equiv 0$ whereas FedPD didn't consider this more general formulation. Just like step 11 (approximately evaluating $\mathrm{prox}_{\eta f_i}$) in FedDR, FedADMM obtains the new local model x_i^{k+1} by inexactly solving (3.12). Note that we do not specify how to (approximately) solve the proximal steps or Langrangian minimization step in (either) algorithm. Various oracles are specified in [248].

Algorithm 3 Federated ADMM Algorithm (FedADMM)

Initialize $x^0, \eta > 0, K,$ and tolerances $\epsilon_{i,0}(i \in [n])$.
Initialize the server with $\bar{x}^0 = x^0$
Initialize all clients with $z_i^0 = 0$ and $x_i^0 = \hat{x}_i^0 = x^0$.
for $k = 0, \ldots, K$ **do**
  Randomly sample $\mathcal{S}_k \subseteq [n]$ with size S.
  $\triangleright$ Client side
  for each client $i \in \mathcal{S}_k$ **do**
    receive $\bar{x}^k$ from the server.
    $x_i^{k+1} \approx \underset{x_i}{arg\,min}\ \mathcal{L}_i\left(x_i, \bar{x}^k, z_i^k\right)$
    $z_i^{k+1} = z_i^k + \eta\left(x_i^{k+1} - \bar{x}^k\right)$ $\quad \diamondsuit$Dual updates
    $\hat{x}_i^{k+1} = x_i^{k+1} + \frac{1}{\eta} z_i^{k+1}$
    send $\Delta\hat{x}_i^k = \hat{x}_i^{k+1} - \hat{x}_i^k$ back to the server
  end for
  $\triangleright$ Server side
  aggregation $\tilde{x}^{k+1} = \tilde{x}^k + \frac{1}{n}\sum_{i \in \mathcal{S}_k} \Delta\hat{x}_i^k$
  update $\bar{x}^{k+1} = \mathrm{prox}_{g/\eta}\left(\tilde{x}^{k+1}\right)$
end for

3.5 Theoretical Analysis

We now present the main theoretical results of the paper. Namely, an equivalence between FedDR and FedADMM. Based on this, we leverage the FedDR convergence results [203] to show that FedADMM converges under partial participation.

We say that two iterative optimization algorithms are "equivalent" if they produce sequences $(x^k)_{k\geq 0}$ and $(y^k)_{k\geq 0}$ such that there exists a unique linear mapping between the two sequences. More general equivalence classes are defined and studied in [251].

Theorem 1. ***(Equivalence between FedDR and FedADMM)*** *Let $(x_i^k, z_i^k, \bar{x}^k)_{k\geq 0}$ be a sequence generated by FedADMM with penalty parameter η, and $(s_i^k, u_i^k, \hat{u}_i^k, \bar{v}^k)$ a sequence generated by FedDR with parameter $\frac{1}{\eta}$. Then FedADMM and FedDR are equivalent.*

Proof. For each triplet $(x_i^k, z_i^k, \bar{x}^k)$ at the k-th iteration of FedADMM with stepsize η, define

$$\begin{cases} s_i^k &= x_i^k - z_i^k/\eta \\ u_i^k &= x_i^k \\ \hat{u}_i^k &= x_i^k + z_i^k/\eta \\ \bar{v}^k &= \bar{x}^k \end{cases} \quad \text{and} \quad \begin{cases} s_i^{k+1} &= x_i^{k+1} - z_i^{k+1}/\eta \\ u_i^{k+1} &= x_i^{k+1} \\ \hat{u}_i^{k+1} &= x_i^{k+1} + z_i^{k+1}/\eta \\ \bar{v}^{k+1} &= \bar{x}^{k+1} \end{cases}$$

Then $(s_i^k, u_i^k, \hat{u}_i^k, \bar{v}^k)$ and $(s_i^{k+1}, u_i^{k+1}, \hat{u}_i^{k+1}, \bar{v}^{k+1})$ satisfy the updating rule of FedDR

$$\begin{cases} s_i^{k+1} &= s_i^k + (\bar{v}^k - u_i^k), \quad \forall i \in \mathcal{S}_k, \\ u_i^{k+1} &= \mathrm{prox}_{rf_i}\left(s_i^{k+1}\right), \quad \forall i \in \mathcal{S}_k, \\ \hat{u}_i^{k+1} &= 2u_i^{k+1} - s_i^{k+1}, \quad \forall i \in \mathcal{S}_k, \\ \bar{v}^{k+1} &= \mathrm{prox}_{rg}(\tfrac{1}{n}\textstyle\sum_{i=1}^n \hat{u}_i^{k+1}), \end{cases}$$

where $r = 1/\eta$ and when $i \notin \mathcal{S}_k$

$$\begin{cases} s_i^{k+1} & = s_i^k, \\ u_i^{k+1} & = u_i^k, \\ \hat{u}_i^{k+1} & = \hat{u}_i^k \end{cases}$$

where the same sampling realizations $\mathcal{S}_k$ are used at each iteration for both algorithm.

We have

$$\begin{aligned} s_i^k + (\bar{v}^k - u_i^k) &= x_i^k - z_i^k/\eta + (\bar{x}^k - x_i^k) \\ &= x_i^{k+1} - z_i^k/\eta + \bar{x}^k - x_i^{k+1} \\ &\overset{(a)}{=} x_i^{k+1} - z_i^{k+1}/\eta = s_i^{k+1} \end{aligned}$$

where (a) is due to the dual updates (line 10) in FedADMM algorithm. Moreover,

$$\begin{aligned} u_i^{k+1} = x_i^{k+1} &= \underset{x_i}{arg\,min}\ \mathcal{L}_i\left(x_i, \bar{x}^k, z_i^k\right) \\ &= \mathrm{prox}_{rf_i}(\bar{x}^k - z_i^k/\eta) \\ &\overset{(b)}{=} \mathrm{prox}_{rf_i}(s_i^{k+1}) \end{aligned}$$

where (b) uses the fact that $\bar{x}^k - z_i^k/\eta = s_i^k + (\bar{v}^k - u_i^k) = s_i^{k+1}$.

Finally, note that

$$\hat{u}_i^{k+1} = 2u_i^{k+1} - s_i^{k+1} = x_i^{k+1} + z_i^{k+1}/\eta,$$

which gives

$$\bar{v}^{k+1} = \bar{x}^{k+1} \overset{(c)}{=} \mathrm{prox}_{rg}\left(\sum_{i=1}^{n}\left(x_i^{k+1} + \frac{1}{\eta}z_i^{k+1}\right)\right) = \mathrm{prox}_{rg}(\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^{k+1}) \tag{3.13}$$

where (c) comes from the FedADMM updating rule (line 11-16 in Alg 3). □

Since we have proved the equivalence of FedDR and FedADMM for arbitrary (nonconvex) problems, FedADMM will directly inherit the convergence properties of FedDR, specifically at rate

$\mathcal{O}(\frac{1}{k})$. The explicit convergence rate of FedADMM is characterized in the following theorem which is a direct application of Theorem 3.1 in [203].

Theorem 2. *Suppose that Assumptions 1, 2, and 3 hold and $\gamma_1, \gamma_2, \gamma_3, \gamma_4 > 0$ are constants. Let $\left(x_i^k, z_i^k, \hat{x}_i^k, \bar{x}^k\right)_{k\geq 0}$ be generated by Alg 3 (FedADMM) using penalty parameter η that satisfies*

$$\eta > \frac{4L\left(1+2\gamma_4\right)}{\sqrt{9-16\gamma_4\left(1+4\gamma_4\right)}-1}.$$

Then when $g \equiv 0$, the following holds

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right] \leq \frac{C_1\left[F\left(x^0\right)-F^\star\right]}{K+1} + \frac{1}{n(K+1)}\sum_{k=0}^{K}\sum_{i=1}^{n}\left(C_2\epsilon_{i,k}^2 + C_3\epsilon_{i,k+1}^2\right)$$

where $\hat{\eta} = 1/\eta$, β, ρ_1, and ρ_2 are defined as

$$\begin{cases} \beta & = \frac{\hat{\mathbf{p}}\left[2-(L\hat{\eta}+1)-2L^2\hat{\eta}^2-4\gamma_4\left(1+L^2\hat{\eta}^2\right)\right]}{2\hat{\eta}(1+\gamma_1)(1+L^2\hat{\eta}^2)} > 0 \\ \rho_2 & = \frac{2(1+\hat{\eta}L)^2}{\gamma_4\hat{\eta}} + \frac{\left(1+\hat{\eta}^2L^2\right)}{\hat{\eta}} \\ & \quad + \frac{\left[2-(L\hat{\eta}+1)-2L^2\hat{\eta}^2-4\gamma_4\left(1+L^2\hat{\eta}^2\right)\right]}{2\hat{\eta}(1+L^2\hat{\eta}^2)\gamma_1} \\ \rho_1 & = \rho_2 + \frac{\left(1+\hat{\eta}^2L^2\right)}{\hat{\eta}} \end{cases}$$

and the constants are

$$C_1 = \frac{2(1+\hat{\eta}L)^2\left(1+\gamma_2\right)}{\hat{\eta}^2\beta},\ C_2 = \rho_1 C_1,\ C_3 = \rho_2 C_1 + \frac{(1+\hat{\eta}L)^2\left(1+\gamma_2\right)}{\hat{\eta}^2\gamma_2}.$$

and $\hat{p} = \min\left\{p_i : i \in [n]\right\} > 0$ in Assumption 3.

Corollary 1. *If the accuracy sequence $\epsilon_{i,k}$ (for all $i \in [n]$ and $k > 0$) at Step 8 in Alg 3 satisfies $\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K+1}\epsilon_{i,k}^2 \leq D$ for a given constant $D > 0$ and all $K \geq 0$. Then, FedADMM needs*

$$K = \left\lfloor \frac{C_1\left[F\left(x^0\right)-F^\star\right] + \left(C_2+C_3\right)D}{\varepsilon^2} \right\rfloor \equiv \mathcal{O}\left(\varepsilon^{-2}\right)$$

iterations to achieve $\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\left\|\nabla f\left(\tilde{x}^k\right)\right\|^2\right] \leq \varepsilon^2$, where $\tilde{x}^K$ is randomly selected from $\{\bar{x}^0, \bar{x}^1, \cdots, \bar{x}^K\}$. In other words, after $K = \mathcal{O}(\varepsilon^{-2})$ iterations, $\tilde{x}^K$ is an ε-stationary solution of problem (3.2) *when $g \equiv 0$.*

Remark 3. *Our convergence analysis can be easily extended to $g \not\equiv 0$, as long as we change the suboptimal condition into the gradient mapping as in [203]. To make $\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K+1}\epsilon_{i,k}^2 \leq D$ hold, interested readers could refer to Remark 3.1 in [203].*

Remark 4. *Although FedADMM is a partial participation version of FedPD when $g \equiv 0$, its communication complexity is still $\mathcal{O}(\varepsilon^{-2})$, which matches the lower bound (up to constant factors) in [248].*

3.6 Numerical Simulations

To demonstrate the equivalence of FedDR and FedADMM, we conduct diverse simulations on both synthetic and real datasets. It is worthwhile to mention that our goal is to show the equivalence of the algorithms, *not* to compare their performance with other algorithms. Performance profiling of FedPD and FedDR can be found in [203, 248]. We have not attempted to optimize any hyperparameters. All the experiments run on Google Colab with default CPU setup.

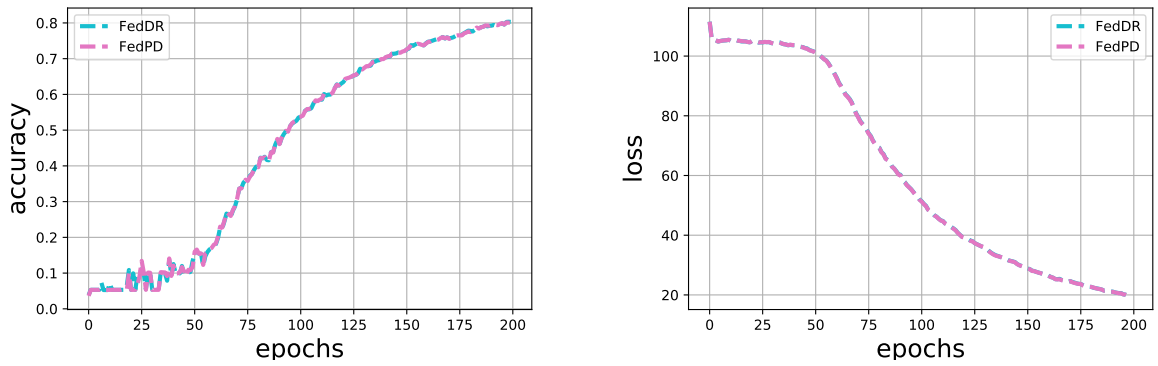
Datasets: We first generate synthetic non-iid datasets by following the same setup as in [174] and denote them as `synthetic-`(α, β). Here α controls how much local models differ from each other and β controls how much the local data at each device differs from that of other devices. We run the experiments by using the unbalanced datasets: `synthetic-(0, 0)`, `synthetic-(0.5, 0.5)` and `synthetic-(1, 1)`. We then compare FedADMM with FedDR on the FEMNIST data set [20]. FEMNIST is a more complex 62-class Federated Extended MNIST dataset. It consists of handwritten characters including: numbers 1-10, 26 upper-and lower-case letters A-Z and a-z from different writers and is also separated by the writers, therefore the dataset is non-iid.

Models and Hyper-parameters: For all the synthetic datasets, we use the model described in [203]: a neutral network with a single hidden layer. The network architecture is $60 \times 32 \times 10$ corresponding

to *input layer*$\times$ *hidden layer* $\times$ *output layer* size. For FEMNIST data, we use the same model as [20], which consists of 2 convolutional layers and two fully connected layers, with 62 neurons in the output layer matching the number of classes in the FEMNIST dataset. For all the experiments, we use $\eta = 1$ and $\alpha = 1$. As in [248], we choose stochastic gradient descent as a local solver with 300 local iterations to solve the step 11 in FedDR and the step 9 in FedADMM. The mini-batch size in calculating the stochastic gradient is 2 and the learning rate is 0.01. We stress that we do not attempt to optimize these parameters.

Implementation: We use the uniform sampling scheme to select the clients in each round. The total number of clients is 30 and we set the number of active clients in each round as 10. To provide a fair comparison, we use the same random seeds across all algorithms.

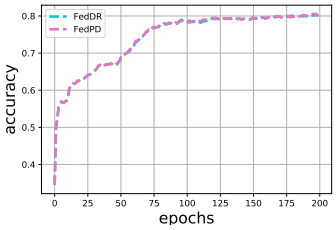
After running multiple experiments on different datasets and models, from figure 3.1 and 3.2, we could observe that the training accuracy and loss of FedDR and FedADMM coincide at each iteration, which verifies our theoretical analysis in section 3.5.



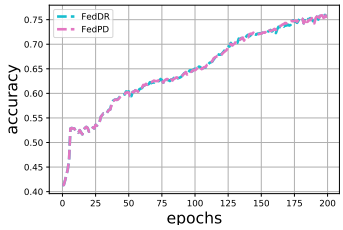
(a) FEMNIST accuracy

(b) FEMNIST loss

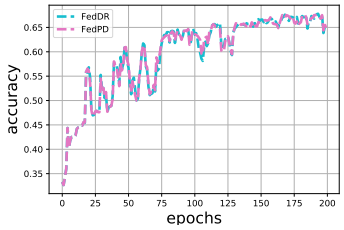
Figure 3.1: Identical performance of FedDR and FedADMM in terms of training accuracy and cross-entropy training loss of FEMNIST dataset



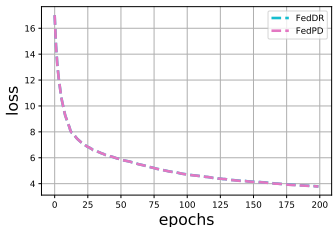
(a) `synthetic(0,0)` accuracy



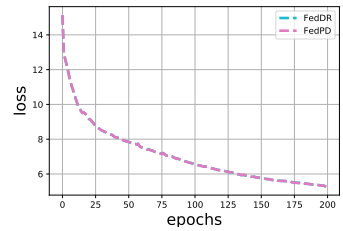
(b) `synthetic(0.5,0.5)` accuracy



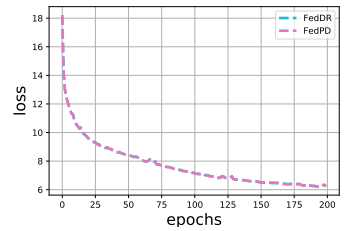
(c) `synthetic(1,1)` accuracy



(d) `synthetic(0,0)` loss



(e) `synthetic(0.5, 0.5)` loss



(f) `synthetic(1,1)` loss

Figure 3.2: Identical performance of FedDR and FedADMM in terms of training accuracy and cross-entropy training loss of synthetic datasets.

3.7 Chapter Summary

We have developed a new federated learning algorithm, FedADMM, for finding stationary points in non-convex composite optimization problems. Current work is focused on incorporating convex constraints into the algorithm, proposing an asynchronous algorithm, asyncFedADMM, and applying it to non-localizable model predictive control problems where communication efficiency is necessary [5].

Chapter 4

Federated Learning for Policy Evaluation

4.1 Introduction

In the popular federated learning (FL) paradigm [95, 133], a set of agents aim to find a common statistical model that explains their collective observations. The motivation to collaborate stems from the fact that if the underlying distributions generating the agents' observations are "similar", then each agent can end up learning a "better" model than if it otherwise used just its own data. This idea has been formalized by the canonical FL algorithm `FedAvg` (and its many variants) where agents communicate local models via a central server while keeping their raw data private. To achieve communication-efficiency - a key consideration in FL - the agents perform multiple local model-updates between two successive communication rounds. There is a rich literature that analyzes the performance of `FedAvg`, focusing primarily on the aspect of *statistical heterogeneity* that originates from differences in the agents' underlying data distributions [23, 84, 89, 135, 137, 150, 226]. Notably, the above works focus on supervised learning problems that are modeled within the framework of distributed optimization. However, for sequential decision-making with multiple agents interacting with *potentially different environments*, little to nothing is known about the effect

of heterogeneity. This is the gap we seek to fill with our work.

The recent survey paper by [157] describes a federated reinforcement learning (FRL) framework which incorporates some of the key ideas from FL in reinforcement learning (RL); applications of FRL in robotics [121], autonomous driving [25], and edge computing [221] are discussed in detail in this paper. As RL algorithms often require many samples to achieve acceptable accuracy, FRL aims to achieve *sample-efficiency* by leveraging information from multiple agents interacting with similar environments. Importantly, as in standard FL, the FRL framework requires agents to keep their personal experiences (e.g., rewards, states, and actions) private, and adhere to stringent communication constraints. While FRL is a promising idea, to model realistic scenarios, one needs to account for the crucial fact that different agents may interact with *non-identical* environments. Indeed, just as statistical heterogeneity is a major challenge in FL, *environmental heterogeneity* is identified as a key open challenge in FRL [157].

To tackle this challenge, we focus on a policy evaluation problem. Our setup involves N agents where each agent interacts with an environment modeled as a MDP. The agents' MDPs share the same state and action space but have different reward functions and state transition kernels, thereby capturing environmental heterogeneity. Each agent seeks to compute the discounted cumulative reward (value function) associated with a common policy μ. Notably, the value functions induced by μ may differ across environments. This leads to the central question we investigate: *Can an agent expedite the process of learning its own value function by leveraging information from potentially different MDPs?* This is a non-trivial question since the effect of combining data from non-identical MDPs is poorly understood.

A typical application of the above FRL setup is that of an autonomous driving system where vehicles in different geographical locations share local models capturing their learned experiences to train a shared model that benefits from the collective exploration data of all vehicles. Although

the vehicles (agents) essentially have the same operations (e.g., steering, braking, accelerating, etc.), they can be exposed to different environments (e.g., road and weather conditions, routes, driving regulations etc.). This is precisely what contributes to environmental heterogeneity.

Refer to [192] for all proofs in this Chapter.

4.1.1 Our Contributions

We study a federated version of the temporal difference (TD) learning algorithm TD(0) [188]. The structure of this algorithm, which we call FedTD(0), is as follows. At each iteration, each agent plays an action according to the policy μ, observes a reward, and transitions to a new state based on its *own* MDP. It then uses TD(0) with linear function approximation to update a local model that approximates its own value function. To (potentially) benefit from other agents' data in a communication-efficient manner, each agent periodically synchronizes with a central server, and performs multiple local updates in between. Notably, as in FL, agents only exchange models but never their personal observations. We perform a comprehensive analysis of FedTD(0) under environmental heterogeneity, and make the following contributions:

Effect of heterogeneity on TD(0) fixed points. Towards understanding the behavior of FedTD(0), we start by asking: *How does heterogeneity in the transition kernels and reward functions of MDPs manifest into differences in the long-term behavior of TD(0) (with linear function approximation) on such MDPs?* Theorem 1 provides an answer by characterizing how perturbing a MDP perturbs the TD(0) fixed point for that MDP. To arrive at this result, we combine results from the perturbation theories of Markov chains and linear equations. Theorem 1 establishes the first perturbation result for TD(0) fixed points, and complements results of a similar flavor in the RL literature such as the *Simulation Lemma* [85].

The Virtual MDP framework. In FL algorithms such as FedAvg, the average of the negative gradients of the agents' loss functions drives the iterates of FedAvg towards the minimizer of a global loss function. In our setting, there is no such global loss function. *So by averaging TD(0) update directions of different MDPs, where do we end up?* To answer this question, we construct a virtual MDP in Section 4.3.2, and characterize several important properties of this fictitious MDP that aid our subsequent analysis. Along the way, we derive a simple yet key result (Proposition 1) pertaining to convex combinations of Markov matrices associated with aperiodic and irreducible Markov chains; this result may be of independent interest.

Analysis under an i.i.d. assumption. To isolate the effect of heterogeneity and build intuition, we start by analyzing FedTD(0) under a standard i.i.d. assumption in the RL literature [13, 34, 38]. After T communication rounds with K local model-updating steps per round, we prove that FedTD(0) guarantees convergence at a rate of $\tilde{O}(1/NKT)$ to a neighborhood of each agent's optimal linear approximation parameter; see Theorem 2. The size of the neighborhood depends on the level of heterogeneity in the agents' MDPs. *The key implication of this result is that in a low-heterogeneity regime, each agent can enjoy an N-fold speed-up in convergence via collaboration.* To prove this result, we introduce a new analysis technique that combines the virtual MDP idea with the optimization interpretation of TD(0) dynamics in [13]. An important benefit of this technique is that it highlights the connections between the dynamics of FedTD(0) and standard FL algorithms, allowing one to leverage existing FL optimization proofs for federated RL.

Bias introduced by Heterogeneity. Our convergence result in Theorem 2 features a bias term due to heterogeneity that cannot be eliminated even by making the step-size arbitrarily small. *Is such a term unavoidable?* We explore this question in Theorem 3 by studying a "steady-state"

deterministic version of FedTD(0). Even for this simple case, we prove that a bias term depending on a natural measure of heterogeneity shows up *inevitably* in the long-term dynamics of FedTD(0). Moreover, unlike the standard FL setting where the effect of heterogeneity manifests itself only when the number of local steps K is strictly greater than 1 [23], the bias term in Theorem 3 persists even when $K = 1$. This reveals a key difference between our setting and federated optimization.

Analysis for the Markovian setting. Our most significant contribution is to provide the *first analysis of a federated RL algorithm (FedTD(0)) that simultaneously accounts for linear function approximation, Markovian sampling, multiple local updates, and heterogeneity.* The effect of heterogeneity coupled with complex temporal correlations makes this setting challenging to analyze. Nonetheless, in Theorem 4, we prove that one can essentially recover the same guarantees as in the i.i.d. setting (Theorem 2). Our result complements the myriad of federated optimization results that account for heterogeneity [89, 226].

We now briefly discuss most directly related work; a detailed description is given in the Appendix.

Related Work. In [38, 122], the authors analyze multi-agent TD learning with linear function approximation over peer-to-peer networks. Neither approach accounts for local steps nor Markovian sampling. Very recently, the authors in [91] do study the effect of Markovian sampling for federated TD learning. However, all of the above papers consider a *homogeneous setting with identical* MDPs for all agents. The only paper we are aware of that performs any theoretical analysis of heterogeneity in FRL is [78]. However, their analysis is limited to the much more simpler tabular setting with no function approximation.

4.2 Model and Problem Formulation

Our RL setting is based on a Markov Decision Process (MDP) [188] defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$, where $\mathcal{S}$ is a finite state space of size n, $\mathcal{A}$ is a finite action space, $\mathcal{P}$ is a set of action-dependent Markov transition kernels, $\mathcal{R}$ is a reward function, and $\gamma \in (0, 1)$ is the discount factor. We consider the problem of evaluating the value function V_μ of a given policy μ, where $\mu : \mathcal{S} \to \mathcal{A}$. The policy μ induces a Markov reward process (MRP) characterized by a transition matrix P_μ, and a reward function R_μ. Under the action of the policy μ at an initial state s, $P_\mu(s, s')$ is the probability of transitioning from state s to state s', and $R_\mu(s)$ is the expected instantaneous reward. The discounted expected cumulative reward obtained by playing policy μ starting from initial state s is:

$$V_\mu(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_\mu(s_t) | s_0 = s\right],$$

where s_t is the state of the Markov chain at time t. From [205], we know that V_μ is the fixed point of the Bellman operator $T_\mu : \mathbb{R}^n \to \mathbb{R}^n$, i.e., $T_\mu V_\mu = V_\mu$, where for any $V \in \mathbb{R}^n$,

$$(T_\mu V)(s) = R_\mu(s) + \gamma \sum_{s' \in \mathcal{S}} P_\mu(s, s') V(s'), \ \forall s \in \mathcal{S}.$$

TD learning with linear function approximation. We consider the setting where the number of states is very large, making it practically infeasible to compute the value function V_μ directly. To mitigate the curse of dimensionality, a common approach [188] is to consider a low-dimensional linear function approximation of the value function V_μ. Let $\{\Phi_k\}_{k=1}^d$ be a set of d linearly independent basis vectors in $\mathbb{R}^n$, and $\Phi \in \mathbb{R}^{n \times d}$ be a matrix with these basis vectors as its columns, i.e., the k-th column of Φ is Φ_k. A parametric approximation $\hat{V}_\theta$ of V_μ in the span of $\{\Phi_k\}_{k=1}^d$ is then given by $\hat{V}_\theta = \Phi\theta$, where $\theta \in \mathbb{R}^d$ is a parameter vector to be learned. Notably, this is tractable since $d \ll n$. We denote the s-th row of Φ by $\phi(s) \in \mathbb{R}^d$, and refer to it as the fixed feature vector

corresponding to state s. We write $\hat{V}_\theta(s) = \phi(s)^\top \theta$ and make the standard assumption [13] that $\|\phi(s)\|^2 \leq 1, \forall s \in \mathcal{S}$.

The objective is to find the best linear approximation of V_μ in the span of $\{\Phi_k\}_{k=1}^d$. More precisely, we seek a parameter vector θ^* that minimizes the distance between $\hat{V}_\theta$ and V_μ (in a suitable sense). When the underlying MDP is *unknown*, one of the most popular techniques to achieve this goal is the classical TD(0) algorithm. TD(0) starts from an initial guess $\theta_0 \in \mathbb{R}^d$. Subsequently, at the t-th iteration, upon playing the given policy μ, a new data tuple $O_t = (s_t, r_t = R_\mu(s_t), s_{t+1})$ comprising of the current state, the instantaneous reward, and the next state is observed. Let us define the TD(0) update direction as

$$g_t(\theta_t) \triangleq \left(r_t + \gamma \phi\left(s_{t+1}\right)^\top \theta_t - \phi\left(s_t\right)^\top \theta_t\right) \phi\left(s_t\right).$$

Using a step-size $\alpha_t \in (0, 1)$, the parameter θ_t is then updated as $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$. Under some mild technical assumptions, it was shown in [205] that the TD(0) iterates converge asymptotically almost surely to a vector θ^*, where θ^* is the unique solution of the projected Bellman equation $\Pi_D T_\mu(\Phi\theta^*) = \Phi\theta^*$. Here, D is a diagonal matrix with entries given by the elements of the stationary distribution π of the Markov matrix P_μ. Furthermore, $\Pi_D(\cdot)$ is the projection operator onto the subspace spanned by $\{\phi_k\}_{k=1}^d$ with respect to the inner product $\langle \cdot, \cdot \rangle_D$.[1]

Objective. We study a multi-agent RL problem where agents interact with similar, but *non-identical* MDPs that share the same state and action space. All agents seek to evaluate the same policy. Our goal is to understand: *Can an agent evaluate the value function of its own MDP in a more sample-efficient way by leveraging data from other agents?* Answering this question is non-trivial since one needs to (i) model heterogeneity in the agents' MDPs; and (ii) understand

[1]We will use $\|\cdot\|_D^2$ to denote the quadratic norm $x^T D x$ induced by the positive definite matrix D, and $\|\cdot\|$ to represent the standard Euclidean norm for vectors and ℓ_2 induced norm for matrices.

the effects of such heterogeneity on the convergence of algorithms that combine information from non-identical MDPs. Existing FL analyses that study statistical heterogeneity in supervised learning fall short of resolving the above issues, since *our problem does not involve minimizing a static loss function*. In the next section, we will formally introduce our setup and the key ideas needed for our subsequent analysis.

4.3 Heterogeneous Federated RL

We consider a federated reinforcement learning setting comprising of N agents that interact with potentially different environments. Agent i's environment is characterized by the following MDP: $\mathcal{M}^{(i)} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}^{(i)}, \mathcal{P}^{(i)}, \gamma \rangle$. While all agents share the same state and action space, the reward functions and state transition kernels of their environments can differ. We focus on a policy evaluation problem where all agents seek to evaluate a common policy μ that induces N Markov reward processes characterized by the tuples $\{P_\mu^{(i)}, R_\mu^{(i)}\}_{i \in [N]}$.[2] Agent i aims to find a linearly parameterized approximation of its value function $V_\mu^{(i)}$. Trivially, agent i can do so without interacting with any other agent by employing the `TD`(0) algorithm. However, the key question we ask is: *By using data from other agents, can it achieve a desired level of approximation with fewer samples relative to when it acts alone?* Naturally, the answer to the above question depends on the level of heterogeneity in the agents' MDPs. Accordingly, inspired by notions of bounded heterogeneity in federated supervised learning [167], we make the following assumptions.

Assumption 1. ***(Markov Kernel Heterogeneity)*** *There exists an $\epsilon > 0$ such that for all agents $i, j \in [N]$, it holds that $|P^{(i)}(s, s') - P^{(j)}(s, s')| \leq \epsilon |P^{(i)}(s, s')|, \forall s, s' \in \mathcal{S}$. Here, for each $i \in [N]$, $P^{(i)}(s, s')$ represents the (s, s')-th element of the matrix $P^{(i)}$.*

[2]For simplicity of notation, we will henceforth drop the dependence of $P^{(i)}$ and $R^{(i)}$ on the policy μ.

Assumption 2. ***(Reward Heterogeneity)*** *There exists an $\epsilon_1 > 0$ such that for all $i, j \in [N]$, it holds that $\|R^{(i)} - R^{(j)}\| \leq \epsilon_1$.*

Clearly, smaller values of ϵ and ϵ_1 capture more similarity in the agents' MDPs. In line with the standard communication architecture in FL [95, 133], suppose all agents can exchange information via a central server. Via such communication, the standard FL task is to find one common model that explains the data of all agents. In a similar spirit, our goal is to find one common parameter θ such that $\hat{V}_\theta = \Phi\theta$ approximates each $V_\mu^{(i)}, i \in [N]$. There is a natural tension here. While using data from multiple agents can help find an approximate model *quickly*, such a model may not *accurately* capture the value function of *any* agent if the agents' MDPs are very dissimilar. *So does more data help or hurt?* It turns out that to answer the above question, we need to carefully understand how the structural heterogeneity assumptions on the MDPs (namely, Assumptions 1 and 2) manifest into differences in the long-term dynamics of TD(0) on these MDPs. In the sequel, we will comprehensively explore this topic.

4.3.1 Impact of Heterogeneity on TD fixed points

Intuitively, if the MRPs induced by a common policy for two different environments are similar, then the long-term behavior of TD(0) on these two MRPs should also be similar. In particular, the TD(0) fixed points of these MRPs should be close. As we shall see later, characterizing this "closeness" in TD(0) fixed points will play a key role in understanding how environmental heterogeneity affects the behavior of a federated TD algorithm. To proceed, we make the following standard assumption.

Assumption 3. *For each $i \in [N]$, the Markov chain induced by the policy μ, corresponding to the state transition matrix $P^{(i)}$, is aperiodic and irreducible.*

The above assumption implies the existence of a unique stationary distribution $\pi^{(i)}$ for each $i \in [N]$; let $D^{(i)}$ be a diagonal matrix with the entries of $\pi^{(i)}$ on its diagonal. For each agent i, we then use θ_i^* to denote the solution of the projected Bellman equation $\Pi_{D^{(i)}} T_\mu^{(i)}(\Phi\theta_i^*) = \Phi\theta_i^*$ for agent i. In words, θ_i^* is the best linear approximation of $V_\mu^{(i)}$ in the span of $\{\phi_k\}_{k=1}^d$. Based on the discussion in Section 4.2, we know that the iterates of TD(0) on agent i's MRP will converge asymptotically (almost surely) to θ_i^*. Our goal is to provide a bound on the gap $\|\theta_i^* - \theta_j^*\|$ as a function of the heterogeneity parameters ϵ and ϵ_1 appearing in Assumptions 1 and 2. The key observation we will exploit is that for each $i \in [N]$, θ_i^* is the unique solution of the linear equation $\bar{A}_i\theta_i^* = \bar{b}_i$, where $\bar{A}_i = \Phi^\top D^{(i)}(\Phi - \gamma P^{(i)}\Phi)$ and $\bar{b}_i = \Phi^\top D^{(i)} R^{(i)}$. For an agent $j \neq i$, viewing $\bar{A}_j$ and $\bar{b}_j$ as perturbed versions of $\bar{A}_i$ and $\bar{b}_i$, we can now appeal to results from the perturbation theory of linear equations [73, Chapter 5.8] to bound $\|\theta_i^* - \theta_j^*\|$. To that end, we first recall a result from the perturbation theory of Markov chains [145] which shows that under Assumption 1, the stationary distributions $\pi^{(i)}$ and $\pi^{(j)}$ are close for any pair $i, j \in [N]$.

Lemma 1. *Suppose Assumption 1 holds. Then, for any pair of agents $i, j \in [N]$, the stationary distributions $\pi^{(i)}$ and $\pi^{(j)}$ satisfy:*

$$\|\pi^{(i)} - \pi^{(j)}\|_1 \leq 2(n-1)\epsilon + \mathcal{O}(\epsilon^2). \tag{4.1}$$

We will now use the bound on $\|\pi^{(i)} - \pi^{(j)}\|_1$ in Lemma 1 to bound $\|\bar{A}_i - \bar{A}_j\|$ and $\|\bar{b}_i - \bar{b}_j\|$. To state our results, we make the standard assumption that for each $i \in [N]$, it holds that $|R^{(i)}(s)| \leq R_{\max}, \forall s \in \mathcal{S}$, i.e., the rewards are uniformly bounded. In [205], it was shown that $-\bar{A}_i$ is a negative definite matrix; thus, there exists some $\delta_1 > 0$ such that $\|\bar{A}_i\| \geq \delta_1$ holds for every agent $i \in [N]$. We also assume that there exists a constant $\delta_2 > 0$ such that $\|\bar{b}_i\| \geq \delta_2, \forall i \in [N]$. We have the following result on the perturbation of TD(0) fixed points.

Theorem 1. ***(Perturbation bounds on TD(0) fixed points)*** *For all $i, j \in [N]$, we have:*

(i) $\|\bar{A}_i - \bar{A}_j\| \leq A(\epsilon) \triangleq \gamma\sqrt{n}\epsilon + (1+\gamma)[2(n-1)\epsilon + \mathcal{O}(\epsilon^2)]$.

(ii) $\|\bar{b}_i - \bar{b}_j\| \leq b(\epsilon, \epsilon_1) \triangleq R_{\max}\,(2(n-1)\epsilon + \mathcal{O}(\epsilon^2)) + \mathcal{O}(\epsilon_1)$.

(iii) Suppose $\exists H > 0$ *such that* $\|\theta_i^*\| \leq H$, $\forall i \in [N]$. *Let* $\kappa(\bar{A}_i)$ *denote the condition number of* $\bar{A}_i$. *Then:*

$$\|\theta_i^* - \theta_j^*\| \leq \Gamma(\epsilon, \epsilon_1) \triangleq \max_{i\in[N]} \left\{ \frac{\kappa(\bar{A}_i)H}{1 - \kappa(\bar{A}_i)\frac{A(\epsilon)}{\delta_1}} \left(\frac{A(\epsilon)}{\delta_1} + \frac{b(\epsilon,\epsilon_1)}{\delta_2} \right) \right\}.$$

Discussion. Theorem 1 reveals how heterogeneity in the rewards and transition kernels of MDPs can be mapped to differences in the limiting behavior of TD(0) on such MDPs from a fixed-point perspective. It formalizes the intuition that if the level of heterogeneity - as captured by ϵ and ϵ_1 - is small, then so is the gap in the TD(0) limit points of the agents' MDPs. This result is novel, and complements similar perturbation results in the RL literature such as the *Simulation Lemma* [85].[3]

In what follows, we will introduce the key concept of a virtual MDP, and build on Theorem 1 to relate properties of this virtual MDP to those of the agents' individual MDPs.

4.3.2 Virtual Markov Decision Process

One of the main goals of our paper is to draw explicit parallels between federated optimization and FRL. Doing so would enable us to apply the rich set of ideas and techniques developed in standard FL to our setting. However, drawing such parallels requires some effort. In a standard FL setting, the goal is to typically minimize a global loss function $f(x) = (1/N)\sum_{i\in[N]} f_i(x)$ composed of the local loss functions of N agents; here, $f_i(x)$ is the local loss function of agent i. In FL, due to heterogeneity in the agents' loss functions, there is a "drift" effect [21, 84]: the local

[3]The simulation lemma tells us that if two MDPs with the same state and action spaces are similar, then so are the value functions induced by a common policy on these MDPs.

iterates of each agent i drift towards the minimizer of $f_i(x)$. However, when the heterogeneity is moderate, the average of the agents' iterates converges towards the minimizer of $f(x)$. To develop an analogous theory for FRL, we need to first answer: *When we average* `TD(0)` *update directions from different MDPs, where does the average* `TD(0)` *update direction lead us?* It is precisely to answer this question that we introduce the concept of a *virtual MDP*.

To model a virtual environment that captures the "average" of the agents' individual environments, we construct MDP $\bar{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \bar{\mathcal{R}}, \bar{\mathcal{P}}, \gamma \rangle$, where

$$\bar{\mathcal{P}} = (1/N) \sum_{i=1}^{N} \mathcal{P}^{(i)}, \text{ and } \bar{\mathcal{R}} = (1/N) \sum_{i=1}^{N} \mathcal{R}^{(i)}.$$

Note that the virtual MDP is a fictitious MDP that we construct solely for the purpose of analysis, and it may not coincide with any of the agents' MDPs, in general.

Properties of the Virtual MDP. When applied to $\bar{\mathcal{M}}$, let the policy μ that we seek to evaluate induce a virtual MRP characterized by the tuple $\{\bar{P}, \bar{R}\}$. It is easy to verify that $\bar{P} = (1/N) \sum_{i=1}^{N} P^{(i)}$, and $\bar{R} = (1/N) \sum_{i=1}^{N} R^{(i)}$. The following result shows how the virtual MRP inherits certain basic properties from the individual MRPs; the result is quite general and may be of independent interest.

Proposition 1. *Let $\{P^{(i)}\}_{i=1}^{N}$ be a set of Markov matrices associated with Markov chains that share the same states, and are each aperiodic and irreducible. Then, for any set of weights $\{w_i\}_{i=1}^{N}$ satisfying $w_i \geq 0, \forall i \in [N]$ and $\sum_{i \in [N]} w_i = 1$, the Markov chain corresponding to the matrix $\sum_{i \in [N]} w_i P^{(i)}$ is also aperiodic and irreducible.*

The above result immediately tells us that the Markov chain corresponding to $\bar{P}$ is aperiodic and irreducible. Thus, there exists an unique stationary distribution $\bar{\pi}$ of this Markov chain; let $\bar{D}$ be the corresponding diagonal matrix. As before, let us define $\bar{A} \triangleq \Phi^\top \bar{D}(\Phi - \gamma \bar{P}\Phi)$, $\bar{b} \triangleq \Phi^\top \bar{D}\bar{R}$, and

use θ^* to denote the solution to the equation $\bar{A}\theta^* = \bar{b}$. Our next result is a consequence of Theorem 1, and characterizes the gap between θ_i^* and θ^*, for each $i \in [N]$.

Proposition 2. *Fix any $i \in [N]$. Using the same definitions as in Theorem 1, we have $\|\bar{A}_i - \bar{A}\| \leq A(\epsilon)$, $\|\bar{b}_i - \bar{b}\| \leq b(\epsilon, \epsilon_1)$ and $\|\theta_i^* - \theta^*\| \leq \Gamma(\epsilon, \epsilon_1)$.*

We will later argue that the federated TD algorithm (to be introduced in Section 4.4) converges to a ball centered around the TD(0) fixed point θ^* of the virtual MRP. Proposition 2 is thus particularly important since it tells us that in a low-heterogeneity regime, by converging close to θ^*, we also converge close to the optimal parameter θ_i^* that minimizes the projected Bellman error for MDP $\mathcal{M}^{(i)}$. This justifies studying the convergence behavior of FedTD(0) on the virtual MRP. We end this section with a result which follows in part from Proposition 1.

Proposition 3. *For the virtual MRP, the following hold: (i) $\lambda_{\max}(\Phi^\top \bar{D}\Phi) \leq 1$; and (ii) $\exists\, \bar{\omega} > 0$ s.t. $\lambda_{\min}(\Phi^\top \bar{D}\Phi) \geq \bar{\omega}$.*

4.4 Federated TD Algorithm

In this section, we describe the FedTD(0) algorithm, a federated version of TD(0). We outline its steps in Algo. 4. The goal of FedTD(0) is to generate a model θ such that $\hat{V}_\theta$ is a good approximation of each agent i's value function $V_\mu^{(i)}$, corresponding to the policy μ. In line with both standard FL algorithms, and also works in MARL/FRL (in homogeneous settings) [38, 91], the agents keep their raw observations (i.e., their rewards, states, and actions) private, and only exchange local models.

FedTD(0) starts from a common initial model and a common starting state for all agents. Subsequently, in each round t, each agent $i \in [N]$ starts from a common model $\bar{\theta}_t$ and uses its local data to perform K local updates of the following form: at each local iteration k, each agent i takes

Algorithm 4 Description of `FedTD(0)`

1: **Input:** Policy μ, local step-size α_l, global step-size $\alpha_g^{(t)}$ at t-th communication round
2: **Initialize:** $\bar{\theta}_0 = \theta_0$ and $s_{0,0}^{(i)} = s_0, \forall i \in [N]$
3: **for** each round $t = 0, \ldots, T-1$ **do**
4: **for** each agent $i \in [N]$ **do**
5: **for** $k = 0, \ldots, K-1$ **do**
6: Agent i initializes $\theta_{t,0}^{(i)} = \bar{\theta}_t$
7: Agent i plays $\mu(s_{t,k}^{(i)})$, observes tuple $O_{t,k}^{(i)} = (s_{t,k}^{(i)}, r_{t,k}^{(i)}, s_{t,k+1}^{(i)})$,
8: and updates local model as $\theta_{t,k+1}^{(i)} = \theta_{t,k}^{(i)} + \alpha_l g_i(\theta_{t,k}^{(i)})$,
9: where $g_i(\theta_{t,k}^{(i)}) \triangleq \left(r_{t,k}^{(i)} + \gamma\phi(s_{t,k+1}^{(i)})^\top \theta_{t,k}^{(i)} - \phi(s_{t,k}^{(i)})^\top \theta_{t,k}^{(i)}\right)\phi(s_{t,k}^{(i)})$
10: **end for**
11: send $\Delta_t^{(i)} = \theta_{t,K}^{(i)} - \bar{\theta}_t$ back to the server
12: **end for**
13: Server computes and broadcasts global model $\bar{\theta}_{t+1} = \Pi_{2,\mathcal{H}}\left(\bar{\theta}_t + \frac{\alpha_g^{(t)}}{N}\sum_{i\in[N]}\Delta_t^{(i)}\right)$
14: **end for**

action $\mu(s_{t,k}^{(i)})$ and observes a data tuple $O_{t,k}^{(i)} = \left(s_{t,k}^{(i)}, r_{t,k}^{(i)}, s_{t+1,k}^{(i)}\right)$ based on its *own* Markov reward process, i.e., $\{P^{(i)}, R^{(i)}\}$; we note here that *observations are independent across agents*. Using its data tuple $O_{t,k}^{(i)}$, each agent i updates its own local model $\theta_{t,k}^{(i)}$ along the direction $g_i(\theta_{t,k}^{(i)})$; see line 7.

Since each agent seeks to benefit from the samples acquired by the other agents, there is intermittent communication via the server. However, such communication needs to be limited as communication-efficiency is a key concern in FL. As such, the agents upload their local models' difference $\Delta_t^{(i)}$ to the server only once every K time-steps (line 11). On the server side, the model differences $\{\Delta_t^{(i)}\}$ are averaged, and a projection is carried out (line 13) to construct a global model $\bar{\theta}_{t+1}$ that is then broadcast to all agents. Here, we use $\Pi_{2,\mathcal{H}}(\cdot)$ to denote the standard Euclidean projection on to a convex compact subset $\mathcal{H} \subset \mathbb{R}^d$ that is assumed to contain each $\theta_i^*, i \in [N]$, and also θ^*. Such a projection step on the server-side ensures that the global models do not blow up,

and is common in the literature on stochastic approximation [15] and RL [13, 38]. Each agent then resumes its local updating process from this global model. We note that the structure of `FedTD(0)` mirrors that of `FedAvg` (and its many variants) where agents perform multiple local model-updates in isolation using their own data (to save communication), and synchronize periodically via a server. From another perspective, the `FedTD(0)` algorithm, which seeks to find the fixed point of the average of the TD update directions, can be grouped into the class of problems that seek to find fixed points using information from different sources [130]. However, there are significant differences in the dynamics of standard FL algorithms and `FedTD(0)`, making it quite challenging to derive finite-time convergence results for the latter. We discuss some of these challenges below.

Challenges in Analysis. First, existing FL analyses are essentially distributed optimization proofs; although our setting bears a cosmetic connection to optimization, federated TD learning does not correspond to minimizing any fixed objective function. Second, unlike the FL setting where the data seen by each agent are drawn i.i.d. from some distribution, the data tuples observed by each agent in `FedTD(0)` are all part of one single Markovian trajectory. This creates complex time-correlations that are challenging to deal with even in a centralized setting with just one agent. Thus, we cannot directly appeal to standard FL proofs. Third, existing analyses in MARL/FRL that go beyond the simple tabular setting all end up assuming that every agent interacts with the *same* MDP, i.e., there is no heterogeneity effect at all to contend with in these works. Concretely, the analysis for `FedTD(0)` we provide in the subsequent sections is unique in that it simultaneously accounts for several key aspects: linear function approximation, Markovian sampling, multiple local updates, and heterogeneity in MDPs.

4.5 Analysis of the I.I.D. Setting

To isolate the effect of heterogeneity and provide key insights regarding our main proof ideas, we will analyze a simpler i.i.d. setting in this section. Specifically, we assume that for each agent $i \in [N]$, the data tuples $\{O_{t,k}^{(i)}\}$ are sampled i.i.d. from the stationary distribution $\pi^{(i)}$ of the Markov matrix $P^{(i)}$. Such an i.i.d assumption is common in the finite-time analysis of RL algorithms [13, 34, 38]. To proceed, for a fixed θ and for each $i \in [N]$, let us define $\bar{g}_i(\theta) \triangleq \mathbb{E}_{O_{t,k}^{(i)} \sim \pi^{(i)}}[g_i(\theta)]$ as the expected TD(0) update direction at iterate θ when the Markov tuple $O_{t,k}^{(i)}$ hits its stationary distribution $\pi^{(i)}$. We make the following standard bounded variance assumption [13]; similar assumptions are also made in FL analyses.

Assumption 4. $\mathbb{E}\|g_i(\theta) - \bar{g}_i(\theta)\|^2 \leq \sigma^2$ *holds for all agents* $i \in [N]$*, in each round* t *and local update* k*, and* $\forall \theta$.

Let H denote the radius of the set $\mathcal{H}$. Also, define $G \triangleq R_{\max} + 2H$ and $\nu = (1-\gamma)\bar{\omega}$, where $\bar{\omega}$ is as in Proposition 3. We can now state our first main result for FedTD(0).

Theorem 2. ***(I.I.D. Setting)*** *There exists a decreasing global step-size sequence* $\{\alpha_g^{(t)}\}$*, a fixed local step-size* α_l*, and a set of convex weights, such that a convex combination* $\tilde{\theta}_T$ *of the global models* $\{\bar{\theta}_t\}$ *satisfies the following for each* $i \in [N]$ *after* T *rounds:*

$$\mathbb{E}\Big\|V_{\tilde{\theta}_T} - V_{\theta_i^*}\Big\|_{\bar{D}}^2 \leq \tilde{\mathcal{O}}\Big(\frac{G^2}{K^2T^2} + \frac{\sigma^2}{\nu^2 NKT} + \frac{\sigma^2}{\nu^4 KT^2} + Q(\epsilon, \epsilon_1)\Big), \tag{4.2}$$

where $Q(\epsilon, \epsilon_1) = \tilde{\mathcal{O}}(\frac{B(\epsilon,\epsilon_1)G}{\nu} + \Gamma^2(\epsilon, \epsilon_1))$*,* $B(\epsilon, \epsilon_1) = H\,(\sqrt{n}\epsilon + 2(n-1)\epsilon + \mathcal{O}(\epsilon^2) + \mathcal{O}(\epsilon_1))$*, and* $\Gamma(\epsilon, \epsilon_1)$ *is as defined in Theorem 1.*

There are several important messages conveyed by Theorem 2 that we now discuss.

Discussion. To parse Theorem 2, let us start by noting that the term $Q(\epsilon, \epsilon_1)$ in Eq. (4.2) captures the effect of heterogeneity; we will comment on this term later. When $T \gg N$, the dominant term among the first three terms in Eq. (4.2) is the $\sigma^2/(\nu^2 NKT)$ term. To appreciate the tightness of this term, we note that in a centralized setting (i.e., when $N = 1$), given access to KT samples, the convergence rate of `TD`(0) is $\sigma^2/(\nu^2 KT)$ [13]. Our analysis thus reveals that by communicating just T times in KT iterations, each agent i can reduce the noise variance σ^2 further by a factor of N, i.e., *achieve a linear speedup w.r.t. the number of agents*. In a low-heterogeneity regime, i.e., when $Q(\epsilon, \epsilon_1)$ is small, we note that by combining data from different MDPs, `FedTD`(0) guarantees *fast* convergence to a model that is a good approximation of each agent's value function; by fast, we imply a N-fold speedup over the rate each agent would have achieved had it not communicated at all. Thus with little communication, `FedTD`(0) quickly provides each agent with a good model that it can then fine-tune for personalization. Theorem 2 is the first result to provide such a guarantee in the context of MARL/FRL, and complements results of a similar flavor in FL [89, 226]. When all the MDPs are identical, $Q(\epsilon, \epsilon_1) = 0$. But when the MDPs are different, should we expect such a heterogeneity term?

To further understand the effect of heterogeneity, it suffices to get rid of all the randomness in our setting. As such, suppose we replace the random `TD`(0) direction $g_i(\theta_{t,k}^{(i)})$ of each agent i in Algo. 4 by its *steady-state* deterministic version $\bar{g}_i(\theta_{t,k}^{(i)}) = \bar{b}_i - \bar{A}_i \theta_{t,k}^{(i)}$, where $\bar{A}_i$ and $\bar{b}_i$ are as defined in Section 4.3.1. This leads to a deterministic version of `FedTD`(0) that we call *mean-path* `FedTD`(0). For simplicity, we assume that there is no projection step in *mean-path* `FedTD`(0). In our next result, we exploit the affine nature of the steady-state `TD`(0) directions to characterize the effect of heterogeneity in the limiting behavior of `FedTD`(0).

Theorem 3. ***(Heterogeneity Bias)*** *Suppose $N = 2$ and $K = 1$. Let the step-size $\alpha = \alpha_g \alpha_l$ be chosen such that $I - \alpha \hat{A}$ is Schur stable, where $\hat{A} = \left(\bar{A}_1 + \bar{A}_2\right)/2$. Define $e_{i,t} \triangleq \bar{\theta}_t - \theta_i^*, i \in \{1, 2\}$.*

The output of mean-path `FedTD`*(0) then satisfies:*

$$\lim_{t\to\infty} e_{1,t} = \frac{1}{2}\hat{A}^{-1}\bar{A}_2(\theta_1^* - \theta_2^*); \quad \lim_{t\to\infty} e_{2,t} = \frac{1}{2}\hat{A}^{-1}\bar{A}_1(\theta_2^* - \theta_1^*). \tag{4.3}$$

Discussion: For the setting described in Theorem 3, the mean-path `FedTD`(0) updates follow the deterministic recursion $\bar{\theta}_{t+1} = (I - \alpha\hat{A})\bar{\theta}_t + \alpha\hat{b}$, where $\hat{b} = (1/2)(\bar{b}_1 + \bar{b}_2)$. This is a discrete-time linear time-invariant system (LTI). The dynamics of this system are stable if and only if the state transition matrix $(I - \alpha\hat{A})$ is Schur stable, justifying the choice of α in Theorem 3. The most important message conveyed by this result is that the gap between the limit point of mean-path `FedTD`(0) and the optimal parameter θ_i^* of either of the two MRPs bears a dependence on *the difference in the optimal parameters of the MRPs - a natural indicator of heterogeneity between the two MRPs*. Furthermore, this term has no dependence on the step-size α, i.e., the effect of the bias introduced by heterogeneity cannot be eliminated by making α arbitrarily small. Aligning with this observation, notice that the heterogeneity term $Q(\epsilon, \epsilon_1)$ in Eq. (4.2) is also step-size independent. The above discussion sheds some light on the fact that a term of the form $Q(\epsilon, \epsilon_1)$ is to be expected in Theorem 2. *Notably, the bias term in Eq.* (4.3) *persists even when the number of local steps is just one, i.e., even when the agents communicate with the server at all time steps.* This is a crucial difference with the standard federated optimization setting where the effect of statistical heterogeneity manifests itself *only* when the number of local steps K is strictly larger than 1 [23, 84, 137].

We end this section with a proof sketch for Theorem 2.

Proof Sketch for Theorem 2. To make a connection to the existing FL optimization proofs, we start with a key observation made in [13]. In this paper, the authors showed that for each $i \in [N]$, the mean-path `TD`(0) direction $\bar{g}_i(\theta)$ acts like a pseudo-gradient and drives the iterates towards θ_i^*.

Unfortunately, however, the average $(1/N)\sum_{i=1}^{N}\bar{g}_i(\theta)$ of the agents' mean-path TD(0) directions may not *exactly* correspond to the mean-path TD(0) direction of any MDP. Nonetheless, using Proposition 2, we prove the following key result that comes to our aid.

Lemma 2. ***(Expected pseudo-gradient heterogeneity)*** *For each $\theta \in \mathcal{H}$, we have:*

$$\left\|\bar{g}(\theta) - \frac{1}{N}\sum_{i=1}^{N}\bar{g}_i(\theta)\right\| \leq B(\epsilon, \epsilon_1), \tag{4.4}$$

where $B(\epsilon, \epsilon_1)$ is as in Theorem 2, and $\bar{g}(\theta)$ is the steady-state expected TD(0) direction of the virtual MDP.

Lemma 2 is crucial to our analysis as it shows that at least in the steady-state, the resulting FedTD(0) update direction can be closely approximated by the mean-path TD(0) direction of the virtual MDP. Furthermore, the latter acts like a pseudo-gradient pointing towards θ^* which is close to each θ_i^* based on Proposition 2. While this reasoning gives us hope, arriving at Eq. (4.2) requires a lot of work as we still need to (i) establish a linear-speedup in reducing the variance σ^2 in the noisy setting; and (ii) analyze a "client-drift" effect for our setting akin to what shows up in FL due to statistical heterogeneity and multiple local steps. In the Appendix, we provide a careful analysis that accounts for each of these issues.

4.6 Analysis of the Markovian Setting

Although the i.i.d. setting we discussed in Section 4.5 helped build a lot of intuition about the dynamics of FedTD(0), our main interest is in analyzing the setting where for each agent $i \in [N]$, the data tuples $\{O_{t,k}^{(i)}\}$ are all part of a *single Markovian trajectory* generated by $P^{(i)}$. The only assumption we will make is that these trajectories are independent across agents, i.e., the agents' observations are independent. Below, we briefly summarize some of the key difficulties that

show up in the analysis for the Markovian setting, and that merit technical innovations on our part. To that end, let us write $g_i(\theta_{t,k}^{(i)})$ more explicitly as $g_i(\theta_{t,k}^{(i)}, O_{t,k}^{(i)})$; this will make certain statistical dependencies more transparent in our subsequent discussion.

Challenges in the Markovian analysis. First, our setting inherits all the difficulties in analyzing Markovian behavior from the centralized case [13]; in particular, for each $i \in [N]$, the parameter sequence $\{\theta_{t,k}^{(i)}\}$ and the data tuples $\{O_{t,k}^{(i)}\}$ are intricately coupled. Second, the synchronization step in `FedTD`(0) creates complex statistical dependencies between the local parameter of any given agent and the past observations of *all* other agents. Third, as in the centralized case, we need to control the gradient bias $(1/NK)\sum_{i=1}^{N}\sum_{k=0}^{K-1}(g_i(\theta_{t,k}^{(i)}, O_{t,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}))$ and bound the gradient norm $\mathbb{E}\|(1/NK)\sum_{i=1}^{N}\sum_{k=0}^{K-1} g_i(\theta_{t,k}^{(i)})\|^2$. However to achieve the $\mathcal{O}(1/NKT)$-type rate, i.e., to prove linear speedup w.r.t. the number of agents N, we need to provide an analog of the variance reduction (i.e., the second term in Eq (4.2)) in the i.i.d. setting, which requires a much more delicate analysis relative to [13], since the observations of each agent i are correlated at different local steps. Indeed, naively bounding terms using the projection radius will not yield the linear speedup property. Finally, we need to control the "client-drift" effect (due to environmental heterogeneity) under the strong coupling between different random variables discussed above.

In our analysis, we will make use of the geometric mixing property of finite-state, aperiodic, and irreducible Markov chains [106]. Specifically, under Assumption 3, for each $i \in [N]$, there exists some $m_i \geq 1$ and $\rho_i \in (0,1)$, such that for all $t \geq 0$ and $0 \leq k \leq K-1$,

$$d_{TV}\left(\mathbb{P}\left(s_{t,k}^{(i)} = \cdot \mid s_{0,0}^{(i)} = s\right), \pi^{(i)}\right) \leq m_i \rho_i^{tK+k}, \forall s \in \mathcal{S},$$

holds, where we use $d_{TV}(P,Q)$ to denote the total-variation distance between two probability measures P and Q. For any $\bar{\epsilon} > 0$, let us define the mixing time for $P^{(i)}$ as $\tau_i^{\mathrm{mix}}(\bar{\epsilon}) \triangleq \min\{t \in \mathbb{N}_0 \mid m_i\rho_i^t \leq \bar{\epsilon}\}$. Finally, let $\tau(\bar{\epsilon}) = \max_{i\in[N]} \tau_i^{\mathrm{mix}}(\bar{\epsilon})$ represent the mixing time corresponding to the Markov chain

that mixes the slowest. As one might expect, and as formalized in our main result below, it is this slowest-mixing Markov chain that dictates certain terms in the convergence rate of `FedTD(0)`.

Theorem 4. *(**Markovian Setting**) There exists a decreasing global step-size sequence $\{\alpha_g^{(t)}\}$, a fixed local step-size α_l, and a set of convex weights, such that a convex combination $\tilde{\theta}_T$ of the global models $\{\bar{\theta}_t\}$ satisfies the following for each agent $i \in [N]$ after T rounds:*

$$\mathbb{E}\left\|V_{\tilde{\theta}_T} - V_{\theta_i^*}\right\|_{\bar{D}}^2 \leq \tilde{\mathcal{O}}\left(\frac{\tau^2 G^2}{K^2 T^2} + \frac{c_{quad}(\tau)}{\nu^2 NKT} + \frac{c_{lin}(\tau)}{\nu^4 KT^2} + Q(\epsilon, \epsilon_1)\right),$$

where $\tau = \lceil \frac{\tau^{\text{mix}}(\alpha_T^2)}{K} \rceil$, $\alpha_T = K\alpha_l \alpha_g^{(T)}$, $c_{quad}(\tau)$ and $c_{lin}(\tau)$ are quadratic and linear functions in τ, respectively, and $Q(\epsilon, \epsilon_1)$ is as defined in Theorem 2.

Discussion: Other than the effect of the mixing time τ which also shows up in a centralized setting [13], the rate in Theorem 4 mirrors that for the i.i.d. case in Theorem 2. *Theorem 4 is significant in that it marks the first comprehensive analysis of environmental heterogeneity in FRL under Markovian sampling.*

Proof Sketch for Theorem 4. As mentioned earlier, we cannot naively use a projection bound of the form $\mathbb{E}\|(1/NK)\sum_{i=1}^N \sum_{k=0}^{K-1} g_i(\theta_{t,k}^{(i)})\|^2 = \mathcal{O}(G^2)$ from the centralized analysis in [13] since the local models may not belong to the set $\mathcal{H}$. More importantly, going down that route will obscure the linear speedup effect. As such, we depart from the analysis techniques in [13, 182] by further decomposing the random TD direction of each agent i as $g_i(\theta_{t,k}^{(i)}) = b_i(O_{t,k}^{(i)}) - A_i(O_{t,k}^{(i)})\theta_{t,k}^{(i)}$. Since $A_i(O_{t,k}^{(i)})$ and $b_i(O_{t,k}^{(i)})$ only depend on the randomness from the Markov chain, and $O_{t,k}^{(i)}$ and $O_{t,k}^{(j)}$ are independent, we can show that the variances of $(1/NK)\sum_{i=1}^N \sum_{k=0}^{K-1} A_i(O_{t,k}^{(i)})$ and $(1/NK)\sum_{i=1}^N \sum_{k=0}^{K-1} b_i(O_{t,k}^{(i)})$ get scaled down by NK (up to higher order terms). Furthermore, to account for the fact that $A_i(O_{t,k}^{(i)})$ and $b_i(O_{t,k}^{(i)})$ differ across agents, we appeal to Lemma 2. Putting these pieces together in a careful manner yields the final rate in Theorem 4.

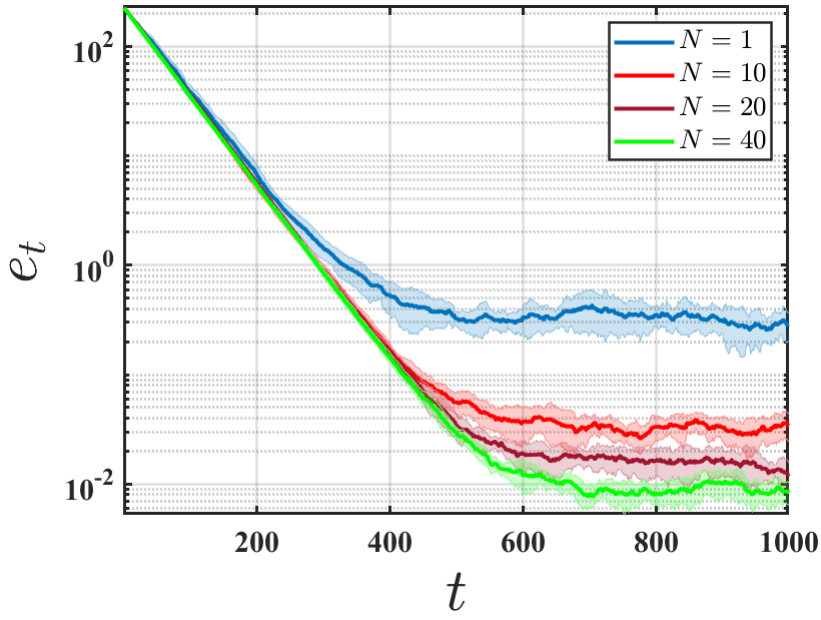


Figure 4.1: Performance of FedTD(0) under Markovian sampling with varying number of agents N. The MDP $\mathcal{M}^{(1)}$ of the first agent is randomly generated with a state space of size $n = 100$. The remaining MDPs are perturbations of $\mathcal{M}^{(1)}$ with the heterogeneity levels $\epsilon = 0.05$ and $\epsilon_1 = 0.1$. We evaluate the convergence in terms of the running error $e_t = \|\bar{\theta}_t - \theta_1^*\|^2$. Complying with theory, increasing N reduces this error. We choose the number of local steps as $K = 10$.

4.7 Chapter Summary

In this work, we have studied the problem of federated reinforcement learning under environmental heterogeneity and explored the question: *Can an agent expedite the process of learning its own value function by using information from agents interacting with different MDPs?* To answer this question, we studied the convergence of a federated TD(0) algorithm with linear function approximation, where N agents under different environments collaboratively evaluate a common policy. The main differences from the existing works are: (i) proposing a new definition of environmental heterogeneity; (ii) characterizing the effect of heterogeneity on TD(0) fixed points; (iii) introducing a virtual MDP to analyze the long-term behavior of the FedTD(0) algorithm;

and (iv) making an explicit connection between federated reinforcement learning and federated supervised learning/optimization by leveraging the virtual MDP. With these elements, we proved that if the environmental heterogeneity between agents' environments is small, then `FedTD`(0) can achieve a linear speedup under both the i.i.d and the Markovian settings, and with multiple local updates.

A few interesting extensions to this work are as follows. First, it is natural to study federated variants of other RL algorithms beyond the `TD`(0) algorithm. Second, it would be interesting to investigate whether the personalization techniques in the traditional FL optimization literature can be applied to solve FedRL problems. Instead of learning a common value function/policy, can we design personalized value functions/policies that might perform better in high-heterogeneity regimes? We leave the exploration of this interesting question as future work.

4.8 Omitted Proofs

4.8.1 Outline

This appendix provides a detailed literature survey, supporting results, and full proofs for all theorems, lemmas, and propositions in the main text. A detailed survey of relevant works is provided in Section 4.8.2. The proofs to Theorem 1, Propositions 1-3, and Lemma 2 are shown in sections 4.8.3, 4.8.4, and 4.8.5 respectively. In Section a), we provide some lemmas that are used in both the i.i.d. and Markovian sampling settings. In section 4.8.7, we introduce some notations which are relevant to the proofs of the main theorems.

Our main result in the i.i.d. sampling regime is proven in Section 4.8.8 and involves several key sub-results involving (amongst other things) a variance reduction result, and bounding the "client-drift" term at each iteration. These results are provided in Section a) and the main result, Theorem 2 is proven in Section b).

The heterogeneity bias theorem, Theorem 3, is proven in Section 4.8.9.

In Section 4.6, several key intermediate steps to proving Theorem 4 are given in subsections a)- a), with the main result being proven in Section b). More simulation results are shown in Section 4.8.11.

4.8.2 Detailed Literature Survey

Federated Learning Algorithms. The literature on algorithmic developments in federated learning is vast; as such, we only cover some of the most relevant/representative works here. The most popularly used FL algorithm, `FedAvg`, was first introduced in [133]. Several works went on to provide a detailed theoretical analysis of `FedAvg` both in the homogeneous case when all clients minimize the same objective function [67, 162, 180, 184, 214, 225], and also in the more challenging heterogeneous setting [69, 88, 89, 93, 113]. In the latter scenario, it was soon realized that `FedAvg` suffers from a "client-drift" effect that hurts its convergence performance [21, 23, 84, 211].

Since then, a lot of effort has gone into improving the convergence guarantees of `FedAvg` via a variety of technical approaches: proximal methods in `FxedProx` [167]; operator-splitting in `FedSplit` [150]; variance-reduction in `Scaffold` [84] and `S-Local-SVRG` [64]; gradient-tracking in `FedLin` [137]; and dynamic regularization in [2]. While these methods improved upon `FedAvg` in various ways, they all fell short of providing any theoretical justification for performing multiple local updates under arbitrary statistical heterogeneity. Very recently, [135] introduced the `ProxSkip` algorithm, and showed that it can indeed lead to communication savings via multiple local steps, despite arbitrary heterogeneity.

Some other approaches to tackling heterogeneous statistical distributions in FL include personalization [36, 47, 72, 190, 191], clustering [61, 76, 169, 185], representation learning [29], and the use of quantiles [100].

Analysis of TD Learning Algorithms. The first work to provide a comprehensive asymptotic analysis of the temporal difference learning algorithm with value function approximation was [205]. In this work, the authors employed the ODE method [16] that is typically used to study

asymptotic convergence rates of stochastic approximation algorithms. Providing finite-time bounds, however, turns out to be a much harder problem. Some early efforts in this direction were [98], [142], [34], and [101]. While these works were able to establish finite-time bounds for linear stochastic approximation algorithms (that subsume the TD learning algorithm), their analysis was limited to the i.i.d. sampling model. For the more challenging Markovian setting, finite-time rates have been recently derived using various perspectives: (i) by making explicit connections to optimization [13]; (ii) by taking a control-theoretic approach and studying the drift of a suitable Lyapunov function [182]; and (iii) by arguing that the mean-path temporal difference direction acts as a "gradient-splitting" of an appropriately chosen function [123]. Each of these interpretations provides interesting new insights into the dynamics of TD algorithms.

4.8.3 Perturbation bounds for TD(0) fixed points

a) Proof of Theorem 1

In this section, we prove the perturbation bounds on TD(0) fixed points shown in Theorem 1. We start by observing that:

$$\begin{aligned}
\|\bar{A}_i - \bar{A}_j\| &= \|\Phi^\top D^{(i)}(\Phi - \gamma P^{(i)}\Phi) - \Phi^\top D^{(j)}(\Phi - \gamma P^{(j)}\Phi)\| \\
&\leq \|\Phi^\top D^{(i)}(\Phi - \gamma P^{(i)}\Phi) - \Phi^\top D^{(i)}(\Phi - \gamma P^{(j)}\Phi) + \\
&\Phi^\top D^{(i)}(\Phi - \gamma P^{(j)}\Phi) - \Phi^\top D^{(j)}(\Phi - \gamma P^{(j)}\Phi)\| \\
&\leq \|\Phi^\top D^{(i)}(\Phi - \gamma P^{(i)}\Phi) - \Phi^\top D^{(i)}(\Phi - \gamma P^{(j)}\Phi)\| \\
&+ \|\Phi^\top D^{(i)}(\Phi - \gamma P^{(j)}\Phi) - \Phi^\top D^{(j)}(\Phi - \gamma P^{(j)}\Phi)\| \\
&\overset{(a)}{\leq} \gamma\|\Phi\|^2\|D^{(i)}\|\|P^{(i)} - P^{(j)}\| + \|\Phi\|^2\|D^{(i)} - D^{(j)}\|\|(I - \gamma P^{(j)})\| \\
&\overset{(b)}{\leq} \gamma\sqrt{n}\epsilon + (1+\gamma)[2(n-1)\epsilon + \mathcal{O}(\epsilon^2)], \qquad (4.5)
\end{aligned}$$

where (a) follows from the triangle inequality. The first term in (b) uses the fact that $\|\Phi\| \leq 1$, $\|D^{(i)}\| \leq 1$, and

$$\|P^{(i)} - P^{(j)}\| \leq \sqrt{n}\|P^{(i)} - P^{(j)}\|_\infty \leq \epsilon\sqrt{n}\|P^{(i)}\|_\infty = \epsilon\sqrt{n},$$

where we use Assumption 1 in the second inequality. The second term in (b) uses the the facts that $\|I - \gamma P^{(j)}\| \leq 1 + \gamma$, $\|D^{(i)} - D^{(j)}\| \leq \|D^{(i)} - D^{(j)}\|_1 \leq \|\pi^{(i)} - \pi^{(j)}\|_1$, along with Lemma 1.

Next, we bound

$$\begin{aligned}
\|\bar{b}_i - \bar{b}_j\| &= \|\Phi D^{(i)}R^{(i)} - \Phi D^{(j)}R^{(j)}\| \\
&\leq \|\Phi D^{(i)}R^{(i)} - \Phi D^{(i)}R^{(j)}\| + \|\Phi D^{(i)}R^{(j)} - \Phi D^{(j)}R^{(j)}\| \\
&\leq \|\Phi\|\|D^{(i)}\|\|R^{(i)} - R^{(j)}\| + \|\Phi\|\|D^{(i)} - D^{(j)}\|\|R^{(j)}\| \\
&\leq \epsilon_1 + R_{\max}\left(2(n-1)\epsilon + \mathcal{O}(\epsilon^2)\right), \qquad (4.6)
\end{aligned}$$

where we use Assumption 2 in the last inequality and follow the same reasoning as we used to bound $\|\bar{A}_i - \bar{A}_j\|$ above.

We are now ready to bound the gap between fixed points as:

$$\frac{\|\theta_i^* - \theta_j^*\|}{\|\theta_i^*\|} \leq \frac{\kappa(\bar{A}_i)}{1 - \kappa(\bar{A}_i)\frac{\|\bar{A}_i - \bar{A}_j\|}{\|\bar{A}_i\|}} \left(\frac{\|\bar{A}_i - \bar{A}_j\|}{\|\bar{A}_i\|} + \frac{\|\bar{b}_i - \bar{b}_j\|}{\|\bar{b}_i\|} \right). \tag{4.7}$$

Here, we leveraged the perturbation theory of linear equations in [73] Section 5.8. Finally, for any $\|\theta_i^*\| \leq H$, we have

$$\|\theta_i^* - \theta_j^*\| \leq \Gamma(\epsilon, \epsilon_1) \triangleq \frac{\kappa(\bar{A}_i)H}{1 - \kappa(\bar{A}_i)\frac{A(\epsilon)}{\delta_1}} \left(\frac{A(\epsilon)}{\delta_1} + \frac{b(\epsilon, \epsilon_1)}{\delta_2} \right),$$

where we used the fact that δ_1 and δ_2 are positive constants that lower bound $\|\bar{A}_i\|$ and $\|\bar{b}_i\|$, respectively.

4.8.4 Properties of the Virtual Markov Decision Process

a) Proof of Proposition 1

Before we prove this proposition, we present the following fact from [154]: *a Markov matrix P is irreducible and aperiodic if and only if there exists a positive integer k such that every entry of the matrix P^k is strictly positive, i.e., $P^k_{s,s'} > 0$, for all $s, s' \in \mathcal{S}$.*

For every Markov matrix $P^{(i)}$, we know that there exists such an integer k_i according to the above fact and Assumption 3 in the paper. Then we define a set $J = \{i \in [N] | w_i > 0\}$. Since $\sum_{i=1}^N w_i = 1$, and $w_i \geq 0$ holds for all $i \in [N]$, we know that J is a non-empty set. If we define $\bar{k} = \min_{i\in[J]}\{k_i\}$ and $j = \arg\min_{i\in[J]}\{k_i\}$, then we have:

$$\left(\sum_{i\in[N]} w_i P^{(i)}\right)^{\bar{k}} = \underbrace{w_j^{\bar{k}}\left(P^{(j)}\right)^{\bar{k}}}_{\text{positive}} + \underbrace{\cdots\cdots}_{\text{nonnegative}}, \tag{4.8}$$

where each entry of $w_j^{\bar{k}}\left(P^{(j)}\right)^{\bar{k}}$ is strictly positive while the other matrices in the summation are non-negative. Thus, we can conclude that the Markov chain associated with the Markov matrix $\sum_{i\in[N]} w_i P^{(i)}$ is also irreducible and aperiodic.

b) Proof of Proposition 2

Following similar arguments as in Theorem 1, we bound $\|\bar{A}_i - \bar{A}\|$:

$$\begin{aligned}
\|\bar{A}_i - \bar{A}\| &= \|\Phi^\top D^{(i)}(\Phi - \gamma P^{(i)}\Phi) - \Phi^\top \bar{D}(\Phi - \gamma\bar{P}\Phi)\| \\
&\overset{(a)}{\leq} \gamma\|\Phi\|^2\|D^{(i)}\|\|P^{(i)} - \bar{P}\| + \|\Phi\|^2\|D^{(i)} - \bar{D}\|\|(I - \gamma\bar{P})\| \\
&\overset{(b)}{\leq} \gamma\sqrt{n}\epsilon + (1+\gamma)[2(n-1)\epsilon + \mathcal{O}(\epsilon^2)] = A(\epsilon),
\end{aligned} \tag{4.9}$$

where inequality (a) follows the same reasoning as (a) in Eq. (4.5), (b) uses the same fact as (b) in Eq. (4.5), and $\|P^{(i)} - \bar{P}\| \leq \frac{1}{N}\sum_{j=1}^N \|P^{(i)} - P^{(j)}\| \leq \epsilon\sqrt{n}$ and $\|D^{(i)} - \bar{D}\| \leq 2(n-1)\epsilon + \mathcal{O}(\epsilon^2)$.

Based on the above facts: (i) $\|\bar{R}\| \leq \frac{1}{N}\sum_{i=1}^{N}\|R^{(i)}\| \leq R_{\max}$, (ii) $\|R^{(i)} - \bar{R}\| \leq \frac{1}{N}\sum_{j=1}^{N}\|R^{(i)} - R^{(j)}\| \leq \epsilon_1$ and (iii) $\|D^{(i)} - \bar{D}\| \leq 2(n-1)\epsilon + \mathcal{O}(\epsilon^2)$, we finish the proof by showing that $\|\bar{b}_i - \bar{b}\| \leq b(\epsilon, \epsilon_1)$. To do so, we follow the same steps as Eq. (4.6), and prove the bound on $\|\theta_i^* - \theta^*\|$ by following the same analysis as Eq. (4.7).

c) Proof of Proposition 3

Since the virtual MDP is an average of the agents' MDPs, i.e., $\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P^{(i)}$, the virtual Markov chain is irreducible and aperiodic from Proposition 1. The maximum eigenvalue of a symmetric positive-semidefinite matrix is a convex function. Then we have $\lambda_{\max}(\Phi^\top \bar{D}\Phi) \leq \sum_{s\in\mathcal{S}} \bar{\pi}(s)\lambda_{\max}\left(\phi(s)\phi(s)^\top\right) \leq \sum_{s\in\mathcal{S}} \bar{\pi}(s) = 1$.

To show that there exists $\omega > 0$ such that $\lambda_{\min}(\Phi^\top \bar{D}\Phi) \geq \omega > 0$, we will establish that $\Phi^\top \bar{D}\Phi$ is a positive-definite matrix. Since Φ is full-column rank, this amounts to showing that $\bar{D}$ is a positive definite matrix. From the definition of $\bar{D}$, establishing positive-definiteness of $\bar{D}$ is equivalent to arguing that every element of the stationary distribution vector $\bar{\pi}$ is strictly positive; here, $\bar{\pi}^\top \bar{P} = \bar{\pi}$. To that end, from Proposition 1, we know that the Markov chain associated with $\bar{P}$ is aperiodic and irreducible. From the Perron-Frobenius theorem [53], we conclude that indeed every entry of $\bar{\pi}$ is strictly positive. If we choose $\omega = \min_{s\in\mathcal{S}}\{\bar{\pi}(s)\} > 0$, we have $\lambda_{\min}(\Phi^\top \bar{D}\Phi) \geq \omega > 0$.

4.8.5 Pseudo-gradient heterogeneity: Proof of Lemma 2

For each $\theta \in \mathcal{H}$, we have:

$$
\begin{aligned}
&\Big\|\bar{g}(\theta) - \frac{1}{N}\sum_{i=1}^{N}\bar{g}_i(\theta)\Big\| = \Big\|\Phi^T\bar{D}(\bar{T}_\mu\Phi\theta - \Phi\theta) - \frac{1}{N}\Big(\sum_{i=1}^{N}\Phi^T D^{(i)}(T_\mu^{(i)}\Phi\theta - \Phi\theta)\Big)\Big\| \\
&\overset{(a)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\Big\|\Phi^T\bar{D}(\bar{T}_\mu\Phi\theta - \Phi\theta) - \Phi^T D^{(i)}(T_\mu^{(i)}\Phi\theta - \Phi\theta)\Big\| \\
&\overset{(b)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\Bigg\|\bar{D}\Big[\frac{1}{N}\sum_{j=1}^{N}R^{(j)} + \gamma\frac{1}{N}\sum_{j=1}^{N}P^{(j)}\Phi\theta - \Phi\theta\Big] - D^{(i)}(T_\mu^{(i)}\Phi\theta - \Phi\theta)\Bigg\| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\Bigg\|\bar{D}\Big[\frac{1}{N}\sum_{j=1}^{N}R^{(j)} + \gamma\frac{1}{N}\sum_{j=1}^{N}P^{(j)}\Phi\theta - \Phi\theta\Big] - \bar{D}(T_\mu^{(i)}\Phi\theta - \Phi\theta) \\
&\quad + \bar{D}(T_\mu^{(i)}\Phi\theta - \Phi\theta) - D^{(i)}(T_\mu^{(i)}\Phi\theta - \Phi\theta)\Bigg\| \\
&\overset{(c)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\Big\|\bar{D}\Big[\frac{1}{N}\sum_{j=1}^{N}R^{(j)} + \gamma\frac{1}{N}\sum_{j=1}^{N}P^{(j)}\Phi\theta - \Phi\theta\Big] - \bar{D}(T_\mu^{(i)}\Phi\theta - \Phi\theta)\Big\| \\
&+ \frac{1}{N}\sum_{i=1}^{N}\Big\|\bar{D}(T_\mu^{(i)}\Phi\theta - \Phi\theta) - D^{(i)}(T_\mu^{(i)}\Phi\theta - \Phi\theta)\Big\| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\Big\|\bar{D}\Big\|\Big\|\frac{1}{N}\sum_{j=1}^{N}R^{(j)} - R^{(i)}\Big\| + \gamma\Big\|\frac{1}{N}\sum_{j=1}^{N}P^{(j)} - P^{(i)}\Big\|\Big\|\Phi\theta\Big\| \\
&+ \frac{1}{N}\sum_{i=1}^{N}\Big\|\bar{D} - D^{(i)}\Big\|\Big\|T_\mu^{(i)}\Phi\theta - \Phi\theta\Big\| \\
&\overset{(d)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\Big\|\frac{1}{N}\sum_{j=1}^{N}R^{(j)} - R^{(i)}\Big\|_2 + \gamma\Big\|\frac{1}{N}\sum_{j=1}^{N}P^{(j)} - P^{(i)}\Big\|\Big\|\Phi\theta\Big\| \\
&+ \frac{1}{N}\sum_{i=1}^{N}\Big\|\bar{D} - D^{(i)}\Big\|\Big\|T_\mu^{(i)}\Phi\theta - \Phi\theta\Big\| \\
&\overset{(e)}{\leq} \Big[\epsilon_1 + \gamma\sqrt{n}\epsilon\|\Phi\theta\| + \Big[2(n-1)\epsilon + \mathcal{O}(\epsilon^2)\Big]\|\Phi\theta\| \\
&\leq H\Big[\mathcal{O}(\epsilon_1) + \gamma\sqrt{n}\epsilon + 2(n-1)\epsilon + \mathcal{O}(\epsilon^2)\Big] = B(\epsilon, \epsilon_1). \qquad (4.10)
\end{aligned}
$$

Inequalities (a) and (c) follow from the triangle inequality, (b) is due to $\|\Phi\| \leq 1$; (d) is due to the fact that $\|\bar{D}\| \leq 1$; and (e) uses the following facts: (i) $\|R^{(i)} - \bar{R}\| \leq \epsilon_1$; (ii) $\|P^{(i)} - P^{(j)}\| \leq \sqrt{n}\|P^{(i)} - P^{(j)}\|_\infty \leq \epsilon\sqrt{n}\|P^{(i)}\|_\infty = \epsilon\sqrt{n}$, which, in turn, follows from the proof of Theorem 1; (iii) $\|D^{(i)} - \bar{D}\| \leq 2(n-1)\epsilon + \mathcal{O}(\epsilon^2)$, which, in turn, follows from the proof of Theorem 1 or Eq (4.5); and (iv) $\|\theta\| \leq H$ for any $\theta \in \mathcal{H}$.

4.8.6 Auxiliary results used in the I.I.D. and Markovian settings

We make repeated use throughout the appendix (often without explicitly stating so) of the following inequalities:

- Given any two vectors $x, y \in \mathbb{R}^d$, for any $\beta > 0$, we have

$$\|x+y\|^2 \leq (1+\beta)\|x\|^2 + \left(1+\frac{1}{\beta}\right)\|y\|^2. \tag{4.11}$$

- Given any two vectors $x, y \in \mathbb{R}^d$, for any $\beta > 0$, we have

$$\langle x, y\rangle \leq \frac{\beta}{2}\|x\|^2 + \frac{1}{2\beta}\|y\|^2. \tag{4.12}$$

  This inequality goes by the name of Young's inequality.

- Given m vectors $x_1, \ldots, x_m \in \mathbb{R}^d$, the following is a simple application of Jensen's inequality:

$$\left\|\sum_{i=1}^m x_i\right\|^2 \leq m\sum_{i=1}^m \|x_i\|^2. \tag{4.13}$$

We prove the following result for the virtual MDP.

Lemma 3. *For any $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$(\theta_2 - \theta_1)^\top [\bar{g}(\theta_1) - \bar{g}(\theta_2)] \geq (1-\gamma)\left\|\hat{V}_{\theta_1} - \hat{V}_{\theta_2}\right\|_{\bar{D}}^2. \tag{4.14}$$

Proof. Consider a stationary sequence of states with random initial state $s \sim \bar{\pi}$ and subsequent state s', which, conditioned on s, is drawn from $\bar{P}(\cdot \mid s)$. Define $\phi \triangleq \phi(s)$ and $\phi' \triangleq \phi(s')$. Define $\chi_1 \triangleq \hat{V}_{\theta_2}(s) - \hat{V}_{\theta_1}(s) = (\theta_2 - \theta_1)^\top \phi$ and $\chi_2 \triangleq \hat{V}_{\theta_2}(s') - \hat{V}_{\theta_1}(s') = (\theta_2 - \theta_1)^\top \phi'$. By stationarity, χ_1 and χ_2 are two correlated random variables with the same same marginal distribution. By definition, $\mathbb{E}[\chi_1^2] = \mathbb{E}[\chi_2^2] = \left\|\hat{V}_{\theta_2} - \hat{V}_{\theta_2}\right\|_{\bar{D}}^2$ since s, s' are drawn from $\bar{\pi}$. And we have,

$$\bar{g}(\theta_1) - \bar{g}(\theta_2) = \mathbb{E}\left[\phi\left(\gamma\phi' - \phi\right)^\top (\theta_1 - \theta_2)\right] = \mathbb{E}\left[\phi\left(\chi_1 - \gamma\chi_2\right)\right].$$

Therefore,

$$\begin{aligned}
(\theta_2 - \theta_1)^\top \left[\bar{g}(\theta_1) - \bar{g}(\theta_2)\right] &= \mathbb{E}\left[\chi_1\left(\chi_1 - \gamma\chi_2\right)\right] \\
&= \mathbb{E}\left[\chi_1^2\right] - \gamma\mathbb{E}\left[\chi_1\chi_2\right] \\
&\geq (1-\gamma)\mathbb{E}\left[\chi_1^2\right] \\
&= (1-\gamma)\left\|\hat{V}_{\theta_2} - \hat{V}_{\theta_2}\right\|_{\bar{D}}^2,
\end{aligned}$$

where we use the Cauchy-Schwartz inequality to conclude $\mathbb{E}[\chi_1\chi_2] \leq \sqrt{\mathbb{E}[\chi_1^2]}\sqrt{\mathbb{E}[\chi_2^2]} = \mathbb{E}[\chi_1^2]$.

□

Lemma 4. *For any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*

$$\|\bar{g}(\theta_1) - \bar{g}(\theta_2)\| \leq 2\left\|\hat{V}_{\theta_1} - \hat{V}_{\theta_2}\right\|_{\bar{D}}. \tag{4.15}$$

Proof. Following the analysis of Lemma 3, we have

$$\begin{aligned}
\|\bar{g}(\theta_1) - \bar{g}(\theta_2)\| &= \|\mathbb{E}\left[\phi\left(\chi_1 - \gamma\chi_2\right)\right]\| \\
&\leq \sqrt{\mathbb{E}\left[\|\phi\|^2\right]}\sqrt{\mathbb{E}\left[\left(\chi_1 - \gamma\chi_2\right)^2\right]} \\
&\leq \sqrt{\mathbb{E}\left[\chi_1^2\right]} + \gamma\sqrt{\mathbb{E}\left[\chi_2^2\right]} \\
&= (1+\gamma)\sqrt{\mathbb{E}\left[\chi_1^2\right]},
\end{aligned} \tag{4.16}$$

where the second inequality is due to $\|\phi\| \leq 1$ and the final equality is due to $\mathbb{E}[\chi_1^2] = \mathbb{E}[\chi_2^2]$. We finish the proof by using the fact that $\mathbb{E}[\chi_1^2] = \left\|\hat{V}_{\theta_2} - \hat{V}_{\theta_2}\right\|_{\bar{D}}^2$ and $1 + \gamma \leq 2$. □

With this Lemma, we next show that the steady-state TD(0) update direction $\bar{g}$ and $\bar{g}_i$ are 2-Lipschitz.

Lemma 5. *(2-Lipschitzness of steady-state TD(0) update direction) For any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*

$$\|\bar{g}(\theta_1) - \bar{g}(\theta_2)\| \leq 2\|\theta_1 - \theta_2\|. \tag{4.17}$$

And for each agent $i \in [N]$, we have

$$\|\bar{g}_i(\theta_1) - \bar{g}_i(\theta_2)\| \leq 2\|\theta_1 - \theta_2\|. \tag{4.18}$$

Proof. From Lemma 4, we can easily conclude that the steady-state TD(0) update direction $\bar{g}$ for the vitual MDP is 2-Lipschitz, i.e.,

$$\|\bar{g}(\theta_1) - \bar{g}(\theta_2)\| \leq 2\|\theta_1 - \theta_2\|, \tag{4.19}$$

based on the fact that $\lambda_{\max}(\Phi^\top \bar{D}\Phi) \leq 1$. We can follow the same reasoning to prove Eq (4.18) since $\|\bar{g}_i(\theta_1) - \bar{g}_i(\theta_2)\| \leq 2\left\|\hat{V}_{\theta_1} - \hat{V}_{\theta_2}\right\|_{D_i}$ holds for each $i \in [N]$ from [13]. □

Next, we prove an analog of the Lipschitz property in Lemma 5 for the random TD(0) update direction of each agent i.

Lemma 6. *(2-Lipschitzness of random TD(0) update direction) For any $\theta_1, \theta_2 \in \mathbb{R}^d$ and $i \in [N]$, we have*

$$\|g_i(\theta_1) - g_i(\theta_2)\| \leq 2\|\theta_1 - \theta_2\|.$$

Proof. In this proof, we will use the fact that the random TD(0) update direction of agent i at the t-th communication round and k-th local update is an affine function of the parameter θ. In

particular, we have $g_i(\theta) = b_i(O_{t,k}^{(i)}) - A_i(O_{t,k}^{(i)})\theta$, where $A_i(O_{t,k}^{(i)}) = \phi(s_{t,k}^{(i)})(\phi^\top(s_{t,k}^{(i)}) - \gamma\phi^\top(s_{t,k+1}^{(i)}))$ and $b_i(O_{t,k}^{(i)}) = r(s_{t,k}^{(i)})\phi(s_{t,k}^{(i)})$. Thus, we have

$$\begin{aligned}
\|g_i(\theta_1) - g_i(\theta_2)\| &= \left\| A_i(O_{t,k}^{(i)})(\theta_1 - \theta_2) \right\| \\
&\leq \left\| A_i(O_{t,k}^{(i)}) \right\| \|\theta_1 - \theta_2\| \\
&\leq \left(\left\| \phi\left(s_{t,k}^i\right) \right\|^2 + \gamma \left\| \phi\left(s_{t,k}^i\right) \right\| \left\| \phi\left(s_{t,k+1}^i\right) \right\| \right) \|\theta_1 - \theta_2\| \\
&\leq 2\|\theta_1 - \theta_2\|,
\end{aligned}$$

where we used that $\|\phi(s)\| \leq 1, \forall s \in \mathcal{S}$ in the last step. □

4.8.7 Notation

For our subsequent analysis, we will use $\mathcal{F}_k^t$ to denote the filtration that captures all the randomness up to the k-th local step in round t. We will also use $\mathcal{F}^t$ to represent the filtration capturing all the randomness up to the end of round $t-1$. With a slight abuse of notation, $\mathcal{F}_{-1}^t$ is to be interpreted as $\mathcal{F}^t$. Based on the description of `FedTD`(0), it should be apparent that for each $i \in [N]$, $\theta_{t,k}^{(i)}$ is $\mathcal{F}_{k-1}^t$-measurable and $\bar{\theta}_t$ is $\mathcal{F}^t$-measurable. Furthermore, we use $\mathbb{E}_t$ to represent the expectation conditioned on all the randomness up to the end of round $t-1$.

For simplicity, we define $\delta_t = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left\|\theta_{t,k}^{(i)} - \bar{\theta}_t\right\|$ and $\Delta_t = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left\|\theta_{t,k}^{(i)} - \bar{\theta}_t\right\|^2$. The latter term is referred to as the *drift term*. Note that $(\delta_t)^2 \leq \Delta_t$ holds for all t via Jensen's inequality. Unless specified otherwise, $\|\cdot\|$ denotes the Euclidean norm.

Step-size: Throughout the paper, we encounter three kinds of step-sizes: local step-size α_l, global step-size α_g, and the effective step-size α. Some of our results will rely on effective step-sizes that decay as a function of the communication round t; we will use $\{\alpha_t\}$ to represent such a decaying effective step-size sequence. While the local step-size α_ℓ will always be held constant, the decay in the effective step-size will be achieved by making the global step-size at the server decay with the communication round. Accordingly, we will use $\{\alpha_g^{(t)}\}$ to represent the decaying global step-size sequence at the server. In what follows, unless specified in the subscript, all the step-sizes appearing in the proofs refer to the effective step-size.

4.8.8 Proof of the i.i.d. setting

a) Auxiliary lemmas for Theorem 2

- Variance reduction

Lemma 7. *(Variance reduction in the i.i.d. setting). In the i.i.d. setting, under Assumption 4, at each round t, we have* $\mathbb{E}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left[g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})\right]\right\|^2 \leq \frac{\sigma^2}{NK}$.

Proof. Define $Y_{t,k}^{(i)} \triangleq g_i(O_{t,k}^{(i)},\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})$. Since $\{O_{t,k}^{(i)}\}$ is drawn i.i.d. over time from its stationary distribution $\pi^{(i)}$, we have $\mathbb{E}[Y_{t,k}^{(i)}] = \mathbb{E}\left[\mathbb{E}[Y_{t,k}^{(i)} \mid \theta_{t,k}^{(i)}]\right] = 0$. As we mentioned before, for each $i \in [N]$, $\theta_{t,k}^{(i)}$ is $\mathcal{F}_{k-1}^t$-measurable. If we condition on $\mathcal{F}_{k-1}^t$, we know that $\theta_{t,k}^{(i)}$ and $\theta_{t,k}^{(j)}$ are deterministic and the only randomness in $Y_{t,k}^{(i)}$ and $Y_{t,k}^{(j)}$ come from $O_{t,k}^{(i)}$ and $O_{t,k}^{(j)}$, which are independent. Therefore, $Y_{t,k}^{(i)}$ and $Y_{t,k}^{(j)}$ are independent conditioned on $\mathcal{F}_{k-1}^t$.

For every $i \neq j \in [N]$, we have

$$\mathbb{E}\left[\left\langle Y_{t,k}^{(i)}, Y_{t,k}^{(j)}\right\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\left\langle Y_{t,k}^{(i)}, Y_{t,k}^{(j)}\right\rangle \mid \mathcal{F}_{k-1}^t\right]\right] \overset{(a)}{=} \mathbb{E}\left[\left\langle \mathbb{E}[Y_{t,k}^{(i)} \mid \mathcal{F}_{k-1}^t], \mathbb{E}[Y_{t,k}^{(j)} \mid \mathcal{F}_{k-1}^t]\right\rangle\right] = 0, \tag{4.20}$$

where (a) follows from the fact that $Y_{t,k}^{(i)}$ and $Y_{t,k}^{(j)}$ are independent conditioned on $\mathcal{F}_{k-1}^t$. For every $k < l$ and $i, j \in [N]$,

$$\mathbb{E}\left[\left\langle Y_{t,k}^{(i)}, Y_{t,l}^{(j)}\right\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\left\langle Y_{t,k}^{(i)}, Y_{t,l}^{(j)}\right\rangle \;\middle|\; \mathcal{F}_{l-1}^t\right]\right] = \mathbb{E}\left[\left\langle Y_{t,k}^{(i)}, \mathbb{E}[Y_{t,l}^{(j)} \mid \mathcal{F}_{l-1}^t]\right\rangle\right] = 0. \tag{4.21}$$

Then,

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left[g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})\right]\right\|^2 &= \mathbb{E}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Y_{t,k}^{(i)}\right\|^2 \\ &= \frac{1}{N^2K^2}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\|Y_{t,k}^{(i)}\|^2 + \frac{2}{N^2K^2}\underbrace{\sum_{i<j}\sum_{k=0}^{K-1}\mathbb{E}[\langle Y_{t,k}^{(i)}, Y_{t,k}^{(j)}\rangle]}_{0}\end{aligned}$$

$$+\frac{2}{N^2K^2}\sum_{i,j=1}^{N}\sum_{k<l}\underbrace{\mathbb{E}[\langle Y_{t,k}^{(i)}, Y_{t,l}^{(j)}\rangle]}_{0}$$
$$\leq \frac{\sigma^2}{NK},$$

where the second equality is due to Eq (4.20) and Eq (4.21) and the last inequality is due to Assumption 4. □

• Per Round Progress

First, we characterize the error decrease at each iteration in the following lemma.

Lemma 8. *(Per Round Progress). If the local step-size α_l satisfies $\alpha_l \leq \frac{(1-\gamma)\bar{\omega}}{48K}$, then the updates of* `FedTD`*(0) with any global step-size α_g satisfy*

$$\mathbb{E}\|\bar{\theta}_{t+1}-\theta^*\|^2 \leq (1+\zeta_1)\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2 + 2\alpha\mathbb{E}\langle\bar{g}(\bar{\theta}_t),\bar{\theta}_t-\theta^*\rangle + 6\alpha^2\mathbb{E}\left\|\bar{g}(\bar{\theta}_t)\right\|^2$$
$$+ 4\alpha^2\left(\frac{1}{\zeta_1}+6\right)\mathbb{E}[\Delta_t] + \frac{2\alpha^2\sigma^2}{NK} + 2\alpha B(\epsilon,\epsilon_1)G + 6\alpha^2B^2(\epsilon,\epsilon_1), \tag{4.22}$$

where ζ_1 is any positive constant, and α is the effective step-size, i.e., $\alpha = K\alpha_l\alpha_g$.

Proof.

$$\mathbb{E}\|\bar{\theta}_{t+1}-\theta^*\|^2 = \mathbb{E}\left\|\Pi_{2,\mathcal{H}}\left(\bar{\theta}_t+\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{t,k}^{(i)})-\theta^*\right)\right\|^2 \quad \text{(updating rule)}$$
$$\leq \mathbb{E}\left\|\bar{\theta}_t+\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{t,k}^{(i)})-\theta^*\right\|^2 \quad \text{(projection is non-expansive)}$$
$$= \mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2 + 2\mathbb{E}\langle\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{t,k}^{(i)}),\bar{\theta}_t-\theta^*\rangle + \mathbb{E}\left\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{t,k}^{(i)})\right\|^2$$
$$= \mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2 + \frac{2\alpha}{NK}\underbrace{\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\langle g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)}),\bar{\theta}_t-\theta^*\rangle}_{\mathcal{C}_1=0}$$
$$+\frac{2\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\langle\bar{g}_i(\theta_{t,k}^{(i)}),\bar{\theta}_t-\theta^*\rangle + \mathbb{E}\left\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{t,k}^{(i)})\right\|^2$$

$$
\begin{aligned}
&= \mathbb{E}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + \frac{2\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\langle \bar{g}_i(\theta_{t,k}^{(i)}), \bar{\theta}_t - \theta^*\rangle + \mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} g_i(\theta_{t,k}^{(i)})\Big\|^2 \\
&\leq \mathbb{E}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + \frac{2\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\langle \bar{g}_i(\theta_{t,k}^{(i)}), \bar{\theta}_t - \theta^*\rangle \\
&+ 2\mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big[g_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)})\Big]\Big\|^2 + 2\mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\bar{g}_i(\theta_{t,k}^{(i)})\Big\|^2 \quad \text{(Young's inequality (6.7))} \\
&\overset{(a)}{\leq} \mathbb{E}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + \frac{2\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\langle \bar{g}_i(\theta_{t,k}^{(i)}), \bar{\theta}_t - \theta^*\rangle + \frac{2\sigma^2}{NK} + 2\mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\bar{g}_i(\theta_{t,k}^{(i)})\Big\|^2 \\
&= \mathbb{E}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + \frac{2\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\langle \bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\bar{\theta}_t) + \bar{g}_i(\bar{\theta}_t) - \bar{g}(\bar{\theta}_t) + \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\rangle \qquad (4.23) \\
&+ 2\mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\bar{g}_i(\theta_{t,k}^{(i)})\Big\|^2 + \frac{2\alpha^2\sigma^2}{NK} \\
&\leq \mathbb{E}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + \frac{2\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\langle \bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\bar{\theta}_t), \bar{\theta}_t - \theta^*\rangle + \frac{2\alpha}{N}\sum_{i=1}^{N}\mathbb{E}\langle \bar{g}_i(\bar{\theta}_t) - \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\rangle \\
&+ 2\alpha\mathbb{E}\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\rangle + 2\mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\bar{g}_i(\theta_{t,k}^{(i)})\Big\|^2 + \frac{2\alpha^2\sigma^2}{NK} \\
&\leq (1+\zeta_1)\mathbb{E}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + \frac{1}{\zeta_1}\mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big[\bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\bar{\theta}_t)\Big]\Big\|^2 + 2\alpha B(\epsilon,\epsilon_1)G \\
&+ 2\alpha\mathbb{E}\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\rangle + 2\mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\bar{g}_i(\theta_{t,k}^{(i)})\Big\|^2 + \frac{2\alpha^2\sigma^2}{NK} \quad \text{(Eq (6.7) and Lemma 2)} \\
&\leq (1+\zeta_1)\mathbb{E}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + \frac{4\alpha^2}{\zeta_1 NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\theta_{t,k}^{(i)} - \bar{\theta}_t\Big\|^2 + 2\alpha B(\epsilon,\epsilon_1)G \\
&+ 2\alpha\mathbb{E}\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\rangle + 2\mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\bar{g}_i(\theta_{t,k}^{(i)})\Big\|^2 + \frac{2\alpha^2\sigma^2}{NK} \quad \text{(2-Lipschitz of } \bar{g}_i \text{ in Lemma 5)} \\
&\leq (1+\zeta_1)\mathbb{E}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + \frac{4\alpha^2}{\zeta_1}\mathbb{E}[\Delta_t] + \frac{2\alpha^2\sigma^2}{NK} + 2\alpha B(\epsilon,\epsilon_1)G \\
&+ 2\alpha\mathbb{E}\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\rangle + 2\mathbb{E}\Big\|\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big[\bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\bar{\theta}_t) + \bar{g}_i(\bar{\theta}_t) - \bar{g}(\bar{\theta}_t) + \bar{g}(\bar{\theta}_t)\Big]\Big\|^2 \\
&\leq (1+\zeta_1)\mathbb{E}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + \frac{4\alpha^2}{\zeta_1}\mathbb{E}[\Delta_t] + \frac{2\alpha^2\sigma^2}{NK} + 2\alpha B(\epsilon,\epsilon_1)G
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\alpha\mathbb{E}\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^* \rangle + 6\mathbb{E}\left\| \frac{\alpha}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\bar{\theta}_t) \right\|^2 \\
&+ 6\mathbb{E}\left\| \frac{\alpha}{N} \sum_{i=1}^{N} \left[\bar{g}_i(\bar{\theta}_t) - \bar{g}(\bar{\theta}_t) \right] \right\|^2 + 6\mathbb{E}\left\| \alpha \bar{g}(\bar{\theta}_t) \right\|^2 \quad \text{(Eq (6.7) and Lemma 2)} \\
&\leq (1+\zeta_1)\mathbb{E}\left\| \bar{\theta}_t - \theta^* \right\|^2 + \frac{4\alpha^2}{\zeta_1}\mathbb{E}[\Delta_t] + \frac{2\alpha^2\sigma^2}{NK} + 2\alpha B(\epsilon, \epsilon_1)G \\
&+ 2\alpha\mathbb{E}\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^* \rangle + 24\alpha^2\mathbb{E}[\Delta_t] \quad \text{(2-Lipschitz of } \bar{g}_i\text{)} \\
&+ 6\alpha^2 B^2(\epsilon, \epsilon_1) + 6\alpha^2\mathbb{E}\left\| \bar{g}(\bar{\theta}_t) \right\|^2 \quad \text{(Eq (6.7))} \\
&= (1+\zeta_1)\mathbb{E}\left\| \bar{\theta}_t - \theta^* \right\|^2 + 2\alpha\mathbb{E}\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^* \rangle + 6\alpha^2\mathbb{E}\left\| \bar{g}(\bar{\theta}_t) \right\|^2 \\
&+ 4\alpha^2 \left(\frac{1}{\zeta_1} + 6 \right) \mathbb{E}[\Delta_t] + \frac{2\alpha^2\sigma^2}{NK} + 2\alpha B(\epsilon, \epsilon_1)G + 6\alpha^2 B^2(\epsilon, \epsilon_1),
\end{aligned} \tag{4.24}
$$

where (a) is due to Lemma 7. Furthermore, the reason why $\mathcal{C}_1 = 0$ is as follows:

$$
\begin{aligned}
\mathcal{C}_1 &= \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbb{E}\langle g_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}), \bar{\theta}_t - \theta^* \rangle \\
&= \sum_{i=1}^{N} \sum_{k=0}^{K-2} \mathbb{E}\langle g_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}), \bar{\theta}_t - \theta^* \rangle + \sum_{i=1}^{N} \mathbb{E}\langle g_i(\theta_{t,K-1}^{(i)}) - \bar{g}_i(\theta_{t,K-1}^{(i)}), \bar{\theta}_t - \theta^* \rangle \\
&= \sum_{i=1}^{N} \sum_{k=0}^{K-2} \mathbb{E}\langle g_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}), \bar{\theta}_t - \theta^* \rangle + \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{E}\left[\langle g_i(\theta_{t,K-1}^{(i)}) - \bar{g}_i(\theta_{t,K-1}^{(i)}), \bar{\theta}_t - \theta^* \rangle \mid \mathcal{F}_{K-1}^t \right] \right] \\
&= \sum_{i=1}^{N} \sum_{k=0}^{K-2} \mathbb{E}\langle g_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}), \bar{\theta}_t - \theta^* \rangle + \sum_{i=1}^{N} \mathbb{E}\left[\left\langle \bar{\theta}_t - \theta^*, \underbrace{\mathbb{E}\left[g_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}) \mid \mathcal{F}_{K-1}^t \right]}_{0} \right\rangle \right] \\
&= \sum_{i=1}^{N} \sum_{k=0}^{K-2} \mathbb{E}\langle g_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}), \bar{\theta}_t - \theta^* \rangle.
\end{aligned}
$$

We can keep repeating this procedure by iteratively conditioning on $\mathcal{F}_{K-2}^t, \cdots, \mathcal{F}_1^t, \mathcal{F}_0^t$. □

- Drift Term Analysis

We now turn to bounding the drift term Δ_t.

Lemma 9. *(Bounded Client Drift) The drift term Δ_t at the t-th round can be bounded as*

$$\mathbb{E}[\Delta_t] = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\left\|\theta_{t,k}^{(i)} - \bar{\theta}_t\right\|^2 \leq 27(\sigma^2 + 3KB^2(\epsilon,\epsilon_1) + 2KG^2)\frac{\alpha^2}{K\alpha_g^2}, \tag{4.25}$$

provided the fixed local step-size α_l satisfies $\alpha_l \leq \min \frac{(1-\gamma)\bar{\omega}}{48K}$.

Proof.

$$\begin{aligned}
&\mathbb{E}\left\|\theta_{t,k}^{(i)} - \bar{\theta}_t\right\|^2 = \mathbb{E}\left\|\theta_{t,k-1}^{(i)} + \alpha_l g_i(\theta_{t,k-1}^{(i)}) - \bar{\theta}_t\right\|^2 \quad \text{(updating rule)}\\
&= \mathbb{E}\left\|\theta_{t,k-1}^{(i)} + \alpha_l \bar{g}_i(\theta_{t,k-1}^{(i)}) - \bar{\theta}_t + \alpha_l\left(g_i(\theta_{t,k-1}^{(i)}) - \bar{g}_i(\theta_{t,k-1}^{(i)})\right)\right\|^2\\
&= \mathbb{E}\left\|\theta_{t,k-1}^{(i)} + \alpha_l \bar{g}_i(\theta_{t,k-1}^{(i)}) - \bar{\theta}_t\right\|^2 + \alpha_l^2\mathbb{E}\left\|g_i(\theta_{t,k-1}^{(i)}) - \bar{g}_i(\theta_{t,k-1}^{(i)})\right\|^2\\
&+ 2\alpha_l \underbrace{\mathbb{E}\left[\mathbb{E}\left\langle g_i(\theta_{t,k-1}^{(i)}) - \bar{g}_i(\theta_{t,k-1}^{(i)}), \theta_{t,k-1}^{(i)} + \alpha_l \bar{g}_i(\theta_{t,k-1}^{(i)}) - \bar{\theta}_t \;\middle|\; \mathcal{F}_{k-1}^t\right\rangle\right]}_{\mathcal{C}_2=0}\\
&\overset{(a)}{\leq} (1+\zeta_2)\mathbb{E}\left\|\theta_{t,k-1}^{(i)} + \alpha_l \bar{g}(\theta_{t,k-1}^{(i)}) - \bar{\theta}_t\right\|^2 + (1+\frac{1}{\zeta_2})\alpha_l^2\mathbb{E}\left\|\bar{g}(\theta_{t,k-1}^{(i)}) - \bar{g}_i(\theta_{t,k-1}^{(i)})\right\|^2\\
&+ \alpha_l^2\mathbb{E}\left\|g_i(\theta_{t,k-1}^{(i)}) - \bar{g}_i(\theta_{t,k-1}^{(i)})\right\|^2\\
&\overset{(b)}{\leq} (1+\zeta_2)(1+\zeta_3)\mathbb{E}\left\|\theta_{t,k-1}^{(i)} + \alpha_l \bar{g}(\theta_{t,k-1}^{(i)}) - \bar{\theta}_t - \alpha_l\bar{g}(\bar{\theta}_t)\right\|^2 + (1+\zeta_2)(1+\frac{1}{\zeta_3})\alpha_l^2\mathbb{E}\left\|\bar{g}(\bar{\theta}_t)\right\|^2\\
&+ (1+\frac{1}{\zeta_2})\alpha_l^2\mathbb{E}\left\|\bar{g}(\theta_{t,k-1}^{(i)}) - \bar{g}(\bar{\theta}_t) + \bar{g}(\bar{\theta}_t) - \bar{g}_i(\bar{\theta}_t) + \bar{g}_i(\bar{\theta}_t) - \bar{g}_i(\theta_{t,k-1}^{(i)})\right\|^2 + \alpha_l^2\sigma^2\\
&\overset{(c)}{\leq} (1+\zeta_2)(1+\zeta_3)\mathbb{E}\left\|\theta_{t,k-1}^{(i)} + \alpha_l \bar{g}(\theta_{t,k-1}^{(i)}) - \bar{\theta}_t - \alpha_l\bar{g}(\bar{\theta}_t)\right\|^2 + (1+\zeta_2)(1+\frac{1}{\zeta_3})\alpha_l^2\mathbb{E}\left\|\bar{g}(\bar{\theta}_t)\right\|^2\\
&+ 3(1+\frac{1}{\zeta_2})\alpha_l^2\mathbb{E}\left\|\bar{g}(\theta_{t,k-1}^{(i)}) - \bar{g}(\bar{\theta}_t)\right\|^2 + 3(1+\frac{1}{\zeta_2})\alpha_l^2\mathbb{E}\left\|\bar{g}(\bar{\theta}_t) - \bar{g}_i(\bar{\theta}_t)\right\|^2\\
&+ 3(1+\frac{1}{\zeta_2})\alpha_l^2\mathbb{E}\left\|\bar{g}_i(\bar{\theta}_t) - \bar{g}_i(\theta_{t,k-1}^{(i)})\right\|^2 + \alpha_l^2\sigma^2\\
&\overset{(d)}{\leq} (1+\zeta_2)(1+\zeta_3)\left[1 - (2\alpha_l(1-\gamma) - 4\alpha_l^2)\bar{\omega}\right]\mathbb{E}\left\|\theta_{t,k-1}^{(i)} - \bar{\theta}_t\right\|^2 + (1+\zeta_2)(1+\frac{1}{\zeta_3})\alpha_l^2\mathbb{E}\left\|\bar{g}(\bar{\theta}_t)\right\|^2\\
&+ 12(1+\frac{1}{\zeta_2})\alpha_l^2\mathbb{E}\left\|\theta_{t,k-1}^{(i)} - \bar{\theta}_t\right\|^2 + 3(1+\frac{1}{\zeta_3})\alpha_l^2B^2(\epsilon,\epsilon_1) + 12(1+\frac{1}{\zeta_3})\alpha_l^2\mathbb{E}\left\|\theta_{t,k-1}^{(i)} - \bar{\theta}_t\right\|^2 + \alpha_l^2\sigma^2\\
&= (1+\zeta_2)(1+\zeta_3)\left[1 - (2\alpha_l(1-\gamma) - 4\alpha_l^2)\bar{\omega} + \frac{24(1+\frac{1}{\zeta_3})\alpha_l^2}{(1+\zeta_2)(1+\zeta_3)}\right]\mathbb{E}\left\|\theta_{t,k-1}^{(i)} - \bar{\theta}_t\right\|^2\\
&+ \underbrace{(1+\zeta_2)(1+\frac{1}{\zeta_3})\alpha_l^2\mathbb{E}\left\|\bar{g}(\bar{\theta}_t)\right\|^2 + 3(1+\frac{1}{\zeta_3})\alpha_l^2B^2(\epsilon,\epsilon_1) + \alpha_l^2\sigma^2}_{\mathcal{D}_1},
\end{aligned}$$

where we used the inequality in Eq (6.6) with any positive constant ζ_2 for (a); for (b), we used Assumption 4 and the same reasoning as Eq (6.6) with any positive constant ζ_3; for (c), we used the inequality in Eq (6.8) to bound the third term; and for (d), we used Lemma 3 and Lemma 4 to bound the first term, the 2-Lipschitz property of $\bar{g}$, $\bar{g}_i$ (i.e., Lemma 5) in the third term and the fifth term, and the gradient heterogeneity bound from Lemma 2 in the fourth term. If we define $\zeta_4 \triangleq (1+\zeta_2)(1+\zeta_3)\left[1-(2\alpha_l(1-\gamma)-4\alpha_l^2)\bar{\omega}+\frac{24(1+\frac{1}{\zeta_3})\alpha_l^2}{(1+\zeta_2)(1+\zeta_3)}\right]$ and define $\mathcal{D}_1$ as above, we have that

$$\mathbb{E}\left\|\theta_{t,k}^{(i)}-\bar{\theta}_t\right\|^2 \leq \zeta_4 \mathbb{E}\left\|\theta_{t,k-1}^{(i)}-\bar{\theta}_t\right\|^2+\mathcal{D}_1. \tag{4.26}$$

Next, we set $\zeta_2=\zeta_3=\frac{1}{K-1}$, $K\geq 2$, and choose the local step-size α_l to satisfy

$$\frac{\alpha_l(1-\gamma)\bar{\omega}}{2}\geq 4\alpha_l^2\bar{\omega} \;\;\&\;\; \frac{\alpha_l(1-\gamma)\bar{\omega}}{2}\geq \frac{24(1+\frac{1}{\zeta_3})\alpha_l^2}{(1+\zeta_2)(1+\zeta_3)},$$

so that $\left[1-(2\alpha_l(1-\gamma)-4\alpha_l^2)\bar{\omega}+\frac{24(1+\frac{1}{\zeta_2})\alpha_l^2}{(1+\zeta_2)(1+\zeta_3)}\right]\leq 1-\alpha_l(1-\gamma)\bar{\omega}$. These inequalities hold when $\alpha_l\leq \min\frac{(1-\gamma)\bar{\omega}}{48K}$. Then, Eq (4.26) becomes

$$\mathbb{E}\left\|\theta_{t,k}^{(i)}-\bar{\theta}_t\right\|^2 \leq (1+\frac{3}{K-1})\left[1-\alpha_l(1-\gamma)\bar{\omega}\right]\mathbb{E}\left\|\theta_{t,k-1}^{(i)}-\bar{\theta}_t\right\|^2+\mathcal{D}_1.$$

If we unroll this recurrence above, using $\theta_{r,0}^{(i)}=\bar{\theta}_t$, we have that

$$\begin{aligned}
&\mathbb{E}\left\|\theta_{t,k}^{(i)}-\bar{\theta}_t\right\|^2 \leq \sum_{s=0}^{k-1}\mathcal{D}_1\left\{\Pi_{j=s+1}^{k-1}(1+\frac{3}{K-1})\left[1-\alpha(1-\gamma)\bar{\omega}\right]\right\}\\
&\overset{(e)}{\leq}\sum_{s=0}^{k-1}\left[\alpha_l^2\sigma^2+3K\alpha_l^2B^2(\epsilon,\epsilon_1),+2\alpha_l^2K\mathbb{E}\left\|\bar{g}(\bar{\theta}_t)\right\|^2\right]\times\Pi_{j=s+1}^{k-1}(1+\frac{3}{K-1})[1-\alpha_l(1-\gamma)\bar{\omega}]\\
&\leq\sum_{s=0}^{k-1}\left[\alpha_l^2\sigma^2+3\alpha_l^2KB^2(\epsilon,\epsilon_1)+2\alpha_l^2K\mathbb{E}\left\|\bar{g}(\bar{\theta}_t)\right\|^2\right](1+\frac{3}{K-1})^{K-1}\Pi_{j=s+1}^{k-1}[1-\alpha_l(1-\gamma)\bar{\omega}]\\
&\overset{(f)}{\leq} 27(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2K\mathbb{E}\left\|\bar{g}(\bar{\theta}_t)\right\|^2)\sum_{s=0}^{k-1}\alpha_l^2\times\underbrace{\Pi_{j=s+1}^{k-1}[1-\alpha(1-\gamma)\bar{\omega}]}_{\leq 1}\\
&\leq 27(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2)K\alpha_l^2 \quad \text{(constant local step-size)}
\end{aligned}$$

where we used the fact that $(1+\zeta_2)(1+\frac{1}{\zeta_3}) \leq 2K$ for (e) and $(1+\frac{3}{K-1})^{K-1} \leq 27$ for (f). we finish the proof by substituting $\alpha_l = \frac{\alpha}{K\alpha_g}$. □

If we incorporate Eq (4.25) into Eq (4.22), we have that

$$\mathbb{E}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2 \leq (1+\zeta_1)\mathbb{E}\left\|\bar{\theta}_r-\theta^*\right\|^2 + 2\alpha\mathbb{E}\langle\bar{g}(\bar{\theta}_r),\bar{\theta}_r-\theta^*\rangle + 6\alpha^2\mathbb{E}\left\|\bar{g}(\bar{\theta}_r)\right\|^2 + 108\frac{\alpha^4}{K\alpha_g^2}(6+\frac{1}{\zeta_1})(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2) + \frac{2\alpha^2\sigma^2}{NK} + 2\alpha B(\epsilon,\epsilon_1)G + 6\alpha^2B^2(\epsilon,\epsilon_1) \tag{4.27}$$

• Parameter Selection

Lemma 10. *Define $\nu \triangleq (1-\gamma)\bar{\omega}$. If we choose any effective step-size $\alpha = K\alpha_g\alpha_l < \frac{(1-\gamma)\bar{\omega}}{96}$, any local step-size $\alpha_l \leq \min\frac{(1-\gamma)\bar{\omega}}{48K}$, and choose the constant $\zeta_1 = \alpha\nu$, the updates of `FedTD(0)` satisfy*

$$\nu_1\mathbb{E}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2 \leq (\frac{1}{\alpha}-\nu_1)\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2 - \frac{1}{\alpha}\mathbb{E}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2 + \underbrace{\frac{2\alpha\sigma^2}{NK}}_{O(\alpha^1)} + \underbrace{\frac{1080\alpha^2}{K\alpha_g^2\nu}(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2)}_{O(\alpha^2)} + \underbrace{2B(\epsilon,\epsilon_1)G+6\alpha B^2(\epsilon,\epsilon_1)}_{\textit{heterogeneity term}}, \tag{4.28}$$

where $\nu_1 = \frac{\nu}{4} = \frac{(1-\gamma)\bar{\omega}}{4}$.

Proof. From Eq (4.27) and $\zeta_1 = \alpha\nu$, we know

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2 \leq (1+\zeta_1)\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2 + 2\alpha\mathbb{E}\langle\bar{g}(\bar{\theta}_t),\bar{\theta}_t-\theta^*\rangle + 6\alpha^2\mathbb{E}\left\|\bar{g}(\bar{\theta}_r)\right\|^2 \\
&+ 108\frac{\alpha^4}{K\alpha_g^2}(6+\frac{1}{\zeta_1})(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2) + \frac{2\alpha^2\sigma^2}{NK} + 2\alpha B(\epsilon,\epsilon_1)G + 6\alpha^2B^2(\epsilon,\epsilon_1) \\
&\leq (1+\alpha\nu-2\alpha\nu)\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2 + 24\alpha^2\mathbb{E}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2 + \frac{2\alpha^2\sigma^2}{NK} \quad \text{(Lemma 3 and 4)} \\
&+ 108\frac{\alpha^4}{K\alpha_g^2}(6+\frac{1}{\alpha\nu})(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2) + 2\alpha B(\epsilon,\epsilon_1)G + 6\alpha^2B^2(\epsilon,\epsilon_1) \\
&\leq (1-\frac{\alpha\nu}{2})\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2 - \frac{\alpha\nu}{2}\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2 + 24\alpha^2\mathbb{E}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2 + \frac{2\alpha^2\sigma^2}{NK}
\end{aligned}$$

$$
\begin{aligned}
&+ 108\frac{\alpha^4}{K\alpha_g^2}(6+\frac{1}{\alpha\nu})(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2)+2\alpha B(\epsilon,\epsilon_1)G+6\alpha^2B^2(\epsilon,\epsilon_1)\\
&\overset{(a)}{\leq}(1-\frac{\alpha\nu}{2})\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2-\frac{\alpha\nu}{2}\mathbb{E}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2+\frac{\alpha\nu}{4}\mathbb{E}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2+\frac{2\alpha^2\sigma^2}{NK}\\
&+ 108\frac{\alpha^4}{K\alpha_g^2}(6+\frac{1}{\alpha\nu})(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2)+2\alpha B(\epsilon,\epsilon_1)G+6\alpha^2B^2(\epsilon,\epsilon_1),
\end{aligned}
$$

where (a) comes from $\lambda_{\max}(\Phi^T\bar{D}\Phi)\leq 1$ and $24\alpha^2\leq 24\alpha\frac{(1-\gamma)\bar{w}}{96}=\frac{\alpha\nu}{4}$. Moving $\mathbb{E}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2$ (on the right-hand side of (a)) to the left hand side of the above inequality yields:

$$
\begin{aligned}
\frac{\alpha\nu}{4}\mathbb{E}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2 &\leq (1-\frac{\alpha\nu}{2})\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2-\mathbb{E}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2+\frac{2\alpha^2\sigma^2}{NK}\\
&+ 108(\frac{6\alpha^4}{K\alpha_g^2}+\frac{\alpha^3}{K\alpha_g^2\nu})(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2)+2\alpha B(\epsilon,\epsilon_1)G+6\alpha^2B^2(\epsilon,\epsilon_1).
\end{aligned}
$$

Dividing by α on both sides of the inequality above and changing ν into ν_1, we have:

$$
\begin{aligned}
\nu_1\mathbb{E}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2 &\leq (\frac{1}{\alpha}-\nu_1)\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2-\frac{1}{\alpha}\mathbb{E}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2+\frac{2\alpha\sigma^2}{NK}\\
&+ 108(\frac{6\alpha^3}{K\alpha_g^2}+\frac{4\alpha^2}{K\alpha_g^2\nu_1})(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2)+2B(\epsilon,\epsilon_1)G+6\alpha B^2(\epsilon,\epsilon_1)\\
&\leq (\frac{1}{\alpha}-\nu_1)\mathbb{E}\left\|\bar{\theta}_t-\theta^*\right\|^2-\frac{1}{\alpha}\mathbb{E}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2+\underbrace{\frac{2\alpha\sigma^2}{NK}}_{O(\alpha^1)}\\
&+ \underbrace{\frac{1080\alpha^2}{K\alpha_g^2\nu_1}(\sigma^2+3KB^2(\epsilon,\epsilon_1)+2KG^2)}_{O(\alpha^2)}+\underbrace{2B(\epsilon,\epsilon_1)G+6\alpha B^2(\epsilon,\epsilon_1)}_{\text{heterogeneity term}},
\end{aligned}
$$

where we used the fact that $\alpha\leq 1$ in the last inequality. □

With these lemmas, we are now ready to prove Theorem 2, which we restate for clarity.

b) Proof of Theorem 2

Given a fixed local step-size $\alpha_l=\frac{1}{2}\frac{(1-\gamma)\bar{\omega}}{48K}$, decreasing effective step-sizes $\alpha_t=\frac{8}{\nu(a+t+1)}=\frac{8}{(1-\gamma)\bar{\omega}(a+t+1)}$, decreasing global step-sizes $\alpha_g^{(t)}=\frac{\alpha_t}{K\alpha_l}$, and weights $w_t=(a+t)$, we have that

$$
\mathbb{E}\left\|V_{\tilde{\theta}_T}-V_{\theta_i^*}\right\|_{\bar{D}}^2\leq\tilde{\mathcal{O}}\left(\frac{G^2}{K^2T^2}+\frac{\sigma^2}{\nu^4KT^2}+\frac{\sigma^2}{\nu^2NKT}+\frac{B(\epsilon,\epsilon_1)G}{\nu}+\Gamma^2(\epsilon,\epsilon_1)\right)\tag{4.29}
$$

holds for any agent $i \in [N]$.

Proof. We take the effective step-size $\alpha_t = \frac{8}{\nu(a+t+1)} = \frac{2}{\nu_1(a+t+1)}$ for $a > 0$. In addition, we define weights $w_t = (a+t)$ and define

$$\tilde{\theta}_T = \frac{1}{W}\sum_{t=1}^{T} w_t \bar{\theta}_t,$$

where $W = \sum_{t=1}^{T} w_t \geq \frac{1}{2}T(a+T)$. By convexity of positive definite quadratic forms ($\lambda_{\min}(\Phi^T \bar{D}\Phi) \geq \bar{\omega} > 0$), we have that

$$\begin{aligned}
\nu_1 \mathbb{E}\left\|V_{\tilde{\theta}_T} - V_{\theta^*}\right\|_{\bar{D}}^2 &\leq \frac{\nu_1}{W}\sum_{t=1}^{T}(a+t)\mathbb{E}\left\|V_{\bar{\theta}_t} - V_{\theta^*}\right\|_{\bar{D}}^2 \\
&\overset{(4.28)}{\leq} \frac{\nu_1(a+1)(a+2)G^2}{2W} + \frac{1}{W}\sum_{t=1}^{T}\left[\frac{2(a+t)\alpha_t}{NK}\sigma^2\right] \\
&+ \frac{1}{W}\sum_{t=1}^{T}\left[\frac{1080(a+t)\alpha_t^2}{K\alpha_g^2\nu_1}(\sigma^2 + 3KB^2(\epsilon,\epsilon_1) + 2KG^2)\right] \\
&+ \frac{1}{W}\sum_{t=1}^{T}(a+t)\left[2B(\epsilon,\epsilon_1)G + 6\alpha_t B^2(\epsilon,\epsilon_1)\right] \\
&\leq \frac{\nu_1(a+1)(a+2)G^2}{2W} + \frac{2\sigma^2}{NKW}\sum_{t=1}^{T}(a+t)\alpha_t \\
&+ \frac{1080(\sigma^2 + 3KB^2(\epsilon,\epsilon_1) + 2KG^2)}{K\alpha_g^2\nu_1 W}\sum_{t=1}^{T}(a+t)\alpha_t^2 + 2B(\epsilon,\epsilon_1)G + \frac{6B^2(\epsilon,\epsilon_1)}{W}\sum_{t=1}^{T}(a+t)\alpha_t \\
&\leq \frac{\nu_1(a+1)(a+2)G^2}{2W} + \frac{4\sigma^2}{\nu_1 NKW}\cdot T \\
&+ \frac{4320(\sigma^2 + 3KB^2(\epsilon,\epsilon_1) + 2KG^2)}{K\alpha_g^2\nu_1^3 W}\cdot(1 + \log(a+T)) + 2B(\epsilon,\epsilon_1)G + \frac{12B^2(\epsilon,\epsilon_1)}{\nu_1 W}\cdot T,
\end{aligned}$$

where we used $\left\|V_{\tilde{\theta}_0} - V_{\theta^*}\right\|_{\bar{D}}^2 \leq G^2$. Dividing by ν_1 on both sides, changing ν_1 into ν, and using $W \geq \frac{T(a+T)}{2}$, we have:

$$\mathbb{E}\left\|V_{\tilde{\theta}_T} - V_{\theta^*}\right\|_{\bar{D}}^2 \leq \tilde{\mathcal{O}}\left(\frac{G^2}{K^2T^2} + \frac{\sigma^2}{\nu^4 KT^2} + \frac{\sigma^2}{\nu^2 NKT} + \frac{B(\epsilon,\epsilon_1)G}{\nu}\right).$$

We finish the proof by using the following inequality: $\mathbb{E}\left\|V_{\tilde{\theta}_T} - V_{\theta_i^*}\right\|_{\bar{D}}^2 \leq 2\mathbb{E}\left\|V_{\tilde{\theta}_T} - V_{\theta^*}\right\|_{\bar{D}}^2 + 2\mathbb{E}\left\|V_{\theta_i^*} - V_{\theta^*}\right\|_{\bar{D}}^2$, in tandem with the third point in Theorem 1. □

4.8.9 Heterogeneity bias: Proof of Theorem 3

In this section, we prove Theorem 3.

Proof of Theorem 3. As θ_1^* and θ_2^* are the TD(0) fixed points of agents 1 and 2, respectively, we have $\theta_1^* = \bar{A}_1^{-1}\bar{b}_1$ and $\theta_2^* = \bar{A}_2^{-1}\bar{b}_2$ from Section 4.3.1. The output of mean-path `FedTD`(0) with $k = 1$ and $\alpha = \alpha_g\alpha_l$ satisfies:

$$
\begin{aligned}
&\bar{\theta}_{t+1} = \bar{\theta}_t + \alpha(-\hat{A}\bar{\theta}_t + \hat{b}) \\
\implies &\bar{\theta}_{t+1} - \theta_1^* = \bar{\theta}_t - \theta_1^* + \alpha(-\hat{A}(\bar{\theta}_t - \theta_1^* + \theta_1^*) + \hat{b}) \\
\implies &e_{1,t+1} = (I - \alpha\hat{A})e_{1,t} - \alpha\hat{A}\theta_1^* + \alpha\hat{b} \\
\implies &e_{1,t+1} = (I - \alpha\hat{A})e_{1,t} - \alpha\left(\frac{\bar{A}_1 + \bar{A}_2}{2}\right)\bar{A}_1^{-1}\bar{b}_1 + \alpha\frac{\bar{b}_1 + \bar{b}_2}{2} \\
\implies &e_{1,t+1} = (I - \alpha\hat{A})e_{1,t} - \alpha\frac{\bar{A}_2\bar{A}_1^{-1}\bar{b}_1}{2} + \alpha\frac{\bar{b}_2}{2} \\
\implies &e_{1,t+1} = (I - \alpha\hat{A})e_{1,t} - \frac{\alpha\bar{A}_2}{2}\left(\bar{A}_1^{-1}\bar{b}_1 - \bar{A}_2^{-1}\bar{b}_2\right) \\
\implies &e_{1,t+1} = \underbrace{(I - \alpha\hat{A})}_{\tilde{\mathcal{A}}}e_{1,t} + \underbrace{\frac{\alpha\bar{A}_2}{2}(\theta_2^* - \theta_1^*)}_{\tilde{\mathcal{Y}}}. \qquad (4.30)
\end{aligned}
$$

Let us now note that $e_{1,t+1} = \tilde{\mathcal{A}}e_{1,t} + \tilde{\mathcal{Y}}$ can be viewed as a discrete-time linear time-invariant (LTI) system where α is chosen s.t. $\tilde{\mathcal{A}}$ is Schur stable, i.e., $|\lambda_{\max}(\tilde{\mathcal{A}})| < 1$. At the t-th iteration, we have:

$$
e_{1,t} = \tilde{\mathcal{A}}^t e_{1,0} + \sum_{k=0}^{t-1}\tilde{\mathcal{A}}^k\tilde{\mathcal{Y}}.
$$

As $t \to \infty$, the small gain theorem tells us that because $\rho(\tilde{\mathcal{A}}) < 1$ (where $\rho(\cdot)$ denotes the spectral radius), $\sum_{k=0}^{t-1}\tilde{\mathcal{A}}^k$ exists and is given by $(I - \tilde{\mathcal{A}})^{-1}$. We can then conclude that

$$
\begin{aligned}
\lim_{t\to\infty} e_{1,t} &= (I - \tilde{\mathcal{A}})^{-1}\tilde{\mathcal{Y}} \\
&= \left(\alpha\hat{A}\right)^{-1}\frac{\alpha\bar{A}_2}{2}(\theta_1^* - \theta_2^*)
\end{aligned}
$$

$$= \frac{1}{2}\hat{A}^{-1}\bar{A}_2\left(\theta_1^* - \theta_2^*\right). \tag{4.31}$$

The limiting expression for $e_{2,t}$ follows the same analysis.

4.8.10 Proof of the Markovian setting

We now turn our attention to proving the main result of the paper, namely, Theorem 4.

a) Outline

As mentioned in the main body, one of the main obstacles to overcome in the analysis is that in general, $\mathbb{E}[(1/N)\sum_{i=1}^{N}(g_i(\theta_{t,k}^{(i)}, O_{t,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}))] \neq 0$. In order to show that a linear speedup is achievable, we first decompose the random TD direction of each agent i as $g_i(\theta_{t,k}^{(i)}) = b_i(O_{t,k}^{(i)}) - A_i(O_{t,k}^{(i)})\theta_{t,k}^{(i)}$ in subsection a) and show that the variances of $(1/NK)\sum_{i=1}^{N}\sum_{k=0}^{K-1} A_i(O_{t,k}^{(i)})$ and $(1/NK)\sum_{i=1}^{N}\sum_{k=0}^{K-1} b_i(O_{t,k}^{(i)})$ get scaled down by NK in subsection a). To decouple the randomness between the parameter $\theta_{t,k}^{(i)}$ and the observations $O_{t,k}^{(i)}$ using the method called *information theoretic control of coupling* in [13], we need to bound $\mathbb{E}\left[\left\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\right\|^2\right]$ in subsection a). As the analysis in the i.i.d. setting and traditional FL, we characterize the drift term, per-iteration error decrease, and parameter selection in subsections a), a) and a), respectively. Finally, we prove Theorem 4 in subsection b).

Additional Notation: Under Assumption 3, for each MDP i, there exists some $m_i \geq 1$ and some $\rho_i \in (0,1)$, such that for all $t \geq 0$ and $0 \leq k \leq K-1$, it holds that

$$d_{TV}\left(\mathbb{P}\left(s_{t,k}^{(i)} = \cdot \mid s_{0,0}^{(i)} = s\right), \pi^{(i)}\right) \leq m_i \rho_i^{tK+k}, \forall s \in \mathcal{S}.$$

Furthermore, we define $\rho = \max_{i\in[N]}\{\rho_i\}$, $m = \max_{i\in[N]}\{m_i\}$.

- Auxiliary lemmas for Theorem 4
- Decomposition Form

The first step in our proof of Theorem 4 is to rewrite agent i's update direction of `FedTD`(0) as:

$$g_i(\theta_{t,k}^{(i)}) = -A_i(O_{t,k}^{(i)})\theta_{t,k}^{(i)} + b_i(O_{t,k}^{(i)})$$

where $A_i(O_{t,k}^{(i)}) = \phi(s_{t,k}^{(i)})(\phi^\top(s_{t,k}^{(i)}) - \gamma\phi^\top(s_{t,k+1}^{(i)}))$ and $b_i(O_{t,k}^{(i)}) = r(s_{t,k}^{(i)})\phi(s_{t,k}^{(i)})$. Note that the steady-state value of $\mathbb{E}[b_i(O_{t,k}^{(i)})]$ is not equal to 0. For convenience, we apply appropriate centering to rewrite g_i as:

$$g_i(\theta_{t,k}^{(i)}) = -A_i(O_{t,k}^{(i)})(\theta_{t,k}^{(i)} - \theta_i^*) + \underbrace{b_i(O_{t,k}^{(i)}) - A_i(O_{t,k}^{(i)})\theta_i^*}_{Z_i(O_{t,k}^{(i)})}. \tag{4.32}$$

Define $Z_i(O_{t,k}^{(i)}) \triangleq b_i(O_{t,k}^{(i)}) - A_i(O_{t,k}^{(i)})\theta_i^*$. As $\bar{g}_i(\theta) \triangleq \mathbb{E}_{O_{t,k}^{(i)} \sim \pi^{(i)}}\left[g_i(\theta)\right]$, we have:

$$\bar{g}_i(\theta_{t,k}^{(i)}) = -\bar{A}_i(\theta_{t,k}^{(i)} - \theta_i^*). \tag{4.33}$$

where $\bar{A}_i = \Phi^\top D^{(i)}(\Phi - \gamma P^{(i)}\Phi)$. Note that $\mathbb{E}_{O_{t,k}^{(i)} \sim \pi^{(i)}}\left[Z_i(O_{t,k}^{(i)})\right]$ equals to 0. Taking into account the definitions above, we establish the following lemmas:

Lemma 11. *(Uniform norm bound) There exist some constants $c_1, c_2, c_3 \geq 0$ such that $\left\|A_i\left(O_{t,k}^{(i)}\right)\right\| \leq c_1 := 1 + \gamma$, $\left\|\bar{A}_i\right\| \leq c_2 := 1 + \gamma$ and $\left\|Z_i\left(O_{t,k}^{(i)}\right)\right\| \leq c_3 := R_{\max} + c_1 H$ holds for all $i \in [N]$.*

Proof. Based on the definition and the fact that $\|\phi(s)\| \leq 1$, we have

$$\left\|A_i\left(O_{t,k}^{(i)}\right)\right\| = \left\|\phi(s_{t,k}^{(i)})(\phi^\top(s_{t,k}^{(i)}) - \gamma\phi^\top(s_{t,k+1}^{(i)}))\right\| \leq \left\|\phi(s_{t,k}^{(i)})\right\|\left\|\phi^\top(s_{t,k}^{(i)}) - \gamma\phi^\top(s_{t,k+1}^{(i)})\right\| \leq 1 + \gamma.$$

Similarly, making use of the fact that $r(s) \leq R_{\max}$ for any $s \in \mathcal{S}$, we apply the same reasoning to conclude that

$$\left\|\bar{A}_i\right\| \leq 1 + \gamma \;\; \& \;\; \left\|Z_i\left(O_{t,k}^{(i)}\right)\right\| \leq R_{\max} + c_1 H$$

. □

Lemma 12. *There exist some constants $L_1, L_2 \geq 0$ such that*

$$\left\|\bar{A}_i - \mathbb{E}\left[A_i\left(O_{t_2,k_2}^{(i)}\right) \mid \mathcal{F}_{k_1}^{t_1}\right]\right\| \leq L_1\rho^{(t_2-t_1)K+k_2-k_1} \;\; \& \;\; \left\|\bar{A}_i - \mathbb{E}_{t_1}\left[A_i\left(O_{t_2,k_2}^{(i)}\right)\right]\right\| \leq L_1\rho^{(t_2-t_1)K+k_2},$$
$$\left\|\mathbb{E}\left[Z_i\left(O_{t_2,k_2}^{(i)}\right) \mid \mathcal{F}_{k_1}^{t_1}\right]\right\| \leq L_2\rho^{(t_2-t_1)K+k_2-k_1} \;\; \& \;\; \left\|\mathbb{E}_{t_1}\left[Z_i\left(O_{t_2,k_2}^{(i)}\right)\right]\right\| \leq L_2\rho^{(t_2-t_1)K+k_2}$$

hold for any $i \in [N]$, $0 \leq k_1, k_2 \leq K-1$ and $t_2 \geq t_1 \geq 0$.

Proof. We have:

$$\begin{aligned}
\left\|\mathbb{E}\left[Z_i\left(O_{t_2,k_2}^{(i)}\right) \mid \mathcal{F}_{k_1}^{t_1}\right]\right\| &= \left\|\mathbb{E}\left[Z_i\left(O_{t_2,k_2}^{(i)}\right) \mid \mathcal{F}_{k_1}^{t_1}\right] - \mathbb{E}_{O_{t_2,k_2}^{(i)}\sim\pi^{(i)}}\left[Z_i\left(O_{t_2,k_2}^{(i)}\right) \mid \mathcal{F}_{k_1}^{t_1}\right]\right\| \\
&= \Bigg\| \sum_{s_{t_2,k_2}^{(i)}, s_{t_2+1,k_2+1}^{(i)}} \Big(\pi^{(i)}(s_{t_2,k_2}^{(i)}) P(s_{t_2+1,k_2+1}^{(i)} \mid s_{t_2,k_2}^{(i)}) \\
&\quad - P(s_{t_2,k_2}^{(i)} = \cdot \mid s_{t_1,k_1}^{(i)}) P(s_{t_2+1,k_2+1}^{(i)} \mid s_{t_2,k_2}^{(i)})\Big) Z_i(O_{t_2,k_2}^{(i)})\Bigg\| \\
&\leq \sum_{s_{t_2,k_2}^{(i)}} \left|\pi^{(i)}(s_{t_2,k_2}^{(i)}) - P(s_{t_2,k_2}^{(i)} = \cdot \mid s_{t_1,k_1}^{(i)})\right| \left\|Z_i(O_{t_2,k_2}^{(i)})\right\| \\
&\overset{(a)}{\leq} \sum_{s_{t_2,k_2}^{(i)}} \left|\pi^{(i)}(s_{t_2,k_2}^{(i)}) - P(s_{t_2,k_2}^{(i)} = \cdot \mid s_{t_1,k_1}^{(i)})\right| (R_{\max} + c_1 H) \\
&= 2(R_{\max} + c_1 H) d_{TV}\left(\mathbb{P}\left(s_{t_2,k_2}^{(i)} = \cdot \mid s_{t_1,k_1}^{(i)} = s\right), \pi^{(i)}\right) \\
&\leq 2(R_{\max} + c_1 H) m_i \rho_i^{(t_2-t_1)K+k_2-k_1}
\end{aligned}$$

where (a) is due to Lemma 11 and the last step follows from Assumption 3. We finish the proof by choosing $L_2 \triangleq \max_{i\in[N]}\{2(R_{\max} + c_1 H) m_i\} = 2c_3 m$. And we follow the same analysis to bound:

$$\begin{aligned}
\left\|\bar{A}_i - \mathbb{E}\left[A_i\left(O_{t_2,k_2}^{(i)}\right) \mid \mathcal{F}_{k_1}^{t_1}\right]\right\| &= \left\|\|\mathbb{E}\left[A_i\left(O_{t_2,k_2}^{(i)}\right) \mid \mathcal{F}_{k_1}^{t_1}\right] - \mathbb{E}_{O_{t_2,k_2}^{(i)}\sim\pi^{(i)}}\left[A_i\left(O_{t_2,k_2}^{(i)}\right) \mid \mathcal{F}_{k_1}^{t_1}\right]\right\| \\
&= \Bigg\| \sum_{s_{t_2,k_2}^{(i)}, s_{t_2+1,k_2+1}^{(i)}} \Big(\pi^{(i)}(s_{t_2,k_2}^{(i)}) P(s_{t_2+1,k_2+1}^{(i)} \mid s_{t_2,k_2}^{(i)}) \\
&\quad - P(s_{t_2,k_2}^{(i)} = \cdot \mid s_{t_1,k_1}^{(i)}) P(s_{t_2+1,k_2+1}^{(i)} \mid s_{t_2,k_2}^{(i)})\Big) A_i(O_{t_2,k_2}^{(i)})\Bigg\| \\
&\leq \sum_{s_{t_2,k_2}^{(i)}} \left|\pi^{(i)}(s_{t_2,k_2}^{(i)}) - P(s_{t_2,k_2}^{(i)} = \cdot \mid s_{t_1,k_1}^{(i)})\right| \left\|A_i(O_{t_2,k_2}^{(i)})\right\| \\
&\overset{(b)}{\leq} 2c_1 d_{TV}\left(\mathbb{P}\left(s_{t_2,k_2}^{(i)} = \cdot \mid s_{t_1,k_1}^{(i)} = s\right), \pi^{(i)}\right) \\
&\leq 2c_1 m_i \rho_i^{(t_2-t_1)K+k_2-k_1}
\end{aligned}$$

We finish the proof by choosing $L_1 \triangleq \max_{i\in[N]}\{2c_1 m_i\} = 2c_1 m$. We employ the same reasoning to prove the remaining three inequalities. □

• Variance Reduction

We are now ready to present the variance reduction Lemma in the Markov setting. The following Lemma establishes an analog of the variance reduction Lemma 7 in the i.i.d. setting. Based on the assumption that trajectories are independent across agents, it is easy to understand that the variance of $(1/NK)\sum_{i=1}^{N}\sum_{k=0}^{K-1} A_i(O_{t,k}^{(i)})$ and $(1/NK)\sum_{i=1}^{N}\sum_{k=0}^{K-1} b_i(O_{t,k}^{(i)})$ can be scaled by the number of agents N. However, it is not obvious that the variances can be scaled by K (the number of local iterations), since the observations of each agent $O_{t,k_1}^{(i)}$ and $O_{t,k_2}^{(i)}$ are correlated at different local steps k_1, k_2. Due to the geometric mixing property of the Markov chain, the correlation between $O_{t,k_1}^{(i)}$ and $O_{t,k_2}^{(i)}$ will geometrically decay after the mixing time. Based on this fact, we show that the variances of $(1/NK)\sum_{i=1}^{N}\sum_{k=0}^{K-1} A_i(O_{t,k}^{(i)})$ and $(1/NK)\sum_{i=1}^{N}\sum_{k=0}^{K-1} b_i(O_{t,k}^{(i)})$ get scaled down by NK with an additional additive, higher order term dependent on the mixing time τ, which is formally stated as follows:

Lemma 13. *(Variance reduction in the Markovian setting) For any $0 < \tau < t$, there exists $d_1, d_2 > 0$ such that:*

$$\mathbb{E}_{t-\tau}\left[\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left[A_i(O_{t,k}^{(i)}) - \bar{A}_i\right]\right\|\right] \leq \frac{d_1}{\sqrt{NK}} + 2L_1\rho^{\tau K}, \tag{4.34}$$

$$\mathbb{E}_{t-\tau}\left[\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left[A_i(O_{t,k}^{(i)}) - \bar{A}_i\right]\right\|^2\right] \leq \frac{d_1^2}{NK} + 4L_1^2\rho^{2\tau K}, \tag{4.35}$$

$$\mathbb{E}_{t-\tau}\left[\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{t,k}^{(i)})\right\|\right] \leq \frac{d_2}{\sqrt{NK}} + 2L_2\rho^{\tau K}, \quad \textit{and} \tag{4.36}$$

$$\mathbb{E}_{t-\tau}\left[\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{t,k}^{(i)})\right\|^2\right] \leq \frac{d_2^2}{NK} + 4L_2^2\rho^{2\tau K}, \tag{4.37}$$

where $d_1 \triangleq \sqrt{(c_1+c_2)^2 + \frac{2(c_1+c_2)L_1\rho}{1-\rho}}$ *and* $d_2 \triangleq \sqrt{c_3^2 + \frac{2c_3L_2\rho}{1-\rho}}$.

Proof.

$$\mathbb{E}_{t-\tau}\left[\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right\|\right] = \mathbb{E}_{t-\tau}\left[\sqrt{\left(\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right)^{\top}\left(\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right)}\right]$$

$$\overset{(a)}{\leq} \sqrt{\mathbb{E}_{t-\tau}\left[\left(\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right)^{\top}\left(\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right)\right]}$$

$$= \Bigg\{\mathbb{E}_{t-\tau}\Bigg[\frac{1}{N^2K^2}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})^{\top}Z_i(O_{t,k}^{(i)}) + \underbrace{\frac{2}{N^2K^2}\sum_{i=1}^{N}\sum_{k<l}Z_i(O_{t,k}^{(i)})^{\top}Z_i(O_{t,l}^{(i)})}_{T_1}$$

$$+ \underbrace{\frac{2}{N^2K^2}\sum_{i<j}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})^{\top}Z_j(O_{t,k}^{(j)})}_{T_2} + \underbrace{\frac{2}{N^2K^2}\sum_{i<j}\sum_{k<l}Z_i(O_{t,k}^{(i)})^{\top}Z_j(O_{t,l}^{(j)})}_{T_3}\Bigg]\Bigg\}^{\frac{1}{2}} \tag{4.38}$$

where (a) is due to the concavity of square root and Jensen's inequality. Furthermore, the term T_1 can be further bounded by:

$$\begin{aligned}
\mathbb{E}_{t-\tau}[T_1] &= \mathbb{E}_{t-\tau}\left[\frac{2}{N^2K^2}\sum_{i=1}^{N}\sum_{k<l}Z_i(O_{t,k}^{(i)})^{\top}Z_i(O_{t,l}^{(i)})\right] \\
&= \mathbb{E}_{t-\tau}\left[\frac{2}{N^2K^2}\sum_{i=1}^{N}\sum_{k<l}Z_i(O_{t,k}^{(i)})^{\top}\mathbb{E}\left[Z_i(O_{t,l}^{(i)}) \mid \mathcal{F}_k^t\right]\right] \\
&\leq \mathbb{E}_{t-\tau}\left[\frac{2}{N^2K^2}\sum_{i=1}^{N}\sum_{k<l}\left\|Z_i(O_{t,k}^{(i)})\right\|\left\|\mathbb{E}\left[Z_i(O_{t,l}^{(i)}) \mid \mathcal{F}_k^t\right]\right\|\right] \quad \text{(Cauchy–Schwarz inequality)} \\
&\leq \mathbb{E}_{t-\tau}\left[\frac{2}{N^2K^2}\sum_{i=1}^{N}\sum_{k<l}c_3L_2\rho^{(l-k)}\right] \quad \text{(Lemma 11 and 12)} \\
&\leq \mathbb{E}_{t-\tau}\left[\frac{2}{N^2K^2}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\sum_{m=1}^{\infty}c_3L_2\rho^{m}\right] \\
&= \frac{2c_3L_2NK}{N^2K^2}\frac{\rho}{1-\rho} = \frac{2c_3L_2\rho}{NK(1-\rho)}.
\end{aligned}$$

And T_2 can be bounded by:

$$\begin{aligned}
\mathbb{E}_{t-\tau}[T_2] &= \frac{2}{N^2K^2}\sum_{i<j}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left[Z_i(O_{t,k}^{(i)})\right]^\top \mathbb{E}_{t-\tau}\left[Z_j(O_{t,k}^{(j)})\right] \ (O_{t,k}^{(i)} \text{ and } O_{t,k}^{(j)} \text{ are independent}) \\
&\leq \frac{2}{N^2K^2}\sum_{i<j}\sum_{k=0}^{K-1} L_2^2\rho^{2\tau K+2k} \ \text{(Lemma 12)} \\
&\leq \frac{2}{K}L_2^2\rho^{2\tau K}.
\end{aligned}$$

Meanwhile, T_3 can be bounded by:

$$\begin{aligned}
\mathbb{E}_{t-\tau}[T_3] &= \frac{2}{N^2K^2}\sum_{i<j}\sum_{k<l}\mathbb{E}_{t-\tau}\left[Z_i(O_{t,k}^{(i)})\right]^\top \mathbb{E}_{t-\tau}\left[Z_j(O_{t,l}^{(j)})\right] \ (O_{t,k}^{(i)} \text{ and } O_{t,l}^{(j)} \text{ are independent}) \\
&\leq \frac{2}{N^2K^2}\sum_{i<j}\sum_{k<l} L_2^2\rho^{2\tau K+k+l} \ \text{(Lemma 12)} \\
&\leq 2L_2^2\rho^{2\tau K}
\end{aligned}$$

Substituting the upper bound of T_1, T_2 and T_3 into Eq (4.38), we have:

$$\begin{aligned}
\mathbb{E}_{t-\tau}\left[\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right\|\right] &\leq \left(\frac{1}{N^2K^2}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left[Z_i(O_{t,k}^{(i)})^\top Z_i(O_{t,k}^{(i)})\right]\right. \\
&\quad \left. + \frac{2c_3L_2\rho}{NK(1-\rho)} + \frac{2}{K}L_2^2\rho^{2\tau K} + 2L_2^2\rho^{2\tau K}\right)^{\frac{1}{2}} \\
&\overset{(a)}{\leq} \sqrt{\frac{NK}{N^2K^2}c_3^2 + \frac{2c_3L\rho}{NK(1-\rho)} + \frac{2}{K}L_2^2\rho^{2\tau K} + 2L_2^2\rho^{2\tau K}} \\
&\leq \sqrt{\frac{1}{NK}\left(c_3^2 + \frac{2c_3L_2\rho}{1-\rho}\right) + 4L_2^2\rho^{2\tau K}} \quad (K \geq 1) \\
&\leq \sqrt{\frac{1}{NK}\left(c_3^2 + \frac{2c_3L_2\rho}{1-\rho}\right)} + \sqrt{4L_2^2\rho^{2\tau K}} \\
&= \sqrt{\frac{1}{NK}\left(c_3^2 + \frac{2c_3L_2\rho}{1-\rho}\right)} + 2L_2\rho^{\tau K}.
\end{aligned}$$

where (a) used the fact that $\left\|Z_i\left(O_{t,k}^{(i)}\right)\right\| \leq c_3$ mentioned in Lemma 11. The proof of other inequalities follows the same reasoning. □

- Bounding $\mathbb{E}\left[\left\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\right\|^2\right]$

Lemma 14. *(Bounding $\|\theta_t - \theta_{t-\tau}\|^2$) Consider $\tau = \lceil \frac{\tau^{\text{mix}}(\alpha_T^2)}{K} \rceil$ and choose the effective step-size*

$$\alpha \leq \min\left\{\frac{1}{30c_4(\tau+1)}, \frac{1}{96c_4^2\tau}, 1\right\}$$

where $c_4 = 3c_1$. For any $t \geq 2\tau$, we have the following bound:

$$\begin{aligned}\mathbb{E}_{t-2\tau}\left[\left\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\right\|^2\right] \leq 8\alpha^2\tau^2c_4^2\mathbb{E}_{t-2\tau}\left[\left\|\bar{\theta}_t - \theta^*\right\|^2\right] + 14\alpha^2\tau^2\frac{d_2^2}{NK} + \frac{52L_2^2\alpha^4\tau}{1-\rho^2} \\ + 4\alpha^2c_4^2\tau\sum_{s=0}^{\tau}E_{t-2\tau}[\Delta_{t-s}] + 3200\alpha^2c_4^2c_1^2\tau^3\Gamma^2(\epsilon,\epsilon_1) + 4\alpha^2c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1).\end{aligned} \tag{4.39}$$

Proof. For any $l \geq 2\tau$, we have

$$\begin{aligned}\left\|\bar{\theta}_{l+1} - \bar{\theta}_l\right\|^2 &= \left\|\Pi_{2,\mathcal{H}}\left(\bar{\theta}_l + \frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{l,k}^{(i)})\right) - \bar{\theta}_l\right\|^2 \\ &\leq \left\|\bar{\theta}_l + \frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{l,k}^{(i)}) - \bar{\theta}_l\right\|^2 \\ &= \alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left[-A_i(O_{l,k}^{(i)})\left(\theta_{l,k}^{(i)} - \theta_i^*\right) + Z_i(O_{l,k}^{(i)})\right]\right\|^2 \\ &\leq 2\alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left[-A_i(O_{l,k}^{(i)})\left(\theta_{l,k}^{(i)} - \theta^*\right) + Z_i(O_{l,k}^{(i)})\right]\right\|^2 \\ &\quad + 2\alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left[-A_i(O_{l,k}^{(i)})\left(\theta^* - \theta_i^*\right)\right]\right\|^2 \\ &\overset{(a)}{\leq} 2\alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left[-A_i(O_{l,k}^{(i)})\left(\theta_{l,k}^{(i)} - \theta^*\right) + Z_i(O_{l,k}^{(i)})\right]\right\|^2 + 2\alpha^2c_1^2\Gamma^2(\epsilon,\epsilon_1) \\ &= 6\alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}A_i(O_{l,k}^{(i)})\left(\theta_{l,k}^{(i)} - \bar{\theta}_l\right)\right\|^2 + 6\alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}A_i(O_{l,k}^{(i)})\left(\bar{\theta}_l - \theta^*\right)\right\|^2 \\ &\quad + 6\alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{l,k}^{(i)})\right\|^2 + 2\alpha^2c_1^2\Gamma^2(\epsilon,\epsilon_1) \\ &\leq 6\alpha^2\left(\frac{c_1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left\|\theta_{l,k}^{(i)} - \bar{\theta}_l\right\|\right)^2 + 6\alpha^2c_1^2\left\|\bar{\theta}_l - \theta^*\right\|^2\end{aligned}$$

$$+ 6\alpha^2 \Big\| \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)}) \Big\|^2 + 2\alpha^2 c_1^2 \Gamma^2(\epsilon, \epsilon_1), \tag{4.40}$$

where (a) comes from the upper bound of fixed points distance in Theorem 1 and the fact that $\left\| A_i \left(O_{t,k}^{(i)} \right) \right\| \leq c_1$ in Lemma 11. Taking square root on both sides of the inequality above, we get:

$$\begin{aligned}
\Big\| \bar{\theta}_{l+1} - \bar{\theta}_l \Big\| &\leq 3 \sqrt{\alpha^2 \left(\frac{c_1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \Big\| \theta_{l,k}^{(i)} - \bar{\theta}_l \Big\| \right)^2} + 3 \sqrt{\alpha^2 c_1^2 \Big\| \bar{\theta}_l - \theta^* \Big\|^2} \\
&+ 3 \sqrt{\alpha^2 \Big\| \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)}) \Big\|^2} + \sqrt{2\alpha^2 c_1^2 \Gamma^2(\epsilon, \epsilon_1)} \\
&\leq \frac{3\alpha c_1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \Big\| \theta_{l,k}^{(i)} - \bar{\theta}_l \Big\| + 3\alpha c_1 \Big\| \bar{\theta}_l - \theta^* \Big\| + 3\alpha \Big\| \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)}) \Big\| + 2\alpha c_1 \Gamma(\epsilon, \epsilon_1).
\end{aligned} \tag{4.41}$$

By using the fact that $\Big\| \bar{\theta}_{l+1} - \theta^* \Big\| \leq \Big\| \bar{\theta}_l - \theta^* \Big\| + \Big\| \bar{\theta}_{l+1} - \bar{\theta}_l \Big\|$, we have:

$$\begin{aligned}
\Big\| \bar{\theta}_{l+1} - \theta^* \Big\| &\leq (1 + 3\alpha c_1) \Big\| \bar{\theta}_l - \theta^* \Big\| \\
&+ \frac{3\alpha c_1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \Big\| \theta_{l,k}^{(i)} - \bar{\theta}_l \Big\| + 3\alpha \Big\| \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)}) \Big\| + 2\alpha c_1 \Gamma(\epsilon, \epsilon_1).
\end{aligned}$$

For simplicity, we define $c_4 \triangleq 3c_1$ and $\delta_l \triangleq \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \Big\| \theta_{l,k}^{(i)} - \bar{\theta}_l \Big\|$. Taking the square on both sides of Eq (4.42), we have:

$$\begin{aligned}
\Big\| \bar{\theta}_{l+1} - \theta^* \Big\|^2 &\leq (1 + \alpha c_4)^2 \Big\| \bar{\theta}_l - \theta^* \Big\|^2 + \alpha^2 c_4^2 \delta_l^2 + 9\alpha^2 \Big\| \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)}) \Big\|^2 + 4\alpha^2 c_1^2 \Gamma^2(\epsilon, \epsilon_1) \\
&\underbrace{+ 6\alpha(1 + \alpha c_4) \Big\| \bar{\theta}_l - \theta^* \Big\| \Big\| \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)}) \Big\|}_{H_1} + \underbrace{2\alpha c_4 (1 + \alpha c_4) \Big\| \bar{\theta}_l - \theta^* \Big\| \delta_l}_{H_2} \\
&\underbrace{+ 6\alpha^2 c_4 \delta_l \Big\| \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)}) \Big\|}_{H_3} + \underbrace{4\alpha^2 c_1 c_4 \delta_l \Gamma(\epsilon, \epsilon_1)}_{H_4} + \underbrace{4\alpha c_1 (1 + \alpha c_4) \Big\| \bar{\theta}_l - \theta^* \Big\| \Gamma(\epsilon, \epsilon_1)}_{H_5} \\
&\underbrace{+ 12\alpha^2 c_1 \Big\| \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)}) \Big\| \Gamma(\epsilon, \epsilon_1)}_{H_6}.
\end{aligned} \tag{4.42}$$

We can further bound H_1 as:

$$\begin{aligned} H_1 &= 6\alpha(1+\alpha c_4)\left\|\bar{\theta}_l - \theta^*\right\|\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\| \\ &= 2\sqrt{3\alpha(1+\alpha c_4)}\left\|\bar{\theta}_l - \theta^*\right\| \cdot \sqrt{3\alpha(1+\alpha c_4)}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\| \\ &\leq 3\alpha(1+\alpha c_4)\left\|\bar{\theta}_l - \theta^*\right\|^2 + 3\alpha(1+\alpha c_4)\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\|^2 \\ &\leq 6\alpha\left\|\bar{\theta}_l - \theta^*\right\|^2 + 6\alpha\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\|^2. \end{aligned} \tag{4.43}$$

where we use the fact $1+\alpha c_4 \leq 2$ in the last inequality. Similary, we can bound H_2 as:

$$H_2 = 2\alpha c_4(1+\alpha c_4)\left\|\bar{\theta}_l - \theta^*\right\|\delta_l \leq 2\alpha\left\|\bar{\theta}_l - \theta^*\right\|^2 + 2\alpha c_4^2\delta_l^2. \tag{4.44}$$

And we bound H_3 as:

$$H_3 = 6\alpha^2 c_4\delta_l\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\| \leq 3\alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\|^2 + 3\alpha^2 c_4^2\delta_l^2. \tag{4.45}$$

For H_4, H_5, H_6, we have:

$$H_4 = 4\alpha^2 c_1 c_4\delta_l\Gamma(\epsilon,\epsilon_1) \leq 2\alpha^2 c_4^2\delta_l^2 + 2\alpha^2 c_1^2\Gamma^2(\epsilon,\epsilon_1),$$

$$H_5 = 4\alpha c_1(1+\alpha c_4)\left\|\bar{\theta}_l - \theta^*\right\|\Gamma(\epsilon,\epsilon_1) \leq 4\alpha\left\|\bar{\theta}_l - \theta^*\right\|^2 + 4\alpha c_1^2\Gamma^2(\epsilon,\epsilon_1),$$

$$H_6 = 12\alpha^2 c_1\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\|\Gamma(\epsilon,\epsilon_1) \leq 6\alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\|^2 + 6\alpha^2 c_1^2\Gamma^2(\epsilon,\epsilon_1),$$

Substituting the upper bound of $H_1, H_2, \ldots, H_6$ into Eq (4.42) and noting that $(1+\alpha c_4)^2 \leq 1+3\alpha c_4$ because $\alpha c_4 \leq 1$, we have:

$$\left\|\bar{\theta}_{l+1} - \theta^*\right\|^2 \leq (1+\alpha(3c_4+12))\left\|\bar{\theta}_l - \theta^*\right\|^2 + (6\alpha^2+2\alpha)c_4^2\delta_l^2$$

$$
\begin{aligned}
&+ (18\alpha^2 + 6\alpha)\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\|^2 + (12\alpha^2 + 4\alpha)c_1^2\Gamma^2(\epsilon,\epsilon_1) \\
&\leq (1+\alpha h_1)\left\|\bar{\theta}_l - \theta^*\right\|^2 + 8\alpha c_4^2\delta_l^2 + 24\alpha\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\|^2 + 16\alpha c_1^2\Gamma^2(\epsilon,\epsilon_1),
\end{aligned} \tag{4.46}
$$

where we denote $h_1 \triangleq 3c_4 + 12$ for simplicity. For any $t-\tau \leq l \leq t$, conditioning on $\mathcal{F}_{t-2\tau}$ on both sides of the above inequality, we have:

$$
\begin{aligned}
\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{l+1} - \theta^*\right\|^2 &\leq (1+\alpha h_1)\,\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_l - \theta^*\right\|^2 + 24\alpha\mathbb{E}_{t-2\tau}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{l,k}^{(i)})\right\|^2 \\
&\quad + 8\alpha c_4^2\mathbb{E}_{t-2\tau}\left[\delta_l^2\right] + \alpha M_3(\epsilon,\epsilon_1) \\
&\leq (1+\alpha h_1)\,\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_l - \theta^*\right\|^2 + 24\alpha\left[\frac{d_2^2}{NK} + 4L_2^2\rho^{2(l-t+2\tau)K}\right] \quad \text{(Lemma 13)} \\
&\quad + 8\alpha c_4^2\mathbb{E}_{t-2\tau}\left[\delta_l^2\right] + \alpha M_3(\epsilon,\epsilon_1) \\
&\overset{(a)}{\leq} (1+\alpha h_1)\,\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_l - \theta^*\right\|^2 + 24\alpha\left[\frac{d_2^2}{NK} + 4L_2^2\alpha^2\rho^{2(l-t+\tau)K}\right] \\
&\quad + 8\alpha c_4^2\mathbb{E}_{t-2\tau}\left[\delta_l^2\right] + \alpha M_3(\epsilon,\epsilon_1) \\
&\leq (1+\alpha h_1)\,\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_l - \theta^*\right\|^2 + \alpha c_t(l) + 8\alpha c_4^2\mathbb{E}_{t-2\tau}\left[\delta_l^2\right] + \alpha M_3(\epsilon,\epsilon_1),
\end{aligned} \tag{4.47}
$$

where we denote $M_3(\epsilon,\epsilon_1) \triangleq 16c_1^2\Gamma^2(\epsilon,\epsilon_1)$ and $c_t(l) = 24\left[\frac{d_2^2}{NK} + 4L_2^2\alpha^2\rho^{2(l-t+\tau)K}\right]$ for simplicity. Inequality (a) is due to $\rho^{2\tau K} \leq \alpha_T^4 \leq \alpha_t^2$. In the following steps, we try to map $\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{l+1} - \theta^*\right\|^2$ to $\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t-\tau} - \theta^*\right\|^2$ for any $t-\tau \leq l \leq t$. By applying Eq (4.47) recursively, we have:

$$
\begin{aligned}
\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{l+1} - \theta^*\right\|^2 &\leq (1+\alpha h_1)^{l+1-t+\tau}\,\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t-\tau} - \theta^*\right\|^2 + \alpha\sum_{k=t-\tau}^{l}(1+\alpha h_1)^{l-k}\left(c_t(k) + M_3(\epsilon,\epsilon_1)\right) \\
&\quad + 8\alpha c_4^2\mathbb{E}_{t-2\tau}\left[\sum_{k=t-\tau}^{l}(1+\alpha h_1)^{l-k}\,\delta_k^2\right] \\
&\overset{(b)}{\leq} (1+\alpha h_1)^{\tau+1}\,\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t-\tau} - \theta^*\right\|^2 + \alpha\underbrace{\sum_{k=t-\tau}^{t}(1+\alpha h_1)^{l-k}\left(c_t(k) + M_3(\epsilon,\epsilon_1)\right)}_{H_7}
\end{aligned}
$$

$$+ 8\alpha c_4^2 \mathbb{E}_{t-2\tau} \underbrace{\left[\sum_{k=t-\tau}^{t} (1+\alpha h_1)^{l-k} \delta_k^2 \right]}_{H_8} \tag{4.48}$$

where (b) is due to $l \leq t$. For H_7, we have:

$$\begin{aligned}
H_7 &\leq \sum_{k=t-\tau}^{t} (1+\alpha h_1)^{t-k} \left(c_t(k) + M_3(\epsilon, \epsilon_1)\right) \quad (l \leq t) \\
&= \sum_{k'=0}^{\tau} (1+\alpha h_1)^{\tau-k'} \left(c_t(k'+t-\tau) + M_3(\epsilon, \epsilon_1)\right) \quad (\text{ changing index } k \text{ into } k' \text{ with } k' = k+\tau-t) \\
&\overset{(a)}{\leq} 24 \sum_{k'=0}^{\tau} (1+\alpha h_1)^{\tau-k'} \left[\frac{d_2^2}{NK} + 4L_2^2 \alpha^2 \rho^{2k'K} + M_3(\epsilon, \epsilon_1) \right] \\
&= 24 \left[\left(\frac{d_2^2}{NK} + M_3(\epsilon, \epsilon_1) \right) \frac{(1+\alpha h_1)^{\tau+1} - 1}{\alpha h_1} + 4L_2^2 \alpha^2 (1+\alpha h_1)^{\tau} \sum_{k'=0}^{\tau} \left(\frac{\rho^{2K}}{1+\alpha h_1} \right)^{k'} \right] \\
&\leq 24 \left[\left(\frac{d_2^2}{NK} + M_3(\epsilon, \epsilon_1) \right) \frac{(1+\alpha h_1)^{\tau+1} - 1}{\alpha h_1} + 4L_2^2 \alpha^2 (1+\alpha h_1)^{\tau} \sum_{k'=0}^{\tau} \rho^{2k'K} \right] \quad (1+\alpha h_1 \geq 1) \\
&\leq 24 \left[\left(\frac{d_2^2}{NK} + M_3(\epsilon, \epsilon_1) \right) \frac{(1+\alpha h_1)^{\tau+1} - 1}{\alpha h_1} + 4L_2^2 \alpha^2 (1+\alpha h_1)^{\tau} \frac{1}{1-\rho^2} \right].
\end{aligned}$$

where (a) is due to the definition of $c_t(k')$. Here we follow the analysis in [91]. Notice that for $x \leq \frac{\log 2}{\tau}$, we have $(1+x)^{\tau+1} \leq 1 + 2x(\tau+1)$. If $\alpha \leq \frac{1}{4h_1\tau} \leq \frac{\log 2}{h_1\tau}$ and $\alpha \leq \frac{1}{2h_1(\tau+1)}$, we have $(1+\alpha h_1)^{\tau+1} \leq 1 + 2\alpha h_1(\tau+1) \leq 2$ and $(1+\alpha h_1)^{\tau} \leq 1 + 2\alpha h_1 \tau \leq 1 + 1/2 \leq 2$. Hence, we have

$$H_7 \leq 24 \left[\left(\frac{d_2^2}{NK} + M_3(\epsilon, \epsilon_1) \right) 2(\tau+1) + \frac{8L_2^2\alpha^2}{1-\rho^2} \right].$$

We apply the similar analysis to bound H_8 as:

$$H_8 = \sum_{k=0}^{\tau} (1+\alpha h_1)^{\tau-k} \delta_{t-\tau+k}^2 \leq \sum_{k=0}^{\tau} (1+\alpha h_1)^{\tau} \delta_{t-\tau+k}^2 \leq \sum_{k=0}^{\tau} (1+2\alpha h_1 \tau) \delta_{t-\tau+k}^2 \leq 2 \sum_{k=0}^{\tau} \delta_{t-k}^2.$$

Substituting the upper bound of H_7 and H_8 into Eq (4.48), we have:

$$\mathbb{E}_{t-2\tau} \left\| \bar{\theta}_{l+1} - \theta^* \right\|^2 \leq 2\mathbb{E}_{t-2\tau} \left\| \bar{\theta}_{t-\tau} - \theta^* \right\|^2 + 24\alpha \left[\left(\frac{d_2^2}{NK} + M_3(\epsilon, \epsilon_1) \right) 2(\tau+1) + \frac{8L_2^2\alpha^2}{1-\rho^2} \right]$$

$$+ 16\alpha c_4^2 \sum_{k=0}^{\tau} \mathbb{E}_{t-2\tau}[\delta_{t-k}^2].$$

Then it is straightforward to bound $\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_l - \theta^*\right\|^2$ as:

$$\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_l - \theta^*\right\|^2 \leq 2\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t-\tau} - \theta^*\right\|^2 + 24\alpha\left[\left(\frac{d_2^2}{NK} + M_3(\epsilon,\epsilon_1)\right)4\tau + \frac{8L_2^2\alpha^2}{1-\rho^2}\right] + 16\alpha c_4^2 \sum_{k=0}^{\tau} \mathbb{E}_{t-2\tau}[\delta_{t-k}^2]. \tag{4.49}$$

Furthermore, based on the triangle inequality, we have:

$$\begin{aligned}\left\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\right\|^2 &\leq \left(\sum_{s=t-\tau}^{t-1}\left\|\bar{\theta}_{s+1} - \bar{\theta}_s\right\|\right)^2 \leq \tau\sum_{s=t-\tau}^{t-1}\left\|\bar{\theta}_{s+1} - \bar{\theta}_s\right\|^2 \\ &\leq \tau\sum_{s=t-\tau}^{t-1}\left[\alpha^2 c_4^2\left\|\bar{\theta}_s - \theta^*\right\|^2 + \alpha^2 c_4^2\delta_s^2 + 6\alpha^2\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{s,k}^{(i)})\right\|^2 + 2\alpha^2 c_1^2\Gamma^2(\epsilon,\epsilon_1)\right]\end{aligned}$$

where the last inequality is due to Eq (4.40) with $c_4 = 3c_1$. If we take the expectation on both sides, we have:

$$\begin{aligned}&\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\right\|^2 \leq \tau\sum_{s=t-\tau}^{t-1}\left[\alpha^2 c_4^2\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_s - \theta^*\right\|^2 + \alpha^2 c_4^2\delta_s^2\right. \\ &\left.+6\alpha^2\mathbb{E}_{t-2\tau}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} Z_i(O_{s,k}^{(i)})\right\|^2 + 2\alpha^2 c_1^2\Gamma^2(\epsilon,\epsilon_1)\right] \\ &\leq \tau\alpha^2 c_4^2\sum_{s=t-\tau}^{t-1}\left[2\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t-\tau} - \theta^*\right\|^2 + 24\alpha\left[\left(\frac{d_2^2}{NK} + M_3(\epsilon,\epsilon_1)\right)4\tau + \frac{8L_2^2\alpha^2}{1-\rho^2}\right]\right. \\ &\left.+16\alpha c_4^2\sum_{k=0}^{\tau}\mathbb{E}_{t-2\tau}[\delta_{t-k}^2]\right] \quad \text{(Eq (4.49))} \\ &+ 6\alpha^2\tau\sum_{s=t-\tau}^{t-1}\left(\frac{d_2^2}{NK} + 4L_2^2\rho^{2(s-t+2\tau)K}\right) \quad \text{(Lemma 13)} \\ &+ \alpha^2 c_4^2\tau\sum_{s=t-\tau}^{t-1}\mathbb{E}_{t-2\tau}[\delta_s^2] + 2\alpha^2 c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1) \\ &\overset{(a)}{\leq} \tau^2\alpha^2 c_4^2\left[2\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t-\tau} - \theta^*\right\|^2 + 96\left(\frac{d_2^2}{NK}\alpha\tau + \frac{2L_2^2\alpha^3}{1-\rho^2}\right)\right]\end{aligned}$$

$$
\begin{aligned}
&+ 6\alpha^2\tau\left[\frac{d_2^2}{NK}\tau + \frac{4L_2^2\alpha^2}{1-\rho^{2K}}\right] + \alpha^2 c_4^2\tau(1+16\alpha\tau c_4^2)\sum_{s=0}^{\tau} E_{t-2\tau}[\delta_{t-s}^2] \\
&+ 96\alpha^2 c_4^2\tau^3 M_3(\epsilon,\epsilon_1) + 2\alpha^2 c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1) \\
&\overset{(b)}{\leq} 2\tau^2\alpha^2 c_4^2\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|^2 + \frac{d_2^2}{NK}\alpha^2\tau^2\left(96\alpha\tau c_4^2+6\right) + \frac{12L_2^2\alpha^4\tau}{1-\rho^2}\left(16\alpha c_4^2\tau+2\right) \\
&+ \alpha^2 c_4^2\tau(1+16\alpha\tau c_4^2)\sum_{s=0}^{\tau} E_{t-2\tau}[\Delta_{t-s}] + 96\alpha^2 c_4^2\tau^3 M_3(\epsilon,\epsilon_1) + 2\alpha^2 c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1) \qquad (4.50)
\end{aligned}
$$

Where we used the fact that $\rho^{2\tau K} \leq \alpha^2$ for (a) and (b), and that $\delta_t^2 \leq \Delta_t$ (via Jensens' inequality) for all $t \geq 0$ in the last inequality. Let us choose α such that $96\alpha\tau c_4^2 + 6 \leq 7$, $16\alpha c_4^2\tau + 2 \leq \frac{13}{6}$ and $1 + 16\alpha\tau c_4^2 \leq 2$, this holds when

$$
\alpha \leq \min\left\{\frac{1}{96\tau c_4^2}, \frac{1}{96 c_4^2\tau}, \frac{1}{16\tau c_4^2}, 1\right\}.
$$

Based on the fact that $\|\bar{\theta}_{t-\tau}-\theta^*\|^2 \leq 2\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\|^2 + 2\|\bar{\theta}_t-\theta^*\|^2$ and the requirement on α, we have

$$
\begin{aligned}
2\alpha^2\tau^2 c_4^2\mathbb{E}_{t-2\tau}\|\bar{\theta}_{t-\tau}-\theta^*\|^2 &\leq 4\alpha^2\tau^2 c_4^2\mathbb{E}_{t-2\tau}\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\|^2 + 4\alpha^2\tau^2 c_4^2\mathbb{E}_{t-2\tau}\|\bar{\theta}_t-\theta^*\|^2 \\
&\overset{(a)}{\leq} 0.5\mathbb{E}_{t-2\tau}\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\|^2 + 4\alpha^2\tau^2 c_4^2\mathbb{E}_{t-2\tau}\|\bar{\theta}_t-\theta^*\|^2 \\
&\overset{(b)}{\leq} \tau^2\alpha^2 c_4^2\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|^2 + \frac{7d_2^2}{2NK}\alpha^2\tau^2 + \frac{13L_2^2\alpha^4\tau}{(1-\rho^2)} \\
&+ \alpha^2 c_4^2\tau\sum_{s=0}^{\tau} E_{t-2\tau}[\Delta_{t-s}] + 48\alpha^2 c_4^2\tau^3 M_3(\epsilon,\epsilon_1) + \alpha^2 c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1) \\
&+ 4\alpha^2\tau^2 c_4^2\mathbb{E}_{t-2\tau}\|\bar{\theta}_t-\theta^*\|^2 \qquad (4.51)
\end{aligned}
$$

where (a) is due to $4\alpha^2\tau^2 c_4^2 \leq 0.5$, and (b) is due to Eq (4.50) and the choice of α. Putting the term $\tau^2\alpha^2 c_4^2\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|^2$ together by rearranging the terms, we have:

$$
\begin{aligned}
\alpha^2\tau^2 c_4^2\mathbb{E}_{t-2\tau}\|\bar{\theta}_{t-\tau}-\theta^*\|^2 &\leq \frac{7d_2^2}{2NK}\alpha^2\tau^2 + \frac{13L_2^2\alpha^4\tau}{(1-\rho^2)} \\
&+ \alpha^2 c_4^2\tau\sum_{s=0}^{\tau} E_{t-2\tau}[\Delta_{t-s}] + 48\alpha^2 c_4^2\tau^3 M_3(\epsilon,\epsilon_1) + \alpha^2 c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1)
\end{aligned}
$$

$$+ 4\alpha^2\tau^2 c_4^2 \mathbb{E}_{t-2\tau}\|\bar{\theta}_t - \theta^*\|^2 \tag{4.52}$$

The proof is completed by substituting this inequality into Eq (4.50) and the definition of $M_3(\epsilon, \epsilon_1)$. Note that we require the effective step-size

$$\alpha \leq \min\Big\{\frac{1}{4h_1\tau}, \frac{1}{2h_1(\tau+1)}, \frac{1}{96c_4^2\tau}, 1\Big\}$$

in this proof, which holds when $\alpha \leq \min\Big\{\frac{1}{30c_4(\tau+1)}, \frac{1}{96c_4^2\tau}, 1\Big\}$ since $c_4 = 3c_1 \geq 1$.

□

• Drift Term Analysis.

Now we bound the drift term as follows:

Lemma 15. *(Bounded Client Drift) If $\alpha_l \leq \frac{1}{2\sqrt{2}c_1(K-1)}$, the drift term satisfies*

$$\mathbb{E}[\Delta_t] = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\theta_{t,k}^{(i)} - \bar{\theta}_t\Big\|^2 \leq \frac{4\alpha^2}{K\alpha_g^2}\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right]. \tag{4.53}$$

Proof.

$$\begin{aligned}
&\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\theta_{t,k}^{(i)} - \bar{\theta}_t\Big\|^2 = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\bar{\theta}_t + \alpha_l\sum_{s=0}^{k-1} g_i(\theta_{t,s}^{(i)}) - \bar{\theta}_t\Big\|^2 \\
&= \alpha_l^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\sum_{s=0}^{k-1} -A_i(O_{t,s}^{(i)})\Big(\theta_{t,s}^{(i)} - \theta_i^*\Big) + Z_i(O_{t,s}^{(i)})\Big\|^2 \\
&\leq 2\alpha_l^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\sum_{s=0}^{k-1} -A_i(O_{t,s}^{(i)})\Big(\theta_{t,s}^{(i)} - \theta_i^*\Big)\Big\|^2 + 2\alpha_l^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\sum_{s=0}^{k-1} Z_i(O_{t,s}^{(i)})\Big\|^2 \\
&\leq 2\alpha_l^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} k\sum_{s=0}^{k-1}\mathbb{E}\Big\|A_i(O_{t,s}^{(i)})\Big(\theta_{t,s}^{(i)} - \theta_i^*\Big)\Big\|^2 + 2\alpha_l^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\sum_{s=0}^{k-1}\mathbb{E}\Big\|Z_i(O_{t,s}^{(i)})\Big\|^2 \\
&+ 2\alpha_l^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\sum_{\substack{s,s'=0\\ s\neq s'}}^{k-1}\mathbb{E}\Big\langle Z_i(O_{t,s}^{(i)}), Z_i(O_{t,s'}^{(i)})\Big\rangle \\
&\leq 2\alpha_l^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} kc_1^2\sum_{s=0}^{k-1}\mathbb{E}\Big\|\theta_{t,s}^{(i)} - \theta_i^*\Big\|^2 + 2\alpha_l^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} kc_3^2 \quad \text{(Lemma 11)}
\end{aligned}$$

$$
\begin{aligned}
&+ 2\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \sum_{\substack{s,s'=0 \\ s \neq s'}}^{k-1} \mathbb{E}\left[\mathbb{E}\Big\langle Z_i(O_{t,s}^{(i)}), Z_i(O_{t,s'}^{(i)}) \Big\rangle \mid \mathcal{F}_s^t \right] \\
&\leq 2\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} k c_1^2 \sum_{s=0}^{k-1} \mathbb{E}\left\| \theta_{t,s}^{(i)} - \theta_i^* \right\|^2 + 2\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} k c_3^2 \\
&+ 2\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \sum_{\substack{s,s'=0 \\ s \neq s'}}^{k-1} \mathbb{E}\left[\Big\langle Z_i(O_{t,s}^{(i)}), \mathbb{E}\left[Z_i(O_{t,s'}^{(i)}) \mid \mathcal{F}_s^t \right] \Big\rangle \right] \\
&\leq 2\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} k c_1^2 \sum_{s=0}^{k-1} \mathbb{E}\left\| \theta_{t,s}^{(i)} - \theta_i^* \right\|^2 + 2\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} k c_3^2 \\
&+ 4\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \sum_{\substack{s,s'=0 \\ s < s'}}^{k-1} \mathbb{E}\left[\left\| Z_i(O_{t,s}^{(i)}) \right\| \left\| \mathbb{E}\left[Z_i(O_{t,s'}^{(i)}) \mid \mathcal{F}_s^t \right] \right\| \right] \\
&\leq 4\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} k c_1^2 \sum_{s=0}^{k-1} \mathbb{E}\left\| \theta_{t,s}^{(i)} - \bar{\theta}_t \right\|^2 + 4\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} k c_1^2 \sum_{s=0}^{k-1} \mathbb{E}\left\| \bar{\theta}_t - \theta_i^* \right\|^2 \quad \text{(Eq (6.6))} \\
&+ 2\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} k c_3^2 + 4\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \sum_{\substack{s,s'=0 \\ s < s'}}^{k-1} c_3 L_2 \rho^{s'-s} \quad \text{(Lemma 12)} \\
&\leq 4\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} k c_1^2 \sum_{s=0}^{k-1} \mathbb{E}\left\| \theta_{t,s}^{(i)} - \bar{\theta}_t \right\|^2 + 4\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} 4k c_1^2 (K-1) H^2 \\
&+ 2\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} k c_3^2 + 4\alpha_l^2 \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \underbrace{\sum_{\substack{s,s'=0 \\ s < s'}}^{k-1} c_3 L_2 \rho^{s'-s}}_{\mathcal{M}_1}
\end{aligned} \tag{4.54}
$$

where we used the property that $\bar{\theta}_t, \theta_i^* \in \mathcal{H}$ in the last inequality, i.e., $\|\bar{\theta}_t\| \leq H^2$ and $\|\theta_i^*\| \leq H^2$.

We now bound $\mathcal{M}_1$ as:

$$
\sum_{\substack{s,s'=0 \\ s < s'}}^{k-1} c_3 L_2 \rho^{s'-s} = c_3 L_2 \sum_{s=0}^{k-1} \sum_{s'=s+1}^{k-1} \rho^{s'-s} = c_3 L_2 \sum_{s=0}^{k-1} \frac{\rho - \rho^{s-s'}}{1-\rho} \leq c_3 L_2 \frac{\rho k}{1-\rho} \tag{4.55}
$$

Define $\mathcal{R}_K \triangleq \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbb{E}\left\| \theta_{t,k}^{(i)} - \bar{\theta}_t \right\|^2$ and note that $\mathcal{R}_K$ is monotonically increasing in K. With

this definition, if we plug in the upper bound of $\mathcal{M}_1$ into Eq (4.54), we have:

$$\begin{aligned}
\mathcal{R}_K &\leq 4\alpha_l^2 \sum_{i=1}^{N}\sum_{k=0}^{K-1} kc_1^2 \sum_{s=0}^{k-1} \mathbb{E}\left\|\theta_{t,s}^{(i)} - \bar{\theta}_t\right\|^2 + 4\alpha_l^2 \sum_{i=1}^{N}\sum_{k=0}^{K-1} 4kc_1^2(K-1)H^2 \\
&\quad + 2\alpha_l^2 \sum_{i=1}^{N}\sum_{k=0}^{K-1} kc_3^2 + 4\alpha_l^2 \sum_{i=1}^{N}\sum_{k=0}^{K-1} c_3 L_2 \frac{\rho k}{1-\rho} \\
&\leq 2\alpha_l^2(K-1)NK\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right] + 4\alpha_l^2 c_1^2(K-1)\sum_{k=1}^{K-1}\underbrace{\sum_{i=1}^{N}\sum_{s=0}^{k-1}\mathbb{E}\left\|\theta_{t,s}^{(i)} - \bar{\theta}_t\right\|^2}_{\mathcal{R}_k} \\
&= 2\alpha_l^2(K-1)NK\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right] + 4\alpha_l^2 c_1^2(K-1)\sum_{k=1}^{K-1}\mathcal{R}_k \qquad (4.56)
\end{aligned}$$

By the monotonicity of $\mathcal{R}_k$, we have

$$\mathcal{R}_K \leq 2\alpha_l^2(K-1)NK\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right] + 4\alpha_l^2 c_1^2(K-1)^2\mathcal{R}_{K-1}$$

Let us choose α_l such that $4\alpha_l^2 c_1^2(K-1)^2 \leq \frac{1}{2}$, i.e., $\alpha_l \leq \frac{1}{2\sqrt{2}c_1(K-1)}$, the following recursion holds:

$$\mathcal{R}_K \leq \frac{1}{2}\mathcal{R}_{K-1} + 2\alpha_l^2(K-1)NK\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right] \qquad (4.57)$$

for all $k \in [K]$. Next, we unroll the recurrence, go back $K-1$ steps and use the fact that $\mathcal{R}_1 = 0$, we have:

$$\begin{aligned}
\mathcal{R}_K &\leq \left\{\sum_{l=1}^{\infty}\left(\frac{1}{2}\right)^l\right\}\left(2\alpha_l^2(K-1)NK\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right]\right) \\
&= 4\alpha_l^2(K-1)NK\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right] \qquad (4.58)
\end{aligned}$$

We finish the proof by dividing NK on both sides and substituting $\alpha_l = \frac{\alpha}{K\alpha_g}$. □

- Per Round Progress

Lemma 16. *(Per Round Progress). If the local step-size* $\alpha_l \leq \frac{1}{2\sqrt{2}c_1(K-1)}$*, and the effective step-size* $\alpha = K\alpha_l\alpha_g$ *satisfies:*

$$\alpha \leq \min\{\frac{\xi_1}{24(c_1+c_2)^2+24\xi_1^2+16}, 1, \frac{\xi_1(c_1+c_2)}{2L_1+8\tau^2c_4^2}, \frac{1}{30c_4(\tau+1)}, \frac{1}{96c_4^2\tau}, \mathcal{X}\},^{4}$$

where

$$\mathcal{X} = \frac{2B(\epsilon,\epsilon_1)G + 3\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1)}{4B^2(\epsilon,\epsilon_1) + 24(c_1+c_2)^2\Gamma^2(\epsilon,\epsilon_1) + 2L_1\Gamma(\epsilon,\epsilon_1)G + 6400c_1^2c_4^2\tau^3\Gamma^2(\epsilon,\epsilon_1) + 8c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1)},$$

and choose $\tau = \lceil \frac{\tau^{\text{mix}}(\alpha_T^2)}{K} \rceil$*, then we have,*

$$\begin{aligned}
\mathbb{E}_{t-2\tau}\|\bar{\theta}_{t+1}-\theta^*\|^2 \leq & \underbrace{(1+32\alpha\xi_1(c_1+c_2))\,\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 + 2\alpha\mathbb{E}_{t-2\tau}\left\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t-\theta^*\right\rangle + 4\alpha^2\mathbb{E}_{t-2\tau}\left\|\bar{g}(\bar{\theta}_t)\right\|^2}_{\textit{Expected progress for the virtual MDP}} \\
& + \underbrace{\frac{9+28\tau^2}{NK}\alpha^2 d_2^2}_{\textit{Linear speedup}} + \underbrace{\alpha^3\left(36L_2^2 + \frac{108\tau}{1-\rho^2}L_2^2 + 4L_1G^2 + 2L_2G\right)}_{\textit{High order terms: } O(\alpha^3)} \\
& + \underbrace{\frac{4\alpha^3}{K\alpha_g^2}(\frac{14}{\xi_1}+14\xi_1)(c_1+c_2)\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 4c_1^2(K-1)H^2\right]}_{\textit{drift term}} \\
& + \underbrace{4\alpha B(\epsilon,\epsilon_1)G + 6\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1)}_{\textit{heterogeneity term}}. \qquad (4.59)
\end{aligned}$$

where ξ_1 *is any universal positive constant.*

Proof. According to the updating rule and the fact that the projection operator is non-expansive, we have:

$$\begin{aligned}
\mathbb{E}_{t-\tau}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2 &= \mathbb{E}_{t-\tau}\left\|\Pi_{2,\mathcal{H}}\left(\bar{\theta}_t + \frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} g_i(\theta_{t,k}^{(i)})\right) - \theta^*\right\|^2 \\
&\leq \mathbb{E}_{t-\tau}\left\|\bar{\theta}_t + \frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1} g_i(\theta_{t,k}^{(i)}) - \theta^*\right\|^2
\end{aligned}$$

[4]This requirement is very easy to satisfy since the denominator in $\mathcal{X}$ is composed by the heterogeneity terms, which is quite small and thereby makes $\mathcal{X}$ large. Overall, the feasible set of the step-sizes is not empty.

$$
\begin{aligned}
=&\mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t-\theta^*\Big\|^2+2\mathbb{E}_{t-\tau}\Big\langle\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\bar{g}_i(\theta_{t,k}^{(i)}),\bar{\theta}_t-\theta^*\Big\rangle\\
&+2\mathbb{E}_{t-\tau}\Big\langle\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}[g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})],\bar{\theta}_t-\theta^*\Big\rangle+\alpha^2\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{t,k}^{(i)})\Big\|^2\\
\leq&\mathbb{E}_{t-\tau}\underbrace{\left\{\Big\|\bar{\theta}_t-\theta^*\Big\|^2+2\Big\langle\frac{\alpha}{N}\sum_{i=1}^{N}\bar{g}_i(\bar{\theta}_t),\bar{\theta}_t-\theta^*\Big\rangle+2\Big\langle\frac{\alpha}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\bar{g}_i(\theta_{t,k}^{(i)})-\bar{g}_i(\bar{\theta}_t),\bar{\theta}_t-\theta^*\Big\rangle\right\}}_{\mathcal{B}_1}\\
&+2\alpha\mathbb{E}_{t-\tau}\underbrace{\Big\langle\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}[g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})],\bar{\theta}_t-\theta^*\Big\rangle}_{\mathcal{B}_2}+\alpha^2\mathbb{E}_{t-\tau}\underbrace{\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{t,k}^{(i)})\Big\|^2}_{\mathcal{B}_3}
\end{aligned}
\tag{4.60}
$$

We now begin to bound the gradient bias term $\mathcal{B}_2$ by decomposing this term into three terms:

$$
\begin{aligned}
&\Big\langle\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}[g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})],\bar{\theta}_t-\theta^*\Big\rangle\\
=&\underbrace{\Big\langle\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}[g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})],\bar{\theta}_t-\bar{\theta}_{t-\tau}\Big\rangle}_{\mathcal{B}_{21}}\\
&+\underbrace{\Big\langle\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}[g_i(\theta_{t,k}^{(i)})-g_i(\theta_{t-\tau,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})+\bar{g}_i(\theta_{t-\tau,k}^{(i)})],\bar{\theta}_{t-\tau}-\theta^*\Big\rangle}_{\mathcal{B}_{22}}\\
&+\underbrace{\Big\langle\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}[g_i(\theta_{t-\tau,k}^{(i)})-\bar{g}_i(\theta_{t-\tau,k}^{(i)})],\bar{\theta}_{t-\tau}-\theta^*\Big\rangle}_{\mathcal{B}_{23}}.
\end{aligned}
\tag{4.61}
$$

Next, we bound $\mathbb{E}_{t-\tau}[\mathcal{B}_{21}]$ as:

$$
\begin{aligned}
&\mathbb{E}_{t-\tau}\Big\langle\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}[g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})],\bar{\theta}_t-\bar{\theta}_{t-\tau}\Big\rangle\leq\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})\Big\|\Big\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\Big\|\\
&\overset{(a)}{=}\mathbb{E}_{t-\tau}\left[\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}(-A_i(O_{t,k}^{(i)})+\bar{A}_i)(\theta_{t,k}^{(i)}-\theta_i^*)+Z_i(O_{t,k}^{(i)})\Big\|\Big\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\Big\|\right]\\
&\leq\mathbb{E}_{t-\tau}\left[\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}(A_i(O_{t,k}^{(i)})-\bar{A}_i)(\theta_{t,k}^{(i)}-\theta_i^*)\Big\|\Big\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\Big\|\right]+\mathbb{E}_{t-\tau}\left[\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\Big\|\Big\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right.
\end{aligned}
$$

$$\leq \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left[\left\|(A_i(O_{t,k}^{(i)})-\bar{A}_i)(\theta_{t,k}^{(i)}-\theta_i^*)\right\|\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|\right]+$$

$$\frac{\alpha}{2}\mathbb{E}_{t-\tau}\left[\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right\|^2\right]+\frac{1}{2\alpha}\mathbb{E}_{t-\tau}\left[\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2\right]$$

$$\overset{(b)}{\leq}\frac{(c_1+c_2)}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\theta_{t,k}^{(i)}-\theta_i^*\right\|\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|+\frac{\alpha}{2}\mathbb{E}_{t-\tau}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right\|^2+\frac{1}{2\alpha}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2$$

$$\leq\frac{\xi_1(c_1+c_2)}{2NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\theta_{t,k}^{(i)}-\theta_i^*\right\|^2+\frac{(c_1+c_2)}{2\xi_1 NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2\quad\text{(Young's inequality (6.7))}$$

$$+\frac{\alpha}{2}\mathbb{E}_{t-\tau}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right\|^2+\frac{1}{2\alpha}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2$$

$$\overset{(c)}{\leq}\frac{3\xi_1(c_1+c_2)}{2NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\theta_{t,k}^{(i)}-\bar{\theta}_t\right\|^2+\frac{3\xi_1(c_1+c_2)}{2NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2$$

$$+\frac{3\xi_1(c_1+c_2)}{2NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\theta^*-\theta_i^*\right\|^2$$

$$+\frac{(c_1+c_2)}{2\xi_1 NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2+\frac{\alpha}{2}\mathbb{E}_{t-\tau}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right\|^2+\frac{1}{2\alpha}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2$$

$$=\frac{3\xi_1(c_1+c_2)}{2NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\theta_{t,k}^{(i)}-\bar{\theta}_t\right\|^2+\frac{3\xi_1(c_1+c_2)}{2}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+\frac{3\xi_1(c_1+c_2)}{2}\Gamma^2(\epsilon,\epsilon_1)$$

$$+\frac{(c_1+c_2)}{2\xi_1}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2+\frac{\alpha}{2}\mathbb{E}_{t-\tau}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\right\|^2+\frac{1}{2\alpha}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2$$

$$\overset{(d)}{\leq}\frac{3\xi_1(c_1+c_2)}{2NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\theta_{t,k}^{(i)}-\bar{\theta}_t\right\|^2+\frac{3\xi_1(c_1+c_2)}{2}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+\frac{3\xi_1(c_1+c_2)}{2}\Gamma^2(\epsilon,\epsilon_1)$$

$$+\frac{c_1+c_2}{2\xi_1}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2+\frac{\alpha}{2}\left[\frac{d_2^2}{NK}+4L_2^2\rho^{2\tau K}\right]+\frac{1}{2\alpha}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2$$

$$=\frac{3\xi_1(c_1+c_2)}{2NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\left\|\theta_{t,k}^{(i)}-\bar{\theta}_t\right\|^2+\frac{3\xi_1(c_1+c_2)}{2}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+\frac{3\xi_1(c_1+c_2)}{2}\Gamma^2(\epsilon,\epsilon_1)$$

$$+\left(\frac{c_1+c_2}{2\xi_1}+\frac{1}{2\alpha}\right)\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2+\frac{\alpha}{2}\left[\frac{d_2^2}{NK}+\underbrace{4L_2^2\rho^{2\tau K}}_{\leq 4L_2^2\alpha^2}\right],\tag{4.62}$$

where (a) is due to $g_i(\theta_{t,k}^{(i)})=-A_i(O_{t,k}^{(i)})(\theta_{t,k}^{(i)}-\theta_i^*)+Z_i(O_{t,k}^{(i)})$, (b) is due to Lemma 12 (the upper

bound of $A_i(O_{t,k}^{(i)})$ and $\bar{A}_i$), (c) is due to Eq (6.8) and (d) is due to Lemma 13.

And we bound $\mathcal{B}_{22}$ as:

$$
\begin{aligned}
\mathcal{B}_{22} &= \Big\langle \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}[g_i(\theta_{t,k}^{(i)}) - g_i(\theta_{t-\tau,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}) + \bar{g}_i(\theta_{t-\tau,k}^{(i)})], \bar{\theta}_{t-\tau} - \theta^* \Big\rangle \\
&\leq \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big\|g_i(\theta_{t,k}^{(i)}) - g_i(\theta_{t-\tau,k}^{(i)}) - \bar{g}_i(\theta_{t,k}^{(i)}) + \bar{g}_i(\theta_{t-\tau,k}^{(i)})\Big\|\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\| \text{ (Cauchy-Schwarz inequality)} \\
&\leq \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big[\Big\|g_i(\theta_{t,k}^{(i)}) - g_i(\theta_{t-\tau,k}^{(i)})\Big\| + \Big\|\bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\theta_{t-\tau,k}^{(i)})\Big\|\Big]\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\| \\
&\overset{(a)}{\leq} \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big[2\Big\|\theta_{t,k}^{(i)} - \theta_{t-\tau,k}^{(i)}\Big\| + 2\Big\|\theta_{t,k}^{(i)} - \theta_{t-\tau,k}^{(i)}\Big\|\Big]\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\| \\
&\leq \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big[4\Big\|\theta_{t,k}^{(i)} - \bar{\theta}_t\Big\| + 4\Big\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\Big\| + 4\Big\|\bar{\theta}_{t-\tau} - \theta_{t-\tau,k}^{(i)}\Big\|\Big]\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\| \text{ (Triangle inequality)} \\
&\leq 4\delta_t\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\| + 4\Big\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\Big\|\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\| + 4\delta_{t-\tau}\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\| \\
&\overset{(b)}{\leq} \frac{2}{\xi_2}\Delta_t + \frac{2}{\xi_2}\Delta_{t-\tau} + (2\xi_2 + 4\xi_2)\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\|^2 + \frac{2}{\xi_2}\Big\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\Big\|^2 \\
&\leq \frac{2}{\xi_2}\Delta_t + \frac{2}{\xi_2}\Delta_{t-\tau} + 12\xi_2\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + (12\xi_2 + \frac{2}{\xi_2})\Big\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\Big\|^2 \quad \text{(Eq 6.8)}
\end{aligned} \tag{4.63}
$$

where (a) is due to the 2-Lipschitz property of steady-state $\bar{g}$ (i.e., Lemma 5) and random direction g_i (i.e., Lemma 6), $\delta_t = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big\|\theta_{t,k}^{(i)} - \bar{\theta}_t\Big\|$ and $\Delta_t \triangleq \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\theta_{t,k}^{(i)} - \bar{\theta}_t\Big\|^2$, and (b) is due to Young's inequality (6.7) with constants ξ_2 and $\delta_t^2 \leq \Delta_t$.

Now, we bound $\mathcal{B}_{23}$ as:

$$
\begin{aligned}
\mathbb{E}_{t-\tau}[\mathcal{B}_{23}] &= \Big\langle \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}[g_i(\theta_{t-\tau,k}^{(i)}) - \bar{g}_i(\theta_{t-\tau,k}^{(i)})], \bar{\theta}_{t-\tau} - \theta^* \Big\rangle \\
&\leq \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\|\Big\|\mathbb{E}_{t-\tau}\Big[g_i(\theta_{t-\tau,k}^{(i)}) - \bar{g}_i(\theta_{t-\tau,k}^{(i)})\Big]\Big\| \text{ (Cauchy-Schwarz inequality)} \\
&= \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big\|\bar{\theta}_{t-\tau} - \theta^*\Big\|\Big\|\mathbb{E}_{t-\tau}\Big[-A_i(O_{t,k}^{(i)})(\theta_{t-\tau,k}^{(i)} - \theta_i^*) + Z_i(O_{t,k}^{(i)}) + \bar{A}_i(\theta_{t-\tau,k}^{(i)} - \theta_i^*)\Big]\Big\|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|\left\{\left\|\mathbb{E}_{t-\tau}(A_i(O_{t,k}^{(i)})-\bar{A}_i)(\theta_{t-\tau,k}^{(i)}-\theta_i^*)\right\|+\left\|\mathbb{E}_{t-\tau}\left[Z_i(O_{t,k}^{(i)})\right]\right\|\right\}\\
&\overset{(a)}{\leq} \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|\left\{L_1\rho^{\tau K+k}\left\|\theta_{t-\tau,k}^{(i)}-\theta_i^*\right\|+L_2\rho^{\tau K+k}\right\}\\
&\leq \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|\left\{L_1\rho^{\tau K+k}\left[\left\|\theta_{t-\tau,k}^{(i)}-\bar{\theta}_{t-\tau}\right\|+\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|+\left\|\theta^*-\theta_i^*\right\|\right]+L_2\rho^{\tau K+k}\right\}\\
&\overset{(b)}{\leq} \alpha^2L_1\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|\delta_{t-\tau}+\alpha^2L_1\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|^2+\alpha^2L_1\Gamma(\epsilon,\epsilon_1)G+\alpha^2L_2G\\
&\leq \alpha^2L_1\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|^2+\alpha^2L_1\Delta_{t-\tau}+\alpha^2L_1\left\|\bar{\theta}_{t-\tau}-\theta^*\right\|^2+\alpha^2L_1\Gamma(\epsilon,\epsilon_1)G+\alpha^2L_2G\\
&\overset{(c)}{\leq} 2\alpha^2L_1G^2+\alpha^2L_2G+\alpha^2L_1\Delta_{t-\tau}+\alpha^2L_1\Gamma(\epsilon,\epsilon_1)G, \qquad (4.64)
\end{aligned}
$$

where (a) is due to Lemma 12, (b) is due to the fact that $\bar{\theta}_{t-\tau},\theta^*\in\mathcal{H}$, which radius is $H\leq\frac{G}{2}$, and $\tau=\lceil\frac{\log_\rho(\alpha_T^2)}{K}\rceil$ (i.e., $\rho^{\tau K}\leq\alpha^2$) and (c) is due to the fact that $\bar{\theta}_{t-\tau},\theta^*\in\mathcal{H}$. Then, the term $\mathcal{B}_2$ can be bounded as:

$$
\begin{aligned}
\mathbb{E}_{t-\tau}[\mathcal{B}_2] &= \mathbb{E}_{t-\tau}[\mathcal{B}_{21}+\mathcal{B}_{22}+\mathcal{B}_{23}]\\
&\leq \frac{3\xi_1(c_1+c_2)}{2}\mathbb{E}_{t-\tau}[\Delta_t]+\frac{3\xi_1(c_1+c_2)}{2}\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+\frac{3\xi_1(c_1+c_2)}{2}\Gamma^2(\epsilon,\epsilon_1)\\
&+\left(\frac{c_1+c_2}{2\xi_1}+\frac{1}{2\alpha}\right)\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2+\frac{\alpha}{2}\left[\frac{d_2^2}{NK}+4L_2^2\rho^{2\tau K}\right]\\
&+\frac{2}{\xi_2}\mathbb{E}_{t-\tau}[\Delta_t]+\frac{2}{\xi_2}\mathbb{E}_{t-\tau}[\Delta_{t-\tau}]+12\xi_2\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+(12\xi_2+\frac{2}{\xi_2})\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2\\
&+2\alpha^2L_1G^2+\alpha^2L_2G+\alpha^2L_1\mathbb{E}_{t-\tau}[\Delta_{t-\tau}]+\alpha^2L_1\Gamma(\epsilon,\epsilon_1)G\\
&\leq \left(\frac{3\xi_1(c_1+c_2)}{2}+12\xi_2\right)\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+\left(\frac{c_1+c_2}{2\xi_1}+\frac{1}{2\alpha}+12\xi_2+\frac{2}{\xi_2}\right)\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t-\bar{\theta}_{t-\tau}\right\|^2\\
&+\left(\frac{3\xi_1(c_1+c_2)}{2}+\frac{2}{\xi_2}\right)\mathbb{E}_{t-\tau}[\Delta_t]+\left(\frac{2}{\xi_2}+\alpha^2L_1\right)\Delta_{t-\tau}+\frac{\alpha}{2}\left[\frac{d_2^2}{NK}+4L_2^2\alpha^2\right]\\
&+2\alpha^2L_1G^2+\alpha^2L_2G+\frac{3\xi_1(c_1+c_2)}{2}\Gamma^2(\epsilon,\epsilon_1)+\alpha^2L_1\Gamma(\epsilon,\epsilon_1)G \qquad (4.65)
\end{aligned}
$$

Next, we bound $\mathcal{B}_3$ as:

$$
\mathbb{E}_{t-\tau}[\mathcal{B}_3]=\mathbb{E}_{t-\tau}\left\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left[g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})+\bar{g}_i(\theta_{t,k}^{(i)})-\bar{g}_i(\bar{\theta}_t)+\bar{g}_i(\bar{\theta}_t)-\bar{g}(\bar{\theta}_t)+\bar{g}(\bar{\theta}_t)\right]\right\|^2
$$

$$
\begin{aligned}
&\leq 4\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big(g_i(\theta_{t,k}^{(i)})-\bar{g}_i(\theta_{t,k}^{(i)})\Big)\Big\|^2+4\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big(\bar{g}_i(\theta_{t,k}^{(i)})-\bar{g}_i(\bar{\theta}_t)\Big)\Big\|^2\\
&+4\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\big(\bar{g}_i(\bar{\theta}_t)-\bar{g}(\bar{\theta}_t)\big)\Big\|^2+4\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\bar{g}(\bar{\theta}_t)\Big\|^2 \quad \text{(Eq 6.8)}\\
&=4\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big[\Big(\bar{A}_i-A_i(O_{t,k}^{(i)})\Big)(\theta_{t,k}^{(i)}-\theta_i^*)+Z_i(O_{t,k}^{(i)})\Big]\Big\|^2\\
&+4\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big(\bar{g}_i(\theta_{t,k}^{(i)})-\bar{g}_i(\bar{\theta}_t)\Big)\Big\|^2+4\mathbb{E}_{t-\tau}\underbrace{\Big\|\frac{1}{N}\sum_{i=1}^{N}\big(\bar{g}_i(\bar{\theta}_t)-\bar{g}(\bar{\theta}_t)\big)\Big\|^2}_{\text{Lemma 2}}+4\mathbb{E}_{t-\tau}\Big\|\bar{g}(\bar{\theta}_t)\Big\|^2\\
&\overset{(a)}{\leq} 8\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\Big(\bar{A}_i-A_i(O_{t,k}^{(i)})\Big)(\theta_{t,k}^{(i)}-\theta_i^*)\Big\|^2+8\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\Big\|^2\\
&+16\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\Big\|\theta_{t,k}^{(i)}-\bar{\theta}_t\Big\|^2+4B^2(\epsilon,\epsilon_1)+4\mathbb{E}_{t-\tau}\Big\|\bar{g}(\bar{\theta}_t)\Big\|^2\\
&\leq \frac{8}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\Big\|\bar{A}_i-A_i(O_{t,k}^{(i)})\Big\|^2\Big\|\theta_{t,k}^{(i)}-\theta_i^*\Big\|^2+8\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\Big\|^2\\
&+16\mathbb{E}_{t-\tau}[\Delta_t]+4B^2(\epsilon,\epsilon_1)+4\mathbb{E}_{t-\tau}\Big\|\bar{g}(\bar{\theta}_t)\Big\|^2\\
&\leq \frac{8(c_1+c_2)^2}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\Big\|\theta_{t,k}^{(i)}-\theta_i^*\Big\|^2+8\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\Big\|^2\\
&+16\mathbb{E}_{t-\tau}[\Delta_t]+4B^2(\epsilon,\epsilon_1)+4\mathbb{E}_{t-\tau}\Big\|\bar{g}(\bar{\theta}_t)\Big\|^2\\
&\leq \frac{24(c_1+c_2)^2}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\Big\|\theta_{t,k}^{(i)}-\bar{\theta}_t\Big\|^2+\frac{24(c_1+c_2)^2}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t-\theta^*\Big\|^2\\
&+\frac{24(c_1+c_2)^2}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}_{t-\tau}\Big\|\theta_i^*-\theta^*\Big\|^2+8\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\Big\|^2\\
&+16\mathbb{E}_{t-\tau}[\Delta_t]+4B^2(\epsilon,\epsilon_1)+4\mathbb{E}_{t-\tau}\Big\|\bar{g}(\bar{\theta}_t)\Big\|^2\\
&=24(c_1+c_2)^2\mathbb{E}_{t-\tau}[\Delta_t]+24(c_1+c_2)^2\mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t-\theta^*\Big\|^2+24(c_1+c_2)^2\Gamma^2(\epsilon,\epsilon_1)\\
&+8\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=0}^{K-1}Z_i(O_{t,k}^{(i)})\Big\|^2+16\mathbb{E}_{t-\tau}[\Delta_t]+4B^2(\epsilon,\epsilon_1)+4\mathbb{E}_{t-\tau}\Big\|\bar{g}(\bar{\theta}_t)\Big\|^2\\
&\overset{(b)}{\leq}(24(c_1+c_2)^2+16)\mathbb{E}_{t-\tau}[\Delta_t]+8(\frac{d_2^2}{NK}+\underbrace{4L_2^2\rho^{\tau K}}_{\leq 4L_2^2\alpha^2})+24(c_1+c_2)^2\mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t-\theta^*\Big\|^2
\end{aligned}
$$

$$+ 4\mathbb{E}_{t-\tau}\Big\|\bar{g}(\bar{\theta}_t)\Big\|^2 + 4B^2(\epsilon,\epsilon_1) + 24(c_1+c_2)^2\Gamma^2(\epsilon,\epsilon_1), \tag{4.66}$$

where (a) is due to 2-Lipschitz of $\bar{g}_i$ (i.e., Lemma 5) and the gradient heterogeneity (i.e., Lemma 2) and (b) is due to Lemma 13.

Next, we bound $\mathcal{B}_1$ as:

$$\begin{aligned}
\mathbb{E}_{t-\tau}[\mathcal{B}_1] &= \mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + 2\mathbb{E}_{t-\tau}\Big\langle \frac{\alpha}{N}\sum_{i=1}^N \bar{g}_i(\bar{\theta}_t), \bar{\theta}_t - \theta^*\Big\rangle + 2\mathbb{E}_{t-\tau}\Big\langle \frac{\alpha}{NK}\sum_{i=1}^N\sum_{k=0}^{K-1} \bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\bar{\theta}_t), \bar{\theta}_t - \theta^*\Big\rangle \\
&\le \mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + 2\alpha\mathbb{E}_{t-\tau}\Big\langle \frac{1}{N}\sum_{i=1}^N \bar{g}_i(\bar{\theta}_t) - \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\Big\rangle + 2\alpha\mathbb{E}_{t-\tau}\Big\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\Big\rangle \\
&\quad + 2\alpha\mathbb{E}_{t-\tau}\Big\langle \frac{1}{NK}\sum_{i=1}^N\sum_{k=0}^{K-1} \bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\bar{\theta}_t), \bar{\theta}_t - \theta^*\Big\rangle \\
&\le \mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + 2\alpha\mathbb{E}_{t-\tau}\Big\|\frac{1}{N}\sum_{i=1}^N \bar{g}_i(\bar{\theta}_t) - \bar{g}(\bar{\theta}_t)\Big\|\Big\|\bar{\theta}_t - \theta^*\Big\| + 2\alpha\mathbb{E}_{t-\tau}\Big\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\Big\rangle \\
&\quad + \frac{\alpha}{\xi_3}\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^N\sum_{k=0}^{K-1}\Big(\bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\bar{\theta}_t)\Big)\Big\|^2 + \alpha\xi_3\mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 \\
&\qquad \text{(Young's inequality Eq (6.7) with constant } \xi_3) \\
&\overset{(a)}{\le} \mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + 2\alpha B(\epsilon,\epsilon_1)G + 2\alpha\mathbb{E}_{t-\tau}\Big\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\Big\rangle \\
&\quad + \frac{\alpha}{\xi_3}\mathbb{E}_{t-\tau}\Big\|\frac{1}{NK}\sum_{i=1}^N\sum_{k=0}^{K-1}\Big(\bar{g}_i(\theta_{t,k}^{(i)}) - \bar{g}_i(\bar{\theta}_t)\Big)\Big\|^2 + \alpha\xi_3\mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 \\
&\overset{(b)}{\le} \mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + 2\alpha B(\epsilon,\epsilon_1)G + 2\alpha\mathbb{E}_{t-\tau}\Big\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\Big\rangle + \frac{4\alpha}{\xi_3}\mathbb{E}_{t-\tau}[\Delta_t] + \alpha\xi_3\mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t - \theta^*\Big\|^2,
\end{aligned} \tag{4.67}$$

where (a) is due to the fact that $\bar{\theta}_t, \theta^* \in \mathcal{H}$ and the gradient heterogeneity; (b) is due to 2-Lipschitz property of function $\bar{g}$ in Lemma 5.

Incorporating the upper of $\mathcal{B}_1$ from Eq (4.67), $\mathcal{B}_2$ from Eq (4.65) and $\mathcal{B}_3$ from Eq (4.66) into Eq (4.60), we have:

$$\mathbb{E}_{t-\tau}\Big\|\bar{\theta}_{t+1} - \theta^*\Big\|^2 \le \mathbb{E}_{t-\tau}\Big\|\bar{\theta}_t - \theta^*\Big\|^2 + 2\alpha\mathbb{E}_{t-\tau}\Big\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\Big\rangle + 4\alpha^2\mathbb{E}_{t-\tau}\Big\|\bar{g}(\bar{\theta}_t)\Big\|^2$$

$$
\begin{aligned}
&+ \left(\alpha\xi_3 + \alpha(3\xi_1(c_1+c_2) + 24\xi_2) + 24\alpha^2(c_1+c_2)^2\right)\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t - \theta^*\right\|^2 \\
&+ \alpha\left(\frac{c_1+c_2}{\xi_1} + \frac{1}{\alpha} + 24\xi_2 + \frac{4}{\xi_2}\right)\mathbb{E}_{t-\tau}\left\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\right\|^2 \\
&+ \frac{9d_2^2}{NK}\alpha^2 + 36L_2^2\alpha^4 + 4\alpha^3L_1G^2 + 2\alpha^3L_2G + \left(\frac{4\alpha}{\xi_2} + 2\alpha^3L_1\right)\Delta_{t-\tau} \\
&+ \alpha\left(\frac{4}{\xi_3} + 3\xi_1(c_1+c_2) + \frac{4}{\xi_2} + \alpha^2\left(24(c_1+c_2)^2 + 16\right)\right)\mathbb{E}_{t-\tau}[\Delta_t] \\
&+ 2\alpha B(\epsilon,\epsilon_1)G + 4\alpha^2B^2(\epsilon,\epsilon_1) + 24\alpha^2(c_1+c_2)^2\Gamma^2(\epsilon,\epsilon_1) \\
&+ 3\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1) + 2\alpha^3L_1\Gamma(\epsilon,\epsilon_1)G
\end{aligned} \tag{4.68}
$$

Conditioned on $\mathcal{F}_{t-2\tau}$ and using Lemma 14 to give an upper bound of $\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t - \bar{\theta}_{t-\tau}\right\|^2$, we have:

$$
\begin{aligned}
&\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t+1} - \theta^*\right\|^2 \leq \mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t - \theta^*\right\|^2 + 2\alpha\mathbb{E}_{t-2\tau}\left\langle\bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\right\rangle + 4\alpha^2\mathbb{E}_{t-2\tau}\left\|\bar{g}(\bar{\theta}_t)\right\|^2 \\
&+ \underbrace{\left(\alpha\xi_3 + \alpha(3\xi_1(c_1+c_2) + 24\xi_2) + 24\alpha^2(c_1+c_2)^2\right)}_{\mathcal{E}_1}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t - \theta^*\right\|^2 \\
&+ \underbrace{\alpha\left(\frac{c_1+c_2}{\xi_1} + \frac{1}{\alpha} + 24\xi_2 + \frac{4}{\xi_2}\right)}_{\mathcal{E}_2}\left\{8\alpha^2\tau^2c_4^2\mathbb{E}_{t-2\tau}\left[\left\|\bar{\theta}_t - \theta^*\right\|^2\right] + 14\alpha^2\tau^2\frac{d_2^2}{NK} + \frac{52L_2^2\alpha^4\tau}{1-\rho^2}\right. \\
&\left.+4\alpha^2c_4^2\tau\sum_{s=0}^{\tau}E_{t-2\tau}[\Delta_{t-s}] + 3200\alpha^2c_1^2c_4^2\tau^3\Gamma^2(\epsilon,\epsilon_1) + 4\alpha^2c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1)\right\} \\
&+ \frac{9d_2^2}{NK}\alpha^2 + 36L_2^2\alpha^4 + 4\alpha^3L_1G^2 + 2\alpha^3L_2G + \left(\frac{4\alpha}{\xi_2} + 2\alpha^3L_1\right)\mathbb{E}_{t-2\tau}[\Delta_{t-\tau}] \\
&+ \underbrace{\alpha\left(\frac{4}{\xi_3} + 3\xi_1(c_1+c_2) + \frac{4}{\xi_2} + \alpha^2\left(24(c_1+c_2)^2 + 16\right)\right)}_{\mathcal{E}_3}\mathbb{E}_{t-2\tau}[\Delta_t] \\
&+ 2\alpha B(\epsilon,\epsilon_1)G + 4\alpha^2B^2(\epsilon,\epsilon_1) + 24\alpha^2(c_1+c_2)^2\Gamma^2(\epsilon,\epsilon_1) \\
&+ 3\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1) + 2\alpha^3L_1\Gamma(\epsilon,\epsilon_1)G
\end{aligned} \tag{4.69}
$$

If we choose step-size α such that $\alpha\mathcal{E}_2 = \alpha\left(\frac{c_1+c_2}{\xi_1} + \frac{1}{\alpha} + 24\xi_2 + \frac{4}{\xi_2}\right) \leq 2$, $\xi_1 = \xi_2 = \xi_3$, $\mathcal{E}_1 = \alpha\xi_3 + \alpha(3\xi_1(c_1+c_2) + 24\xi_2) + 24\alpha^2(c_1+c_2)^2 \leq 28\alpha\xi_1(c_1+c_2) + 24\alpha^2(c_1+c_2)^2 \leq 30\alpha\xi_1(c_1+c_2)$

$(c_1, c_2 > 1)$ and $\mathcal{E}_3 = \frac{4}{\xi_3} + 3\xi_1(c_1+c_2) + \frac{4}{\xi_2} + \alpha^2\left(24(c_1+c_2)^2+16\right) \leq (\frac{9}{\xi_1} + 9\xi_1)(c_1+c_2)$, i.e.,

$$\alpha \leq \frac{1}{\left(\frac{c_1+c_2}{\xi_1} + 24\xi_2 + \frac{4}{\xi_2}\right)} = \frac{\xi_1}{(c_1+c_2+24\xi_1^2+4)}$$
$$\alpha \leq \min\{\frac{\xi_1}{12(c_1+c_2)}, 1, \frac{(\frac{5}{\xi_1}+5\xi_1)(c_1+c_2)}{24(c_1+c_2)^2+16}\},$$

which is sufficient to hold when $\alpha \leq \min\{\frac{\xi_1}{24(c_1+c_2)^2+24\xi_1^2+16}, 1\}$, then we have:

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2 \leq \mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 + 2\alpha\mathbb{E}_{t-2\tau}\left\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t-\theta^*\right\rangle + 4\alpha^2\mathbb{E}_{t-2\tau}\left\|\bar{g}(\bar{\theta}_t)\right\|^2 \\
&+ 30\alpha\xi_1(c_1+c_2)\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 \\
&+ 2\left\{8\alpha^2\tau^2c_4^2\mathbb{E}_{t-2\tau}\left[\left\|\bar{\theta}_t-\theta^*\right\|^2\right] + 14\alpha^2\tau^2\frac{d_2^2}{NK} + \frac{52L_2^2\alpha^4\tau}{1-\rho^2}\right. \\
&\left.+4\alpha^2c_4^2\tau\sum_{s=0}^{\tau}E_{t-2\tau}[\Delta_{t-s}] + 3200\alpha^2c_1^2c_4^2\tau^3\Gamma^2(\epsilon,\epsilon_1) + 4\alpha^2c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1)\right\} \\
&+ \frac{9d_2^2}{NK}\alpha^2 + 36L_2^2\alpha^4 + 4\alpha^3L_1G^2 + 2\alpha^3L_2G + \left(\frac{4\alpha}{\xi_2} + 2\alpha^3L_1\right)\mathbb{E}_{t-2\tau}[\Delta_{t-\tau}] \\
&+ \alpha\left(\frac{4}{\xi_3} + 3\xi_1(c_1+c_2) + \frac{4}{\xi_2} + \alpha^2\left(24(c_1+c_2)^2+16\right)\right)\mathbb{E}_{t-2\tau}[\Delta_t] \\
&+ 2\alpha B(\epsilon,\epsilon_1)G + 4\alpha^2B^2(\epsilon,\epsilon_1) + 24\alpha^2(c_1+c_2)^2\Gamma^2(\epsilon,\epsilon_1) \\
&+ 3\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1) + 2\alpha^3L_1\Gamma(\epsilon,\epsilon_1)G \\
&\leq \mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 + 2\alpha\mathbb{E}_{t-2\tau}\left\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t-\theta^*\right\rangle + 4\alpha^2\mathbb{E}_{t-2\tau}\left\|\bar{g}(\bar{\theta}_t)\right\|^2 \\
&+ \left(30\alpha\xi_1(c_1+c_2) + 16\alpha^2\tau^2c_4^2\right)\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 \\
&+ \frac{9+28\tau^2}{NK}\alpha^2d_2^2 + 36\left(1+\frac{3\tau}{1-\rho^2}\right)L_2^2\alpha^4 + 4\alpha^3L_1G^2 + 2\alpha^3L_2G \\
&+ \left(\frac{4\alpha}{\xi_1} + 2\alpha^3L_1\right)\mathbb{E}_{t-2\tau}[\Delta_{t-\tau}] + \alpha(\frac{9}{\xi_1}+9\xi_1)(c_1+c_2)\mathbb{E}_{t-2\tau}[\Delta_t] + 8\alpha^2c_4^2\tau\sum_{s=0}^{\tau}E_{t-2\tau}[\Delta_{t-s}] \\
&+ 2\alpha B(\epsilon,\epsilon_1)G + 4\alpha^2B^2(\epsilon,\epsilon_1) + 24\alpha^2(c_1+c_2)^2\Gamma^2(\epsilon,\epsilon_1) \\
&+ 3\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1) + 2\alpha^3L_1\Gamma(\epsilon,\epsilon_1)G \\
&+ 6400\alpha^2c_1^2c_4^2\tau^3\Gamma^2(\epsilon,\epsilon_1) + 8\alpha^2c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1) \qquad (4.70)
\end{aligned}$$

if we choose the step-size α such that the high order $O(\alpha^2)$ terms are dominanted by the first order terms $O(\alpha)$, i.e., $4\alpha^2B^2(\epsilon,\epsilon_1)+24\alpha^2(c_1+c_2)^2\Gamma^2(\epsilon,\epsilon_1)+2\alpha^3L_1\Gamma(\epsilon,\epsilon_1)G+6400\alpha^2c_1^2c_4^2\tau^3\Gamma^2(\epsilon,\epsilon_1)+8\alpha^2c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1)\leq 2\alpha B(\epsilon,\epsilon_1)G+3\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1)$, i.e.,

$$\alpha\leq\min\{\frac{2B(\epsilon,\epsilon_1)G+3\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1)}{4B^2(\epsilon,\epsilon_1)+24(c_1+c_2)^2\Gamma^2(\epsilon,\epsilon_1)+2L_1\Gamma(\epsilon,\epsilon_1)G+6400c_1^2c_4^2\tau^3\Gamma^2(\epsilon,\epsilon_1)+8c_1^2\tau^2\Gamma^2(\epsilon,\epsilon_1)},1\}$$

we have:

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2\leq\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+2\alpha\mathbb{E}_{t-2\tau}\left\langle\bar{g}(\bar{\theta}_t),\bar{\theta}_t-\theta^*\right\rangle+4\alpha^2\mathbb{E}_{t-2\tau}\left\|\bar{g}(\bar{\theta}_t)\right\|^2\\
&+\left(30\alpha\xi_1(c_1+c_2)+16\alpha^2\tau^2c_4^2\right)\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2\\
&+\frac{9+28\tau^2}{NK}\alpha^2d_2^2+36\left(1+\frac{3\tau}{1-\rho^2}\right)L_2^2\alpha^4+4\alpha^3L_1G^2+2\alpha^3L_2G\\
&+\left(\frac{4\alpha}{\xi_1}+2\alpha^3L_1\right)\mathbb{E}_{t-2\tau}[\Delta_{t-\tau}]+\alpha(\frac{9}{\xi_1}+9\xi_1)(c_1+c_2)\mathbb{E}_{t-2\tau}[\Delta_t]+8\alpha^2c_4^2\tau\sum_{s=0}^{\tau}E_{t-2\tau}[\Delta_{t-s}]\\
&+4\alpha B(\epsilon,\epsilon_1)G+6\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1)
\end{aligned}\tag{4.71}$$

With Lemma (15), we have the upper bound of $\mathbb{E}_{t-2\tau}[\Delta_t]$, $\mathbb{E}_{t-2\tau}[\Delta_{t-\tau}]$ and $\tau\sum_{s=0}^{\tau}E_{t-2\tau}[\Delta_{t-s}]$. Then we have:

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2\leq\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+2\alpha\mathbb{E}_{t-2\tau}\left\langle\bar{g}(\bar{\theta}_t),\bar{\theta}_t-\theta^*\right\rangle+4\alpha^2\mathbb{E}_{t-2\tau}\left\|\bar{g}(\bar{\theta}_t)\right\|^2\\
&\quad+\underbrace{\left(30\alpha\xi_1(c_1+c_2)+16\alpha^2\tau^2c_4^2\right)}_{\mathcal{E}_4}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2\\
&\quad+\frac{9+28\tau^2}{NK}\alpha^2d_2^2+\alpha^3\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)\\
&\quad+\frac{4\alpha^2}{K\alpha_g^2}\underbrace{\left(\frac{4\alpha}{\xi_1}+2\alpha^3L_1+\alpha(\frac{9}{\xi_1}+9\xi_1)(c_1+c_2)+8\alpha^2c_4^2\tau^2\right)}_{\mathcal{E}_5}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+4c_1^2(K-1)H^2\right]\\
&\quad+4\alpha B(\epsilon,\epsilon_1)G+6\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1)
\end{aligned}\tag{4.72}$$

If we choose step-size such that $\mathcal{E}_4=30\alpha\xi_1(c_1+c_2)+16\alpha^2\tau^2c_4^2\leq 32\alpha\xi_1(c_1+c_2)$ and $\mathcal{E}_5=$

$\frac{4\alpha}{\xi_1} + 2\alpha^3 L_1 + \alpha(\frac{9}{\xi_1} + 9\xi_1)(c_1+c_2) + 8\alpha^2 c_4^2\tau^2 \leq \alpha(\frac{14}{\xi_1} + 14\xi_1)(c_1+c_2)$, i.e.,

$$\alpha \leq \min\{\frac{\xi_1(c_1+c_2)}{8\tau^2 c_4^2}, 1, \frac{(\frac{1}{\xi_1}+\xi_1)(c_1+c_2)}{2L_1 + 8c_4^2\tau^2}\},$$

which is sufficient to hold when $\alpha \leq \frac{\xi_1(c_1+c_2)}{2L_1+8\tau^2 c_4^2}$, then we have:

$$\begin{aligned} E_{t-2\tau}\left\|\bar{\theta}_{t+1} - \theta^*\right\|^2 \leq\ & \mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t - \theta^*\right\|^2 + 2\alpha\mathbb{E}_{t-2\tau}\left\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\right\rangle + 4\alpha^2\mathbb{E}_{t-2\tau}\left\|\bar{g}(\bar{\theta}_t)\right\|^2 \\ & + 32\alpha\xi_1(c_1+c_2)\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t - \theta^*\right\|^2 \\ & + \frac{9+28\tau^2}{NK}\alpha^2 d_2^2 + \alpha^3\left(36L_2^2 + \frac{108\tau}{1-\rho^2}L_2^2 + 4L_1G^2 + 2L_2G\right) \\ & + \frac{4\alpha^3}{K\alpha_g^2}(\frac{14}{\xi_1} + 14\xi_1)(c_1+c_2)\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right] \\ & + 4\alpha B(\epsilon,\epsilon_1)G + 6\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1). \end{aligned} \tag{4.73}$$

□

• Parameter Selection

With Lemma 16, we have:

$$\begin{aligned} \mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t+1} - \theta^*\right\|^2 \leq\ & (1 + 32\alpha\xi_1(c_1+c_2))\,\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t - \theta^*\right\|^2 + 2\alpha\mathbb{E}_{t-2\tau}\left\langle \bar{g}(\bar{\theta}_t), \bar{\theta}_t - \theta^*\right\rangle + 4\alpha^2\mathbb{E}_{t-2\tau}\left\|\bar{g}(\bar{\theta}_t)\right\|^2 \\ & + \frac{9+28\tau^2}{NK}\alpha^2 d_2^2 + \alpha^3\left(36L_2^2 + \frac{108\tau}{1-\rho^2}L_2^2 + 4L_1G^2 + 2L_2G\right) \\ & + \frac{4\alpha^3}{K\alpha_g^2}(\frac{14}{\xi_1} + 14\xi_1)(c_1+c_2)\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right] \\ & + 4\alpha B(\epsilon,\epsilon_1)G + 6\alpha\xi_1(c_1+c_2)\Gamma^2(\epsilon,\epsilon_1). \end{aligned} \tag{4.74}$$

Proposition 4. *If α satisfies the requirement as Lemma 16, choose $\xi_1 = \frac{(1-\gamma)\bar{\omega}}{32(c_1+c_2)}$ and $\tau = \lceil\frac{\tau^{\text{mix}}(\alpha_T^2)}{K}\rceil$, we have:*

$$\nu_1\mathbb{E}_{t-2\tau}\left\|V_{\bar{\theta}_t} - V_{\theta^*}\right\|_{\bar{D}}^2 \leq (\frac{1}{\alpha} - \nu_1)\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t - \theta^*\right\|^2 - \frac{1}{\alpha}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t+1} - \theta^*\right\|^2 + \frac{9+28\tau^2}{NK}\alpha d_2^2$$

$$+\alpha^2\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)$$
$$+\frac{\alpha^2c_6}{K}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right]+4B(\epsilon,\epsilon_1)G+\nu_1\Gamma^2(\epsilon,\epsilon_1) \tag{4.75}$$

where $\nu_1=\frac{\nu}{4}=\frac{(1-\gamma)\bar{\omega}}{4}$ *and* $c_6\triangleq\frac{4}{\alpha_g^2}(\frac{14}{\xi_1}+14\xi_1)(c_1+c_2)$.

Proof. Incorporating $\xi_1=\frac{(1-\gamma)\bar{\omega}}{32(c_1+c_2)}$, $c_6\triangleq\frac{4}{\alpha_g^2}(\frac{14}{\xi_1}+14\xi_1)(c_1+c_2)$ and $6\xi_1(c_1+c_2)\leq\nu_1$ into Eq (4.74), we have

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2\leq\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+2\alpha\mathbb{E}_{t-2\tau}\left\langle\bar{g}(\bar{\theta}_t),\bar{\theta}_t-\theta^*\right\rangle+4\alpha^2\mathbb{E}_{t-2\tau}\left\|\bar{g}(\bar{\theta}_t)\right\|^2\\
&+\alpha(1-\gamma)\bar{\omega}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2\\
&+\frac{9+28\tau^2}{NK}\alpha^2d_2^2+\alpha^3\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)\\
&+\frac{\alpha^3c_6}{K}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right]\\
&+4\alpha B(\epsilon,\epsilon_1)G+\alpha\nu_1\Gamma^2(\epsilon,\epsilon_1)\\
&\overset{(a)}{\leq}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2-2\alpha(1-\gamma)\bar{\omega}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+16\alpha^2\mathbb{E}_{t-2\tau}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2\\
&+\alpha(1-\gamma)\bar{\omega}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2\\
&+\frac{9+28\tau^2}{NK}\alpha^2d_2^2+\alpha^3\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)\\
&+\frac{\alpha^3c_6}{K}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right]\\
&+4\alpha B(\epsilon,\epsilon_1)G+\alpha\nu_1\Gamma^2(\epsilon,\epsilon_1)\\
&=\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2-\frac{\alpha(1-\gamma)\bar{\omega}}{2}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2\\
&-\frac{\alpha(1-\gamma)\bar{\omega}}{2}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2+16\alpha^2\mathbb{E}_{t-2\tau}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2\\
&+\frac{9+28\tau^2}{NK}\alpha^2d_2^2+\alpha^3\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)\\
&+\frac{\alpha^3c_6}{K}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right]
\end{aligned}\tag{4.76}$$

$$
\begin{aligned}
&+ 4\alpha B(\epsilon,\epsilon_1)G + \alpha\nu_1\Gamma^2(\epsilon,\epsilon_1)\\
&\leq \mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 - \frac{\alpha(1-\gamma)\bar{\omega}}{2}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2\\
&- \frac{\alpha(1-\gamma)\bar{\omega}}{2}\mathbb{E}_{t-2\tau}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2 + 16\alpha^2\mathbb{E}_{t-2\tau}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2\\
&+ \frac{9+28\tau^2}{NK}\alpha^2 d_2^2 + \alpha^3\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)\\
&+ \frac{\alpha^3 c_6}{K}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right]\\
&+ 4\alpha B(\epsilon,\epsilon_1)G + \alpha\nu_1\Gamma^2(\epsilon,\epsilon_1)\\
&\overset{(b)}{\leq} \mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 - \frac{\alpha(1-\gamma)\bar{\omega}}{2}\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 - \frac{\alpha(1-\gamma)\bar{\omega}}{4}\mathbb{E}_{t-2\tau}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2\\
&+ \frac{9+28\tau^2}{NK}\alpha^2 d_2^2 + \alpha^3\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)\\
&+ \frac{\alpha^3 c_6}{K}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right]\\
&+ 4\alpha B(\epsilon,\epsilon_1)G + \alpha\nu_1\Gamma^2(\epsilon,\epsilon_1)\\
&\leq (1-2\alpha\nu_1)\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 - \alpha\nu_1\mathbb{E}_{t-2\tau}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2 + \frac{9+28\tau^2}{NK}\alpha^2 d_2^2\\
&+ \alpha^3\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)\\
&+ \frac{\alpha^3 c_6}{K}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right] + 4\alpha B(\epsilon,\epsilon_1)G + \alpha\nu_1\Gamma^2(\epsilon,\epsilon_1)
\end{aligned}
\tag{4.77}
$$

where (a) is due to Lemma 3 and the selection of parameter; (b) is due to $16\alpha^2 \leq \frac{\alpha(1-\gamma)\bar{\omega}}{4}$. Rearranging the terms and using the fact $1-2\alpha\nu_1 \leq 1-\alpha\nu_1$, we have:

$$
\begin{aligned}
\alpha\nu_1\mathbb{E}_{t-2\tau}\left\|V_{\bar{\theta}_t}-V_{\theta^*}\right\|_{\bar{D}}^2 \leq{}& (1-\alpha\nu_1)\mathbb{E}_{t-2\tau}\left\|\bar{\theta}_t-\theta^*\right\|^2 - \mathbb{E}_{t-2\tau}\left\|\bar{\theta}_{t+1}-\theta^*\right\|^2 + \frac{9+28\tau^2}{NK}\alpha^2 d_2^2\\
&+ \alpha^3\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)\\
&+ \frac{\alpha^3 c_6}{K}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right] + 4\alpha B(\epsilon,\epsilon_1)G + \alpha\nu_1\Gamma^2(\epsilon,\epsilon_1)
\end{aligned}
\tag{4.78}
$$

Then we finish the proof by dividing α on both sides. □

With these Lemmas, we are now ready to prove Theorem 4.

b) Proof of Theorem 4.

Given a fixed local step-size $\alpha_l \leq \frac{1}{4\sqrt{2}c_1(K-1)}$, decreasing effective step-sizes $\alpha_t = \frac{8}{\nu(a+t+1)} = \frac{8}{(1-\gamma)\bar{\omega}(a+t+1)}$, decreasing global step-sizes $\alpha_g^{(t)} = \frac{\alpha_t}{K\alpha_l}$ and weights $w_t = (a+t)$, we have:

$$\mathbb{E}\left\|V_{\tilde{\theta}_T} - V_{\theta_i^*}\right\|_{\bar{D}}^2 \leq \tilde{\mathcal{O}}\left(\frac{\tau^2 G^2}{K^2T^2} + \frac{c_{\text{quad}}(\tau)}{\nu^2 NKT} + \frac{c_{\text{lin}}(\tau)}{\nu^4 KT^2} + \frac{B(\epsilon,\epsilon_1)G}{\nu} + \Gamma^2(\epsilon,\epsilon_1)\right) \tag{4.79}$$

Proof. We take the step-size $\alpha_t = \frac{8}{\nu(a+t+1)} = \frac{2}{\nu_1(a+t+1)}$ for $a > 0$. In addition, we define weights $w_t = (a+t)$ and define

$$\tilde{\theta}_T = \frac{1}{W}\sum_{t=1}^{T} w_t \bar{\theta}_t$$

where $W = \sum_{t=1}^{T} w_t \geq \frac{1}{2}T(a+T)$. By convexity of positive definite quadratic forms ($\lambda_{\min}(\Phi^T \bar{D}\Phi) \geq \bar{\omega} > 0$), we have

$$\begin{aligned}
&\nu_1\mathbb{E}\left\|V_{\tilde{\theta}_T} - V_{\theta^*}\right\|_{\bar{D}}^2 \leq \frac{\nu_1}{W}\sum_{t=1}^{T}(a+t)\mathbb{E}\left\|V_{\bar{\theta}_t} - V_{\theta^*}\right\|_{\bar{D}}^2 \\
&\leq \frac{\nu_1}{W}\sum_{t=1}^{2\tau-1}(a+t)\mathbb{E}\left\|V_{\bar{\theta}_t} - V_{\theta^*}\right\|_{\bar{D}}^2 + \frac{\nu_1}{W}\sum_{t=2\tau}^{T}(a+t)\mathbb{E}\left\|V_{\bar{\theta}_t} - V_{\theta^*}\right\|_{\bar{D}}^2 \\
&\leq \nu_1\frac{(2\tau-1)(a+2\tau-1)G^2}{W} + \frac{\nu_1}{W}\sum_{t=2\tau}^{T}(a+t)\mathbb{E}\left\|V_{\bar{\theta}_t} - V_{\theta^*}\right\|_{\bar{D}}^2 \\
&\overset{(4.75)}{\leq} \nu_1\frac{(2\tau-1)(a+2\tau-1)G^2}{W} + \frac{\nu_1(a+2\tau)(a+2\tau+1)G^2}{2W} \\
&+ \frac{1}{W}\sum_{t=2\tau}^{T}\left[\frac{(9+28\tau^2)d_2^2}{NK}(a+t)\alpha_t + (a+t)\alpha_t^2\left(36L_2^2 + \frac{108\tau}{1-\rho^2}L_2^2 + 4L_1G^2 + 2L_2G\right)\right] \\
&+ \frac{1}{W}\sum_{t=2\tau}^{T}\frac{(a+t)\alpha^2 c_6}{K}\left[c_3^2 + \frac{2c_3L_2\rho}{1-\rho} + 8c_1^2(K-1)H^2\right]
\end{aligned}$$

$$+\frac{1}{W}\sum_{t=2\tau}^{T}\left[4(a+t)B(\epsilon,\epsilon_1)G+(a+t)\nu_1\Gamma^2(\epsilon,\epsilon_1)\right] \tag{4.80}$$

where $\left\|V_{\bar{\theta}_{2\tau}}-V_{\theta^*}\right\|_{\bar{D}}^2 \leq G^2$. We know that $\frac{1}{W}\sum_{t=2\tau}^{T}(a+t)\alpha_t^2 \leq \frac{1}{W}\sum_{t=1}^{T}(a+t)\frac{4}{\nu_1^2(a+t)^2} \leq \frac{8\log(a+T)}{\nu_1^2T^2}$ and that $\frac{1}{W}\sum_{t=2\tau}^{T}(a+t)\alpha_t \leq \frac{4}{\nu_1 T}$. Plugging in these inequalities into Eq (4.80), we have:

$$\begin{aligned}
&\nu_1\mathbb{E}\left\|V_{\bar{\theta}}-V_{\theta^*}\right\|_{\bar{D}}^2 \leq \frac{3\nu_1(a+2\tau)(a+2\tau+1)G^2}{2W}+\frac{4(9+28\tau^2)d_2^2}{\nu_1 NKT}\\
&+\frac{8\log(a+T)}{\nu_1^2T^2}\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)\\
&+\frac{8c_6\log(a+T)}{\nu_1^2T^2K}\left[c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right]\\
&+4B(\epsilon,\epsilon_1)G+\nu_1\Gamma^2(\epsilon,\epsilon_1)\\
&=\frac{3\nu_1(a+2\tau)(a+2\tau+1)G^2}{2W}+\frac{4(9+28\tau^2)d_2^2}{\nu_1 NKT}\\
&+\frac{8\log(a+T)}{\nu_1^2T^2K}\underbrace{\left[K\left(36L_2^2+\frac{108\tau}{1-\rho^2}L_2^2+4L_1G^2+2L_2G\right)+c_6\left(c_3^2+\frac{2c_3L_2\rho}{1-\rho}+8c_1^2(K-1)H^2\right)\right]}_{c_{\text{lin}}(\tau)}\\
&+4B(\epsilon,\epsilon_1)G+\nu_1\Gamma^2(\epsilon,\epsilon_1)
\end{aligned} \tag{4.81}$$

where $c_{\text{quad}}(\tau)=4d_2^2(9+28\tau^2)$. Dividing ν_1 on the both sides, changing ν_1 into ν ($\nu=(1-\gamma)\bar{\omega}$) and noting that $c_6=\frac{4}{\alpha_g^2}(\frac{14}{\xi_1}+14\xi_1)(c_1+c_2)=\mathcal{O}(\frac{1}{\nu})$, we have:

$$\mathbb{E}\left\|V_{\tilde{\theta}_T}-V_{\theta^*}\right\|_{\bar{D}}^2 \leq \tilde{\mathcal{O}}\left(\frac{\tau^2G^2}{K^2T^2}+\frac{c_{\text{quad}}(\tau)}{\nu^2NKT}+\frac{c_{\text{lin}}(\tau)}{\nu^4KT^2}+\frac{B(\epsilon,\epsilon_1)G}{\nu}+\Gamma^2(\epsilon,\epsilon_1)\right). \tag{4.82}$$

We finish the proof by using the inequality, $\mathbb{E}\left\|V_{\tilde{\theta}_T}-V_{\theta_i^*}\right\|_{\bar{D}}^2 \leq 2\mathbb{E}\left\|V_{\tilde{\theta}_T}-V_{\theta^*}\right\|_{\bar{D}}^2+2\mathbb{E}\left\|V_{\theta_i^*}-V_{\theta^*}\right\|_{\bar{D}}^2$ and combining with the third point in Theorem 1. □

4.8.11 Additional Simulation Results

a) Simulation results for the I.I.D. setting

In this subsection, we provide numerical results for `FedTD`(0) under the i.i.d. sampling setting to verify the theoretical results of Theorem 2. In particular, the MDP $\mathcal{M}^{(1)}$ of the first agent is randomly generated with a state space of size $n = 100$. The remaining MDPs are perturbations of $\mathcal{M}^{(1)}$ with the heterogeneity levels $\epsilon = 0.1$ and $\epsilon_1 = 0.1$. The number of local steps is chosen as $K = 20$. We evaluate the convergence in terms of the running error $e_t = \|\bar{\theta}_t - \theta_1^*\|^2$. Each experiment is run 10 times. We plot the mean and standard deviation across the 10 runs in Figure 4.2.

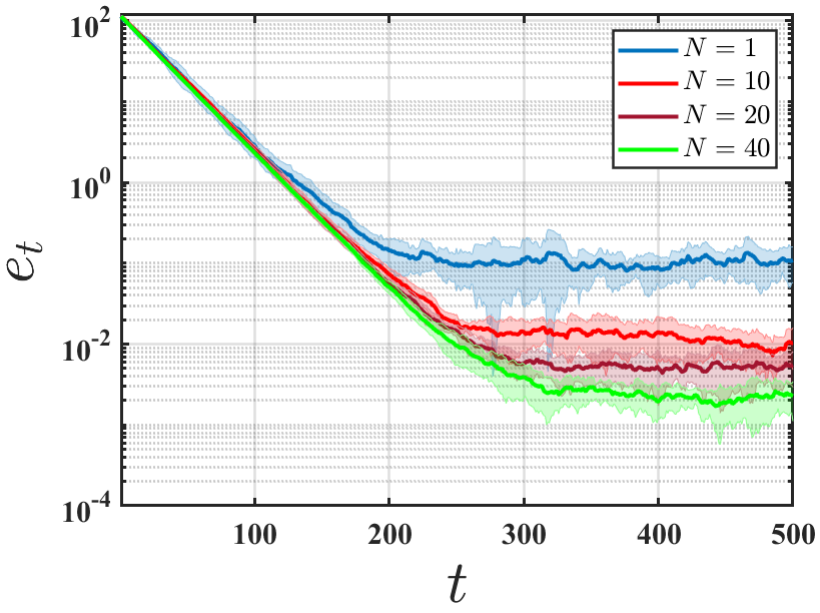


Figure 4.2: Performance of `FedTD`(0) with i.i.d. sampling with varying number of agents N. Solid lines denote the mean and shaded regions indicate the standard deviation over ten runs.

As shown in Fig 4.2, `FedTD`(0) converges for all choices of N. Larger values of N decreases the error, which is consistent with our theoretical analysis in Theorem 2.

b) Simulation results for the Markovian setting

In this subsection, we provide numerical results for `FedTD`(0) under the Markovian sampling setting to verify the theoretical results of Theorem 4. Here we generate all MDPs in the same way as the i.i.d setting and choose the number of local steps as $K = 20$. All the remaining parameters are kept the same as those in the subsection a).

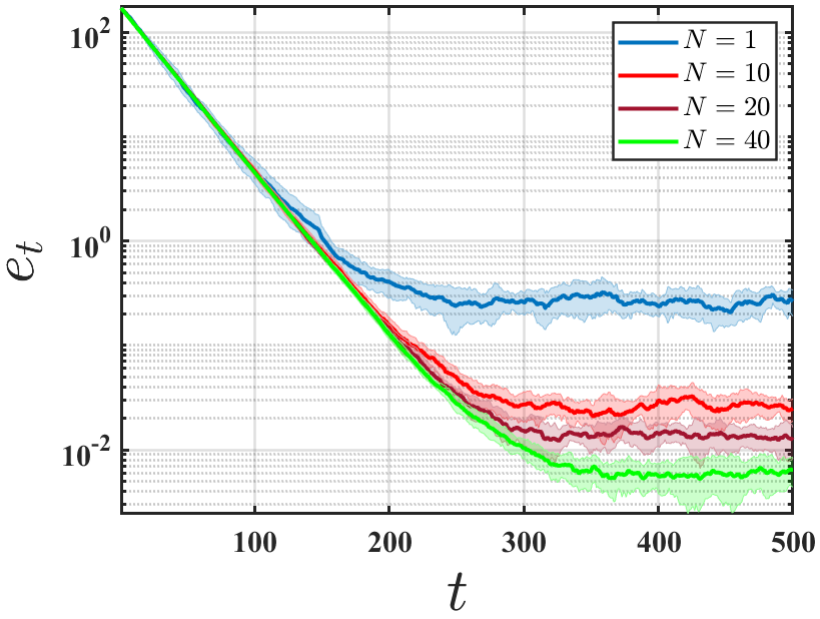


Figure 4.3: Performance of `FedTD`(0) with the Markovian sampling with varying number of agents N. Solid lines denote the mean and shaded regions indicate the standard deviation over ten runs.

As shown in Fig 4.3, `FedTD`(0) converges for all choices of N. Larger values of N decreases the error, which is consistent with our theoretical analysis in Theorem 4.

Chapter 5

Federated Learning for Policy Optimization

5.1 Introduction

Recently, there has been a lot of interest applying Federated Learning (FL) algorithms to reinforcement learning (RL) problems in order to solve complex sequential decision-making tasks [26, 79, 120, 158, 222]. Federated reinforcement learning (FRL) has been widely applied as it provides the following advantages: First, FRL protects each agent's privacy by only allowing the model to be shared between the server and agent, while keeping the raw data localized. Secondly, by sharing the model with the server, FRL can reduce the sample complexity and produce a better policy than if each agent learns individually with its own limited data. However, existing work in the FRL framework is limited to either multiple agents interacting with the same environment [48, 92] *or* multiple agents with distinct, *yet similar* environments [79, 192, 228]. It remains an open problem to formally characterize how FRL performs when multiple agents from completely different environments, i.e., with arbitrarily large heterogeneity levels, are allowed to collaborate. In this chapter, we provide an answer to the following question: *what is the best achievable sample complexity when considering severely heterogeneous environments*?

We focus on developing FRL algorithms that compute an optimal universal policy that ensures uniformly good performance for N agents, despite their operation in disparate environments. The motivation for a shared policy stems from practical applications necessitating uniform approaches for distinct agents. For instance, Spotify, a leading audio streaming company, intends to design a uniform pricing plan that suits the listening habits of all users. Given the substantial variations in listening habits among users, establishing a pricing strategy that aligns with the preferences of all users is of great importance. Similarly, autonomous vehicles navigating diverse settings like urban streets, rural areas, and highways must adapt to varied challenges. A uniform policy that adjusts to this environmental heterogeneity ensures consistent, safe decision-making across all terrains, highlighting the need for robust algorithms capable of handling dynamic driving conditions efficiently. Moreover, a universally optimal policy could serve as a foundational model that can be individually fine-tuned, a concept that has gained a lot of attention in meta- and few-shot RL research [52, 165, 238]. This approach underscores the broader necessity of designing a uniform and adaptable policy for heterogeneous settings.

In this chapter, the environment heterogeneity refers to the fact that each agent has a different reward function, state transition kernel, or initial state distribution, while they share common state and action spaces. Notably, compared with the existing work [79, 192], we do not assume that all the environments are similar, i.e., environmental heterogeneity does not need to be bounded by small constants. Instead, we consider a more general setting where the magnitude of heterogeneity can be arbitrary. With this setup, we aim to answer the following question:

> *Is it possible to design a provably efficient FRL algorithm which can accommodate arbitrary levels of environmental heterogeneity among agents?*

We answer this question affirmatively. Our main contributions are listed below.

• **New momentum-powered federated reinforcement learning algorithms:** We propose two new algorithms FEDSVRPG-M and FEDHAPG-M for solving heterogeneous FRL problems (formally specified in Eq. (5.3)). Leveraging momentum, we prove that our algorithms, even with constant local step-sizes, converge to the exact stationary point of the heterogeneous FRL problem, *regardless of the magnitude of environment heterogeneity*. This stands in contrast to the state-of-the-art work, which only show convergence to a ball around the stationary point whose radius depends on the environmental heterogeneity levels. Importantly, our results hold even when different notions of environment heterogeneity are considered such as the heterogeneity in Markov decision processes (MDPs) or policy advantage heterogeneity [228].

• **State-of-the-art convergence rates:** By integrating variance-reduction techniques and curvature information into the policy gradient estimation, our algorithms achieve sample-efficiency improvement over prior work [48]. In particular, we reduce the sample complexity from $\mathcal{O}\left(\epsilon^{-\frac{5}{3}}/N^{\frac{2}{3}}\right)$ to $\mathcal{O}\left(\epsilon^{-\frac{3}{2}}/N\right)$ when finding the ϵ-approximate first order stationary point[1] (ϵ-FOSP) [143]. When only a single agent is included, i.e., $N = 1$, our results align with the best known sample complexity of $\mathcal{O}\left(\varepsilon^{-\frac{3}{2}}\right)$ from [49].

• **Practical algorithm structures:** Our algorithms are easy to implement because: (1) *Constant local step-sizes*. This feature reduces the amount of algorithm tuning. In contrast, many FL optimization algorithms [84, 218, 235] require diminishing local step-sizes preset according to complex schedules in order to counteract the effects of heterogeneity. (2) *Sampling one trajectory per local iteration*. This means our algorithms can address the challenge of poor sample efficiency in RL. Unlike existing variance-reduced policy gradient (PG) algorithms for the single agent setting [58, 147, 231], our approach avoids the need for large batch sizes during certain iterations.

[1]Finding a parameter θ such that $\|\nabla J(\theta)\|^2 \leq \epsilon$, where J is defined in Eq. (5.3). Note that in work such as [49, 175], the notion $\|\nabla J(\theta)\|^2 \leq \epsilon^2$ is applied instead.

Table 5.1: Comparision of the results for policy-based methods in FRL. LU and HETER denote the multiple local updates and environment heterogeneity, respectively.

ALGORITHM	CONVERGENCE	SPEEDUP	LU	HETER
PAVG [79]	finite but inexact	No speedup	✓	✓
FEDKL [228]	asymptotic	No speedup	✗	✓
FEDPG-BR [48]	finite and exact	Sublinear: $N^{\frac{2}{3}}$	✗	✗
FAPI [227]	asymptotic and inexact	No speedup	✗	✓
FEDSVRPG-M (Ours)	finite and exact	Linear: N	✓	✓
FEDHAPG-M (Ours)	finite and exact	Linear: N	✓	✓

(3) *Accommodating multiple local updates*. With this feature, our algorithms become more suitable for real-world applications, where communication latency causes serious bottlenecks.

• **Linear speedup:** Analysis of FEDSVRPG-M and FEDHAPG-M shows that they can converge N-times faster than the scenario where each agent learns a policy on its own. Essentially, by adopting the FL approach, the sample complexity of our algorithms can be linearly scaled by the number of agents N, i.e., collaboration always helps. To our knowledge, *we are the first to achieve a linear speedup for finding a stationary point of FRL problems using policy-based methods*. Importantly, the linear speedup is established even when considering multiple local updates and without making any assumptions about environment heterogeneity. Compared to prior work, our result outperforms that of [48, 79], which at best achieves sublinear speedup, see Table 5.1.

Refer to [210] for all proofs in this Chapter.

5.2 Background and Preliminaries

5.2.1 Relative Work

Federated RL A comprehensive overview of techniques and open problems in FRL was offered by [158]. Much of the work in FRL has focused on developing federated versions of value-based methods [92, 192, 224]. Notably, [92] and [224] established the benefits of FL in terms of linear speedup, assuming all agents operate in *identical* environment. Wang et al. [192] introduced the FEDTD(0) algorithm to address the FRL problem with *distinct yet similar* environments demonstrated linear speed up was achievable. On the other hand, [243] proposed the FEDSARSA algorithm to solve the on-policy FRL problem, but it is applicable only in similar environments. Another major area of FRL research studies federated policy-based algorithms [48, 79, 103, 105, 193, 228]. However, [48] only consider uniform environments and only one local update step. While [228] explored diverse environments, they only showed an asymptotic convergence. Most relevant to our work, [79] studied *heterogeneous* environments. Nevertheless, the algorithms from [79] were saddled with a non-vanishing convergence error. This non-vanishing error depended on the environmental heterogeneity levels. Note that none of these papers investigated the FRL problems with *arbitrary environment heterogeneity*. To bridge this gap, our proposed algorithms, FEDSVRPG-M and FEDHAPG-M, utilize policy-based techniques and can converge *exactly*. See Table 5.1 for a comparison of our results with the existing work in FRL policy-based methods.

Federated Learning. Federated learning (FL) is a machine learning approach where a model is trained across multiple clients. Each client runs several iterations of a learning algorithm on its own dataset. Periodically, clients send their local models to the server. The server aggregates the models and then broadcasts the resulting model to all clients and the process repeats. By performing multiple local updates with its own data, FL can substantially reduce communication costs. Our proposed

algorithms align with the structure of standard FL algorithms such as FEDAVG [132]: an agent performs multiple local updates (using SGD) between two communication rounds. Nonetheless, such local updates will introduce *"client-drift"* problems [24, 84, 211], presenting a key challenge in FL regarding the trade-off between communication cost and model accuracy. Additionally, handling data that is not identically distributed across devices, affecting both data modeling and convergence analysis, presents another challenge. These challenges are further amplified in the context of FRL.

5.2.2 Centralized Reinforcement Learning

A centralized reinforcement learning task[2] is generally modeled as a discrete-time Markov Decision Process (MDP): $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space and ρ denotes the initial state distribution. Here, $\mathcal{P}\left(s' \mid s, a\right)$ denotes the probability that the agent transitions from the state s to s' when taking the action $a \in \mathcal{A}$. The discount factor is $\gamma \in (0, 1)$, and $\mathcal{R}(s, a) : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the reward function for taking action a at state s for some constant $R_{\max} > 0$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a mapping from the state space $\mathcal{S}$ to the probability distribution over the action space $\mathcal{A}$.

Under any stationary policy, the agent can collect a trajectory

$$\tau \triangleq \left\{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}, s_H\right\},$$

which is the collection of state-action pairs, where H is the maximum length of all trajectories. Once a trajectory τ is obtained, a cumulative discounted reward can be observed; $\mathcal{R}(\tau) \triangleq \sum_{h=0}^{H-1} \gamma^h \mathcal{R}\left(s_h, a_h\right)$.

[2]To distinguish from the federated setting, we refer to the single-agent case as centralized RL or when it's clear from context, simply reinforcement learning.

5.2.3 Policy Gradients

Given finite state and action spaces, the policy $\pi(a|s)$ can be stored in a $|\mathcal{S}|\times|\mathcal{A}|$ table. However, in practice, both the state and action spaces are large and the tabular approach becomes intractable. Alternatively, the policy is parameterized by an unknown parameter $\theta \in \mathbb{R}^d$, the resulting policy is denoted by π_θ. Given the initial distribution ρ, $p(\tau \mid \theta)$ denotes the probability distribution over trajectory τ, which can be calculated as

$$p(\tau \mid \theta) = \rho\left(s_0\right) \prod_{h=0}^{H-1} \pi_\theta\left(a_h \mid s_h\right) \mathcal{P}\left(s_{h+1} \mid s_h, a_h\right).$$

The goal of RL is to find the optimal policy parameter θ that maximizes the expected discounted trajectory reward:

$$\max_{\theta\in\mathbb{R}^d} J(\theta) \triangleq \mathbb{E}_{\tau\sim p(\tau|\theta)}[\mathcal{R}(\tau)] = \int \mathcal{R}(\tau)p(\tau \mid \theta)d\tau. \tag{5.1}$$

Note that the underlying distribution p in Eq. (5.1) depends on the variable θ which varies through the whole optimization procedure. This property, referred to as *non-obliviousness*, highlights a unique challenge in RL and creates a notable distinction from supervised learning problems, where the distribution p is stationary.

To deal with the *non-oblivious* and *non-convex* problem (5.1), a standard approach is to use the policy gradient (PG) method [189, 223]. PG takes the first-order derivative of the objective (5.1) where $\nabla J(\theta)$ can be expressed as

$$\int \mathcal{R}(\tau)\nabla p(\tau \mid \theta)d\tau = \mathbb{E}_{\tau\sim p(\tau|\theta)}[\nabla \log p(\tau \mid \theta)\mathcal{R}(\tau)].$$

Then, the policy θ can be optimized by running gradient ascent-based algorithms. However, since the distribution $p(\tau \mid \theta)$ is unknown, it is impossible to calculate the full gradient. To address this issue, stochastic gradient ascent is typically used, producing a sequence of the form:

$$\theta \leftarrow \theta + \eta \cdot \frac{1}{B}\sum_{i=1}^{B} g(\tau_i \mid \theta)$$

where $\eta > 0$ denotes the stepsize, B is the number of trajectories, and $g(\tau_i \mid \theta)$ is an estimate of the full gradient $\nabla J(\theta)$ using the trajectory τ_i. The most common unbiased estimators of PG are REINFORCE [223] and GPOMDP [10]. In this paper, $g(\tau \mid \theta)$ is defined as

$$g(\tau \mid \theta) = \sum_{t=0}^{H-1} \left(\sum_{h=t}^{H-1} \gamma^h \mathcal{R}\left(s_h, a_h\right) \right) \nabla \log \pi_\theta \left(a_t \mid s_t\right).$$

Importance Sampling Since problem 5.1 is *non-oblivious*, we have

$$\mathbb{E}_{\tau \sim p(\tau|\theta)} \left[g(\tau \mid \theta) - g\left(\tau \mid \theta'\right) \right] \neq \nabla J(\theta) - \nabla J\left(\theta'\right)$$

. To address this issue of distribution shift, we introduce an importance sampling (IS) weight, denoted by

$$w\left(\tau \mid \theta', \theta\right) \triangleq \frac{p\left(\tau \mid \theta'\right)}{p(\tau \mid \theta)} = \prod_{h=0}^{H-1} \frac{\pi_{\theta'}\left(a_h \mid s_h\right)}{\pi_\theta\left(a_h \mid s_h\right)}. \tag{5.2}$$

With the definition of the IS weight, we can ensure that

$$\mathbb{E}_{\tau \sim p(\tau|\theta)} \left[g(\tau \mid \theta) - w\left(\tau \mid \theta', \theta\right) g\left(\tau \mid \theta'\right) \right] = \nabla J(\theta) - \nabla J\left(\theta'\right).$$

5.3 Problem Formulation

We are now ready to characterize heterogeneity in our N-agent FRL problem. Environmental heterogeneity is modeled by allowing each agent to have its own state transition kernel $\mathcal{P}^{(i)}$, reward function $\mathcal{R}^{(i)}$, or the initial state distribution $\rho^{(i)}$. However, all agents share the same state and action space. These environments are characterized by the MDPs, $\mathcal{M}_i = \left\langle \mathcal{S}, \mathcal{A}, \mathcal{R}^{(i)}, \mathcal{P}^{(i)}, \gamma, \rho^{(i)} \right\rangle$, for $i = 1, \cdots, N$.

The objective of FRL is to enable N agents to collaboratively learn a common policy function or a value function that uniformly performs well across all environments. To preserve privacy,

agents are not allowed to exchange their raw observations (i.e., their rewards, states, or actions). In particular, we consider solving the following optimization problem:

$$\begin{aligned} \max_{\theta} \quad & \left\{ J(\theta) \triangleq \frac{1}{N} \sum_{i=1}^{N} J_i(\theta) \right\} \\ \text{where} \quad & J_i(\theta) \triangleq \mathbb{E}\left[\sum_{h=0}^{H-1} \gamma^h \mathcal{R}^{(i)}\left(s_h, a_h\right) \mid s_0 \sim \rho^{(i)}, \right. \\ & \left. a_h \sim \pi_\theta\left(\cdot \mid s_h\right), s_{h+1} \sim \mathcal{P}^{(i)}\left(\cdot \mid s_h, a_h\right) \right]. \end{aligned} \tag{5.3}$$

Objective. For solving the optimization problem (5.3), we aim to find the ϵ-FOSP, i.e., a parameter θ such that $\|\nabla J(\theta)\|^2 \leq \varepsilon$. There exists work that leverages the "gradient domination" condition [3, 37, 49, 124] for finding a global optimal policy in the centralized RL setting. The gradient domination condition is useful as it guarantees that every stationary policy is globally optimal. However, as shown in Zeng et al. [241], we cannot expect this condition to hold in general for FL or multi-agent problems. Specifically, even if a single performance function, $J_i(\theta)$, satisfies the "gradient domination" condition, the average function $J(\theta) = \frac{1}{N}\sum_{i=1}^{N} J_i(\theta)$ might not. [241] resolved this issue by introducing strong assumptions into the problem. For instance, Assumption 2 in their paper requires that the joint states between the environments are equally explored, which is difficult to verify in real-world applications.

Difference in the problem setup. Our setting is more general than existing work [79, 192]. In our work, each MDP can have a distinct initial state distribution, a feature not addressed in [79]. Furthermore, our framework does not require the bounded heterogeneity assumption of [192] and thus can handle arbitrary environment heterogeneity.

5.4 Algorithms

To solve problem (5.3), we present two federated momentum-based algorithms: FEDSVRPG-M and FEDHAPG-M. FEDSVRPG-M is based on a variance reduction method, while FEDHAPG-M leverages a fast Hessian-aided technique. Since FEDSVRPG-M only uses the first-order information (gradient), it is computationally cheaper than FEDHAPG-M, which aims to approximate second-order information (Hessians). Conversely, FEDHAPG-M, with its use of second-order information, is more robust than FEDSVRPG-M.

In the centralized RL setting, momentum-based PG methods [75, 240] are proposed to reduce the variance of stochastic gradients. In contrast, our algorithms integrate momentum within a federated context, achieving dual benefits: it not only accelerates the convergence and stabilizes oscillations, but also mitigates the impact of environment heterogeneity. Consequently, our algorithms can exactly converge to the ϵ-FOSP of problem (5.3), no matter how large the environment heterogeneity is. This represents a significant improvement upon [79, 227], which only show the convergence to the neighborhood around the stationary point of problem. The size of the neighborhood in their papers is determined by the environment heterogeneity.

5.4.1 FEDSVRPG-M

We now describe the federated stochastic variance-reduced PG with momentum algorithm (FEDSVRPG-M for short). We outline its steps in Algorithm 5.

FEDSVRPG-M initializes all agents and the server with a common model θ_0. In Algorithm 5, we use the superscript (i) to index the i-th agent and the subscript r and k to denote the r-th communication round and k-th local iteration. In each communication round r, each agent $i \in [N]$ is initiated from a common model θ_r and samples a single trajectory from its own environment to

perform K local iterations. At each local iteration k, instead of using PG, FEDSVRPG-M uses the following momentum-based variance-reduced stochastic PG estimator:

$$u_{r,k}^{(i)} = \beta g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left[u_r + g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}\right)\right], \tag{5.4}$$

where $\beta \in (0,1]$ and $w^{(i)}$ is the importance sampling weight, which is defined as:

$$w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) \triangleq \frac{p^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}\right)}{p^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right)}.$$

When $\beta = 1$, Eq. (5.4) reduces to the stochastic PG direction. When $\beta = 0$, it reduces to the variance-reduced PG direction. Notably, compared to the IS-MBPG algorithm of [75] for the centralized RL setting, the local updating rule in Algorithm 5 differs in that we estimate the PG directions locally, $\theta_{r,k}^{(i)}$, and globally θ_{r-1}, instead of two consecutive local policies. Furthermore, FEDSVRPG-M only requires constant local step-sizes, in contrast to the decreasing step-sizes in [75]. Moreover, FEDSVRPG-M only samples *one trajectory* per iterate, i.e., not does not require very large batch sizes, which is often necessary for centralized variance-reduced PG methods [231, 240]. For more discussion on the variance-reduced PG-type algorithms, we refer readers to [58].

A notable feature of FEDSVRPG-M is communication efficiency and data locality. To save the communication costs and preserve privacy, all agents upload their local model's difference $\Delta_r^{(i)}$, instead of the raw trajectories, to the server only after K local iterations (line 10). Following this step, the server aggregates all the differences to update the global model θ_{r+1} using the global step-size λ and then broadcasts it to all agents. Note that FEDSVRPG-M follows the same structure of the vanilla FEDAVG and achieves the same communication cost per communication round as FEDAVG.

Comparison with prior work. Note that the algorithms in [48] require the server to own its own environment (an MDP). They utilized the variance-reduced PG method for updating global models on the server side and applied the stochastic PG method to update the local model *only once* on the agent side. In contrast, our algorithms eliminate the need for the server to own its environment, enhancing its applicability in real-world scenarios. This is crucial as, in numerous cases, the server may function as a third-party entity without access to the environment.

Challenges. Most importantly, our algorithms *accommodate multiple local updates*, a crucial step for reducing the communication costs in FL. Thus, it is important for us to mitigate the common *"client-drift"* problems due to heterogeneity among agents. Notably, even for the standard FL algorithms in the supervised setting, it takes a substantial effort for the FL community to tackle this problem, such as FEDPROX [110], FEDNOVA [216], SCAFFOLD [84] and FEDLIN [136]. This challenge is further exacerbated in FRL, where the *non-oblivious* nature of problems makes it uncertain whether the bounded gradient heterogeneity assumption, commonly employed in FL optimization literature, remains applicable. Consequently, achieving a balance between communication cost and convergence rate is challenging. We analyze the performance of FEDSVRPG-M in Section 5.5.

5.4.2 FEDHAPG-M

Recently, HAPG [175] has been proposed for the centralized RL to reduce the sample complexity from $\mathcal{O}\left(1/\epsilon^4\right)$ to $\mathcal{O}\left(1/\epsilon^3\right)$ to obtain the ϵ-FOSP. The main success of HAPG comes from that it utilizes the stochastic approximation of the second-order policy differential. While HAPG uses curvature information, the computation cost of HAPG is still *linear* per iteration with respect to the parameter dimension d, as it avoids computing the Hessian explicitly.

We now provide a federated variant of HAPG; Federated Hessian Aided Policy Gradient

Algorithm 5 Description of FEDSVRPG-M

Input: initial model $\theta_{-1} = \theta_0$, gradient estimate u_0, local step-size η, global step-size λ and momentum β.

for $r = 0, 1, \ldots, R-1$ **do**

▷ Agent side

for each agent $i \in [N]$ **do**

Initial local model $\theta_{r,0}^{(i)} = \theta_r$

for $k = 0, 1, \ldots, K-1$ **do**

Sample a trajectory $\tau_{r,k}^{(i)} \sim p^{(i)}\left(\tau \mid \theta_{r,k}^{(i)}\right)$ and compute $u_{r,k}^{(i)}$ using Eq. (5.4).

Update local model $\theta_{r,k+1}^{(i)} = \theta_{r,k}^{(i)} + \eta u_{r,k}^{(i)}$

end for

Send $\Delta_r^{(i)} = \theta_{r,K}^{(i)} - \theta_r$ to the server

end for

▷ Server side

Aggregate $u_{r+1} = \frac{1}{\eta NK} \sum_{i=1}^{N} \Delta_r^{(i)}$

Update global model $\theta_{r+1} = \theta_r + \lambda u_{r+1}$

end for

with Momentum (FEDHAPG-M). As discussed in FEDSVRPG-M, the usage of momentum in FEDHAPG-M primarily serves to offer an "anchoring" direction that encodes PG estimates from all agents. Consequently, it eliminates the need for bounded environment heterogeneity assumption in existing FRL literature [79, 192, 228]. Moreover, FEDHAPG-M employs a second-order approximation instead of computing the difference between two consecutive stochastic gradients. As a result, FEDHAPG-M obtains an improved sample complexity akin to that of FEDSVRPG-M.

Note that FEDHAPG-M follows the same structure of the vanilla FEDAVG and FEDSVRPG-M, differing only in the local update procedure. In FEDHAPG-M, we replace the local update direction in FEDAVG with a variant of HAPG, see line $7 \sim 9$ in Algorithm 6. It is worth

noting that the uniform sampling step in line 7 guarantees that $\Lambda_{r,k}^{(i)}$ is an unbiased estimator of $\nabla J(\theta_{r,k}^{(i)}) - \nabla J(\theta_{r-1})$. To estimate the term $\Lambda_{r,k}^{(i)}$, as in [55, 175], we first assume that the function $J_i(\theta)$ is twice differentiable for all $i \in [N]$. Then we compute it as:

$$\begin{aligned}\Lambda_{r,k}^{(i)} \triangleq & \left\langle \nabla \log p\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right), v_{r,k}^{(i)} \right\rangle g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right) \\ & + \nabla \left\langle g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right), v_{r,k}^{(i)} \right\rangle \end{aligned} \tag{5.5}$$

where $v_{r,k}^{(i)} \triangleq \theta_{r,k}^{(i)} - \theta_{r-1}$. The variable θ_{r-1} represents the last-iterate global policy maintained in the server. As mentioned in [49], the computation of the second term in Eq (5.5) can be simplified through via automatic differentiation of the scalar quantity $g\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right)$. Thus, the computation cost of FEDHAPG-M does not increase and remains at $\mathcal{O}(Hd)$.

Discussion. Same as FEDSVRPG-M, FEDHAPG-M enjoys the following favorable features: (1) Only sampling one trajectory per local iteration; (2) No need for the server to have its own environment; (3) Multiple local updates. Such features were not simultaneously addressed in [48, 227].

5.5 Convergence Analysis

First, we introduce some standard assumptions.

Assumption 4. *Let $\pi_\theta^{(i)}(a \mid s)$ be the policy of the i-th agent at state s. There exist constants $G, M > 0$ such that the log-density of the policy function satisfies*

$$\left\|\nabla_\theta \log \pi_\theta^{(i)}(a \mid s)\right\| \leq G, \quad \left\|\nabla_\theta^2 \log \pi_\theta^{(i)}(a \mid s)\right\|_2 \leq M,$$

for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$ and $i \in [N]$.

Algorithm 6 Description of FEDHAPG-M

Input: initial model $\theta_{-1} = \theta_0$ and gradient estimate u_0, local step-size η, global step-size λ and momentum β.

for $r = 0, \cdots, R-1$ **do**

$\triangleright$ Agent side

for each agent $i \in [N]$ **do**

Initial local model $\theta_{r,0}^{(i)} = \theta_r$

for $k = 0, \cdots, K-1$ **do**

Choose α uniformly at random from $[0, 1]$, and compute $\theta_{r,k}^{(i)}(\alpha) = \alpha\theta_{r-1} + (1-\alpha)\theta_{r,k}^{(i)}$

Sample a trajectory $\tau_{r,k}^{(i)}$ from the density $p^{(i)}\left(\tau \mid \theta_{r,k}^{(i)}(\alpha)\right)$ and compute $u_{r,k}^{(i)} = \beta w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}, \theta_{r,k}^{(i)}(\alpha)\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left[u_r + \Lambda_{r,k}^{(i)}\right]$, where $\Lambda_{r,k}^{(i)}$ can be computed by using Eq. (5.5)

Update local model $\theta_{r,k+1}^{(i)} = \theta_{r,k}^{(i)} + \eta u_{r,k}^{(i)}$

end for

Send $\Delta_r^{(i)} = \theta_{r,K}^{(i)} - \theta_r$ back to the server

end for

$\triangleright$ Server side

Aggregate $u_{r+1} = \frac{1}{\eta NK}\sum_{i=1}^{N} \Delta_r^{(i)}$

Update global model $\theta_{r+1} = \theta_r + \lambda u_{r+1}$

end for

Assumption 5. *For each agent $i \in [N]$, the variance of stochastic gradient $g_i(\tau \mid \theta)$ is bounded, i.e., there exists a constant $\sigma > 0$, for all policies π_θ such that* $\mathrm{Var}(g_i(\tau \mid \theta)) = \mathbb{E}\|g_i(\tau \mid \theta) - \nabla J_i(\theta)\|^2 \leq \sigma^2$.

Assumption 6. *For each agent $i \in [N]$, the variance of importance sampling weight $w^{(i)}(\tau \mid \theta_1, \theta_2)$ is bounded, i.e., there exists a constant $W > 0$ such that*

$$\mathrm{Var}\left(w^{(i)}(\tau \mid \theta_1, \theta_2)\right) \leq W$$

holds for any $\theta_1, \theta_2 \in \mathbb{R}^d$ *and* $\tau \sim p^{(i)}(\cdot \mid \theta_2)$.

Assumption 4, 5 and 6 are commonly made in the convergence analysis of PG algorithms and their variance-reduced variants [124, 147, 175, 231]. They can be easily verified for Gaussian policies [33, 147, 153]. With these assumptions, we are ready to present the convergence guarantees for our FEDSVRPG-M algorithms.

Theorem 5. *(*FEDSVRPG-M*) Under Assumption 4–6, let* $u_0 = \frac{1}{NB}\sum_{i=1}^{N}\sum_{b=1}^{B} g_i\left(\tau_b^{(i)}|\theta_0\right)$ *with* $B = \left\lceil \frac{K}{R\beta^2} \right\rceil$ *and* $\left\{\tau_b^{(i)}\right\}_{b=1}^{B} \overset{iid}{\sim} p^{(i)}(\tau|\theta_0)$. *There exists a constant local step-size* η*, a proper global step-size* λ *and momentum coefficient* β*, such that the output of* FEDSVRPG-M *after* R *rounds satisfies:*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] \lesssim \left(\frac{\bar{L}\Delta\sigma}{NKR}\right)^{2/3} + \frac{\bar{L}\Delta}{R} \tag{5.6}$$

where $\Delta \triangleq J\left(\theta^*\right) - J(\theta_0)$, $G_0 \triangleq \frac{1}{N}\sum_{i=1}^{N}\left\|\nabla J_i\left(\theta_0\right)\right\|^2$.

Note that $\bar{L}$ in Theorem 5 is a constant depending on the constants $G, M, W, H, R_{\max}$ and $\frac{1}{(1-\gamma)^2}$. See Appendix for details. The notation $\lesssim$ denotes that inequalities hold up to some numeric number.

Comparison with prior work in FRL. FEDSVRPG-M surpasses all existing results in FRL in convergence, as shown in Table 5.1. Specifically, the results in Theorem 6 from [79] achieve only inexact convergence to a suboptimal solution, depending on the heterogeneity levels among N agents. In contrast, FEDSVRPG-M exactly converges to the ϵ-FOSP of Problem (5.3), with no heterogeneity term observed in Eq. (5.6). [48] exclusively considered the homogeneous environment. However, their results are limited to the sublinear result. i.e., the stationary point optimality can be scaled by $N^{\frac{2}{3}}$. In contrast, the dominant term $\left(\frac{\bar{L}\Delta\sigma}{NKR}\right)^{2/3}$ in the right-hand side of FiEq. (5.6) demonstrates that our algorithm provides a N-fold linear speedup over the single-agent scenario. Unique to our algorithm is the fact that this speed up is agnostic to the heterogeneity levels, unlike [224] and [192] which obtain a speedup in the no and low heterogeneity regimes respectively.

Table 5.2: Impact of environment heterogeneity κ and momentum coefficient β. We evaluate FEDSVRPG-M with various κ and various momentum coefficient β in $\{0.1, 0.2, 0.5, 0.8\}$. The baseline method is denoted by $\beta = 1$. Larger κ denotes larger environment heterogeneity. Each setting was run with 16,000 random seeds.

	RANDOM MDPS					
	$\kappa = 0$	$\kappa = 0.2$	$\kappa = 0.4$	$\kappa = 0.6$	$\kappa = 0.8$	$\kappa = 1.0$
$\beta = 0.1$	$\mathbf{8.013_{\pm 0.07}}$	$\mathbf{7.957_{\pm 0.07}}$	$\mathbf{7.968_{\pm 0.06}}$	$\mathbf{7.961_{\pm 0.06}}$	$\mathbf{7.964_{\pm 0.07}}$	$\mathbf{7.981_{\pm 0.06}}$
$\beta = 0.2$	$7.876_{\pm 0.06}$	$7.877_{\pm 0.06}$	$7.851_{\pm 0.06}$	$7.837_{\pm 0.06}$	$7.841_{\pm 0.06}$	$7.824_{\pm 0.07}$
$\beta = 0.5$	$7.561_{\pm 0.07}$	$7.208_{\pm 0.06}$	$7.529_{\pm 0.07}$	$7.525_{\pm 0.06}$	$7.536_{\pm 0.07}$	$7.525_{\pm 0.06}$
$\beta = 0.8$	$7.211_{\pm 0.07}$	$7.203_{\pm 0.07}$	$7.201_{\pm 0.06}$	$7.192_{\pm 0.06}$	$7.193_{\pm 0.06}$	$7.184_{\pm 0.06}$
$\beta = 1.0$	$6.965_{\pm 0.07}$	$6.951_{\pm 0.06}$	$6.955_{\pm 0.06}$	$6.936_{\pm 0.06}$	$6.940_{\pm 0.06}$	$6.937_{\pm 0.07}$

Comparison with prior work in RL. Compared to the centralized RL, i.e., $N = 1$, FEDSVRPG-M exhibits a convergence rate of $\mathcal{O}\left(1/(KR)^{\frac{2}{3}}\right)$, which aligns with the near-optimal convergence rate in [49]. In contrast, [75], utilizing diminishing step-sizes, achieves a slower convergence rate of $\mathcal{O}\left(\log(KR)/(KR)^{\frac{2}{3}}\right)$.

Comparison with prior work in FL optimization. To appreciate the tightness of our results, we note that our results align with the state-of-the-art convergence rates [28, 77] in the FL optimization literature. However, our results are established for a more complex RL setting. In contrast to the supervised learning scenario, where the distribution of τ is fixed over all iterations, our problem is *non-oblivious*. Furthermore, FEDSVRPG-M allows for the constant local step-sizes. In contrast, many FL optimization algorithms [92, 235] require the decreasing local step-sizes to mitigate heterogeneity among agents.

Now, we analyze the convergence of FEDHAPG-M.

Theorem 6. *(*FEDHAPG-M*) Under Assumption 4–6, choose the same u_0 as Theorem 5. There exists a constant local step-size η, a proper global step-size λ and momentum coefficient β, such that the output of* FEDHAPG-M *after R rounds satisfies*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] \lesssim \left(\frac{\hat{L}\Delta\sigma}{NKR}\right)^{2/3} + \frac{\hat{L}\Delta}{R} \tag{5.7}$$

where $\Delta \triangleq J\left(\theta^*\right) - J(\theta_0)$, $G_0 \triangleq \frac{1}{N}\sum_{i=1}^{N}\left\|\nabla J_i\left(\theta_0\right)\right\|^2$

From Theorem 6, we remark that FEDHAPG-M enjoys the same worst-case convergence rate, i.e., $\mathcal{O}(1/(NKR)^{2/3})$, as FEDSVRPG-M, except for the differences in the constant $\hat{L}$ and parameter selection. Interested readers are referred to Appendix for details.

Based on Theorem 5 and 6, we can now translate the convergence results to the total sample complexity of each agent, which is shown in the following corollary.

Corollary 1. *Under Assumption 4–6, the sample complexity of* FEDSVRPG-M *and* FEDHAPG-M *is $\mathcal{O}\left(\epsilon^{-\frac{3}{2}}/N\right)$ per agent to find an ϵ-FOSP.*

5.6 Experiments

We first use tabular environments to verify our theories on the proposed FEDSVRPG-M algorithm. It is important to note that FEDHAPG-M algorithm can not be assessed in the tabular setting due to the objective function $J_i(\theta)$ not being twice differentiable. We then evaluate both FEDSVRPG-M and FEDHAPG-M's performance on MuJoCo [197] with a deep RL extension. The baseline algorithm is the PAVG algorithm [79].

Tabular Case. We evaluate the performance of our algorithms in the environment of random MDPs, where both state transitions and reward functions are generated randomly. We use the same

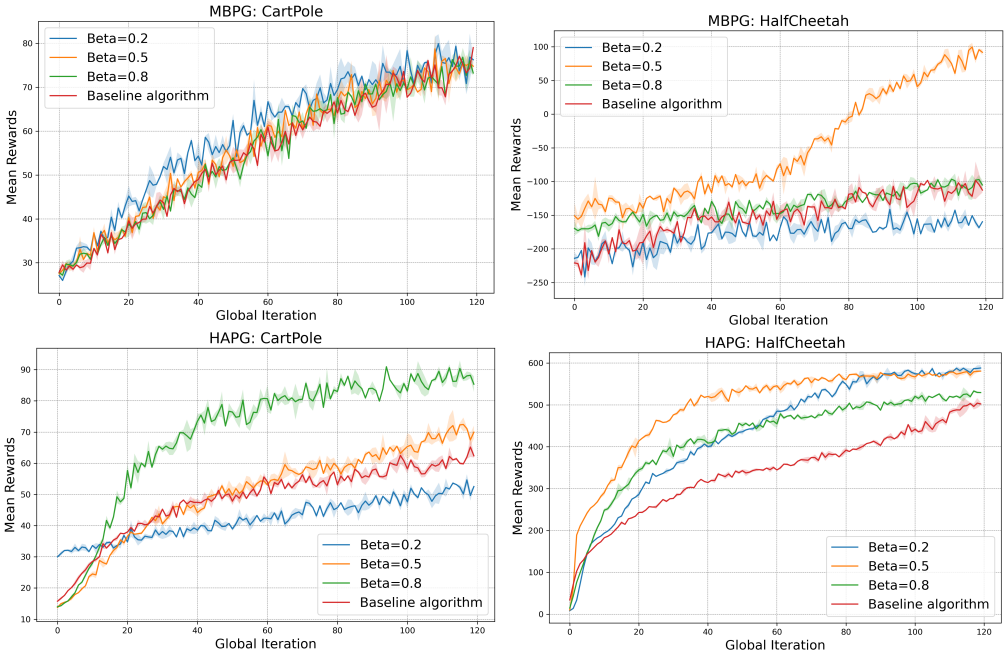


Figure 5.1: Mean rewards over global iterations for the CartPole and HalfCheetah tasks: (**Top**): FEDSVRPG-M; (**Bottom**): FEDHAPG-M.

method as [79] to control the environment heterogeneity. First, we randomly sample a nominal state transition kernel $\mathcal{P}_0$ and then generate the environments $\left\{\mathcal{P}^{(i)} = \kappa\mathcal{P}_i + (1-\kappa)\mathcal{P}_0\right\}_{i=1}^{N}$. Each entry of the kernels $\{\mathcal{P}_i\}_{i=1}^{N}$ are uniformly sampled between 0 and 1 and then normalized. Then, we can evaluate the impact of environment heterogeneity by varying κ. We compare the performance of FEDSVRPG-M with the existing baseline algorithm (PAVG). The results are shown in Table 5.2. The performance is measured by the average performance function in Eq. (5.3). We observe that FEDSVRPG-M with $\beta = 0.1$ outperforms the baseline algorithm ($\beta = 1$). Furthermore, the performance of FEDSVRPG-M is agnostic to the environment heterogeneity level κ. These trends are expected and consistent with theoretical analysis in Sec. 5.5.

Deep RL Case. We evaluate the performance of our algorithms across two benchmark RL tasks: CartPole and HalfCheetah. While CartPole is a classic control task with discrete actions, HalfCheetah represents a continuous RL task. Both are widely recognized tasks in the MuJoCo

simulation environment [197]. Comprehensive details of the experimental setups can be found in the appendix. To introduce environment heterogeneity, we change the initial state distribution parameters in both tasks. We use Categorical Policy for CartPole, and Gaussian Policy for HalfCheetah. All policies are parameterized by the fully connected neural network which has two hidden layers and a hyperbolic tangent activation function. The hidden layers neural network sizes are 32 for Gaussian policies and 8 for Categorical policies. In Figure 5.1, we show how the mean rewards change over the global iterations for our proposed algorithms and baseline algorithm. In both tasks, as the number of iterations increases, all algorithms exhibit a rising trend in mean rewards. There exist a $\beta \neq 1$ that our proposed algorithms outperform the baseline algorithm. In particular, FEDSVRPG-M exhibits optimal performance at $\beta = 0.2$ for CartPole and $\beta = 0.5$ for HalfCheetah. In contrast, FEDHAPG-M performs optimally with $\beta = 0.8$ for CartPole and $\beta = 0.5$ for HalfCheetah. FEDHAPG-M, which uses second-order information, shows smaller variance than FEDSVRPG-M, as indicated by the narrower color-shaded regions in the figure. Overall, our algorithms demonstrated superior performance compared to the baseline. See Appendix in [210] for more experiments evaluating the linear speedup in the number of agents N.

5.7 Chapter Summary

We introduced FEDSVRPG-M and FEDHAPG-M, overcoming the limitation of bounded environment heterogeneity assumed in prior FRL research. Our results demonstrate the best known convergence for these algorithms and highlight the benefits of collaboration in FRL, even in scenarios with conflicting rewards across different environments. In the future, we plan to focus on algorithms that facilitate downstream fine-tuning or personalization, aiming to discover each MDP's optimal policy through FRL, rather than seeking a universally optimal policy.

5.8 Omitted Proofs

5.8.1 Notation

We denote $\mathcal{F}_0 = \emptyset$ and $\mathcal{F}_{r,k}^{(i)} := \sigma\left(\left\{\theta_{r,j}^{(i)}\right\}_{0\leq j\leq k} \cup \mathcal{F}_r\right)$ and $\mathcal{F}_{r+1} := \sigma\left(\cup_i \mathcal{F}_{r,K}^{(i)}\right)$ for all $r \geq 0$ where $\sigma(\cdot)$ indicates the σ-algebra. Let $\mathbb{E}_r[\cdot] := \mathbb{E}\left[\cdot \mid \mathcal{F}_r\right]$ be the expectation, conditioned on the filtration $\mathcal{F}_r$, with respect to the random variables $\left\{\tau_{r,k}^{(i)}\right\}_{1\leq i\leq N, 0\leq k<K}$ in the r-th iteration. Moreover, we use $\mathbb{E}[\cdot]$ to denote the global expectation over all randomness in algorithms. For all $r \geq 0$, we define the following notations to simpify the proof:

$$
\begin{aligned}
\Sigma_r &:= \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right) - u_{r+1}\right\|^2\right], \\
\mathcal{D}_r &:= \frac{1}{NK}\sum_i \sum_k \mathbb{E}\left[\left\|\theta_{r,k}^{(i)} - \theta_r\right\|\right]^2, \\
c_{r,k}^{(i)} &:= \mathbb{E}\left[\theta_{r,k+1}^{(i)} - \theta_{r,k}^{(i)} \mid \mathcal{F}_{r,k}^{(i)}\right], \\
\mathcal{M}_r &:= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\left\|c_{r,0}^{(i)}\right\|^2\right].
\end{aligned}
$$

Throughout the appendix, we denote

$$
\Delta := J\left(\theta^*\right) - J(\theta_0),\ G_0 := \frac{1}{N}\sum \left\|\nabla J_i\left(\theta_0\right)\right\|^2,\ \theta_{-1} := \theta_0 \ \text{ and } \ \Sigma_{-1} := \mathbb{E}\left[\left\|\nabla J\left(\theta_0\right) - u_0\right\|^2\right].
$$

and θ^* denotes the optimal policy of the optimization problem (3).

5.8.2 Useful Lemmas and Inequalities

We make repeated use throughout the appendix (often without explicitly stating so) of the following inequalities:

- Given any two vectors $x, y \in \mathbb{R}^d$, for any $a > 0$, we have

$$
\|x + y\|^2 \leq (1 + a)\|x\|^2 + \left(1 + \frac{1}{a}\right)\|y\|^2. \tag{5.8}
$$

- Given any two vectors $x, y \in \mathbb{R}^d$, for any constant $a > 0$, we have

$$\langle x, y \rangle \leq \frac{a}{2}\|x\|^2 + \frac{1}{2a}\|y\|^2. \tag{5.9}$$

  This inequality goes by the name of Young's inequality.

- Given m vectors $x_1, \ldots, x_m \in \mathbb{R}^d$, the following is a simple application of Jensen's inequality:

$$\left\| \sum_{i=1}^{m} x_i \right\|^2 \leq m \sum_{i=1}^{m} \|x_i\|^2 . \tag{5.10}$$

Proposition 1. *(Proposition 5.2 in [232]) Under Assumption 1, both $J(\theta)$ and $\{J_i(\theta)\}_{i=1}^N$ are L-smooth with $L = HR_{\max}\left(M + HG^2\right)/(1-\lambda)$. In addition, for all $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*

$$\|g_i(\tau \mid \theta_1) - g_i(\tau \mid \theta_2)\|_2 \leq L_g \|\theta_1 - \theta_2\|_2$$

and $\|g_i(\tau \mid \theta)\|_2 \leq C_g$ for all $\theta \in \mathbb{R}^d$ and $i \in [N]$, where $L_g = HMR_{\max}/(1-\lambda), C_g = HGR_{\max}/(1-\lambda)$.

Lemma 17. *If $\lambda L \leq \frac{1}{24}$, the following inequality holds for all $r \geq 0$:*

$$\mathbb{E}\left[J\left(\theta_{r+1}\right)\right] \geq \mathbb{E}\left[J\left(\theta_r\right)\right] + \frac{11\lambda}{24}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] - \frac{13\lambda}{24}\Sigma_r.$$

Proof. Since J is L-smooth, we have

$$\begin{aligned} J\left(\theta_{r+1}\right) &\geq J\left(\theta_r\right) + \left\langle \nabla J\left(\theta_r\right), \theta_{r+1} - \theta_r \right\rangle - \frac{L}{2}\left\|\theta_{r+1} - \theta_r\right\|^2 \\ &= J\left(\theta_r\right) + \lambda\left\|\nabla J\left(\theta_r\right)\right\|^2 + \lambda\left\langle \nabla J\left(\theta_r\right), u_{r+1} - \nabla J\left(\theta_r\right)\right\rangle - \frac{L\lambda^2}{2}\left\|u_{r+1}\right\|^2 . \end{aligned}$$

where we use the fact that $\theta_{r+1} = \theta_r + \eta u_{r+1}$. By using Young's inequality, we have

$$\begin{aligned}
&J\left(\theta_{r+1}\right) \\
\geq &J\left(\theta_r\right) + \frac{\lambda}{2}\left\|\nabla J\left(\theta_r\right)\right\|^2 - \frac{\lambda}{2}\left\|\nabla J\left(\theta_r\right) - u_{r+1}\right\|^2 - L\lambda^2\left(\left\|\nabla J\left(\theta_r\right)\right\|^2 + \left\|\nabla J\left(\theta_r\right) - u_{r+1}\right\|^2\right) \\
\geq &J\left(\theta_r\right) + \frac{11\lambda}{24}\left\|\nabla J\left(\theta_r\right)\right\|^2 - \frac{13\lambda}{24}\left\|\nabla J\left(\theta_r\right) - u_{r+1}\right\|^2,
\end{aligned}$$

where the last inequality holds due to $\lambda L \leq \frac{1}{24}$. Taking the global expectation completes the proof. □

Lemma 18. *(Lemma 6.1 in [232]) Under Assumptions 4 and 6, we have*

$$\operatorname{Var}\left(w^{(i)}(\tau \mid \theta_1, \theta_2)\right) \leq C_w \left\|\theta_1 - \theta_2\right\|^2$$

holds for any $\theta_1, \theta_2 \in \mathcal{R}^d$ and any $i \in [N]$, where $C_\omega = H\left(2HG^2 + M\right)(W+1)$.

5.8.3 Federated Stochastic Variance-Reduced Policy Gradient with Momentum

According to the updating rule of FEDSVRPG-M, we have

$$\mathbb{E}[u_{r+1}] = \frac{1}{NK}\sum_{i,k}\mathbb{E}\left[\nabla J_i(\theta_{r,k}^{(i)}) + (1-\beta)\left(u_r - \nabla J_i(\theta_r)\right)\right].$$

Lemma 19. *If $\lambda \leq \sqrt{\frac{16\beta NK}{\widetilde{L_2}^2}}$, we have*

$$\Sigma_r \leq (1 - \frac{8\beta}{9})\Sigma_{r-1} + \frac{\widetilde{L_1}^2}{\beta}\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} + 18\lambda^2\frac{\widetilde{L_2}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2$$

holds for $r \geq 1$, *where* $\widetilde{L_1}^2 := L^2 + 24C_wC_g^2 + 6L_g^2$ *and* $\widetilde{L_2}^2 := L_g^2 + 2C_wC_g^2$. *When* $r = 0$, *we have*

$$\Sigma_0 \leq (1-\beta)\Sigma_{-1} + \frac{\widetilde{L_1}^2}{\beta}\mathcal{D}_0 + \frac{3\beta^2\sigma^2}{NK}$$

Proof.

$$\begin{aligned}
\Sigma_r &= \mathbb{E}\left[\|u_{r+1} - \nabla J(\theta_r)\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{NK}\sum_{i,k}\beta g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left[u_r + g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}\right)\right]\right.\right. \\
&\left.\left. -\nabla J(\theta_r)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|(1-\beta)(u_r - \nabla J(\theta_{r-1})) + \frac{1}{NK}\sum_{i,k}\left[g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - \nabla J(\theta_r)\right]\right.\right. \\
&\left.\left. -(1-\beta)\left[\frac{1}{NK}\sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}\right) - \nabla J(\theta_{r-1})\right]\right\|^2\right] \\
&= (1-\beta)^2\Sigma_{r-1} + \underbrace{2\mathbb{E}\left[\left\langle (1-\beta)(u_r - \nabla J(\theta_{r-1})), \frac{1}{NK}\sum_{i,k}\left[\nabla J_i(\theta_{r,k}^{(i)}) - \nabla J(\theta_r)\right]\right\rangle\right]}_{T_1} \\
&+ \underbrace{\begin{array}{c}\mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\left[g_i\left(\tau_{r,k}^{(i)}|\theta_{r,k}^{(i)}\right) - \nabla J(\theta_r)\right]\right. \\ \left. -(1-\beta)\left[\frac{1}{NK}\sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)}|\theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)}|\theta_{r-1}\right) - \nabla J(\theta_{r-1})\right]\right\|^2\end{array}}_{T_2}
\end{aligned}$$

Using Young's inequality to bound T_1, we have

$$\begin{aligned}
T_1 &\leq \beta(1-\beta)^2\mathbb{E}\left\|u_r - \nabla J(\theta_{r-1})\right\|^2 + \frac{1}{\beta}\mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\nabla J_i(\theta_{r,k}^{(i)}) - \nabla J(\theta_r)\right\|^2 \\
&\leq \beta(1-\beta)^2\Sigma_{r-1} + \frac{L^2}{\beta}\underbrace{\frac{1}{NK}\sum_{i,k}\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2}_{\mathcal{D}_r}
\end{aligned} \tag{5.11}$$

Further bounding T_2, we have

$$T_2 \leq \mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\left(g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_r\right)\right)\right.$$

$$
\begin{aligned}
&+\beta\left(\frac{1}{NK}\sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_r\right) - \nabla J(\theta_r)\right) \\
&+(1-\beta)\Big(\frac{1}{NK}\sum_{i,k}\Big(w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_r) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1})\Big) \\
&-\nabla J(\theta_r) + \nabla J(\theta_{r-1})\Big)\Big\|^2 \\
&\leq 3\,\underbrace{\mathbb{E}\Big\|\frac{1}{NK}\sum_{i,k}\Big(g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_r\right)\Big)\Big\|^2}_{T_{21}} + 3\frac{\beta^2\sigma^2}{NK} \\
&+3(1-\beta)^2\,\underbrace{\mathbb{E}\left[\left\|\frac{1}{NK}\sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_r) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1})\right\|^2\right]}_{T_{22}}
\end{aligned}
\tag{5.12}
$$

where we use the Young's inequality in the last equality and the fact that $\mathbb{E}[\|X-\mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$ holds for any random variable X.

To precede, we continue to bound T_{21} and have that

$$
\begin{aligned}
T_{21} &= \mathbb{E}\Big\|\frac{1}{NK}\sum_{i,k}\Big(g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_r\right)\Big)\Big\|^2 \\
&\leq 2\mathbb{E}\Big\|\frac{1}{NK}\sum_{i,k}\Big(1 - w^{(i)}(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)})\Big) g_i(\tau_{r,k}^{(i)} \mid \theta_r)\Big\|^2 \\
&+ 2\mathbb{E}\Big\|\frac{1}{NK}\sum_{i,k}\Big[g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - g_i(\tau_{r,k}^{(i)} \mid \theta_r)\Big]\Big\|^2 \\
&\leq \frac{2C_w C_g^2}{NK}\sum_{i,k}\mathbb{E}\Big\|\theta_{r,k}^{(i)} - \theta_r\Big\|^2 + 2\frac{L_g^2}{NK}\sum_{i,k}\mathbb{E}\Big\|\theta_{r,k}^{(i)} - \theta_r\Big\|^2 \\
&= (2C_w C_g^2 + 2L_g^2)\mathcal{D}_r
\end{aligned}
\tag{5.13}
$$

where we use the fact that $\|g^{(i)}(\tau \mid \theta)\|_2 \leq C_g$ for all $\theta \in \mathbb{R}^d$ and $i \in [N]$.

To bound T_{22}, we have

$$
T_{22} = \mathbb{E}\left[\left\|\frac{1}{NK}\sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_r) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1})\right\|^2\right]
$$

$$
\begin{aligned}
&\leq 3\mathbb{E}\left[\left\|\frac{1}{NK}\sum_{i,k}\left[w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right)-1\right] g_i(\tau_{r,k}^{(i)} \mid \theta_r)\right\|^2\right] \\
&+ 3\frac{1}{N^2K^2}\sum_{i,k}\mathbb{E}\left\|g_i(\tau_{r,k}^{(i)} \mid \theta_r) - g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1})\right\|^2 \\
&+ 3\mathbb{E}\left[\left\|\frac{1}{NK}\sum_{i,k}\left[w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right)-1\right] g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1})\right\|^2\right] \\
&\leq 3C_g^2C_w\frac{1}{N^2K^2}\sum_{i,k}\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2 + 3\frac{L_g^2}{NK}\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2 + 3C_g^2C_w\frac{1}{N^2K^2}\sum_{i,k}\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_{r-1}\right\|^2 \\
&\leq 6C_g^2C_w\frac{1}{NK}\mathcal{D}_r + \frac{3L_g^2+6C_wC_g^2}{NK}\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2 \qquad (5.14)
\end{aligned}
$$

Combining the upper bound of T_{21} and T_{22} (i.e., (5.13) and (5.14)) into T_2 in Eq. (5.12), we have

$$
T_2 \leq (24C_wC_g^2+6L_g^2)\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} + 9(1-\beta)^2\frac{L_g^2+2C_wC_g^2}{NK}\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2 \qquad (5.15)
$$

Therefore, for $r \geq 1$, we have

$$
\begin{aligned}
\Sigma_r &\leq (1-\beta)\Sigma_{r-1} + \frac{L^2+24C_wC_g^2+6L_g^2}{\beta}\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} + 9(1-\beta)^2\frac{L_g^2+2C_wC_g^2}{NK}\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2 \qquad (5.16) \\
&\leq (1-\beta)\Sigma_{r-1} + \frac{L^2+24C_wC_g^2+6L_g^2}{\beta}\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} \\
&+ 18\lambda^2\frac{L_g^2+2C_wC_g^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 + 18\lambda^2\frac{L_g^2+2C_wC_g^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})-u_r\right\|^2 \\
&= \left(1-\beta+18\lambda^2\frac{L_g^2+2C_wC_g^2}{NK}\right)\Sigma_{r-1} + 18\lambda^2\frac{L_g^2+2C_wC_g^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \\
&+ \frac{L^2+24C_wC_g^2+6L_g^2}{\beta}\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} \qquad (5.17)
\end{aligned}
$$

By choosing λ such that $18\lambda^2\frac{L_g^2+2C_wC_g^2}{NK} \leq \frac{8\beta}{9}$, which holds when $\lambda \leq \sqrt{\frac{16\beta NK}{L_g^2+2C_wC_g^2}}$, we have

$$
\Sigma_r \leq (1-\frac{8\beta}{9})\Sigma_{r-1} + \frac{L^2+24C_wC_g^2+6L_g^2}{\beta}\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} + 18\lambda^2\frac{L_g^2+2C_wC_g^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \qquad (5.18)
$$

holds for $r > 0$. When $r = 0$, we have that

$$\Sigma_0 \leq (1-\beta)\Sigma_{-1} + \frac{L^2 + 24C_wC_g^2 + 6L_g^2}{\beta}\mathcal{D}_0 + \frac{3\beta^2\sigma^2}{NK} \tag{5.19}$$

which can be derived from Eq.(5.16). □

Lemma 20. *(Bounding drift-term) If the local step-size satisfies* $\eta \leq \min\{\frac{L}{32e^2\widetilde{L_3}^2K}, \frac{1}{KL}\}$*, the drift-term can be upper bounded as:*

$$\mathcal{D}_r \leq 4eK^2\mathcal{M}_r + (16\eta^4K^4L^2 + 8\eta^2K)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right)$$

where $\widetilde{L_3}^2 := 2C_wC_g^2 + 2L_g^2$.

Proof. Define $c_{r,k}^{(i)} := -\eta\left(\nabla J_i(\theta_{r,k}^{(i)}) + (1-\beta)(u_r - \nabla J_i(\theta_{r-1}))\right)$. For any $1 \leq j \leq k-1 \leq K-2$, we have:

$$\begin{aligned}\mathbb{E}\left\|c_{r,j}^{(i)} - c_{r,j-1}^{(i)}\right\|^2 &\leq \eta^2L^2\mathbb{E}\left\|\theta_{r,j}^{(i)} - \theta_{r,j-1}^{(i)}\right\|^2 \\ &= \eta^2L^2\left(\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2 + \mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j}^{(i)} - \theta_{r,j-1}^{(i)} \mid \mathcal{F}_{r,j-1}^{(i)}\right]\right]\right).\end{aligned} \tag{5.20}$$

where we use the bias-variance decomposition in the last inequality.

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j}^{(i)} - \theta_{r,j-1}^{(i)} \mid \mathcal{F}_{r,j-1}^{(i)}\right]\right] \\ &= \eta^2\mathbb{E}\Bigg\|g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - \nabla J_i(\theta_{r,j-1}^{(i)}) \\ &- (1-\beta)\left[w^{(i)}\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}, \theta_{r,j-1}^{(i)}\right)g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}\right) - \nabla J_i(\theta_{r-1})\right]\Bigg\|^2 \\ &= \eta^2\mathbb{E}\Bigg\|\beta\left[g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - \nabla J_i(\theta_{r,j-1}^{(i)})\right] \\ &+ (1-\beta)\left[g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - w^{(i)}\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}, \theta_{r,j-1}^{(i)}\right)g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}\right)\right.\end{aligned}$$

$$
\begin{aligned}
&- \left(\nabla J_i(\theta_{r,j-1}^{(i)}) - \nabla J_i(\theta_{r-1})\right)\Big]\Big\|^2 \\
&\leq 2\eta^2\beta^2\sigma^2 \\
&+ 2\eta^2(1-\beta)^2 \underbrace{\mathbb{E}\left\| g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - w^{(i)}\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}, \theta_{r,j-1}^{(i)}\right) g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}\right)\right\|^2}_{T_3} \qquad (5.21)
\end{aligned}
$$

where Eq.(5.21) holds due to the Young's inequality and the fact that $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$.

To precede, we bound T_3 as

$$
\begin{aligned}
T_3 &= \mathbb{E}\left\| g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - w^{(i)}\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}, \theta_{r,j-1}^{(i)}\right) g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}\right)\right\|^2 \\
&\leq 2\mathbb{E}\left\| \left(1 - w^{(i)}(\tau_{r,j-1}^{(i)} \mid \theta_r, \theta_{r,j-1}^{(i)})\right) g_i(\tau_{r,j-1}^{(i)} \mid \theta_r)\right\|^2 \\
&+ 2\mathbb{E}\left\| g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - g_i(\tau_{r,j-1}^{(i)} \mid \theta_{r-1})\right\|^2 \\
&\leq 2C_w C_g^2 \mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_{r-1}\right\|^2 + 2L_g^2 \mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_{r-1}\right\|^2 \\
&= (2C_w C_g^2 + 2L_g^2)\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_{r-1}\right\|^2 \qquad (5.22)
\end{aligned}
$$

where we use the fact that $\|g^{(i)}(\tau \mid \theta)\|_2 \leq C_g$ for all $\theta \in \mathbb{R}^d$ and $i \in [N]$.

With the upper bound of T_3 and $\widetilde{L_3}^2 := 2C_w C_g^2 + 2L_g^2$, we have

$$
\begin{aligned}
&\mathbb{E}\left\| c_{r,j}^{(i)} - c_{r,j-1}^{(i)}\right\|^2 \leq \eta^2 L^2 \left(\mathbb{E}\left\| c_{r,j-1}^{(i)}\right\|^2 + 2\eta^2\beta^2\sigma^2 + 2\eta^2(1-\beta)^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_{r-1}\right\|^2\right) \\
&\leq \eta^2 L^2 \left(\mathbb{E}\left\| c_{r,j-1}^{(i)}\right\|^2 + 2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_r\right\|^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right). \qquad (5.23)
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\mathbb{E}\left\| c_{r,j}^{(i)}\right\|^2 \leq (1 + \frac{1}{q})\mathbb{E}\left\| c_{r,j-1}^{(i)}\right\|^2 + (1+q)\mathbb{E}\left\| c_{r,j}^{(i)} - c_{r,j-1}^{(i)}\right\|^2 \\
&\leq (1 + \frac{2}{q})\mathbb{E}\left\| c_{r,j-1}^{(i)}\right\|^2 + (1+q)\eta^2 L^2 \left(2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_r\right\|^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) \\
&\qquad (5.24)
\end{aligned}
$$

where we use the fact that $\eta L \leq \frac{1}{K} \leq \frac{1}{q+1}$ and let $q = k-1$. By unrolling this recurrence, we have

$$\begin{aligned}
\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 &\leq (1+\frac{2}{k-1})^j \mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + k\eta^2 L^2 \sum_{i=0}^{j-1}(2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2)\Pi_{j'=i+1}^{j-1}(1+\frac{2}{k-1}) \\
&\quad + k\eta^2 L^2 \sum_{s=0}^{j-1}(4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r,s}^{(i)}-\theta_r\right\|^2)\Pi_{j'=s+1}^{j-1}(1+\frac{2}{k-1}) \\
&\leq (1+\frac{2}{k-1})^{k-1}\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + k\eta^2 L^2 \sum_{i=0}^{k-1}(2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2)(1+\frac{2}{k-1})^{k-1} \\
&\quad + k\eta^2 L^2 \sum_{j'=0}^{j-1}(4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2)(1+\frac{2}{k-1})^{k-1} \qquad (5.25)
\end{aligned}$$

Based on the inequality $(1+\frac{2}{K-1}^{k-1}) \leq e^2 \leq 8$, we have

$$\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 \leq e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + 8k^2\eta^4L^2\left(2\beta^2\sigma^2 + 4\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) + 4e^2k\eta^4L^2\widetilde{L_3}^2\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2 \qquad (5.26)$$

By Lemma A.3, we have

$$\begin{aligned}
\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2 &\leq 2\mathbb{E}\left\|\sum_{j=0}^{k-1}c_{r,j}^{(i)}\right\|^2 + 2\sum_{j=0}^{k-1}\mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j+1}^{(i)}-\theta_{r,j}^{(i)} \mid \mathcal{F}_{r,j}^{(i)}\right]\right] \\
&\overset{(a)}{\leq} 2k\sum_{j=0}^{k-1}\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 + 2\sum_{j=0}^{k-1}\left(2\beta^2\eta^2\sigma^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) \qquad (5.27)
\end{aligned}$$

where (a) is due to Eq.(5.21) and Eq.(5.22). Plugging Eq.(5.26) into Eq.(5.27), we have

$$\begin{aligned}
&\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2 \leq \\
&2k\sum_{j=0}^{k-1}\left\{e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + 8k^2\eta^4L^2\left(2\beta^2\sigma^2 + 4\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) + 4e^2k\eta^4L^2\widetilde{L_3}^2\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2\right\} \\
&+ 2\sum_{j=0}^{k-1}\left(2\beta^2\eta^2\sigma^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) \qquad (5.28)
\end{aligned}$$

Summing up the above equation over $k = 0, \cdots, K-1$, we have

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2 \leq \sum_{k=0}^{K-1}\left\{2k^2e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + 16k^4\eta^4L^2\left(2\beta^2\sigma^2 + 4\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)\right\}$$

$$
\begin{aligned}
&+ \sum_{k=0}^{K-1} 8e^2k^2\eta^4L^2\widetilde{L_3}^2 \sum_{j=0}^{k-1}\sum_{j'=0}^{j-1} \mathbb{E}\left\|\theta_{r,j'}^{(i)} - \theta_r\right\|^2 \\
&+ \sum_{k=0}^{K-1} \left(4k\beta^2\eta^2\sigma^2 + 8k\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2 + 8\eta^2\widetilde{L_3}^2 \sum_{j=0}^{k-1}\mathbb{E}\left\|\theta_{r,j}^{(i)} - \theta_r\right\|^2\right) \\
&\leq 2eK^3\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + (8\eta^4K^5L^2 + 4\eta^2K^2)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) \\
&+ K^2\sum_{k=0}^{K-1} 8e^2\eta^4L^2\widetilde{L_3}^2 \sum_{j=0}^{K-1}\sum_{j'=0}^{K-1} \mathbb{E}\left\|\theta_{r,j'}^{(i)} - \theta_r\right\|^2 + \sum_{k=0}^{K-1} 8\eta^2\widetilde{L_3}^2 \sum_{j=0}^{K-1}\mathbb{E}\left\|\theta_{r,j}^{(i)} - \theta_r\right\|^2 \\
&= 2eK^3\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + (8\eta^4K^5L^2 + 4\eta^2K^2)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) \\
&+ (8e^2\eta^4K^4L^2\widetilde{L_3}^2 + 8\eta^2\widetilde{L_3}^2K)\sum_{j=0}^{K-1}\mathbb{E}\left\|\theta_{r,j}^{(i)} - \theta_r\right\|^2 \qquad (5.29)
\end{aligned}
$$

With the choice of step-size η satisfying $8e^2\eta^4K^4L^2\widetilde{L_3}^2 + 8\eta^2\widetilde{L_3}^2K \leq \frac{1}{2}$, after some rearrangement, we have

$$
\frac{1}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 \leq 2eK^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + (8\eta^4K^4L^2 + 4\eta^2K)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right)
$$

In summary, we can bound the drift-term as

$$
\mathcal{D}_r \leq 4eK^2\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2}_{\mathcal{M}_r} + (16\eta^4K^4L^2 + 8\eta^2K)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right)
$$

□

Lemma 21. *If* $\lambda L \leq \frac{1}{24}$ *and* $\eta^2\left[\frac{289}{72}(1-\beta)^2 + 8e(\lambda\beta LR)^2\right] \leq \frac{\beta^2}{288eK^2\widetilde{L_1}^2}$, *we have*

$$
\sum_{r=0}^{R-1}\mathcal{M}_r = \frac{1}{N}\sum_{r=0}^{R-1}\sum_{i=1}^{N}\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 \leq \frac{\beta^2}{288eK^2\widetilde{L_1}^2}\sum_{r=-1}^{R-2}\left(\Sigma_r + \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\right) + 4\eta^2\beta^2eRG_0. \qquad (5.30)
$$

where $G_0 := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla J_i\left(\theta_0\right)\right\|^2\right]$ *and* $\widetilde{L_1}^2$ *is defined in Lemma 19.*

Proof. Recall that $c_{r,0}^{(i)} := -\eta \left(\nabla J_i(\theta_r) + (1-\beta)(u_r - \nabla J_i(\theta_{r-1}))\right)$. Then, it is straightforward to have

$$\begin{aligned}
\left\|c_{r,0}^{(i)}\right\|^2 &\leq 2\eta^2 \left((1-\beta)^2 \|u_r\|^2 + \|\nabla J_i(\theta_r) - (1-\beta)\nabla J_i(\theta_{r-1})\|^2\right) \\
&\leq 2\eta^2(1-\beta)^2 \|u_r\|^2 + 4\eta^2(1-\beta)^2 \|\nabla J_i(\theta_r) - \nabla J_i(\theta_{r-1})\|^2 + 4\eta^2\beta^2 \|\nabla J_i(\theta_r)\|^2 \\
&\leq 2\eta^2(1-\beta)^2 \left(1 + 2(\lambda L)^2\right) \|u_r\|^2 + 4\eta^2\beta^2 \|\nabla J_i(\theta_r)\|^2 \\
&\overset{(a)}{\leq} \frac{289}{144}\eta^2(1-\beta)^2 \|u_r\|^2 + 4\eta^2\beta^2 \|\nabla J_i(\theta_r)\|^2 . \qquad (5.31)
\end{aligned}$$

where (a) is due to the choice of λ such that $\lambda L \leq \frac{1}{24}$.

Using the Young's inequality, we have that for any $\zeta > 0$,

$$\begin{aligned}
\mathbb{E}\left[\|\nabla J_i(\theta_r)\|^2\right] &\leq (1+\zeta)\mathbb{E}\left[\|\nabla J_i(\theta_{r-1})\|^2\right] + \left(1 + \frac{1}{\zeta}\right)\mathbb{E}\|\nabla J_i(\theta_r) - \nabla J_i(\theta_{r-1})\|^2 \\
&\leq (1+\zeta)\mathbb{E}\left[\|\nabla J_i(\theta_{r-1})\|^2\right] + \left(1 + \frac{1}{\zeta}\right) L^2 \mathbb{E}\|\theta_r - \theta_{r-1}\|^2 \\
&\leq (1+\zeta)\mathbb{E}\left[\|\nabla J_i(\theta_{r-1})\|^2\right] + 2\left(1 + \frac{1}{\zeta}\right)(\lambda L)^2 \left(\mathbb{E}\|u_r - \nabla J(\theta_{r-1})\|^2 + \mathbb{E}\|\nabla J(\theta_r)\|^2\right) \\
&= (1+\zeta)\mathbb{E}\left[\|\nabla J_i(\theta_{r-1})\|^2\right] + 2\left(1 + \frac{1}{\zeta}\right)(\lambda L)^2 \left(\Sigma_{r-1} + \mathbb{E}\|\nabla J(\theta_r)\|^2\right)
\end{aligned}$$

By unrolling the recursive bound, we have

$$\mathbb{E}\left[\|\nabla J_i(\theta_r)\|^2\right] \leq (1+\zeta)^r \mathbb{E}\left[\|\nabla J_i(\theta_0)\|^2\right] + \frac{2}{\zeta}(\lambda L)^2 \sum_{j=0}^{r-1} \left(\Sigma_j + \mathbb{E}\left[\|\nabla J(\theta_j)\|^2\right]\right)(1+\zeta)^{r-j}$$

By choosing $\zeta = \frac{1}{r}$, we have

$$\mathbb{E}\left[\|\nabla J_i(\theta_r)\|^2\right] \leq e\mathbb{E}\left[\|\nabla J_i(\theta_0)\|^2\right] + 2e(r+1)(\lambda L)^2 \sum_{j=0}^{r-1} \left(\Sigma_j + \mathbb{E}\left[\|\nabla J(\theta_j)\|^2\right]\right) \qquad (5.32)$$

Summing up Eq. (5.31) over $r = 0, 1, \cdots, R-1$ and then averaing Eq. (5.31) over all $i \in N$, we have

$$\sum_{r=0}^{R-1} \mathcal{M}_r \leq \sum_{r=0}^{R-1} \mathbb{E}\left[\frac{289}{144}\eta^2(1-\beta)^2\|u_r\|^2 + 4\eta^2\beta^2 \frac{1}{N}\sum_{i=1}^{N} \|\nabla J_i(\theta_r)\|^2\right]$$

$$
\begin{aligned}
&\leq \sum_{r=0}^{R-1} \frac{289}{72}\eta^2(1-\beta)^2\left(\Sigma_{r-1}+\mathbb{E}[\|\nabla J(\theta_{r-1})\|^2]\right)\\
&\overset{(b)}{+} 4\eta^2\beta^2\sum_{r=0}^{R-1}\left(\frac{e}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla J_i\left(\theta_0\right)\right\|^2\right]+2e(r+1)(\lambda L)^2\sum_{j=0}^{r-1}\left(\Sigma_j+\mathbb{E}\left[\left\|\nabla J\left(\theta_j\right)\right\|^2\right]\right)\right)
\end{aligned}
\tag{5.33}
$$

$$
\begin{aligned}
&\leq \frac{289}{72}\eta^2(1-\beta)^2\sum_{r=0}^{R-1}\left(\Sigma_{r-1}+\mathbb{E}\left[\left\|\nabla J\left(\theta_{r-1}\right)\right\|^2\right]\right)\\
&+4\eta^2\beta^2\left(eRG_0+2e(\lambda LR)^2\sum_{r=0}^{R-2}\left(\Sigma_r+\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\right)\right)\\
&\overset{(c)}{\leq} \frac{\beta^2}{288eK^2\widetilde{L_1}^2}\sum_{r=-1}^{R-2}\left(\Sigma_r+\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\right)+4\eta^2\beta^2eRG_0.
\end{aligned}
$$

where (b) is due to the upper bound of $\mathbb{E}\left[\left\|\nabla J_i\left(\theta_r\right)\right\|^2\right]$ in Eq.(5.32) and (c) is due to the choice of η such that $\eta^2\left[\frac{289}{72}(1-\beta)^2+8e(\lambda\beta LR)^2\right]\leq\frac{\beta^2}{288eK^2\widetilde{L_1}^2}$. □

- Proof of Theorem 5

Theorem 7. *(Complete version of Theorem 5) Under Assumptions 4–6, by setting*

$$u_0=\frac{1}{NB}\sum_{i=1}^{N}\sum_{b=1}^{B}g_i\left(\tau_b^{(i)}|\theta_0\right)$$

with $\left\{\tau_b^{(i)}\right\}_{b=1}^{B}\overset{iid}{\sim}p^{(i)}(\tau|\theta_0)$ *and choosing* $\beta=\min\left\{1,\left(\frac{NK\bar{L}^2\Delta^2}{\sigma^4R^2}\right)^{1/3}\right\}$, $\lambda=\min\left\{\frac{1}{24\bar{L}},\sqrt{\frac{\beta NK}{162\bar{L}^2}}\right\}$, $B=\left\lceil\frac{K}{R\beta^2}\right\rceil$, *and*

$$\eta K\bar{L}\lesssim\min\left\{\left(\frac{\bar{L}\Delta}{G_0\lambda\bar{L}R}\right)^{1/2},\left(\frac{\beta}{N}\right)^{1/2},\left(\frac{\beta}{NK}\right)^{1/4}\right\}$$

in Algorithm 5, then the output of FEDSVRPG-M *after* R *rounds satisfies:*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\lesssim\left(\frac{\bar{L}\Delta\sigma}{NKR}\right)^{2/3}+\frac{\bar{L}\Delta}{R},\tag{5.34}$$

where $\bar{L} := \max\{L, \widetilde{L_1}, \widetilde{L_2}, \widetilde{L_3}\}$ *and* $L, \widetilde{L_1}, \widetilde{L_2}, \widetilde{L_3}$ *are defined in Proposition 1, Lemma 19 and Lemma 20, respectively.*

Proof. Based on Lemma 19, we have for any $r \geq 1$

$$\begin{aligned}
\Sigma_r &\leq (1-\frac{8\beta}{9})\Sigma_{r-1} + \frac{\widetilde{L_1}^2}{\beta}\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} + 18\lambda^2\frac{\widetilde{L_2}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \\
&\leq (1-\frac{8\beta}{9})\Sigma_{r-1} + 18\lambda^2\frac{\widetilde{L_2}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 + \frac{3\beta^2\sigma^2}{NK} \\
&+ \frac{\widetilde{L_1}^2}{\beta}\left(4eK^2\mathcal{M}_r + (16\eta^4K^4L^2 + 8\eta^2K)\right)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) \qquad (5.35)
\end{aligned}$$

where the last inequality is due to Lemma 20. When $r = 0$, we have

$$\Sigma_0 \leq (1-\beta)\Sigma_{-1} + \frac{3\beta^2\sigma^2}{NK} + \frac{\widetilde{L_1}^2}{\beta}\left(4eK^2\mathcal{M}_0 + (16\eta^4K^4L^2 + 8\eta^2K)\right)\beta^2\sigma^2$$

Summing up the above equation over r from 0 to $R-1$, we have

$$\begin{aligned}
\sum_{r=0}^{R-1}\Sigma_r &\leq \left(1-\frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\Sigma_r + \frac{18(\lambda\widetilde{L_2})^2}{NK}\mathbb{E}\left[\sum_{r=0}^{R-2}\left\|\nabla J\left(\theta_r\right)\right\|^2\right] + \frac{3\beta^2\sigma^2}{NK}R \\
&+ \frac{\widetilde{L_1}^2}{\beta}\left(4eK^2\sum_{r=0}^{R-1}\mathcal{M}_r + 8(\eta K)^2\left(2(\eta KL)^2 + \frac{1}{K}\right)\left(R\beta^2\sigma^2 + 2L^2\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\theta_r - \theta_{r-1}\right\|^2\right]\right)\right)
\end{aligned}$$

By incorporating Lemma 21 into the inequality above, we have

$$\begin{aligned}
\sum_{r=0}^{R-1}\Sigma_r &\leq \left(1-\frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\Sigma_r + \frac{18(\lambda\widetilde{L_2})^2}{NK}\mathbb{E}\left[\sum_{r=0}^{R-2}\left\|\nabla J\left(\theta_r\right)\right\|^2\right] + \frac{3\beta^2\sigma^2}{NK}R \\
&+ \frac{\widetilde{L_1}^2}{\beta}8(\eta K)^2\left(2(\eta KL)^2 + \frac{1}{K}\right)\left(R\beta^2\sigma^2 + 2L^2\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\theta_r - \theta_{r-1}\right\|^2\right]\right) \\
&+ \frac{\widetilde{L_1}^2}{\beta}4eK^2\left\{\frac{\beta^2}{288eK^2\widetilde{L_1}^2}\sum_{r=-1}^{R-2}\left(\Sigma_r + \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\right) + 4\eta^2\beta^2eRG_0\right\} \\
&\leq \left[1 - \frac{8\beta}{9} + \frac{\beta}{72} + \frac{32(\eta K\widetilde{L_1})^2}{\beta}(2(\eta KL)^2 + \frac{1}{K})(\lambda L)^2\right]\sum_{r=-1}^{R-2}\Sigma_r \\
&+ \left[\frac{18(\lambda\widetilde{L_2})^2}{NK} + \frac{32(\eta K\widetilde{L_1})^2}{\beta}(2(\eta KL)^2 + \frac{1}{K})(\lambda L)^2 + \frac{\beta}{72}\right]\sum_{r=-1}^{R-2}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]
\end{aligned}$$

$$+\left[8\beta\widetilde{L_1}^2(\eta K)^2(2(\eta KL)^2+\frac{1}{K})+\frac{3\beta^2}{NK}\right]R\sigma^2+16\beta(e\eta K\widetilde{L_1})^2RG_0 \tag{5.36}$$

Where the last inequality is derived by $\|\theta_r-\theta_{r-1}\|^2\leq 2\lambda^2\left(\|\nabla J\left(\theta_{r-1}\right)\|^2+\|u_r-\nabla J\left(\theta_{r-1}\right)\|^2\right)$.

We require the following inequalities to hold,

$$\begin{cases}\frac{32(\eta K\widetilde{L_1})^2}{\beta}(2(\eta KL)^2+\frac{1}{K})(\lambda L)^2\leq\frac{\beta}{18}\\ 8\widetilde{L_1}^2(\eta K)^2(2(\eta KL)^2+\frac{1}{K})\leq\frac{\beta^2}{NK}\\ \lambda\widetilde{L_2}\leq\sqrt{\frac{\beta NK}{162}}.\end{cases} \tag{5.37}$$

Then, we have that

$$\begin{aligned}\sum_{r=0}^{R-1}\Sigma_r&\leq\left[1-\frac{8\beta}{9}+\frac{\beta}{72}+\frac{\beta}{18}\right]\sum_{r=-1}^{R-2}\Sigma_r+\left[\frac{\beta}{9}+\frac{\beta}{18}+\frac{\beta}{72}\right]\sum_{r=-1}^{R-2}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right]\\&+\left[\frac{\beta^2}{NK}+\frac{3\beta^2}{NK}\right]R\sigma^2+16\beta(e\eta K\widetilde{L_1})^2RG_0\\&\leq(1-\frac{7\beta}{9})\sum_{r=-1}^{R-2}\Sigma_r+\frac{2\beta}{9}\sum_{r=-1}^{R-2}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right]+\frac{4R\beta^2\sigma^2}{NK}+16\beta(e\eta K\widetilde{L_1})^2RG_0\end{aligned}$$

After some rearrangement, we have

$$\sum_{r=0}^{R-1}\Sigma_r\leq\frac{9}{7\beta}\Sigma_{-1}+\frac{2}{7}\sum_{r=-1}^{R-2}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right]+\frac{36R\beta\sigma^2}{7NK}+\frac{144}{7}(e\eta K\widetilde{L_1})^2RG_0$$

Based on Lemma 17, we have

$$\frac{1}{\lambda}\mathbb{E}[J(\theta_R)-J(\theta_0)]\geq\frac{2}{7}\sum_{r=0}^{R-1}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right]-\frac{1}{35\beta}\Sigma_{-1}-\frac{39R\beta\sigma^2}{14NK}-\frac{78}{7}(e\eta K\widetilde{L_1})^2RG_0$$

Notice that $u_0=\frac{1}{NB}\sum_i\sum_{b=1}^{B}g_i\left(\tau_b^{(i)}|\theta_0\right)$ implies $\Sigma_{-1}=\mathbb{E}\|u_0-\nabla J(\theta_0)\|^2\leq\frac{\sigma^2}{NB}\leq\frac{\beta^2\sigma^2R}{NK}$.

Define $\bar{L}:=\max\{L,\widetilde{L_1},\widetilde{L_2},\widetilde{L_3}\}$ and after some rearrangement, we have

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right]&\lesssim\frac{\bar{L}\Delta}{\lambda\bar{L}R}+\frac{\Sigma_{-1}}{\beta R}+(\eta K\widetilde{L_1})^2G_0+\frac{\beta\sigma^2}{NK}\\&\overset{(a)}{\lesssim}\frac{\bar{L}\Delta}{\lambda\bar{L}R}+\frac{\beta\sigma^2}{NK}\end{aligned}$$

$$\overset{(b)}{\lesssim} \frac{\bar{L}\Delta}{R} + \frac{\bar{L}\Delta}{\sqrt{\beta NK}} + \frac{\beta\sigma^2}{NK}$$
$$\overset{(c)}{\lesssim} \frac{\bar{L}\Delta}{R} + \left(\frac{\bar{L}\Delta\sigma}{NKR}\right)^{2/3}$$

where (a) is due to the fact $\eta K\bar{L} \lesssim \left(\frac{\bar{L}\Delta}{G_0\lambda LR}\right)^{\frac{1}{2}}$; For (b), it holds because $\lambda\bar{L} \leq \min\{\frac{1}{24}, \sqrt{\frac{\beta NK}{162}}\}$; For (c), it holds because $\beta = \min\left\{1, \left(\frac{NK\bar{L}^2\Delta^2}{\sigma^4R^2}\right)^{1/3}\right\}$. □

5.8.4 Federated Hessian Aided Policy Gradient with Momentum

According to the updating rule of FEDHAPG-M, we can rewrite $\Lambda_{r,k}^{(i)}$ as

$$\Lambda_{r,k}^{(i)} = \left(\nabla \log p^{(i)} \left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha) \right)^T v_{r,k}^{(i)} \right) \nabla \Phi_i \left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha) \right) + \nabla^2 \Phi_i \left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha) \right) v_{r,k}^{(i)} \tag{5.38}$$

where $\Phi_i(\tau \mid \theta) = \sum_{h=0}^{H-1} \sum_{i=h}^{H-1} \lambda^i \mathcal{R}^{(i)}(s_i, a_i) \log \pi_\theta(a_h, s_h)$ and $v_{r,k}^{(i)} = \theta_{r,k}^{(i)} - \theta_{r-1}$. Note that

$$\mathbb{E}_{\alpha \sim U[0,1], \tau \sim p^{(i)}\left(\tau \mid \theta_{r,k}^{(i)}(\alpha)\right)} \left[\Lambda_{r,k}^{(i)} \right] = \nabla J \left(\theta_{r,k}^{(i)} \right) - \nabla J \left(\theta_{r-1} \right).$$

Moreover, we have $\Lambda_{r,k}^{(i)} := \hat{\nabla}_i^2 \left(\theta_{r,k}^{(i)}(\alpha), \tau_{r,k}^{(i)} \right) v_{r,k}^{(i)}$ where

$$\begin{aligned} \hat{\nabla}_i^2 \left(\theta_{r,k}^{(i)}(\alpha), \tau_{r,k}^{(i)} \right) =& \nabla \Phi_i \left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha) \right) \nabla \log p^{(i)} \left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha) \right)^T \\ &+ \nabla^2 \Phi_i \left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha) \right). \end{aligned}$$

and $\mathbb{E}_{\tau \sim p^{(i)}\left(\tau \mid \theta_{r,k}^{(i)}(\alpha)\right)} \left[\hat{\nabla}^2 \left(\theta_{r,k}^{(i)}(\alpha), \tau \right) \right] = \nabla^2 J_i \left(\theta_{r,k}^{(i)}(\alpha) \right)$.

Proposition 2. *(Lemma 4.1 in [175]) Under Assumption 4, we have for all θ and $i \in [N]$*

$$\left\| \hat{\nabla}_i^2(\theta, \tau) \right\|^2 \leq \frac{H^2 G^4 R_{\max}^2 + M^2 R^2}{(1-\lambda)^4} = \widetilde{L_4}^2.$$

where τ is a trajectory sampled according to $p^{(i)}(\tau|\theta)$.

Lemma 22. *If the step-size satisfies $\lambda \leq \sqrt{\frac{\beta NK}{72 \widetilde{L_4}^2}}$, we have*

$$\Sigma_r \leq (1 - \frac{8\beta}{\beta}) \Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta} \mathcal{D}_r + \frac{2\beta^2 \sigma^2}{NK} + \frac{8\lambda^2 \widetilde{L_4}^2}{NK} \mathbb{E} \left\| \nabla J(\theta_{r-1}) \right\|^2 \tag{5.39}$$

holds for $r \geq 1$. When $r = 0$, we have

$$\Sigma_r \leq (1 - \frac{8\beta}{\beta}) \Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta} \mathcal{D}_r + \frac{2\beta^2 \sigma^2}{NK}. \tag{5.40}$$

Proof.

$$
\begin{aligned}
\Sigma_r &= \mathbb{E}\left[\|u_{r+1} - \nabla J(\theta_r)\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{NK}\sum_{i,k}\beta w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}, \theta_{r,k}^{(i)}(\alpha)\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left[u_r + \Lambda_{r,k}^{(i)}\right] - \nabla J(\theta_r)\right\|^2\right] \\
&= \mathbb{E}\Bigg[\Bigg\|(1-\beta)(u_r - \nabla J(\theta_{r-1})) \\
&+ \frac{1}{NK}\sum_{i,k}\left\{\beta w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}, \theta_{r,k}^{(i)}(\alpha)\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left(\Lambda_{r,k}^{(i)} + \nabla J(\theta_{r-1}\right) - \nabla J(\theta_{r,k}^{(i)})\right\} \\
&+ \frac{1}{NK}\sum_{i,k}\left[\nabla J(\theta_{r,k}^{(i)}) - \nabla J(\theta_r)\right]\Bigg\|^2\Bigg] \\
&= (1-\beta)^2\Sigma_{r-1} + \underbrace{\mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\left[\nabla J(\theta_{r,k}^{(i)}) - \nabla J(\theta_r)\right]\right\|^2}_{H_1} \\
&+ \underbrace{2\mathbb{E}\left[\left\langle(1-\beta)(u_r - \nabla J(\theta_{r-1})), \frac{1}{NK}\sum_{i,k}\left[\nabla J_i(\theta_{r,k}^{(i)}) - \nabla J(\theta_r)\right]\right\rangle\right]}_{H_2} \\
&+ \underbrace{\mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\left\{\beta w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}, \theta_{r,k}^{(i)}(\alpha)\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left(\Lambda_{r,k}^{(i)} + \nabla J(\theta_{r-1})\right) - \nabla J(\theta_{r,k}^{(i)})\right\}\right\|^2}_{H_3}
\end{aligned}
\tag{5.41}
$$

To precede, we bound H_1 as

$$
\begin{aligned}
H_1 &= \mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\left[\nabla J(\theta_{r,k}^{(i)}) - \nabla J(\theta_r)\right]\right\|^2 \\
&\leq \frac{L^2}{NK}\sum_{i,k}\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 = L^2\mathcal{D}_r
\end{aligned}
\tag{5.42}
$$

Using Young's inequality to bound H_2, we have

$$
H_2 \leq \beta(1-\beta)^2\mathbb{E}\left\|u_r - \nabla J(\theta_{r-1})\right\|^2 + \frac{1}{\beta}\mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\nabla J_i(\theta_{r,k}^{(i)}) - \nabla J(\theta_r)\right\|^2
$$

$$\leq \beta(1-\beta)^2\Sigma_{r-1} + \frac{L^2}{\beta}\underbrace{\frac{1}{NK}\sum_{i,k}\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2}_{\mathcal{D}_r} \tag{5.43}$$

For H_3, we bound it as

$$\begin{aligned}
H_3 &= \mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\left\{\beta w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}, \theta_{r,k}^{(i)}(\alpha)\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left(\Lambda_{r,k}^{(i)} + \nabla J(\theta_{r-1})\right) - \nabla J(\theta_{r,k}^{(i)})\right\}\right\|^2 \\
&\leq 2\beta^2\frac{\sigma^2}{NK} + 2(1-\beta)^2\frac{1}{N^2K^2}\sum_{i,k}\mathbb{E}\left\|\Lambda_{r,k}^{(i)} + \nabla J(\theta_{r-1}) - \nabla J(\theta_{r,k}^{(i)})\right\|^2 \\
&\overset{(a)}{\leq} \frac{2\beta^2\sigma^2}{NK} + 2(1-\beta)^2\frac{1}{N^2K^2}\sum_{i,k}\mathbb{E}\left\|\Lambda_{r,k}^{(i)}\right\|^2 \\
&\overset{(b)}{=} \frac{2\beta^2\sigma^2}{NK} + 2(1-\beta)^2\frac{1}{N^2K^2}\sum_{i,k}\mathbb{E}\left\|\hat{\nabla}^2\left(\theta_{r,k}^{(i)}, \tau_{r,k}^{(i)}\right) v_{r,k}^{(i)}\right\|^2 \overset{(b)}{\leq} \frac{2\beta^2\sigma^2}{NK} + 2(1-\beta)^2\frac{1}{N^2K^2}\sum_{i,k}\widetilde{L_4}^2\mathbb{E}\left\|v_{r,k}^{(i)}\right\|^2 \\
&\leq \frac{2\beta^2\sigma^2}{NK} + 4(1-\beta)^2\widetilde{L_4}^2\underbrace{\frac{1}{NK}\sum_{i,k}\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2}_{\mathcal{D}_r} + 4(1-\beta)^2\frac{\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2
\end{aligned} \tag{5.44}$$

where we use the fact that $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$ for (a); for (b), it holds due to Proposition 2.

Plugging the upper bound of H_1 (Eq. (5.42)), H_2(Eq. (5.43)) and H_3 (Eq. (5.44))into Eq.(5.41), we have

$$\begin{aligned}
\Sigma_r &\leq (1-\beta)\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + 4\frac{\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2 \\
&= (1-\beta)\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + 4\frac{\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|u_r\right\|^2 \\
&\leq (1-\beta)\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + 8\frac{\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|u_r - \nabla J(\theta_{r-1})\right\|^2 + 8\frac{\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \\
&\overset{(a)}{\leq} (1 - \frac{8\beta}{\beta})\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + \frac{8\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2
\end{aligned} \tag{5.45}$$

where (a) is due to the choice of λ such that $\frac{8\lambda^2\widetilde{L_4}^2}{NK} \leq \frac{\beta}{9}$, which holds when $\lambda \leq \sqrt{\frac{\beta NK}{72\widetilde{L_4}^2}}$. □

Lemma 23. *(Bounding drift-term) If the local step-size satisfies* $\eta \leq \min\{\frac{L}{32e^2\widetilde{L_4}^2K}, \frac{1}{KL}\}$*, the drift-term can be upper bounded as:*

$$\mathcal{D}_r \leq 4eK^2\mathcal{M}_r + (16\eta^4K^4L^2 + 8\eta^2K)\left(\beta^2\sigma^2 + 2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right)$$

Proof. Define $c_{r,k}^{(i)} := -\eta\left(\nabla J_i(\theta_{r,k}^{(i)}) + (1-\beta)(u_r - \nabla J_i(\theta_{r-1}))\right)$. For any $1 \leq j \leq k-1 \leq K-2$, we have:

$$\begin{aligned}\mathbb{E}\left\|c_{r,j}^{(i)} - c_{r,j-1}^{(i)}\right\|^2 &\leq \eta^2L^2\mathbb{E}\left\|\theta_{r,j}^{(i)} - \theta_{r,j-1}^{(i)}\right\|^2 \\ &= \eta^2L^2\left(\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2 + \mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j}^{(i)} - \theta_{r,j-1}^{(i)} \mid \mathcal{F}_{r,j-1}^{(i)}\right]\right]\right). \end{aligned} \tag{5.46}$$

where we use the bias-variance decomposition in the last inequality. To precede, we bound the variance term as:

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j}^{(i)} - \theta_{r,j-1}^{(i)} \mid \mathcal{F}_{r,j-1}^{(i)}\right]\right] \\ &= \eta^2\mathbb{E}\Bigg\|\beta\left[w^{(i)}\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}, \theta_{r,j-1}^{(i)}(\alpha)\right)g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - \nabla J_i(\theta_{r,j-1}^{(i)})\right] \\ &+ (1-\beta)\left[\Lambda_{r,j-1}^{(i)} - \nabla J_i(\theta_{r,j-1}^{(i)}) + \nabla J_i(\theta_{r-1})\right]\Bigg\|^2 \\ &\leq 2\eta^2\beta^2\sigma^2 + 2\eta^2(1-\beta)^2\mathbb{E}\left\|\Lambda_{r,j-1}^{(i)} - \nabla J_i(\theta_{r,j-1}^{(i)}) + \nabla J_i(\theta_{r-1})\right\|^2 \\ &\overset{(a)}{\leq} 2\eta^2\beta^2\sigma^2 + 2\eta^2(1-\beta)^2\mathbb{E}\left\|\Lambda_{r,j-1}^{(i)}\right\|^2 \\ &\leq 2\eta^2\beta^2\sigma^2 + 2\eta^2(1-\beta)^2\mathbb{E}\left\|\hat{\nabla}_i^2\left(\theta_{r,j-1}^{(i)}, \tau_{r,j-1}^{(i)}\right)v_{r,j-1}^{(i)}\right\|^2 \\ &\leq 2\eta^2\beta^2\sigma^2 + 4\eta^2(1-\beta)^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_r\right\|^2 + 4\eta^2(1-\beta)^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2 \end{aligned} \tag{5.47}$$

where we use the fact that $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$ for (a). Plugging the upper bound of variance into Eq.(5.46), we have

$$\mathbb{E}\left\|c_{r,j}^{(i)} - c_{r,j-1}^{(i)}\right\|^2 \leq \eta^2L^2\left(\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2 + 2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_r\right\|^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right).$$

Then for any $1 \leq j \leq k-1 \leq K-2$, we have

$$\begin{aligned}
&\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 \leq (1+\frac{1}{q})\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2 + (1+q)\mathbb{E}\left\|c_{r,j}^{(i)} - c_{r,j-1}^{(i)}\right\|^2 \\
&\leq (1+\frac{2}{q})\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2 + (1+q)\eta^2 L^2\left(2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_r\right\|^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right)
\end{aligned} \tag{5.48}$$

where we use the fact that $\eta L \leq \frac{1}{K} \leq \frac{1}{q+1}$ and let $q = k-1$. By unrolling this recurrence, for any $1 \leq j \leq k-1 \leq K-2$, we have

$$\begin{aligned}
\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 &\leq (1+\frac{2}{k-1})^j\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + k\eta^2L^2\sum_{i=0}^{j-1}(2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2)\Pi_{j'=i+1}^{j-1}(1+\frac{2}{k-1}) \\
&\quad + k\eta^2L^2\sum_{s=0}^{j-1}(4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,s}^{(i)} - \theta_r\right\|^2)\Pi_{j'=s+1}^{j-1}(1+\frac{2}{k-1}) \\
&\leq (1+\frac{2}{k-1})^{k-1}\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + k\eta^2L^2\sum_{i=0}^{k-1}(2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2)(1+\frac{2}{k-1})^{k-1} \\
&\quad + k\eta^2L^2\sum_{j'=0}^{j-1}(4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j'}^{(i)} - \theta_r\right\|^2)(1+\frac{2}{k-1})^{k-1}
\end{aligned} \tag{5.49}$$

Based on the inequality $(1+\frac{2}{K-1}^{k-1}) \leq e^2 \leq 8$, we have

$$\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 \leq e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + 8k^2\eta^4L^2\left(2\beta^2\sigma^2 + 4\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) + 4e^2k\eta^4L^2\widetilde{L_4}^2\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)} - \theta_r\right\|^2 \tag{5.50}$$

By Lemma A.3, we have

$$\begin{aligned}
\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 &\leq 2\mathbb{E}\left\|\sum_{j=0}^{k-1}c_{r,j}^{(i)}\right\|^2 + 2\sum_{j=0}^{k-1}\mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j+1}^{(i)} - \theta_{r,j}^{(i)} \mid \mathcal{F}_{r,j}^{(i)}\right]\right] \\
&\overset{(a)}{\leq} 2k\sum_{j=0}^{k-1}\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 + 2\sum_{j=0}^{k-1}\left(2\beta^2\eta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j}^{(i)} - \theta_r\right\|^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right)
\end{aligned} \tag{5.51}$$

where (a) is due to Eq.(5.47). Plugging Eq.(5.50) into Eq.(5.51), we have

$$\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 \leq$$

$$2k\sum_{j=0}^{k-1}\left\{e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2+8k^2\eta^4L^2\left(2\beta^2\sigma^2+4\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)+4e^2k\eta^4L^2\widetilde{L_4}^2\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2\right\}$$
$$+2\sum_{j=0}^{k-1}\left(2\beta^2\eta^2\sigma^2+4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2+4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) \tag{5.52}$$

Summing up the above equation over $k=0,\cdots,K-1$, we have

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2 &\leq \sum_{k=0}^{K-1}\left\{2k^2e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2+16k^4\eta^4L^2\left(2\beta^2\sigma^2+4\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)\right\}\\
&+\sum_{k=0}^{K-1}8e^2k^2\eta^4L^2\widetilde{L_4}^2\sum_{j=0}^{k-1}\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2\\
&+\sum_{k=0}^{K-1}\left(4k\beta^2\eta^2\sigma^2+8k\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2+8\eta^2\widetilde{L_4}^2\sum_{j=0}^{k-1}\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2\right)\\
&\leq 2eK^3\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2+(8\eta^4K^5L^2+4\eta^2K^2)\left(\beta^2\sigma^2+2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)\\
&+K^2\sum_{k=0}^{K-1}8e^2\eta^4L^2\widetilde{L_4}^2\sum_{j=0}^{K-1}\sum_{j'=0}^{K-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2+\sum_{k=0}^{K-1}8\eta^2\widetilde{L_4}^2\sum_{j=0}^{K-1}\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2\\
&=2eK^3\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2+(8\eta^4K^5L^2+4\eta^2K^2)\left(\beta^2\sigma^2+2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)\\
&+(8e^2\eta^4K^4L^2\widetilde{L_4}^2+8\eta^2\widetilde{L_4}^2K)\sum_{j=0}^{K-1}\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2\end{aligned} \tag{5.53}$$

With the choice of step-size η satisfying $8e^2\eta^4K^4L^2\widetilde{L_4}^2+8\eta^2\widetilde{L_4}^2K\leq\frac{1}{2}$, after some rearrangement, we have

$$\frac{1}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2\leq 2eK^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2+(8\eta^4K^4L^2+4\eta^2K)\left(\beta^2\sigma^2+2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) \tag{5.54}$$

In summary, we can bound the drift-term as

$$\mathcal{D}_r\leq 4eK^2\mathcal{M}_r+(16\eta^4K^4L^2+8\eta^2K)\left(\beta^2\sigma^2+2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) \tag{5.55}$$

□

Lemma 24. *If* $\lambda L \leq \frac{1}{24}$ *and* $\eta^2 \left[\frac{289}{72}(1-\beta)^2 + 8e(\lambda\beta LR)^2\right] \leq \frac{\beta^2}{288eK^2\left(2L^2+\widetilde{4L_4}^2\right)}$, *we have*

$$\sum_{r=0}^{R-1}\mathcal{M}_r = \frac{1}{N}\sum_{r=0}^{R-1}\sum_{i=1}^{N}\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 \leq \frac{\beta^2}{288eK^2\left(2L^2+\widetilde{4L_4}^2\right)}\sum_{r=-1}^{R-2}\left(\Sigma_r + \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\right) + 4\eta^2\beta^2 eRG_0. \tag{5.56}$$

where $G_0 := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla J_i\left(\theta_0\right)\right\|^2\right]$.

Proof. The proof is the same as that of Lemma 21. □

- Proof of Theorem 6

Theorem 8. *(Complete version of Theorem 6) Under Assumption 4–6, by setting* $u_0 = \frac{1}{NB}\sum_{i=1}^{N}\sum_{b=1}^{B} g_i\left(\tau_b^{(i)}|\theta_0\right)$ *with* $\left\{\tau_b^{(i)}\right\}_{b=1}^{B} \overset{iid}{\sim} p^{(i)}(\tau|\theta_0)$ *and choosing* $\beta = \min\left\{1, \left(\frac{NK\hat{L}^2\Delta^2}{\sigma^4R^2}\right)^{1/3}\right\}$, $\lambda = \min\left\{\frac{1}{24\hat{L}}, \sqrt{\frac{\beta NK}{72\hat{L}^2}}\right\}$, $B = \left\lceil\frac{K}{R\beta^2}\right\rceil$, *and*

$$\eta K\hat{L} \lesssim \min\left\{\left(\frac{\hat{L}\Delta}{G_0\lambda\hat{L}R}\right)^{1/2}, \left(\frac{\beta}{N}\right)^{1/2}, \left(\frac{\beta}{NK}\right)^{1/4}\right\}$$

in Algorithm 6, then the output of FEDHAPG-M *after* R *rounds satisfies*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] \lesssim \left(\frac{\hat{L}\Delta\sigma}{NKR}\right)^{2/3} + \frac{\hat{L}\Delta}{R} \tag{5.57}$$

where $\hat{L} := \sqrt{2L^2 + \widetilde{4L_4}^2}$ *and* $L, \widetilde{L_4}$ *are defined in Proposition 1 and Proposition 2, respectively.*

Proof. Based on Lemma 22, we have for any $r \geq 1$

$$\Sigma_r \leq (1-\frac{8\beta}{\beta})\Sigma_{r-1} + \frac{2L^2+4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + \frac{8\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2$$

$$\leq (1-\frac{8\beta}{\beta})\Sigma_{r-1} + \frac{2\beta^2\sigma^2}{NK} + \frac{8\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \tag{5.58}$$

$$+ \frac{2L^2+4\widetilde{L_4}^2}{\beta}\left[4eK^2\mathcal{M}_r + (16\eta^4K^4L^2+8\eta^2K)\left(\beta^2\sigma^2 + 2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)\right] \tag{5.59}$$

where the last inequality is due to Lemma 23. When $r=0$, we have

$$\Sigma_0 \leq (1-\beta)\Sigma_{-1} + \frac{2\beta^2\sigma^2}{NK} + \frac{2L^2+4\widetilde{L_4}^2}{\beta}\left[4eK^2\mathcal{M}_0 + (16\eta^4K^4L^2+8\eta^2K)\right]\beta^2\sigma^2$$

Summing up the above equation over r from 0 to $R-1$, we have

$$\begin{aligned}\sum_{r=0}^{R-1}\Sigma_r \leq& \left(1-\frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\Sigma_r + \frac{8(\lambda\widetilde{L_4})^2}{NK}\mathbb{E}\left[\sum_{r=0}^{R-2}\left\|\nabla J\left(\theta_r\right)\right\|^2\right] + \frac{2\beta^2\sigma^2}{NK}R \\ &+ \frac{2L^2+4\widetilde{L_4}^2}{\beta}\left[4eK^2\sum_{r=0}^{R-1}\mathcal{M}_r + 8(\eta K)^2(2(\eta KL)^2+\frac{1}{K})\right]\left(R\beta^2\sigma^2 + 2L^2\sum_{r=0}^{R-1}\mathbb{E}\left\|\theta_r-\theta_{r-1}\right\|^2\right)\end{aligned}$$

By incorporating Lemma 24 into the inequality above, we have

$$\begin{aligned}\sum_{r=0}^{R-1}\Sigma_r \leq& \left(1-\frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\Sigma_r + \frac{8(\lambda\widetilde{L_4})^2}{NK}\mathbb{E}\left[\sum_{r=0}^{R-2}\left\|\nabla J\left(\theta_r\right)\right\|^2\right] + \frac{2\beta^2\sigma^2}{NK}R \\ &+ \frac{2L^2+4\widetilde{L_4}^2}{\beta}8(\eta K)^2\left(2(\eta KL)^2+\frac{1}{K}\right)\left(R\beta^2\sigma^2 + 2L^2\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\theta_r-\theta_{r-1}\right\|^2\right]\right) \\ &+ \frac{2L^2+4\widetilde{L_4}^2}{\beta}4eK^2\left\{\frac{\beta^2}{288eK^2\left(2L^2+4\widetilde{L_4}^2\right)}\sum_{r=-1}^{R-2}\left(\Sigma_r + \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\right) + 4\eta^2\beta^2eRG_0\right\} \\ &\leq \left[1-\frac{8\beta}{9}+\frac{\beta}{72}+\frac{32(\eta K)^2(2L^2+4\widetilde{L_4})^2}{\beta}(2(\eta KL)^2+\frac{1}{K})(\lambda L)^2\right]\sum_{r=-1}^{R-2}\Sigma_r \\ &+ \left[\frac{8(\lambda\widetilde{L_4})^2}{NK}+\frac{32(\eta K)^2(2L^2+4\widetilde{L_4})^2}{\beta}(2(\eta KL)^2+\frac{1}{K})(\lambda L)^2+\frac{\beta}{72}\right]\sum_{r=-1}^{R-2}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] \\ &+ \left[8\beta\left(2L^2+4\widetilde{L_4}^2\right)(\eta K)^2(2(\eta KL)^2+\frac{1}{K})+\frac{2\beta^2}{NK}\right]R\sigma^2 + 16\beta\left(2L^2+4\widetilde{L_4}^2\right)(e\eta K)^2RG_0\end{aligned} \tag{5.60}$$

Where the last inequality is derived by $\left\|\theta_r-\theta_{r-1}\right\|^2 \leq 2\lambda^2\left(\left\|\nabla J\left(\theta_{r-1}\right)\right\|^2 + \left\|u_r-\nabla J\left(\theta_{r-1}\right)\right\|^2\right)$.

Note that $\hat{L}^2 = 2L^2 + \widetilde{4L_4}^2$. We require the following inequalities to hold,

$$\begin{cases} \frac{32(\eta K\hat{L})^2}{\beta}(2(\eta KL)^2 + \frac{1}{K})(\lambda L)^2 \leq \frac{\beta}{18} \\ 8\hat{L}^2(\eta K)^2(2(\eta KL)^2 + \frac{1}{K}) \leq \frac{\beta^2}{NK} \\ \lambda\hat{L} \leq \sqrt{\frac{\beta NK}{72}}. \end{cases} \tag{5.61}$$

Then, we have that

$$\begin{aligned} \sum_{r=0}^{R-1}\Sigma_r &\leq \left[1 - \frac{8\beta}{9} + \frac{\beta}{72} + \frac{\beta}{18}\right]\sum_{r=-1}^{R-2}\Sigma_r + \left[\frac{\beta}{9} + \frac{\beta}{18} + \frac{\beta}{72}\right]\sum_{r=-1}^{R-2}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right] \\ &\quad + \left[\frac{\beta^2}{NK} + \frac{2\beta^2}{NK}\right]R\sigma^2 + 16\beta(e\eta K\widetilde{L_1})^2 RG_0 \\ &\leq (1 - \frac{7\beta}{9})\sum_{r=-1}^{R-2}\Sigma_r + \frac{2\beta}{9}\sum_{r=-1}^{R-2}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right] + \frac{4R\beta^2\sigma^2}{NK} + 16\beta(e\eta K\hat{L})^2 RG_0 \end{aligned}$$

After some rearrangement, we have

$$\sum_{r=0}^{R-1}\Sigma_r \leq \frac{9}{7\beta}\Sigma_{-1} + \frac{2}{7}\sum_{r=-1}^{R-2}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right] + \frac{36R\beta\sigma^2}{7NK} + \frac{144}{7}(e\eta K\hat{L})^2 RG_0$$

Based on Lemma 17, we have

$$\frac{1}{\lambda}\mathbb{E}[J(\theta_R) - J(\theta_0)] \geq \frac{2}{7}\sum_{r=0}^{R-1}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right] - \frac{1}{35\beta}\Sigma_{-1} - \frac{39R\beta\sigma^2}{14NK} - \frac{78}{7}(e\eta K\hat{L})^2 RG_0$$

Notice that $u_0 = \frac{1}{NB}\sum_i\sum_{b=1}^{B} g_i\left(\tau_b^{(i)}|\theta_0\right)$ implies $\Sigma_{-1} = \mathbb{E}\|u_0 - \nabla J(\theta_0)\|^2 \leq \frac{\sigma^2}{NB} \leq \frac{\beta^2\sigma^2 R}{NK}$. After some rearrangement, we have

$$\begin{aligned} \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\|\nabla J\left(\theta_r\right)\|^2\right] &\lesssim \frac{\hat{L}\Delta}{\lambda\hat{L}R} + \frac{\Sigma_{-1}}{\beta R} + (\eta K\hat{L})^2 G_0 + \frac{\beta\sigma^2}{NK} \\ &\overset{(a)}{\lesssim} \frac{\hat{L}\Delta}{\lambda\hat{L}R} + \frac{\beta\sigma^2}{NK} \\ &\overset{(b)}{\lesssim} \frac{\hat{L}\Delta}{R} + \frac{\hat{L}\Delta}{\sqrt{\beta NK}} + \frac{\beta\sigma^2}{NK} \\ &\overset{(c)}{\lesssim} \frac{\hat{L}\Delta}{R} + \left(\frac{\hat{L}\Delta\sigma}{NKR}\right)^{2/3} \end{aligned}$$

where (a) is due to the fact $\eta K \hat{L} \lesssim \left(\frac{\hat{L}\Delta}{G_0 \lambda \tilde{L} R}\right)^{\frac{1}{2}}$; For (b), it holds because $\lambda \hat{L} \leq \min\{\frac{1}{24}, \sqrt{\frac{\beta NK}{72}}\}$; For (c), it holds because $\beta = \min\left\{1, \left(\frac{NK\hat{L}^2\Delta^2}{\sigma^4 R^2}\right)^{1/3}\right\}$. □

5.8.5 Additional Experiments and Implementation Details

a) Details of Tabular Case.

Random MDPs consist of $N = 20$ environments. In each MDP, both the state and action spaces have a size of 5. We choose $R_{\max} = 1$. The discounted factor λ is 0.9. The state transition kernel is generated randomly (element-wisely Bernoulli distributed). The number of local updates is set as $K = 32$. Additionally, the local step-size is chosen to be $\eta = 0.05$.

b) Details of DRL Case

Experiments Setup We adopted a local step-size of 0.75 and a global step-size of 0.6. We experimented with momentum coefficients, denoted as β, ranging from 0.2, 0.5, to 0.8. Additional parameters were set as follows: $N = 5$, $R_{\max} = 120$, and $K = 10$. All experiments are conducted in a host machine that is equipped with an Intel(R) Core(TM) i9-10900X CPU that operates at a base frequency of 3.70GHz. This processor boasts 10 cores and 20 threads, with a maximum turbo frequency of 4300 MHz. It has a total of 125GB of RAMA and 4 NVIDIA GeForce RTX 2080 GPU, compatible with CUDA Version 11.0. The source code is provided in the supplementary materials.

Experimental Environments The **CartPole** environment, often referred to as the "inverted pendulum" problem, is a classic task in the field of reinforcement learning. In this environment, a pole is attached to a cart, which moves along a frictionless track. The primary objective is to balance the pole upright by moving the cart left or right, without the pole falling over or the cart moving too far off the track. At the start of the experiment, the pole is slightly tilted, and the goal is to prevent it from falling over by applying force to the cart. The environment provides a reward at each time step for keeping the pole upright. The episode ends when the pole tilts beyond a certain

angle from the vertical or the cart moves out of a defined boundary on the track.

The **HalfCheetah** environment is another popular benchmark in reinforcement learning, especially within the continuous control domain. It's designed to emulate the challenges of agile and efficient locomotion. The agent in this environment is a two-dimensional, simplified robotic model inspired by the anatomy of a cheetah, albeit it only represents the "half" body, often from the waist down, thus the name "HalfCheetah." The robotic agent comprises multiple joints and segments, representing the limbs of the cheetah. The primary goal in the HalfCheetah environment is to control and coordinate the movements of these joints to make the robot run as fast as possible on a flat surface. At each timestep, the agent receives a reward based on how fast it's moving forward minus a small cost for the actions taken (to prevent erratic behaviors). The challenge lies in efficiently propelling the HalfCheetah forward, optimizing for speed and stability.

The **Walker** environment is a more complex task that simulates a bipedal agent which needs to learn to walk. Unlike CartPole, where the challenge is to balance a single pole, the Walker environment involves controlling multiple joints and limbs of a simulated agent to achieve locomotion. The agent receives rewards based on its forward movement and is penalized for falling or performing awkward movements. More information about these environments can be found in **(author?)** [197].

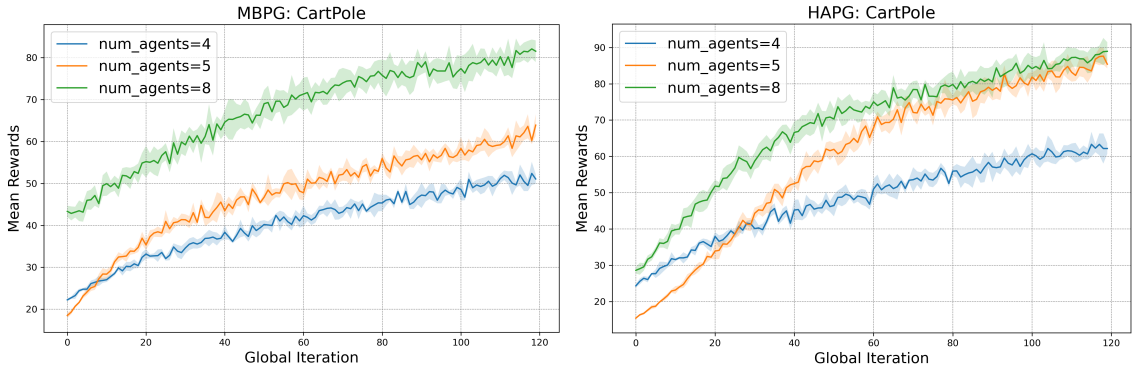


Figure 5.2: Mean rewards over global iterations for the CartPole task under different values of N (agent number): (**Left**): FEDSVRPG-M; (**Right**): FEDHAPG-M. The shaded areas represent the variance of rewards. Complying with theory, increasing N will increase the rewards. For both algorithms, the local step-size η is 0.05, global step-size λ satisfies $\lambda = \eta K$ and the number of local updates K is 10.

Ablation Study on Agent Number N. We further provide the ablation study of our FEDSVRPG-M and FEDHAPG-M algorithms on N (agent number). With large N, environment heterogeneity level increases. We choose $\beta = 0.2$ to train policies in the ablation study. Figure 5.2 illustrates how different N values ($N = 4, 5$, and 8) influence the average rewards in the CartPole task as the number of iterations increases. We find that all policies with larger N values report better performance throughout the iterations. The color-shaded regions indicate the variance in rewards. Such phenomenon observed in Figure 5.2 complies with our theoretical analysis about linear speedup.

Experiments on FEDHAPG-M Algorihtm The table 5.3 presents the mean testing rewards and variances for the policies trained by the FedHAPG-M algorithm with various β values and the baseline algorithm [79] across two tasks: CartPole and Walker. For both tasks, the FedHAPG-M algorithm with $\beta = 0.8$ outperforms the other configurations in terms of mean rewards.

Table 5.3: Mean Rewards and Variances of Policy Trained by FEDHAPG-M with Different Beta Values and Baseline Algorithm

Algorithms	CartPole	Walker
FEDHAPG-M with $\beta = 0.2$	83.46 ± 7.92	130.93 ± 7.72
FEDHAPG-M with $\beta = 0.5$	86.54 ± 12.99	287.14 ± 72.26
FEDHAPG-M with $\beta = 0.8$	**86.58** ± 11.21	**301.57** ± 28.04
Baseline algorithm	85.92 ± 12.17	299.69 ± 3.02

Chapter 6

Federated Learning for Control

6.1 Introduction

In recent years, there has been significant progress in the application of model-free reinforcement learning (RL) methods to fields such as video games [138] and robotic manipulation [107, 159, 196]. Although RL has shown impressive results in simulation, it often suffers from poor sample complexity, thereby limiting its effectiveness in real-world applications [42]. To resolve the sample complexity issue and accelerate the learning process, federated learning (FL) has emerged as a popular paradigm [96, 132], where multiple similar agents collaboratively learn a common model without sharing their raw data. The incentive for collaboration arises from the fact that these agents are "similar" in some sense and hence end up learning a "superior" model than if they were to learn alone. In the RL setting, Federated Reinforcement Learning (FRL) aims to learn a common value function [192] or produce a better policy from multiple RL agents interacting with similar environments. In the recent survey paper [158], FRL has empirically shown great success in reducing the sample complexity for autonomous driving [116, 139], IoT devices [117], and resource management in networking [104, 237].

Lately, there has been a lot of interest in applying RL techniques to classical control problems such as the Linear Quadratic Regulaor (LQR) problem [6]. In the standard control setting, the dynamical model of the system is known and one seeks to obtain a controller that stabilizes the closed-loop system and provides optimal performance. RL approaches such as policy gradient [189, 223] (which we pursue here) differ in that they are "model-free", i.e., a control policy is obtained despite not having access to the model of the dynamics. Despite the lack of convexity in even simple problems, policy gradient (PG) methods have been shown to be globally convergent for certain structured settings such as the LQR problem [50]. While this is promising, a major challenge in applying PG methods is that in general, one does not have access to *exact deterministic* policy gradients. Instead, one relies on estimating such gradients via sampling based approaches. This typically leads to noisy gradients that can suffer from high variance. As such, reducing the variance in PG estimates to achieve "good performance" may end up requiring several samples.

Motivation. The main premise of this paper is to draw on ideas from the FL literature to alleviate the high sample-complexity burden of PG methods [4, 124, 219], with the focus being on model-free control. As a motivating example, consider a fleet of identical robots produced by the same manufacturer. Each robot can collect data from its own dynamics and learn its own optimal policy using, for instance, PG methods. Since the fleet of robots shares similar dynamics, and more data can potentially lead to improved policy performance (via more accurate PG estimates), it is natural to ask: *Can a robot accelerate the process of learning its own optimal policy by leveraging the data of the other robots in the fleet?* The answer is not as obvious as one might expect since in reality, it is unlikely that any two robots will have *exactly* the same underlying dynamics, i.e., *heterogeneity in system dynamics is inevitable.* The presence of such heterogeneity makes the question posed above both interesting and non-trivial. In particular, when the heterogeneity across agents' dynamics is large, leveraging data from other agents might degrade the performance of a

single agent. Indeed, large heterogeneity may make it impossible to learn a common stabilizing policy[1]. Moreover, even when such a stabilizing policy exists, it may deviate from each agent's local optimal policy, rendering poor performance and discouraging participation in the FL process. Thus, to understand whether more data[2] helps or hurts, it is crucial to characterize the effects of heterogeneity in the federated control setting.

With this aim in mind, we study a multi-agent *model-free* LQR problem based on policy gradient methods. Specifically, there are M agents in our setup, each with its own *distinct yet similar* linear time-invariant (LTI) dynamics. Inspired by the typical objective in FL, our goal is to find a common policy which can minimize the average of the LQR costs of all the agents. With this setup, we seek to answer the following questions.

Q1. *Is this common policy stabilizing for all the systems? If so, under what conditions?*

Q2. *How far is the learned common policy from each agent's locally optimal policy?*

Q3. *Can an agent use the common policy as an initial guess to fine-tune and learn its own optimal policy faster (i.e., with fewer overall samples) than if it acted alone?*

Challenges: There are several challenges to answering the above questions. First, even for the single agent setting, the policy gradient-based LQR problem is non-convex, and requires a fairly intricate analysis [50]. Second, a key distinction relative to standard federated supervised learning stems from the need to maintain *stability* – this problem is amplified in the heterogeneous multi-agent scenario we consider. It remains an open problem to design an algorithm ensuring that policies are simultaneously stabilizing for each distinct system. Third, to reduce the communication cost, FL algorithms rely on the agents performing multiple local update steps between successive

[1]See Section 6.9.3 for more details on the underlying intuition and necessity behind the low heterogeneity regime.

[2]In accordance with both FL & FRL frameworks, the agents in our problem do not exchange their private data (e.g., rewards, states, etc.). Instead, each agent only transmits its policy gradient.

communication rounds. When agents have non-identical loss functions, these local steps lead to a "client-drift" effect where each agent drifts towards its own local minimizer [21, 23]. While several works in FL have investigated this phenomenon [2, 64, 84, 88, 90, 100, 110, 113, 135, 136, 150, 217], *the effect of "client-drift" on stability remains completely unexplored*. Unless accounted for, such drift effects can potentially produce unstable controllers for some systems.

Our Contributions: In response to the above challenges, we propose a policy gradient method called `FedLQR` to solve the (model-based *and* model-free) federated LQR problem, and provide a rigorous finite-time analysis of its performance that accounts for the interplay between system heterogeneity, multiple local steps, client-drift effects, and stability. Our specific contributions in this regard are as follows.

• **Iterative stability guarantees.** We show via a careful inductive argument that under suitable requirements on the level of heterogeneity across systems, the learning rate schedule can be designed to ensure that `FedLQR` provides a stabilizing controller at every iteration for *all* systems. Theorem 9 provides a proof in the model-based setting, and Theorem 10 provides the model-free result.

• **Bounded policy gradient heterogeneity in the LQR problem.** We prove in Lemma 27 that, for each pair of agents $i, j \in [M]$, the policy gradient direction (in the model-based setting) of agent i is close to that of agent j, if their dynamics are similar (i.e., Definition 1). *This is the first result to observe and characterize this bounded gradient heterogeneity phenomenon in the multi-agent LQR setting.*

• **Quantifying the gap between the `FedLQR`'s output and each system's optimal policy.** Building on Lemma 27, we prove that when the agents' dynamics are similar, the common policy returned by `FedLQR` is close to each agent's optimal policy; see Theorem 9. In other words, we can leverage the federated formulation to help each agent find its own optimal policy up to some accuracy depending on the level of heterogeneity. Our work is the first to provide a result of this

flavor.

• **Linear speedup.** As our main contribution, we prove that in the model-free setting, `FedLQR` converges to a solution that is in a neighborhood of each agent's optimal policy, *using M-times fewer samples* relative to when each agent just uses its own data (see Theorem 10). The radius of this neighborhood captures the level of heterogeneity across the agents' dynamics. *The key implication of this result is that in a low-heterogeneity regime, `FedLQR` (in the model-free setting) reduces the sample-complexity by a factor of M w.r.t. the centralized setting [50, 129], highlighting the benefit of collaboration.*[3] Simply put, `FedLQR` enables each agent to *quickly* find an *approximate* locally optimal policy; as in standard FL [31], the agent can then use this policy as an initial guess to *fine-tune* based on its own data.

In summary, we provide a new theoretical framework that quantitatively *characterizes the interplay between the price of heterogeneity and the benefit of collaboration* for model-free control.

Refer to [193] for all proofs in this Chapter.

6.2 Background and Preliminaries

6.2.1 Related Work

There has been a line of work [50, 65, 71, 80, 82, 129, 140] that explores various RL algorithms for solving the model-free LQR problem. However, their analysis is limited to the single-agent setting. Most recently, [163] solves the model-free LQR tracking problem in a federated manner and achieves a linear convergence speedup with respect to the number of agents. However, they consider a simplified setting where all agents follow the *same* dynamics. As such, the stability analysis of [163] follows from arguments for the centralized setting. In sharp contrast, to establish

[3]Throughout this paper, we use the terms "centralized" and "single-agent" interchangeably.

the linear speedup for `FedLQR`, we need to address the key technical challenges arising from the effect of heterogeneity and local steps on the stability of distinct systems. This requires new analysis tools that we develop. For related work on multi-agent RL (that do not specifically look at the control setting) we point the reader to [118, 244] and the references therein. Below we highlight the topics and the corresponding relevant work related to our problem setting.

- **Policy Gradient (PG):** The policy gradient (PG) approach is a fundamental component of the success of reinforcement learning (RL) and plays a crucial role in policy optimization (PO). This approach directly optimizes the policy to improve system-level performances through gradient ascent steps. The concept of policy optimization has been influential in RL [189] with some well-known algorithms such as `REINFORCE` [223], trust-region policy optimization `TRPO` [171], actor-critic methods [94], and proximal policy optimization `PPO` [172]. We highlight an important difference between standard MDP models and control models in RL. In control, one requires the policy to provide closed-loop stability, i.e., all trajectories of the system must converge for a given policy. In contrast, convergence in the MDP setting requires irreducibly and aperiodicity properties that are assumed *before* a policy is selected. As a result, the control task is significantly more challenging.

  The extensive body of literature on policy optimization for reinforcement learning (RL) and its adaptability to the model-free setting paves the way for leveraging policy gradient methods in the pursuit of learning optimal control policies for classical control problems [74, 152]. Despite the non-convex nature of the formulation involved in policy gradient methods, recent work [50, 65, 71, 80, 82, 102, 129, 140, 152] has demonstrated global convergence in solving the model-free LQR problem via policy gradient methods. This convergence is achieved due to certain properties of the quadratic cost function inherent in the LQR problem as introduced

in [50]. In contrast to the aforementioned work, which exclusively focus on the centralized control setting, our paper offers convergence guarantees for the multi-agent setting. In this context, each agent follows similar, but not identical, dynamics, thereby distinguishing it from the simpler scenario in [163].

- **Model-free Linear Quadratic Control:** The linear quadratic regulator (LQR) problem is a well-studied classical control problem that has gained significant attention due to its wide applicability and its role as a baseline for more complex control strategies [6]. Recently, to address the non-convex nature of the policy gradient LQR, [187] has proposed convexifying the corresponding optimal control problem to efficiently solve the model-based LQR problem via policy gradient. Furthermore, the model-free LQR has attracted considerable interest after [50] provided guarantees on the global convergence of policy gradient methods for both model-based and model-free LQR settings. This breakthrough paved the way for subsequent works [65, 71, 80, 82, 102, 129, 140, 152] that analyze convergence guarantees and sample complexity in the context of the model-free LQR problem. Notably, [35] characterizes the sample complexity of the LQR problem.

  Another line of work explores certainty equivalent control [131, 178], providing regret bounds to demonstrate the quality of the designed linear quadratic regulator in terms of the accuracy of the estimated system model. However, the key distinction between these works and the present paper lies in the consideration of multiple and heterogeneous systems. Moreover, [131, 178] use the regret framework, which is different from the PAC learning-based framework [51] exploited in our paper.

6.2.2 Notation

Given a set of matrices $\{S^{(i)}\}_{i=1}^{M}$, we denote $||S||_{\max} := \max_i ||S^{(i)}||$, and $||S||_{\min} := \min_i ||S^{(i)}||$. All vector norms are Euclidean and matrix norms are spectral, unless otherwise stated.

6.3 Problem Formulation

Classical control approaches aim to design optimal controllers from a well-defined dynamical system model. The model-based LQR is a well-studied problem that admits a convex solution. In this work, we consider the LQR problem but in the *model-free setting*. Moreover, we consider a *federated model-free LQR problem* in which there are M agents, each with their own distinct but "similar' dynamics. Our goal is to collaboratively learn an optimal controller that minimizes an average quadratic cost. We seek to characterize the optimality of our solution as a function of the "difference" across the agent's dynamics. In what follows, we formally describe our problem of interest.

Federated LQR: Consider a system with M agents. Associated with each agent is a linear time-invariant (LTI) dynamical system of the form

$$x_{t+1}^{(i)} = A^{(i)} x_t^{(i)} + B^{(i)} u_t^{(i)}, \quad x_0^{(i)} \sim \mathcal{D}, \quad i = 1, \dots, M,$$

with $A^{(i)} \in \mathbb{R}^{n_x \times n_x}$, $B^{(i)} \in \mathbb{R}^{n_x \times n_u}$. We assume each initial state $x_0^{(i)}$ is randomly generated from the same distribution $\mathcal{D}$. In the single-agent setting, the optimal LQR control policy $u_t^{(i)} = -K_i^* x_t^{(i)}$ for each agent is given by the solution to

$$\begin{aligned} K_i^* = \arg\min_K \left\{ C^{(i)}(K) := \mathcal{E}\left[\sum_{t=0}^{\infty} x_t^{(i)\top} Q x_t^{(i)} + u_t^{(i)\top} R u_t^{(i)} \right] \right\} \\ \text{s.t.} \quad x_{t+1}^{(i)} = A^{(i)} x_t^{(i)} + B^{(i)} u_t^{(i)},\ u_t^{(i)} = -K x_t^{(i)},\ x_0^{(i)} \sim \mathcal{D}, \end{aligned} \tag{6.1}$$

where $Q \in \mathcal{R}^{n_x \times n_x}$ and $R \in \mathcal{R}^{n_u \times n_u}$ are known positive definite matrices. In our federated setting, the objective is to find an optimal common policy $\{u_t\}_{t=0}^{\infty}$ to minimize the average cost of all the agents $C_{\text{avg}}(K) := \frac{1}{M}\sum_{i=1}^{M} C^{(i)}(K)$ *without knowledge of the system dynamics*, i.e., $(A^{(i)}, B^{(i)})$. Classical results [6] from optimal control theory show that, given the system matrices $A^{(i)}$, $B^{(i)}$, Q and R, the optimal policy can be written as a linear function of the current state. Thus, we consider a common policy of the form $u_t^{(i)} = -Kx_t^{(i)}$. The objective of the federated LQR problem can be written as:

$$K^* = \arg\min_K \left\{ C_{\text{avg}}(K) := \frac{1}{M}\sum_{i=1}^{M} \mathcal{E}\left[\sum_{t=0}^{\infty} x_t^{(i)\top} Q x_t^{(i)} + u_t^{(i)\top} R u_t^{(i)}\right]\right\}$$
$$\text{s.t.} \quad x_{t+1}^{(i)} = A^{(i)}x_t^{(i)} + B^{(i)}u_t^{(i)}, \quad u_t^{(i)} = -Kx_t^{(i)}, x_0^{(i)} \sim \mathcal{D}. \tag{6.2}$$

The rationale for finding K^* is as follows. Intuitively, when all agents have similar dynamics, K^* will be close to each K_i^*. Thus, K^* will serve to provide a good common initial guess from which each agent i can then fine-tune/personalize (using only its own data) to converge *exactly* to its own locally optimal controller K_i^* . The key here is that the initial guess K^* can be obtained *quickly* by using the *collective data* of all the agents. We will formalize this intuition in Theorem 10.

We make the standard assumption that for each agent, $(A^{(i)}, B^{(i)})$ is stabilizable. In addition, we make the following assumption on the distribution of the initial state:

Assumption 7. *Let* $\mu := \sigma_{\min}\left(\mathbb{E}_{x_0^{(i)}\sim\mathcal{D}} x_0^{(i)} x_0^{(i)\top}\right)$ *and assume* $\mu > 0$. *For each* $i \in [M]$, *the initial state* $x_0^{(i)} \sim \mathcal{D}$ *and distribution* $\mathcal{D}$ *satisfy*

$$\mathcal{E}_{x_0^{(i)}\sim\mathcal{D}}[x_0^{(i)}] = 0, \;\; \mathcal{E}_{x_0^{(i)}\sim\mathcal{D}}[x_0^{(i)} x_0^{(i)\top}] \succ \mu\mathcal{I}_{d_x}, \text{ and } \|x_0^{(i)}\| \leq H \text{ almost surely.}$$

We quantify the heterogeneity in the agent's dynamics through the following definition:

Definition 1. *(Bounded system heterogeneity) There exist positive constants* ϵ_1 *and* ϵ_2 *such that*

$$\max_{i,j\in[M]} \|A^{(i)} - A^{(j)}\| \leq \epsilon_1, \text{ and } \max_{i,j\in[M]} \|B^{(i)} - B^{(j)}\| \leq \epsilon_2.$$

We assume that ϵ_1 and ϵ_2 are finite. Similar bounded heterogeneity assumptions are commonly made in FL [84, 90, 160]. However, unlike typical FL works where one directly imposes heterogeneity assumptions on the agents gradients, in our setting, we need to carefully characterize how heterogeneity in the system parameters $(A^{(i)}, B^{(i)})$ translates to differences in the policy gradients; see Lemma 27. Before providing our solution to the federated LQR problem, we first recap existing results on model-free LQR in the single-agent setting.

The single-agent setting: When there is only one agent, i.e., $M = 1$, let us denote the system matrix as (A, B). If (A, B) is known, the optimal controller K^* can be computed by solving the discrete-time Algebraic Riccati Equation (ARE) [6].

Strikingly, [50] show that policy gradient methods can find the globally optimal LQR policy K^* despite the non-convexity of the problem. Tthe policy gradient of the LQR problem can be expressed as:

$$\nabla C(K) = 2E_K\Sigma_K = 2\left(\left(R + B^\top P_K B\right)K - B^\top P_K A\right)\Sigma_K,$$

where P_K is the positive definite solution to the Lyapunov equation: $P_K = Q + K^\top RK + (A - BK)^\top P_K(A - BK)$, $E_K := \left(R + B^\top P_K B\right)K - B^\top P_K A$, and $\Sigma_K := \mathbb{E}_{x_0\sim\mathcal{D}}\sum_{t=0}^{\infty} x_t x_t^\top$. The policy gradient method $K \leftarrow K - \eta\nabla C(K)$ will find the global optimal LQR policy, i.e., $K \to K^*$, provided that $\mathcal{E}_{x_0\sim\mathcal{D}}[x_0 x_0^\top]$ is full rank and an initial stabilizing policy is used. When the model is unknown, the analysis technique employed by [50] is to construct near-exact gradient estimates from reward samples and show that the sample complexity of such a method is bounded polynomially in the parameters of the problem.

In contrast to the single-agent setting, the heterogeneous, multi-agent scenario we consider here is considerably more difficult to analyze. First, designing an algorithm satisfying the iterative stability guarantees becomes a complex task. Second, since each agent in the system has its own

unique dynamics and gradient estimates, it can be difficult to aggregate these directions in a manner that ensures the updating direction moves toward the average optimal policy K^*. Nonetheless, in the sequel, we will overcome these challenges and provide a finite-time analysis of `FedLQR`.

6.4 Necessity of the Low Heterogeneity Requirement

In our main theorems, we require certain bounds on the parameters ϵ_1 and ϵ_2 that define the heterogeneity of the M dynamical systems we work with. Here, we point out that, unlike standard federated learning settings, these bounds are *necessary* for convergence. From a control and dynamical systems viewpoint, these bounds are perhaps intuitive: if the systems are too different, then there is no reason to believe there exists a stabilizing controller, i.e., there is no solution to the problem (6.2). In what follows, we will formalize this point. To do so, let us define an "instance" of our FedLQR problem via a parameter M that characterizes the number of agents/systems and the set of corresponding system matrices $\{A^{(i)}, B^{(i)}\}_{i\in[M]}$.[4]

We now prove a couple of simple impossibility results. Our first result shows that even when the input matrices are identical across agents, heterogeneity in the state transition matrices can lead to the non-existence of simultaneously stabilizing controllers, thereby rendering the FedLQR problem infeasible.

Proposition 5. *There exists an instance of the FedLQR problem with $M = 2$ and $\epsilon_2 = 0$, such that if $\epsilon_1 > 2$, then it is impossible to find a common linear state-feedback gain K that simultaneously stabilizes all systems.*

[4]Although technically the cost matrices Q and R are also part of a FedLQR problem formulation, they are not needed to establish the necessity of a low-heterogeneity requirement. As such, we do not include them here in our definition of an instance.

Proof: Consider an instance with just two scalar systems defined by:

$$x_{t+1}^{(1)} = \alpha x_t^{(1)} + u_t^{(1)} \quad \text{and} \quad x_{t+1}^{(2)} = -\alpha x_t^{(2)} + u_t^{(2)},$$

for some $\alpha > 0$. By simple inspection, note that in this case $\epsilon_1 = 2\alpha$ and $\epsilon_2 = 0$. Thus, $\epsilon_1 > 2 \Rightarrow \alpha > 1$. Now for a controller $u_t^{(i)} = -kx_t^{(i)}$ to stabilize both systems, the spectral radius conditions are $|\alpha - k| < 1$ and $|\alpha + k| < 1$. Trivially, there exists no gain k that satisfies both these requirements when $\alpha > 1$. This completes the proof.

To complement the above result, we now show that the effect of heterogeneity is not just limited to the state transition matrices. In particular, even when the state transition matrices are identical across agents, (arbitrarily small) heterogeneity in the input matrices can also lead to the non-existence of simultaneously stabilizing control gains. We formalize this below.

Proposition 6. *There exists an instance of the FedLQR problem with $M = 2$ and $\epsilon_1 = 0$, such that if $\epsilon_2 > 0$, then it is impossible to find a common linear state-feedback gain K that simultaneously stabilizes all systems.*

Proof: Consider an instance with two scalar systems defined by:

$$x_{t+1}^{(1)} = x_t^{(1)} + \beta u_t^{(1)} \quad \text{and} \quad x_{t+1}^{(2)} = x_t^{(2)} - \beta u_t^{(2)},$$

for some β. By simple inspection, note that in this case $\epsilon_1 = 0$ and $\epsilon_2 = 2\beta$. Thus, $\epsilon_2 > 0 \Rightarrow \beta > 0$. Now for a controller $u_t^{(i)} = -kx_t^{(i)}$ to stabilize both systems, the spectral radius conditions are $|1 - \beta k| < 1$ and $|1 + \beta k| < 1$. Trivially, there exists no gain k that satisfies both these requirements when $\beta > 0$. This concludes the proof.

The above example suggests that in certain settings, we can tolerate no heterogeneity whatsoever in the input matrices. More generally, the main take-home message from this section is that the requirement of a "low-heterogeneity regime" is *fundamental* to the problem and not merely an artifact of our analysis.

6.5 The `FedLQR` algorithm

In this section, we introduce our algorithm `FedLQR`, formally described by Algorithm 7, to solve for K^* in (6.2) . First, we impose the following assumption regarding the algorithm's initial condition K_0:

Assumption 8. *We can access an initial stabilizing controller, K_0, which stabilizes all systems $\{(A^{(i)}, B^{(i)})\}_{i=1}^M$, i.e., the spectral radius $\rho(A^{(i)} - B^{(i)}K_0) < 1$ holds for all $i \in [M]$.*

Algorithm description: At a high level, `FedLQR` follows the standard FL algorithmic template: a server first initializes a global policy, K_0, which it sends to the agents. Each agent proceeds to execute multiple PG updates using their local data. Once the local training is finished, agents transmit their model update to the server. The server aggregates the models and broadcasts an averaged model to the clients. The process repeats until a termination criterion is met. Prototypical FL algorithms that adhere to this structure include, for instance, `FedAvg` [90] and `FedProx` [110].

With this template in mind, we now dive into the details: `FedLQR` initializes the server and all agents with $K_{0,0}^{(i)} = K_0$ – a controller that stabilizes all agent's dynamics. In each round n, starting from a common global policy K_n, each agent i independently samples n_s trajectories from its own system at each local iteration l and performs approximate policy gradient updates using the zeroth-order optimization procedure [50] which we denote `ZO`; see line 7. For clarity, we present the explicit steps of using the zeroth-order method to estimate the true gradient in Algorithm 8, which will be discussed shortly. Between every communication round, each agent updates their local policy L times. Such an L is chosen to balance between the benefit of information sharing and the cost of communication. After L local iterations, each agent i uploads its local policy difference $\Delta_n^{(i)}$ (line 10) to the server. Once all differences are received, the server averages these differences $\{\Delta_n^{(i)}\}$ (line 12) to construct a new global policy K_{n+1}. The whole process is repeated N times.

Algorithm 7 Model-free Federated Policy Learning for the LQR (`FedLQR`)

Input: initial policy K_0, local step-size η_l, and global step-size η_g
Initialize the server with K_0 and η_g
for $n = 0, \dots, N-1$ **do**
  for each system $i \in [M]$ **do**
    for $l = 0, \dots, L-1$ **do**
      Agent i initializes $K_{n,0}^{(i)} = K_n$
      Agent i estimates $\widehat{\nabla} C^{(i)}(K_{n,l}^{(i)}) = \texttt{ZO}(K_{n,l}^{(i)}, i)$ and updates the local policy as
        $K_{n,l+1}^{(i)} = K_{n,l}^{(i)} - \eta_l \widehat{\nabla} C^{(i)}(K_{n,l}^{(i)})$
    end for
    Agent i sends $\Delta_n^{(i)} = K_{n,L}^{(i)} - K_n$ back to the server
  end for
  Server computes and broadcasts the global model $K_{n+1} = K_n + \frac{\eta_g}{M} \sum_{i=1}^{M} \Delta_n^{(i)}$
end for

Zeroth-order optimization [32, 144] provides a method of optimization that only requires oracle access to the function being optimized. Here, we briefly describe the details of our zeroth-order gradient estimation step[5] in Algorithm 8. To get a gradient estimator at a given policy K, we sample trajectories from the i-th system n_s times. At each time s, we use the perturbed policy $\widehat{K}_s$ (line 3) and a randomly generated initial point $x_0 \sim \mathcal{D}$ to simulate the i-th closed-loop system for τ steps. Thus, we can approximately calculate the cost by adding the stage cost from the first τ time steps on this trajectory (line 4), and then estimating the gradient as in line 6.

Discussion of Assumption 8: Assumption 8 is commonly adopted in the LQR [4, 35, 50, 163] and robust control literature [18, 40, 126]. In addition, there exist efficient ways to find such a stabilizing policy K_0; [18, 152, 249] each address the model-based setting, while [81] address this

[5] See Appendix in [193] for more details on zeroth-order optimization.

problem in the RL setting of heterogeneous multi-agent systems, and [102] in the single-agent, model-free setting. Moreover, it is well-known that the sample complexity of finding an initial stabilizing policy only adds a logarithmic factor to that for solving the LQR problem [140, 249].

Challenges in `FedLQR` analysis: Although `FedLQR` is similar in spirit to `FedAvg` [112, 132] (in the supervised learning setting), it is significantly more difficult to analyze the convergence of `FedLQR` for the following reasons.

- First, the problem we study is non-convex. Unlike most existing non-convex FL optimization results [84] which only guarantee convergence to stationary points, our work investigates whether `FedLQR` can find a globally optimal policy.

- Second, standard convergence analyses in FL [84, 110, 132, 216] rely on a "bounded gradient-heterogeneity" assumption. For the LQR problem, it is not clear a priori whether similar bounded policy gradient dissimilarity still holds. In fact, this is something we prove in Lemma 27.

- Third, the randomness in FL usually comes from only one source: the data obtained by each agent are drawn i.i.d. from some distribution; we call this *randomness from samples*. However, in `FedLQR`, there are three distinct sources of randomness: *sample randomness, initial condition randomness, and randomness from the smoothing matrices*. To reason about these different forms of randomness (that are intricately coupled), we provide a careful martingale-based analysis.

- Finally, we need to determine whether the solution given by `FedLQR` is meaningful, i.e., to decide whether the policy generated at each (local or global) iteration will stabilize all the systems.

To tackle these difficulties, we first define a stability region in our setting comprising of M heterogeneous systems as:

Definition 2. *(The stabilizing set) The stabilizing set is defined as* $\mathcal{G}^0 := \cap_{i=1}^{M} \mathcal{G}^{(i)}$ *where*

$$\mathcal{G}^{(i)} := \{K : C^{(i)}(K) - C^{(i)}(K_i^*) \leq \beta \left(C^{(i)}(K_0) - C^{(i)}(K_i^*)\right)\}.$$

As in [129], $\mathcal{G}^0$ is defined as the intersection of sub-level sets containing points K whose cost gap is at most β times the initial cost gap for all systems. It was shown in [74] that this is a compact set. Each sub-level set corresponds to a cost gap to agent i's optimal policy K_i^*, which is at most β times the initial cost gap $C^{(i)}(K_0) - C^{(i)}(K_i^*)$. Note that β can be any positive finite constant. Since any finite cost function indicates that K is a stabilizing controller, we conclude that any $K \in \mathcal{G}^0$ stabilizes all the systems. Following from Assumption 8, there exists a constant β such that $\mathcal{G}^0$ is nonempty. Moreover, it is worth remarking that the LQR cost function in the single-agent setting is *coercive*. That is, the cost acts as a barrier function, ensuring that the policy gradient update remains within the feasible stabilizing set $\mathcal{G}^{(i)}$. By defining the stabilizing set $\mathcal{G}^0$ as above, the cost function $C^{(i)}(K)$ retains its coerciveness on $\mathcal{G}^0$ for the federated setting considered in this paper.

In order to solve the federated LQR problem and provide convergence guarantees for `FedLQR`, we first need to recap some favorable properties of the LQR problem in the single-agent setting that enables PG to find the globally optimal policy.

6.5.1 Background on the centralized LQR using PG

In the single-agent setting, it was shown that policy gradient methods (i.e., model-free) can produce the global optimal policy despite the LQR problem being non-convex [50]. We summarize the properties that make this possible and which we also exploit in our analysis.

Lemma 25. *(Local Cost and Gradient Smoothness) Suppose* K' *is such that* $\|K' - K\| \leq h_\Delta(K) <$

Algorithm 8 Zeroth-order gradient estimation (ZO)

1: **Input:** K, number of trajectories n_s, trajectory length τ, smoothing radius r, dimensions n_x and n_u, system index i.

2: **for** $s = 1, \ldots, n_s$ **do**

3: Sample a policy $\widehat{K}_s = K + U_s$, with U_s drawn uniformly at random over matrices with (Frobenius) norm r.

4: Simulate the i-th system for τ steps starting from $x_0 \sim \mathcal{D}$ using policy $\widehat{K}_s$. Let $\widehat{C}_s$ be the empirical estimate:

$$\widehat{C}_s = \sum_{t=1}^{\tau} c_t, \quad \text{where} \quad c_t := x_t^\top \left(Q + \widehat{K}_s^\top R \widehat{K}_s \right) x_t$$

5: **end for**

6: **Return** the estimate:

$$\widehat{\nabla} C(K) = \frac{1}{n_s} \sum_{s=1}^{n_s} \frac{n_x n_u}{r^2} \widehat{C}_s U_s.$$

∞. *Then, the cost and gradient function satisfy:*

$$|C(K') - C(K)| \leq h_{cost}(K) \|K' - K\|,$$

$$\|\nabla C(K') - \nabla C(K)\| \leq h_{grad}(K) \|\Delta\| \text{ and } \|\nabla C(K') - \nabla C(K)\|_F \leq h_{grad}(K) \|\Delta\|_F,$$

respectively, where $h_\Delta(K)$, $h_{cost}(K)$ *and* $h_{grad}(K)$ *are positive scalars depending on* $C(K)$.

Lemma 26. *(Gradient Domination) Let* K^* *be an optimal policy. Then,*

$$C(K) - C(K^*) \leq \frac{\|\Sigma_{K^*}\|}{4\mu^2 \sigma_{\min}(R)} \|\nabla C(K)\|_F^2$$

holds for any stabilizing controller K*, i.e., any* K *satisfying the spectral radius* $\rho(A - BK) < 1$.

For simplicity, we skip the explicit expressions in these lemmas for $h_\Delta(K)$, $h_{\text{cost}}(K)$, and $h_{\text{grad}}(K)$ as functions of the parameters of the LQR problem. Interested readers are referred to the Appendix for full details. With Definition 2 of the stabilizing set in hand, we can define the

following quantities:

$$\bar{h}_{\text{grad}} := \sup_{K \in \mathcal{G}^0} h_{\text{grad}}(K),\ \bar{h}_{\text{cost}} := \sup_{K \in \mathcal{G}^0} h_{\text{cost}}(K),\ \text{and}\ \underline{h}_{\Delta} := \inf_{K \in \mathcal{G}^0} h_{\Delta}(K).$$

With these quantities, we can transform the *local* properties of the LQR problem discussed in Lemmas 25–26 into properties that hold over the *global* stabilizing set $\mathcal{G}^0$. For convenience, we use letters with bar such as $\bar{h}_{\text{grad}}$ to denote the global parameters. We are now ready to present our main results of `FedLQR` in the next section.

6.6 Main results

To analyze the performance of `FedLQR` in the model-free case, we first need to examine its behavior in the model-based case. Although this is not our end goal, these results are of independent interest.

6.6.1 Model-based setting

When $(A^{(i)}, B^{(i)})$ are available, exact gradients can be computed, and so the ZO scheme is no longer needed. In this case, the updating rule of `FedLQR` reduces to

$$K_{n+1} = K_n - \frac{\eta}{ML} \sum_{i=1}^{M} \sum_{l=0}^{L-1} \nabla C(K_{n,l}^{(i)}),$$

where $\eta := L\eta_g\eta_l$. Intuitively, if two systems are similar, i.e., satisfy Assumption 1, their exact policy gradient directions should not differ too much. We formalize this intuition as follows.

Lemma 27. ***(Policy gradient heterogeneity)*** *For any $i, j \in [M]$ and $K \in \mathcal{G}^0$, we have:*

$$||\nabla C^{(i)}(K) - \nabla C^{(j)}(K)|| \leq \epsilon_1 h_{het}^1(K) + \epsilon_2 h_{het}^2(K), \tag{6.3}$$

where $h^1_{het}(K)$ and $h^2_{het}(K)$ are positive bounded functions depending on the parameters of the LQR problem.[6]

By Lemma 27, if K belongs to a bounded set, the right-hand side of Eq. (6.3) is of the order $\mathcal{O}(\epsilon_1 + \epsilon_2)$. In other words, the exact gradient direction of agent i can be well-approximated by the gradient direction of agent j when the heterogeneity constants ϵ_1 and ϵ_2 are small. This justifies why it is beneficial to use other agents' data under the low-heterogeneity setting. Moreover, we can immediately conclude that the exact update direction of our `FedLQR` algorithm is also close to each agent's policy gradient direction based on Lemma 27. This fact is crucial for analyzing the convergence of `FedLQR` since we can map the convergence of `FedLQR` to that of the centralized LQR problem (with only one agent). However, Lemma 27 alone is not sufficient to provide the final guarantees since we still need to consider the impact of multiple local updates and stability concerns with heterogeneous systems. Nevertheless, by overcoming these difficulties, we establish the convergence of `FedLQR` in the model-based setting as follows:

Theorem 9. ***(Optimality in each agent's cost function)*** *When the heterogeneity level satisfies*[7] $(\epsilon_1 \bar{h}^1_{het} + \epsilon_2 \bar{h}^2_{het})^2 \leq \bar{h}^3_{het}$ *, there exist constant step-sizes η_g and η_l such that `FedLQR` enjoys the following performance guarantees over N rounds:*

$$C^{(i)}(K_N) - C^{(i)}(K_i^*) \leq \left(1 - \frac{\eta \mu^2 \sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)^N (C^{(i)}(K_0) - C^{(i)}(K_i^*)) + c_{uni,1} \times \mathcal{B}(\epsilon_1, \epsilon_2),$$

with $\mathcal{B}(\epsilon_1, \epsilon_2) := \frac{\upsilon \left\|\Sigma_{K_i^*}\right\|}{4\mu^2 \sigma_{\min}(R)} (\epsilon_1 h^1_{het} + \epsilon_2 h^2_{het})^2$, where $\bar{h}^{1,2}_{het} := \sup_{K \in \mathcal{G}^0} h^{1,2}_{het}(K)$, $\upsilon := \min\{n_x, n_u\}$, and $c_{uni,1}$ is a universal constant. Moreover, we have $K_n \in \mathcal{G}^0$ for all $n = 0, \cdots, N$.

[6]For simplicity, we write $h^1_{\text{het}}, h^2_{\text{het}}$ as a function of only K since only K changes during the iterations while other parameters remain fixed.

[7]The notation $\bar{h}^3_{\text{het}}$ is a positive scalar depending on the parameters of the LQR problem; see Appendix in [193] for full details.

Main Takeaways: Theorem 9 reveals that the output K_n of `FedLQR` can stabilize all M systems at each round n. However, `FedLQR` can only converge to a ball of radius $\mathcal{B}(\epsilon_1, \epsilon_2)$ around each system's optimal controller K_i^*, regardless of the choice of the step-sizes. The term $\mathcal{B}(\epsilon_1, \epsilon_2)$ captures the effect of heterogeneity and becomes zero when each agent follows the same system dynamics, i.e., $\epsilon_1 = \epsilon_2 = 0$. When there is no heterogeneity, the convergence rate matches the rate of the centralized setting [50] up to a constant factor. But, since there is no noise introduced by the zeroth-order gradient estimate, there is no expectation of obtaining a benefit from collaboration. Nonetheless, understanding the model-based setting provides valuable insights for exploring the model-free setting. With this theorem, now we are ready to provide the convergence guarantees for our algorithm `FedLQR` in solving Eq. (6.2). Next, we will elaborate these guarantees in the following corollary.

Corollary 2. ***(Optimality in average cost function)*** *When the heterogeneity level satisfies*[8] $(\epsilon_1 \bar{h}_{het}^1 + \epsilon_2 \bar{h}_{het}^2)^2 \leq \bar{h}_{het}^3$*, after N rounds, `FedLQR` enjoys the following optimality gap in average cost function across all M agents:*

$$C_{avg}(K_N) - C_{avg}(K^*) \leq \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)\sigma_{\min}(Q)}{\bar{C}_{\max}}\right)^N \sup_{i\in[M]} (C^{(i)}(K_0) - C^{(i)}(K_i^*)) + c_{uni,1} \times \mathcal{B}(\epsilon_1, \epsilon_2)$$

where $\bar{C}_{\max} := \sup_{K\in\mathcal{G}^0, i\in[M]} C^{(i)}(K)$ and ν is as defined in Theorem 9.

The main message conveyed by Corollary 2 is that `FedLQR` can converge to a ball around the average optimal controller K^* with a linear convergence rate. The size of the ball depends on the system heterogeneity level, i.e., ϵ_1 and ϵ_2. Combining Theorem 9 and Corollary 2, we infer

[8]The notation $\bar{h}_{\text{het}}^3$ is a positive scalar depending on the parameters of the LQR problem; see Appendix in [193] for full details.

that FedLQR not only approximates each system's optimal controller K_i^* but also approximately converges toward the average optimal controller K^* when the underlying M systems are close. The primary distinction between converging to K_i^* and K^* lies in the linear convergence rate. Compared to converging to K_i^*, where the linear converge rate depends only on system i's parameter $\left\|\Sigma_{K_i^*}\right\|$, the rate in converging to K^* depends on all systems' parameters, i.e., $\left\{\left\|\Sigma_{K_i^*}\right\|\right\}_{i=1}^N$. Note that these parameters $\left\{\left\|\Sigma_{K_i^*}\right\|\right\}_{i=1}^N$ are bounded by a universal upper bound $\frac{\bar{C}_{\max}}{\sigma_{\min}(Q)}$ in Corollary 2. See appendix in [193] for a comprehensive proof.

How to ensure FedLQR's stability? We briefly discuss our proof technique for ensuring the iterative stability guarantees. The main idea is to leverage an inductive argument. We start from a stabilizing global policy $K_n \in \mathcal{G}^0$. We aim to show that the next global policy K_{n+1} is stabilizing. This is achieved by demonstrating that K_{n+1} can reduce each system's cost function compared to K_n. To achieve this goal, we take the following steps: (1) at each iteration, initiate from the globally stabilizing controller computed at the previous iterate, (2) determine a small global step-size such that inequalities in Section 6.5.1 can be applied; (3) use Lemma 27 to provide a descent direction to reduce each system's cost function; (4) bound the drift term $\frac{1}{ML}\sum_{i=1}^M\sum_{l=0}^{L-1}\|K_{n,l}^{(i)} - K_n\|^2$. Step (4) can be accomplished using a small local step-size η_l such that each local policy is a small perturbation of the global policy K_n. Equipped with these results, we are ready to present our main results of the model-free setting.

6.6.2 Model-free setting

We now analyze FedLQR's convergence in the model-free setting, where the policy gradient steps are approximately computed using zeroth-order optimization (Algorithm 8), without knowing the true dynamics, i.e., $A^{(i)}, B^{(i)}$ are not available and so $\nabla C^{(i)}(K^{(i)})$ can't be directly computed) . The key point in this setting is to bound the gap between the estimated gradient and the true gradient.

In the centralized setting [50], the gap can [can be made arbitrarily accurate with enough trajectory samples n_s, sufficiently long trajectory length τ, and small smoothing radius r.

We aim to achieve a sample complexity reduction for each agent by utilizing data from other similar but non-identical systems with the help of the server. This presents a significant challenge, as averaging gradient estimates from multiple agents may not necessarily reduce the variance even for homogeneous systems due to the high correlation between local gradient estimates. This challenge is compounded in our case as the gradient estimates are not only *correlated* but also come from *non-identical systems*. As a result, the variance reduction and sample complexity reduction for the `FedLQR` algorithm is not obvious a priori. After addressing these challenges using a martingale-type analysis, we show that one can establish variance reduction for our our setting as well. This is formalized in the next result:

Lemma 28. ***(Variance Reduction Lemma)*** *Suppose the smoothing radius r and trajectory length τ from Algorithm 8 satisfy $r \leq h_r\left(\frac{\epsilon}{4}\right)$ and $\tau \geq h_\tau\left(\frac{r\epsilon}{4n_x n_u}\right)$, respectively.*[9] *Moreover, suppose the sample size satisfies:*[10]

$$n_s \geq \frac{h_{\textit{sample,trunc}}\left(\frac{\epsilon}{4}, \frac{\delta}{ML}, \frac{H^2}{\mu}\right)}{ML}. \tag{6.4}$$

Then, when $K_n \in \mathcal{G}^0$, with probability $1-\delta$, the estimated gradients satisfy:

$$\left\| \frac{1}{ML} \sum_{i=1}^{M} \sum_{l=0}^{L-1} \left[\widehat{\nabla} C^{(i)}(K_{n,l}^{(i)}) - \nabla C^{(i)}(K_{n,l}^{(i)}) \right] \right\|_F \leq \epsilon.$$

The most important information conveyed by our variance reduction lemma is that each agent at each local step only needs to sample $\frac{1}{ML}$ fraction of samples required in the centralized setting. Notably, this lemma plays an important role in showing that `FedLQR` can help improve the sample

[9]The notation h_r, h_τ, $h_{\text{sample,trunc}}$ and h'_r in Lemma 28 and Theorem 10 are polynomial functions of the LQR problem, depending on ϵ. For simplicity, we defer their definition to the Appendix.

[10]For the convenience of comparison with existing literature, we use the same notation as [50, 65].

efficiency. Equipped with Lemma 28, we now present the main convergence guarantees for FedLQR:

Theorem 10. ***(Model-free)*** *Suppose the trajectory length satisfies* $\tau \geq h_\tau\left(\frac{r\epsilon'}{4n_x n_u}\right)$, *the smoothing radius satisfies* $r \leq h'_r\left(\frac{\epsilon'}{4}\right)$, *and the sample size of each agent* n_s *satisfies Eq.* (6.4) *with* $\epsilon' = \sqrt{\frac{c_{uni,3}\mu^2\sigma_{\min}(R)}{4\left\|\Sigma_{K_i^*}\right\|}\cdot\epsilon}$. *When the heterogeneity level satisfies* $(\epsilon_1\bar{h}^1_{het}+\epsilon_2\bar{h}^2_{het})^2 \leq \bar{h}^3_{het}$, *then, given any* $\delta\in(0,1)$, *with probability* $1-\delta$, *there exist constant step-sizes* η_g *and* η_l, *which are independent of* ϵ', *such that* FedLQR *enjoys the following performance guarantees:*

1. ***(Stability of the global policy)*** *The global policy at each round* n *is stabilizing, i.e.,* $K_n \in \mathcal{G}^0$;
2. ***(Stability of the local policies)*** *The local policies satisfy* $K^{(i)}_{n,l} \in \mathcal{G}^0$ *for all* i *and* l;
3. ***(Convergence rate)*** *After* $N \geq \frac{c_{uni,4}\left\|\Sigma_{K_i^*}\right\|}{\eta\mu^2\sigma_{\min}(R)}\log\left(\frac{2(C^{(i)}(K_0)-C^{(i)}(K_i^*))}{\epsilon'}\right)$ *rounds, we have*

$$C^{(i)}(K_N) - C^{(i)}(K_i^*) \leq \epsilon' + c_{uni,2}\times\mathcal{B}(\epsilon_1,\epsilon_2), \forall i\in[M], \tag{6.5}$$

where $c_{uni,2}, c_{uni,3}, c_{uni,4}$ *are universal constants and* $\mathcal{B}(\epsilon_1,\epsilon_2)$ *is defined in Theorem 9.*

This theorem establishes the finite-time convergence guarantees for FedLQR. The first two points in Theorem 10 provide the iterative stability guarantees of FedLQR, i.e., the trajectories of FedLQR will always stay inside the stabilizing set $\mathcal{G}^0$. The third point implies that when heterogeneity is small, i.e., $\mathcal{B}(\epsilon_1,\epsilon_2)$ is negligible, FedLQR converges to each system's optimal policy with a linear speedup w.r.t. the number of agents M, which we discuss further next.

Discussion: For a fixed desired precision ϵ, we denote N to be the number of rounds such that the first term ϵ' in Eq (6.5) is smaller than ϵ. In what follows, we focus on analyzing the total sample complexity of FedLQR for each agent, which can be calculated by $N\times L\times n_s$. Note that N, in our case, is in the same order as the centralized setting. However, in terms of the sample

size n_s requirement at each local step, it is only a $\frac{1}{ML}$-fraction of that needed in the centralized setting, as presented in the variance reduction Lemma 28. Therefore, in a low-heterogeneity regime, where $\mathcal{B}(\epsilon_1, \epsilon_2)$ is negligible, our `FedLQR` *algorithm* reduces the sample complexity of learning the optimal LQR policy by $\tilde{\mathcal{O}}(\frac{1}{M})$ of the centralized setting [50, 129].[11] Specifically, `FedLQR` improves the sample cost required by each agent from $\tilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ to $\tilde{\mathcal{O}}(\frac{1}{M\epsilon^2})$ up to a small heterogeneity bias term. This result is highly desirable since the number of agents in FL is usually large; leading to a significant speedup due to collaboration.

It is important to mention that our results also capture the cost of federation embedded in the term $\mathcal{B}(\epsilon_1, \epsilon_2)$. That is when two systems exhibit significant differences from each other, leveraging data across them may not be beneficial in finding a common stabilizing policy that applies to both. *In a summary, Eq.* (6.4)–(6.5) *provide an explicit interplay between the price of heterogeneity and the benefit of collaboration*. The trade-off in Theorem 10 is explored in the simulation study presented in the next section.

6.7 Numerical Results

The following section describes the experimental setup and results when applying `FedLQR` in the model-free setting.[12]

6.7.1 System Generation

Numerical experiments are conducted to illustrate and evaluate the effectiveness of `FedLQR` (Algorithm 7). The simulations involve different and unstable dynamical systems described by

[11] In [50], the sample complexity of policy gradient method is $\tilde{\mathcal{O}}(\frac{1}{\epsilon^4})$, this was later improved to $\tilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ by [129]. We compare our results to the refined analysis in [129].

[12] Code can be downloaded from `https://github.com/jd-anderson/FedLQR`

discrete-time linear time-invariant (LTI) models, as in (6.3), where each system has $n_x = 3$ states and $n_u = 3$ inputs. To generate different systems while respecting the bounded heterogeneity assumption (Assumption 1), the following steps are followed:

1. Given nominal system matrices (A_0, B_0), generate random variables $\gamma_1^{(i)} \sim \mathcal{U}(0, \epsilon_1)$ and $\gamma_2^{(i)} \sim \mathcal{U}(0, \epsilon_2)$, $\forall i \in [M]$, with ϵ_1 and ϵ_2 being predefined dissimilarity parameters.

2. The random variables generated above are combined with modification masks $Z_1 \in \mathbb{R}^{3\times3}$ and $Z_2 \in \mathbb{R}^{3\times3}$ to generate the different systems matrices $(A^{(i)}, B^{(i)})$ for all $i \in [M]$.

3. The systems $(A^{(i)}, B^{(i)})$ for $0 < i \leq M$ are then constructed by perturbing the nominal systems according to: $A^{(i)} = A_0 + \gamma_1^{(i)} Z_1$ and $B^{(i)} = B_0 + \gamma_2^{(i)} Z_2$, where Z_1 and Z_2 are defined in step 2.

4. The nominal matrices are included in the set of generated systems as $(A^{(1)}, B^{(1)}) = (A_0, B_0)$.

In particular, we consider

$$A_0 = \begin{bmatrix} 1.20 & 0.50 & 0.40 \\ 0.01 & 0.75 & 0.30 \\ 0.10 & 0.02 & 1.50 \end{bmatrix}, \quad B_0 = I_3, \quad Q = 2I_3, \text{ and } R = \frac{1}{2} I_3,$$

for the nominal system matrices and cost matrices respectively.

The optimal controller for the nominal system $(A^{(1)}, B^{(1)})$ is

$$K_1^* = \begin{bmatrix} 1.0056 & 0.4293 & 0.3570 \\ 0.0262 & 0.6239 & 0.2657 \\ 0.1003 & 0.0298 & 1.2960 \end{bmatrix},$$

and was obtained by solving the discrete algebraic Riccati equation (DARE).

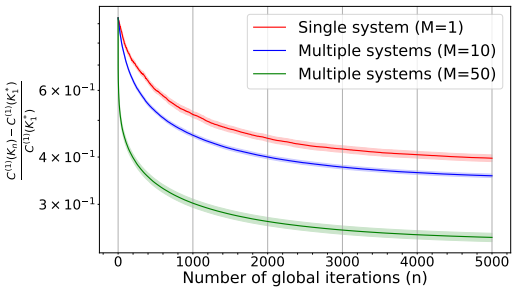
6.7.2 Algorithm Parameters

For the gradient estimation step in the zeroth-order algorithm (Algorithm 8), we set the initial state for cost computation as a random sample from a standard normal distribution, denoted as $\mathcal{D} \overset{d}{=} \mathcal{N}(0, I_3)$, for all systems $i \in [M]$. Additionally, we consider $n_s = 5$ trajectories, where each trajectory has a rollout length of $\tau = 15$, and we set the smoothing radius $r = 0.1$ for the zeroth-order gradient estimation.

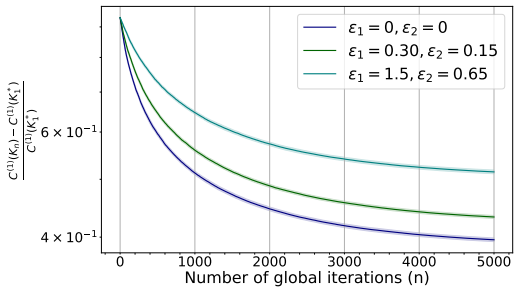
Throughout our simulations, we consider the following initial stabilizing controller $K_0 = 1.62I_3$ (Line 1 in Algorithm 7). Note that although the control action $u_t^{(i)} = -K_0 x_t^{(i)}$ may not be optimal for any of the M systems. For example, the suboptimality of K_0 applied to the nominal system is evidenced by its cost of $C^{(1)}(K_0) = 18.4049$, compared to the optimal cost of $C^{(1)}(K_1^*) = 9.5220$, when computed from an initial state $x_0^{(1)} = [1 \;\; 1 \;\; 1]^\top$ and time horizon $T = 500$. However, it is important to note that K_0 is still stabilizing all M systems. Note that we will use K_0 as the initial controller for all of the experiments in this paper.

6.7.3 Experiments

To assess the performance of `FedLQR`, we evaluate the normalized gap between the current cost $C^{(1)}(K_n)$ of the nominal system when using the common stabilizing controller K_n and its corresponding optimal cost $C^{(1)}(K_1^*)$. This metric is represented as $\frac{C^{(1)}(K_n) - C^{(1)}(K_1^*)}{C^{(1)}(K_1^*)}$ for each global iteration $n \in [N]$. In our experiments, we set the step sizes as $\eta_g = 1 \times 10^{-2}$, with an adaptive decrease of 0.05% per global iteration, and $\eta = 1 \times 10^{-4}$, and employ a single local iteration $L = 1$ for each communication round between the systems and the server. Further details regarding other parameters, such as the number of systems M, heterogeneity levels (ϵ_1, ϵ_2), and modification masks Z_1 and Z_2, will be provided in the figures and the subsequent discussion.



(a) Varying M with $\epsilon_1 = \epsilon_2 = 0.5$.



(b) Varying ϵ_1, ϵ_2, with $M = 10$.

Figure 6.1: Gap between the current and optimal cost with respect to the number of global iterations.

Figures 6.1-(a,b) present the normalized distance between the current cost associated with the common stabilizing controller and the optimal cost for the nominal system, plotted with respect to the number of global iterations. These figures demonstrate the impact of varying the number of systems M and the heterogeneity parameters (ϵ_1, ϵ_2) on the convergence and performance of Algorithm 7.

In Figure 6.1-(a), we specifically investigate the effect of the number of systems M participating in the collaboration to compute a common controller K^* on the convergence of our algorithm. In this analysis, we set the heterogeneity parameters as $\epsilon_1 = 0.5$ and $\epsilon_2 = 0.5$ and consider modification masks $Z_1 = Z_2 = I_3$. The figure reveals a noticeable reduction in the gap between the current and optimal cost as the number of participating systems M increases. This numerical result aligns with our theoretical findings, which indicate that the number of samples required to achieve reliable estimation for the cost function's gradient can be scaled down with the number of systems participating in the collaboration. Consequently, as the number of systems involved increases, there is a considerable reduction in the gap between the common computed controller and the optimal one.

Figure 6.1-(b) illustrates the influence of the heterogeneity parameters (ϵ_1, ϵ_2) on the convergence

rate of Algorithm 7. In this analysis, we set the number of systems as $M = 10$, and the modification masks $Z_1 = \texttt{diag}([3.5 \;\; 1 \;\; 0.1])$ and $Z_2 = \texttt{diag}([1.5 \;\; 0.1 \;\; 1])$. Consistent with our theoretical findings, we observe that an increase in the dissimilarity among the systems results in a significant gap between the common and optimal controller. This discrepancy arises due to the additive effect of system heterogeneity on the convergence rate of our algorithm, as elaborated in Theorem 10.

6.8 Chapter Summary and Future Work

We investigated the problem of learning a common and optimal LQR policy with the objective of minimizing an average quadratic cost. The primary focus of this paper was to thoroughly examine and provide comprehensive answers to the following questions: (i) Is the learned common policy stabilizing for all agents? (ii) How close is the learned common policy to each agent's own optimal policy? (iii) Can each agent learn its own optimal policy faster by leveraging data from all agents? To address these questions, we proposed a federated and model-free approach, `FedLQR`, where M heterogenous systems collaborate to learn a common and optimal policy while keeping the system's data *private*. Our analysis tackles numerous technical challenges, including system heterogeneity, multiple local gradient descent updates, and stability. We have demonstrated that `FedLQR` produces a common policy that stabilizes all systems and converges to the optimal policy (Theorem 10) of each agent up to a heterogeneity bias term. Furthermore, `FedLQR` achieves a reduction in sample complexity proportional to the number of participating agents M (Lemma 28). We also have provided numerical results to effectively showcase and evaluate the performance of our `FedLQR` approach in a model-free setting. Future work will address the assumption of requiring full-state information to extend our results to the Linear Quadratic Gaussian (LQG) problem in a federated setting. We are currently investigating data-driven and system-theoretic metrics for heterogeneity, as

well as personalization-based methods to mitigate the impact of heterogeneity on the performance of the proposed approach.

6.9 Omitted Proofs

6.9.1 SUPPLEMENTARY ROADMAP

This appendix is organized as follows. Section 6.9.2 offers a comprehensive and detailed overview of the relevant literature related to this paper. Section 6.9.3 discusses the underlying intuition and necessity behind the low heterogeneity regime. In Section 6.9.4, we present numerical results that illustrate and evaluate the performance of the proposed `FedLQR` algorithm (Algorithm 7). Sections 6.9.5 and 6.9.6 present important auxiliary norm inequalities and lemmas that play a key role in proving the main results of this paper. The proof of our main results related to the model-based setting is provided in Section 6.9.7, while Section 6.9.9 is dedicated to the corresponding results in the model-free setting. Additional details on the zeroth-order optimization method are provided in Section 6.9.8.

a) Notation Recap

For convenience we briefly recap and summarize our notation. We use $\|S\|_{max}$ to denote the maximum spectral norm taken over the family of matrices $S^{(1)}, \ldots, S^{(M)}$. All norms for matrices and vectors are spectral and Euclidean respectively, unless otherwise stated. The integer sequence $1, 2, \ldots, N$ is denoted as $[N]$. The spectral radius of a square matrix is denoted by $\rho(\cdot)$.

6.9.2 RELATED WORK

This section provides a more detailed and comprehensive literature survey on the key topics closely related to the subject matter of this paper. We aim to explore and summarize the main ideas presented in the existing literature pertaining to federated learning (FL), policy gradient (PG), federated reinforcement learning (FRL), as well as model-based and model-free linear quadratic

Symbol	Meaning
M	number of systems
L	number of local updates (counter: l)
N	number of rounds of averaging (counter: n)
K_n	averaged controller at round n
K_i^*	optimal controller for system $(A^{(i)}, B^{(i)})$
$K_{n,l}^{(i)}$	controller for system i after l local iterations and n averaging rounds

control.

- **Federated Learning (FL):**

  In this work, we employ the federated learning (FL) paradigm to facilitate collaborative learning among systems without the need to share raw data with other participants or a server [14, 96, 97, 132]. Despite FL being a relatively recent creation, it has already garnered significant attention and boasts a wealth of literature. Below we highlight work that is most relevant to our problem setting.

  Federated averaging (`FedAvg`) stands as the pioneering and most widely adopted algorithm in FL. Originally proposed by McMahan et al. in [132], `FedAvg` has demonstrated its effectiveness in homogeneous settings [68, 161, 181, 183, 215] where all participating clients aim to minimize the same objective function. However, ensuring convergence guarantees for

FedAvg becomes notably more challenging in the presence of heterogeneity [70, 87, 90, 112], thus necessitating additional assumptions on the gradient and Hessian dissimilarity bounds [84, 87, 112, 115]. This difficulty arises primarily due to a "client-drift" effect, which is inherent to the FedAvg algorithm and has a detrimental impact on its convergence performance [22, 24]. As a result of the challenges posed by FedAvg, several alternative algorithms have been proposed to address its limitations. Notable examples of these algorithms include FedProx [110], Scaffold [84], FedSplit [149], FedDR [203], FedADMM [211], FedLin [136], and S-Local-SVRG [64]. Each of them introduces unique techniques and modifications to the original FedAvg algorithm, aiming to enhance convergence guarantees while handling communication cost concerns, statistical heterogeneity, client dropout, and sample complexity more effectively.

Applying federated learning (FL) to control systems introduces a novel research direction that comes with its own set of challenges. Control systems exhibit unique characteristics, such as non-iid and non-isotropic data, as well as system instability, which arise due to the dynamic nature of the systems. These characteristics pose specific challenges when attempting to leverage data from multiple systems for tasks such as system identification [212] or control synthesis [163].

Although [163] addresses the model-free LQR tracking problem in a federated manner, it focuses on a significantly simpler scenario where all agents follow identical dynamics (i.e., no heterogeneity). In contrast, our present work introduces new analysis techniques to achieve linear speedup in FedLQR when dealing with heterogeneous dynamical systems and multiple local updates per communication round.

- **Policy Gradient (PG):**

The policy gradient (PG) approach is a fundamental component of the success of reinforcement learning (RL) and plays a crucial role in policy optimization (PO). This approach directly optimizes the policy to improve system-level performances through gradient ascent steps. The concept of policy optimization has been influential in RL [127, 128, 189] with some well-known algorithms such as `REINFORCE` [223], trust-region policy optimization `TRPO` [171], actor-critic methods [94], and proximal policy optimization `PPO` [172]. We highlight an important difference between standard MDP models and control models in RL. In control, one requires the policy to provide closed-loop stability, i.e., all trajectories of the system must converge for a given policy. In contrast, convergence in the MDP setting requires irreducibly and aperiodicity properties that are assumed *before* a policy is selected. As a result, the control task is significantly more challenging.

The extensive body of literature on policy optimization for reinforcement learning (RL) and its adaptability to the model-free setting paves the way for leveraging policy gradient methods in the pursuit of learning optimal control policies for classical control problems [74, 152]. Despite the non-convex nature of the formulation involved in policy gradient methods, recent work [50, 65, 71, 80, 82, 102, 129, 140, 152] has demonstrated global convergence in solving the model-free LQR problem via policy gradient methods. This convergence is achieved due to certain properties of the quadratic cost function inherent in the LQR problem as introduced in [50]. In contrast to the aforementioned work, which exclusively focus on the centralized control setting, our paper offers convergence guarantees for the multi-agent setting. In this context, each agent follows similar, but not identical, dynamics, thereby distinguishing it from the simpler scenario in [163].

- **Federated Reinforcement Learning (FRL):**

The flexibility of policy gradient methods in the model-free RL setting has paved the way for a relatively recent research direction known as federated reinforcement learning (FRL), which aims to address practical implementation challenges of RL through the use of federated learning [158]. FRL focuses on learning a common value function [46, 192] or improving the policy by leveraging multiple RL agents interacting with similar environments. The empirical evidence presented in the survey paper [158] demonstrates the significant success of FRL in reducing sample complexity across various applications such as autonomous driving [116], IoT devices [117], resource management in networking [237], and communication efficiency [59]. However, it is important to note that existing recent works in this field do not specifically tackle the challenge of finding a common and stabilizing optimal policy that is suitable for all RL agents in a heterogeneous setting.

- **Model-free Linear Quadratic Control:**

  The linear quadratic regulator (LQR) problem is a well-studied classical control problem that has gained significant attention due to its wide applicability and its role as a baseline for more complex control strategies [6]. Recently, to address the non-convex nature of the policy gradient LQR, [187] has proposed convexifying the corresponding optimal control problem to efficiently solve the model-based LQR problem via policy gradient. Furthermore, the model-free LQR has attracted considerable interest after [50] provided guarantees on the global convergence of policy gradient methods for both model-based and model-free LQR settings. This breakthrough paved the way for subsequent works [65, 71, 80, 82, 102, 129, 140, 152] that analyze convergence guarantees and sample complexity in the context of the model-free LQR problem. Notably, [35] characterizes the sample complexity of the LQR problem.

  Another line of work explores certainty equivalent control [131, 178], providing regret bounds

to demonstrate the quality of the designed linear quadratic regulator in terms of the accuracy of the estimated system model. However, the key distinction between these works and the present paper lies in the consideration of multiple and heterogeneous systems. Moreover, [131, 178] use the regret framework, which is different from the PAC learning-based framework [51] exploited in our paper.

6.9.3 NECESSITY OF THE LOW HETEROGENEITY REQUIREMENT

In our main theorems, we require certain bounds on the parameters ϵ_1 and ϵ_2 that define the heterogeneity of the M dynamical systems we work with. Here, we point out that, unlike standard federated learning settings, these bounds are *necessary* for convergence. From a control and dynamical systems viewpoint, these bounds are perhaps intuitive: if the systems are too different, then there is no reason to believe there exists a stabilizing controller, i.e., there is no solution to the problem (6.2). In what follows, we will formalize this point. To do so, let us define an "instance" of our FedLQR problem via a parameter M that characterizes the number of agents/systems and the set of corresponding system matrices $\{A^{(i)}, B^{(i)}\}_{i\in[M]}$.[13]

We now prove a couple of simple impossibility results. Our first result shows that even when the input matrices are identical across agents, heterogeneity in the state transition matrices can lead to the non-existence of simultaneously stabilizing controllers, thereby rendering the FedLQR problem infeasible.

Proposition 7. *There exists an instance of the FedLQR problem with $M = 2$ and $\epsilon_2 = 0$, such that if $\epsilon_1 > 2$, then it is impossible to find a common linear state-feedback gain K that simultaneously*

[13]Although technically the cost matrices Q and R are also part of a FedLQR problem formulation, they are not needed to establish the necessity of a low-heterogeneity requirement. As such, we do not include them here in our definition of an instance.

stabilizes all systems.

Proof: Consider an instance with just two scalar systems defined by:

$$x_{t+1}^{(1)} = \alpha x_t^{(1)} + u_t^{(1)} \quad \text{and} \quad x_{t+1}^{(2)} = -\alpha x_t^{(2)} + u_t^{(2)},$$

for some $\alpha > 0$. By simple inspection, note that in this case $\epsilon_1 = 2\alpha$ and $\epsilon_2 = 0$. Thus, $\epsilon_1 > 2 \Rightarrow \alpha > 1$. Now for a controller $u_t^{(i)} = -kx_t^{(i)}$ to stabilize both systems, the spectral radius conditions are $|\alpha - k| < 1$ and $|\alpha + k| < 1$. Trivially, there exists no gain k that satisfies both these requirements when $\alpha > 1$. This completes the proof.

To complement the above result, we now show that the effect of heterogeneity is not just limited to the state transition matrices. In particular, even when the state transition matrices are identical across agents, (arbitrarily small) heterogeneity in the input matrices can also lead to the non-existence of simultaneously stabilizing control gains. We formalize this below.

Proposition 8. *There exists an instance of the FedLQR problem with $M = 2$ and $\epsilon_1 = 0$, such that if $\epsilon_2 > 0$, then it is impossible to find a common linear state-feedback gain K that simultaneously stabilizes all systems.*

Proof: Consider an instance with two scalar systems defined by:

$$x_{t+1}^{(1)} = x_t^{(1)} + \beta u_t^{(1)} \quad \text{and} \quad x_{t+1}^{(2)} = x_t^{(2)} - \beta u_t^{(2)},$$

for some β. By simple inspection, note that in this case $\epsilon_1 = 0$ and $\epsilon_2 = 2\beta$. Thus, $\epsilon_2 > 0 \Rightarrow \beta > 0$. Now for a controller $u_t^{(i)} = -kx_t^{(i)}$ to stabilize both systems, the spectral radius conditions are $|1 - \beta k| < 1$ and $|1 + \beta k| < 1$. Trivially, there exists no gain k that satisfies both these requirements when $\beta > 0$. This concludes the proof.

The above example suggests that in certain settings, we can tolerate no heterogeneity whatsoever in the input matrices. More generally, the main take-home message from this section is that the

requirement of a "low-heterogeneity regime" is *fundamental* to the problem and not merely an artifact of our analysis.

6.9.4 NUMERICAL RESULTS

The following section describes the experimental setup and results when applying `FedLQR` in the model-free setting.[14]

a) System Generation

Numerical experiments are conducted to illustrate and evaluate the effectiveness of `FedLQR` (Algorithm 7). The simulations involve different and unstable dynamical systems described by discrete-time linear time-invariant (LTI) models, as in (6.3), where each system has $n_x = 3$ states and $n_u = 3$ inputs. To generate different systems while respecting the bounded heterogeneity assumption (Assumption 1), the following steps are followed:

1. Given nominal system matrices (A_0, B_0), generate random variables $\gamma_1^{(i)} \sim \mathcal{U}(0, \epsilon_1)$ and $\gamma_2^{(i)} \sim \mathcal{U}(0, \epsilon_2)$, $\forall i \in [M]$, with ϵ_1 and ϵ_2 being predefined dissimilarity parameters.

2. The random variables generated above are combined with modification masks $Z_1 \in \mathbb{R}^{3\times3}$ and $Z_2 \in \mathbb{R}^{3\times3}$ to generate the different systems matrices $(A^{(i)}, B^{(i)})$ for all $i \in [M]$.

3. The systems $(A^{(i)}, B^{(i)})$ for $0 < i \leq M$ are then constructed by perturbing the nominal systems according to: $A^{(i)} = A_0 + \gamma_1^{(i)} Z_1$ and $B^{(i)} = B_0 + \gamma_2^{(i)} Z_2$, where Z_1 and Z_2 are defined in step 2.

4. The nominal system matrices are included in the set of generated systems as $(A^{(1)}, B^{(1)}) = (A_0, B_0)$.

[14] Code can be downloaded from `https://github.com/jd-anderson/FedLQR`

In particular, we consider

$$A_0 = \begin{bmatrix} 1.20 & 0.50 & 0.40 \\ 0.01 & 0.75 & 0.30 \\ 0.10 & 0.02 & 1.50 \end{bmatrix}, \quad B_0 = I_3, \quad Q = 2I_3, \quad R = \frac{1}{2} I_3,$$

for the nominal system matrices and cost matrices respectively. The optimal controller for the nominal system $(A^{(1)}, B^{(1)})$ is

$$K_1^* = \begin{bmatrix} 1.0056 & 0.4293 & 0.3570 \\ 0.0262 & 0.6239 & 0.2657 \\ 0.1003 & 0.0298 & 1.2960 \end{bmatrix},$$

and was obtained by solving the discrete algebraic Riccati equation (DARE).

b) Algorithm Parameters

For the gradient estimation step in the zeroth-order algorithm (Algorithm 8), we set the initial state for cost computation as a random sample from a standard normal distribution, denoted as $\mathcal{D} \overset{d}{=} \mathcal{N}(0, I_3)$, for all systems $i \in [M]$. Additionally, we consider $n_s = 5$ trajectories, where each trajectory has a rollout length of $\tau = 15$, and we set the smoothing radius $r = 0.1$ for the zeroth-order gradient estimation.

Throughout our simulations, we consider the following initial stabilizing controller $K_0 = 1.62 I_3$ (Line 1 in Algorithm 7). Note that although the control action $u_t^{(i)} = -K_0 x_t^{(i)}$ may not be optimal for any of the M systems. For example, the suboptimality of K_0 applied to the nominal system is evidenced by its cost of $C^{(1)}(K_0) = 18.4049$, compared to the optimal cost of $C^{(1)}(K_1^*) = 9.5220$, when computed from an initial state $x_0^{(1)} = [1 \;\; 1 \;\; 1]^\top$ and time horizon $T = 500$. However, it is important to note that K_0 is still stabilizing all M systems. Note that we will use K_0 as the initial controller for all of the experiments in this paper.

c) Experiments

To assess the performance of `FedLQR`, we evaluate the normalized gap between the current cost $C^{(1)}(K_n)$ of the nominal system when using the common stabilizing controller K_n and its corresponding optimal cost $C^{(1)}(K_1^*)$. This metric is represented as $\frac{C^{(1)}(K_n)-C^{(1)}(K_1^*)}{C^{(1)}(K_1^*)}$ for each global iteration $n \in [N]$. In our experiments, we set the step sizes as $\eta_g = 1 \times 10^{-2}$, with an adaptive decrease of 0.05% per global iteration, and $\eta = 1 \times 10^{-4}$, and employ a single local iteration $L = 1$ for each communication round between the systems and the server. Further details regarding other parameters, such as the number of systems M, heterogeneity levels (ϵ_1, ϵ_2), and modification masks Z_1 and Z_2, will be provided in the figures and the subsequent discussion.

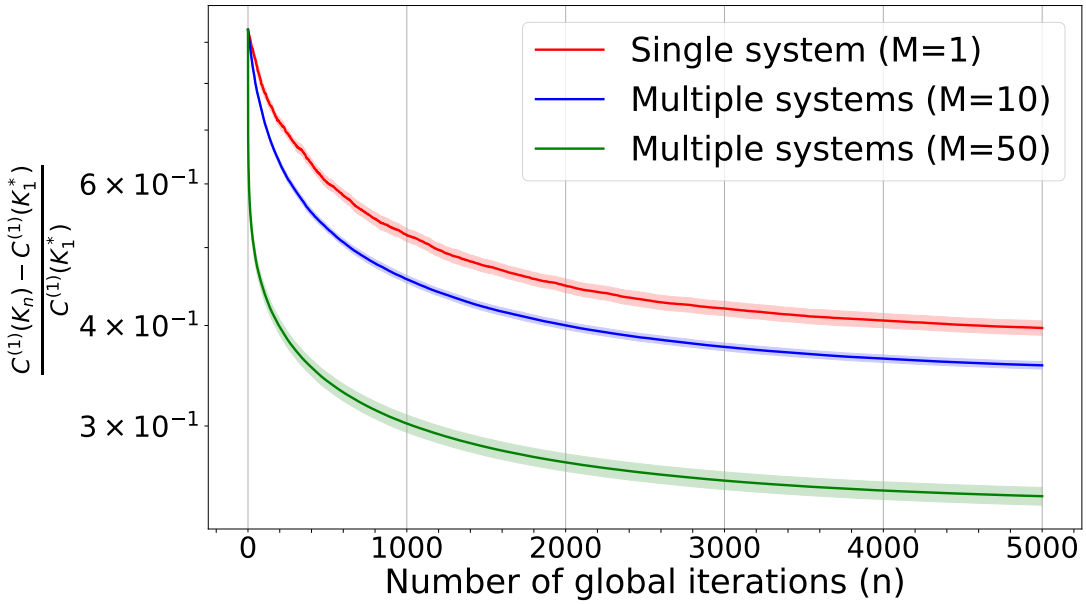


Figure 6.2: Gap between the current and optimal cost with respect to the number of global iterations. Varying the number of systems for a fixed heterogeneity level $\epsilon_1 = 0.5$, $\epsilon_2 = 0.5$.

Figures 6.2 and 6.3 present the normalized distance between the current cost associated with the

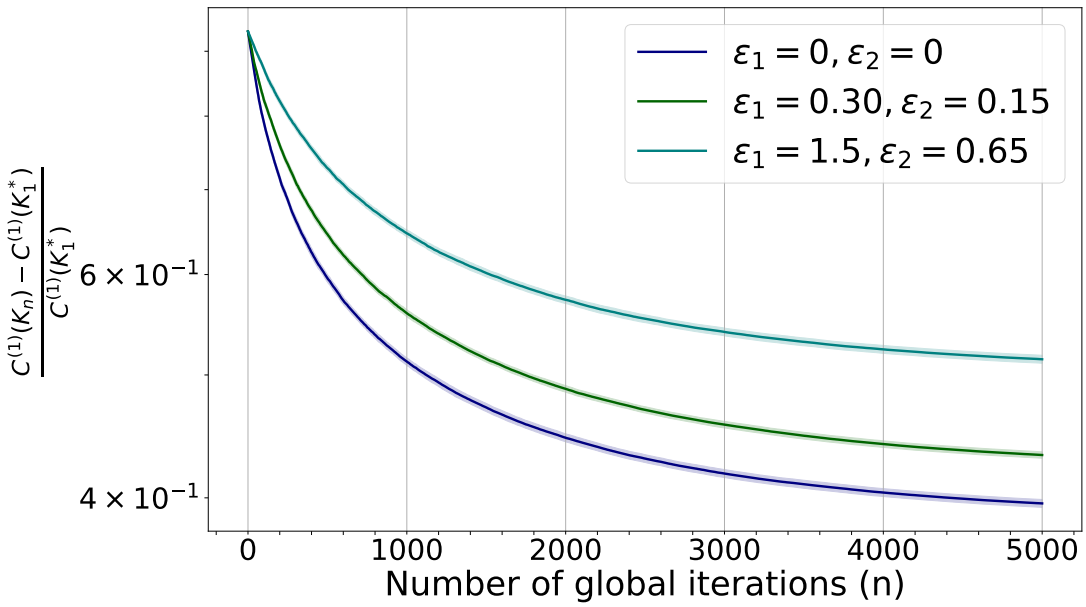


Figure 6.3: Gap between the current and optimal cost with respect to the number of global iterations. Varying the heterogeneity level among the systems, with a fixed number of systems $M = 10$.

common stabilizing controller and the optimal cost for the nominal system, plotted with respect to the number of global iterations. These figures demonstrate the impact of varying the number of systems M and the heterogeneity parameters (ϵ_1, ϵ_2) on the convergence and performance of Algorithm 7.

In Figure 6.2, we specifically investigate the effect of the number of systems M participating in the collaboration to compute a common controller K^* on the convergence of our algorithm. In this analysis, we set the heterogeneity parameters as $\epsilon_1 = 0.5$ and $\epsilon_2 = 0.5$ and consider modification masks $Z_1 = Z_2 = I_3$. The figure reveals a noticeable reduction in the gap between the current and optimal cost as the number of participating systems M increases. This numerical result aligns with our theoretical findings, which indicate that the number of samples required to achieve reliable estimation for the cost function's gradient can be scaled down with the number of systems participating in the collaboration. Consequently, as the number of systems involved increases, there is a considerable reduction in the gap between the common computed controller and the optimal one.

Figure 6.3 illustrates the influence of the heterogeneity parameters (ϵ_1, ϵ_2) on the convergence rate of Algorithm 7. In this analysis, we set the number of systems as $M = 10$, and the modification masks $Z_1 = \texttt{diag}([3.5 \;\; 1 \;\; 0.1])$ and $Z_2 = \texttt{diag}([1.5 \;\; 0.1 \;\; 1])$. Consistent with our theoretical findings, we observe that an increase in the dissimilarity among the systems results in a significant gap between the common and optimal controller. This discrepancy arises due to the additive effect of system heterogeneity on the convergence rate of our algorithm, as elaborated in Theorem 10.

6.9.5 Useful Norm Inequalities

- Given any two matrices A, B of the same dimensions, for any $\xi > 0$, we have

$$\|A+B\|_F^2 \leq (1+\xi)\|A\|_F^2 + \left(1+\frac{1}{\xi}\right)\|B\|_F^2. \tag{6.6}$$

- Given any two matrices A, B of the same dimensions, for any $\xi > 0$, we have

$$\langle A, B\rangle \leq \frac{\xi}{2}\|A\|_F^2 + \frac{1}{2\xi}\|B\|_F^2. \tag{6.7}$$

  This inequality goes by the name of Young's inequality.

- Given m matrices $A_1, \ldots, A_m$ of the same dimensions, the following is a simple application of Jensen's inequality:

$$\begin{aligned} \left\|\sum_{i=1}^m A_i\right\|^2 &\leq m\sum_{i=1}^m \|A_i\|^2, \\ \left\|\sum_{i=1}^m A_i\right\|_F^2 &\leq m\sum_{i=1}^m \|A_i\|_F^2. \end{aligned} \tag{6.8}$$

- Given any two vectors $x, y \in \mathbb{R}^d$, for any constant $\zeta > 0$, we have

$$\|x+y\|^2 \leq (1+\zeta)\|x\|^2 + \left(1+\frac{1}{\zeta}\right)\|y\|^2. \tag{6.9}$$

- Given any two vectors $x, y \in \mathbb{R}^d$, for any constant $\zeta > 0$, we have

$$\langle x, y\rangle \leq \frac{\zeta}{2}\|x\|^2 + \frac{1}{2\zeta}\|y\|^2. \tag{6.10}$$

6.9.6 Useful Lemmas and Constants

Lemma 29. *For each $i \in [M]$, we have that:*

$$||\Sigma_K^{(i)}|| \leq \frac{C^{(i)}(K)}{\sigma_{min}(Q)}, \quad ||P_K^{(i)}|| \leq \frac{C^{(i)}(K)}{\mu}. \tag{6.11}$$

Proof: The proof of this lemma is explained in detail in the proof of Lemma 13 of the supplemental materials in [50]. □

Lemma 30. *(Uniform bounds for $\nabla C(K)$ and $||K||$) For each agent $i \in [M]$, the gradient $\nabla C^{(i)}(K)$ and $\|K\|$ can be bounded as follows:*

$$\|\nabla C^{(i)}(K)\| \leq \|\nabla C^{(i)}(K)\|_F \leq h_1(K) \quad \textit{and} \quad \|K\| \leq h_2(K),$$

where $h_1(K)$, and $h_2(K)$ are some positive scalars depending on the function $C(K)$.

Proof: In this Lemma, $h_1(K)$, and $h_2(K)$ are the functions defined as:

$$h_0(K) := \sqrt{\frac{\|R_K\|_{\max}\left(C_{\max}(K) - C_{\min}(K)\right)}{\mu}},$$
$$h_1(K) := \frac{C_{\max}(K)h_0(K)}{\sigma_{\min}(Q)}, \quad h_2(K) := \frac{h_0(K) + \left\|B^\top P_K A\right\|_{\max}}{\sigma_{\min}(R)},$$

where $\|R_K\|_{\max} := \max_i \left\|R + B^{(i)\top} P_K^{(i)} B^{(i)}\right\|$. By using Lemma 13 of [50], we have

$$\begin{aligned}\|\nabla C^{(i)}(K)\|^2 &\leq \operatorname{Tr}\left(\Sigma_K^{(i)} E_K^{(i)\top} E_K^{(i)} \Sigma_K^{(i)}\right) \leq \left\|\Sigma_K^{(i)}\right\|^2 \operatorname{Tr}\left(E_K^{(i)\top} E_K^{(i)}\right) \\ &\leq \left(\frac{C^{(i)}(K)}{\sigma_{\min}(Q)}\right)^2 \operatorname{Tr}\left(E_K^{(i)\top} E_K^{(i)}\right).\end{aligned}$$

By Lemma 11 of [50], we obtain

$$\operatorname{Tr}\left(E_K^{(i)\top} E_K^{(i)}\right) \leq \frac{\left\|R + B^{(i)\top} P_K^{(i)} B^{(i)}\right\| \left(C^{(i)}(K) - C^{(i)}\left(K_i^*\right)\right)}{\mu},$$

which proves the first claim:

$$\|\nabla C^{(i)}(K)\| \leq \frac{C^{(i)}(K)}{\sigma_{\min}(Q)} \sqrt{\frac{\left\|R + B^{(i)\top} P_K^{(i)} B^{(i)}\right\| \left(C^{(i)}(K) - C^{(i)}\left(K_i^*\right)\right)}{\mu}}$$

$$\leq \frac{C_{\max}(K)}{\sigma_{\min}(Q)} \sqrt{\frac{\left\| R + B^{(i)\top} P_K^{(i)} B^{(i)} \right\|_{\max} \left(C_{\max}(K) - C_{\min}(K) \right)}{\mu}}.$$

On the other hand, by exploiting Lemma 11 of [50] we can also write

$$\begin{aligned}
\|K\| &\leq \left\| \left(R + B^{(i)\top} P_K^{(i)} B^{(i)} \right)^{-1} \right\| \left\| \left(R + B^{(i)\top} P_K^{(i)} B^{(i)} \right) K \right\| \\
&\leq \frac{1}{\sigma_{\min}(R)} \left\| \left(R + B^{(i)\top} P_K^{(i)} B^{(i)} \right) K \right\| \\
&\leq \frac{1}{\sigma_{\min}(R)} \left(\left\| \left(R + B^{(i)\top} P_K^{(i)} B^{(i)} \right) K - B^{(i)\top} P_K^{(i)} A^{(i)} \right\| + \left\| B^{(i)\top} P_K^{(i)} A^{(i)} \right\| \right) \\
&= \frac{\left\| E_K^{(i)} \right\|}{\sigma_{\min}(R)} + \frac{\left\| B^{(i)\top} P_K^{(i)} A^{(i)} \right\|}{\sigma_{\min}(R)} \\
&\leq \frac{\sqrt{\mathrm{Tr}\left(E_K^{(i)\top} E_K^{(i)} \right)}}{\sigma_{\min}(R)} + \frac{\left\| B^{(i)\top} P_K^{(i)} A^{(i)} \right\|}{\sigma_{\min}(R)} \\
&= \frac{\sqrt{\left(C^{(i)}(K) - C^{(i)}\left(K_i^* \right) \right) \left\| R + B^{(i)\top} P_K^{(i)} B^{(i)} \right\|}}{\sqrt{\mu} \sigma_{\min}(R)} + \frac{\left\| B^{(i)\top} P_K^{(i)} A^{(i)} \right\|}{\sigma_{\min}(R)},
\end{aligned}$$

which completes the proof for the second claim. □

It is worth noting that the local cost and gradient smoothness, and gradient domination properties in Lemma 25 and Lemma 26 not only hold for the single-agent setting but also hold for the multi-agent setting. Moreover, we will make use of the following matrix Martingale concentration inequality:

Lemma 31. *(Rectangular Matrix Freedman [204]). Consider a matrix martingale*

$$\{Y_k : k = 0, 1, 2, \ldots\}$$

whose values are matrices with dimension $d_1 \times d_2$, and let $\{X_k : k = 1, 2, 3, \ldots\}$ be the difference sequence. Assume that the difference sequence is uniformly bounded:

$$\|X_k\| \leq R \quad \textit{almost surely} \quad \textit{for } k = 1, 2, 3, \ldots.$$

Define two predictable quadratic variation processes for this martingale:

$$W_{\text{col},k} := \sum_{j=1}^{k} \mathbb{E}_{j-1}\left(X_j X_j^*\right) \text{ and}$$

$$W_{\text{row},k} := \sum_{j=1}^{k} \mathbb{E}_{j-1}\left(X_j^* X_j\right) \quad \text{for } k = 1, 2, 3, \ldots$$

Then, for all $t \geq 0$ *and* $\sigma^2 > 0$,

$$\mathbb{P}\left\{\exists k \geq 0 : \|Y_k\| \geq t \text{ and } \max\left\{\|W_{col,k}\|, \|W_{row,k}\|\right\} \leq \sigma^2\right\} \leq (d_1 + d_2) \cdot \exp\left\{-\frac{-t^2/2}{\sigma^2 + Rt/3}\right\}.$$

a) Proof of Lemma 25

Proof: In this proof, we aim to show

$$\left|C^{(i)}(K') - C^{(i)}(K)\right| \leq h_{\text{cost}}(K)\|K' - K\|,$$

$$\left\|\nabla C^{(i)}(K') - \nabla C^{(i)}(K)\right\| \leq h_{\text{grad}}(K)\|\Delta\| \text{ and } \left\|\nabla C^{(i)}(K') - \nabla C^{(i)}(K)\right\|_F \leq h_{\text{grad}}(K)\|\Delta\|_F,$$

hold for all agents $i \in [M]$ and K' satisfying $\|K' - K\| \leq h_\Delta(K) < \infty$.

The term $h_\Delta(K)$ is the polynomial defined as

$$h_\Delta(K) := \frac{\sigma_{\min}(Q)\mu}{4||B||_{\max} C_{\max}(K)\left(\|A - BK\|_{\max} + 1\right)},$$

the term $h_{\text{cost}}(K)$ and $h_{\text{grad}}(K)$ are defined as

$$h_{\text{cost}}(K) := \frac{4\,\text{Tr}(\Sigma_0)\, C_{\max}(K)\|R\|}{\mu\sigma_{\min}(Q)}\left(\|K\| + \frac{h_\Delta(K)}{2} + \|B\|_{\max}\|K\|^2(\|A - BK\|_{\max} + 1)\frac{C_{\max}(K)}{\mu\sigma_{\min}(Q)}\right)$$

$$\begin{aligned} h_{\text{grad}}(K) := 4\left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)}\right)\Big[&\|R\| + \|B\|_{\max}(\|A\|_{\max} + ||B||_{\max}(\|K\| + h_\Delta(K))) \\ &\times \left(\frac{h_{\text{cost}}(K) C_{\max}(K)}{\text{Tr}(\Sigma_0)}\right) + \|B\|_{\max}^2 \frac{C_{\max}(K)}{\mu}\Big] \\ &+ 8\left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)}\right)^2 \left(\frac{\|B\|_{\max}(\|A - BK\|_{\max} + 1)}{\mu}\right) h_0(K). \end{aligned}$$

For the single-agent (i.e., $M = 1$) setting, the proof is explained in detail in the proof of Lemma 24 and Lemma 25 of the supplemental materials in [50]. For the multi-agent setting (i.e., $M > 1$), we can complete the proof by taking the maximum over the clients $i \in [M]$ of all the system-dependent parameters, such as $\|B\|_{\max}$. □

b) Proof of Lemma 26

Proof: For the single-agent (i.e., $M = 1$) setting, the proof is explained in the proof of Lemma 11 of the supplemental materials in [50]. For the multi-agent setting (i.e., $M > 1$), it is easy to see that

$$C^{(i)}(K) - C^{(i)}\left(K_i^*\right) \leq \frac{\left\|\Sigma_{K_i^*}\right\|}{4\mu^2\sigma_{\min}(R)}\|\nabla C^{(i)}(K)\|_F^2$$

holds for any stabilizing controller K and any agent $i \in [M]$. □

6.9.7 The model-based setting

We first introduce the following operators on a symmetric matrix X,

$$\begin{aligned}\mathcal{T}_K^{(i)}(X) &:= \sum_{t=0}^{\infty}(A^{(i)} - B^{(i)}K)^t X \left[(A^{(i)} - B^{(i)}K)^\top\right]^t, \\ \mathcal{F}_K^{(i)}(X) &:= (A^{(i)} - B^{(i)}K)X(A^{(i)} - B^{(i)}K)^\top. \end{aligned} \tag{6.12}$$

We also define the induced norms of $\mathcal{T}$ and $\mathcal{F}$ as

$$\|\mathcal{T}_K\| = \sup_X \frac{\|\mathcal{T}_K(X)\|}{\|X\|}, \quad \|\mathcal{F}_K\| = \sup_X \frac{\|\mathcal{F}_K(X)\|}{\|X\|}.$$

Lemma 32. *When $(A^{(i)} - B^{(i)}K)$ has spectral radius smaller than 1, we have*

$$\mathcal{T}_K^{(i)} = \left(\mathrm{I} - \mathcal{F}_K^{(i)}\right)^{-1}$$

holds for each $i \in [M]$.

Proof: The proof is explained in detailed in the proof of Lemma 18 in [50]. □

Lemma 33. *If*[15]

$$\left\|\mathcal{T}_K^{(i)}\right\| \left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\| \leq \frac{1}{2} \tag{6.13}$$

holds for any system $i, j \in [M]$, then we have

$$\begin{aligned}\left\|\left(\mathcal{T}_K^{(i)} - \mathcal{T}_K^{(j)}\right)(X)\right\| &\leq 2\left\|\mathcal{T}_K^{(i)}\right\| \left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\| \left\|\mathcal{T}_K^{(i)}(X)\right\| \\ &\leq 2\left\|\mathcal{T}_K^{(i)}\right\|^2 \left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\| \| X\|.\end{aligned}$$

[15]This lemma has a similar flavor to that of Lemma 20 in [50]. It is worthwhile to mention that the inequality (6.13) imposes certain conditions on heterogeneity. Note that the constant $\frac{1}{2}$ can be changed into any finite constant. Thus, this heterogeneity requirement can be subsumed by that in Eq.(6.21).

Proof: Define $\mathcal{A} = \mathrm{I} - \mathcal{F}_K^{(i)}$, and $\mathcal{B} = \mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}$. In this case $\mathcal{A}^{-1} = \mathcal{T}_K^{(i)}$ and $(\mathcal{A}-\mathcal{B})^{-1} = \mathcal{T}_K^{(j)}$. Hence, the condition $\left\|\mathcal{T}_K^{(i)}\right\| \left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\| \leq \frac{1}{2}$ translates to the condition $\|\mathcal{A}^{-1}\| \, \|\mathcal{B}\| \leq \frac{1}{2}$.

First, we observe that

$$\left(\mathcal{A}^{-1} - (\mathcal{A}-\mathcal{B})^{-1}\right)(X) = \left(\mathrm{I} - \left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right)\left(\mathcal{A}^{-1}(X)\right) = \left(\mathrm{I} - \left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right)\left(\mathcal{T}_K^{(i)}(X)\right), \tag{6.14}$$

where $f \circ g$ denotes the composition $f(g(x))$. Since $\left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1} = \mathrm{I} + \mathcal{A}^{-1} \circ \mathcal{B} \circ \left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}$, we have:

$$\left\|\left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right\| \leq 1 + \left\|\mathcal{A}^{-1} \circ \mathcal{B}\right\| \left\|\left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right\| \leq 1 + \frac{1}{2}\left\|\left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right\| \tag{6.15}$$

Now rearranging terms in Eq.(6.15), we obtain $\left\|\left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right\| \leq 2$. Therefore, we have

$$\begin{aligned}\left\|\mathrm{I} - \left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right\| = \left\|\mathcal{A}^{-1} \circ \mathcal{B} \circ \left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right\| &\leq \left\|\mathcal{A}^{-1}\right\| \|\mathcal{B}\| \left\|\left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right\| \\ &\leq 2\left\|\mathcal{A}^{-1}\right\| \|\mathcal{B}\|,\end{aligned}$$

and so

$$\left\|\mathrm{I} - \left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right\| \leq 2\left\|\mathcal{A}^{-1}\right\| \|\mathcal{B}\| = 2\left\|\mathcal{T}_K^{(i)}\right\| \left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\|. \tag{6.16}$$

Then, we have

$$\begin{aligned}\left\|\left(\mathcal{T}_K^{(i)} - \mathcal{T}_K^{(j)}\right)(X)\right\| &= \left\|\left(\mathcal{A}^{-1} - (\mathcal{A}-\mathcal{B})^{-1}\right)(X)\right\| \\ &\overset{(a)}{\leq} \left\|\left(\mathrm{I} - \left(\mathrm{I} - \mathcal{A}^{-1} \circ \mathcal{B}\right)^{-1}\right)\right\| \left\|\mathcal{T}_K^{(i)}(X)\right\| \\ &\overset{(b)}{\leq} 2\left\|\mathcal{T}_K^{(i)}\right\| \left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\| \left\|\mathcal{T}_K^{(i)}(X)\right\| \\ &\leq 2\left\|\mathcal{T}_K^{(i)}\right\| \left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\| \left\|\mathcal{T}^{(i)}\right\| \|X\|,\end{aligned}$$

where (a) is due to Eq.(6.14) and (b) is due to Eq.(6.16). This completes the proof of Lemma 33. □.

a) Proof of Lemma 27

Proof: First, we know that $\nabla C^{(i)}(K)$ and $\nabla C^{(j)}(K)$ are given by,

$$\nabla C^{(i)}(K) = 2E_K^{(i)}\Sigma_K^{(i)}, \quad \text{and} \quad \nabla C^{(j)}(K) = 2E_K^{(j)}\Sigma_K^{(j)}$$

where,

$$E_K^{(i)} = (R + B^{(i)\top}P_K^{(i)}B^{(i)})K - B^{(i)\top}P_K^{(i)}A^{(i)},$$

and

$$\Sigma_K^{(i)} =_{x_0^{(i)}\sim D} \sum_{t=0}^{\infty} x_t^{(i)}x_t^{(i)\top}.$$

Thus, we can write,

$$\begin{aligned}||\nabla C^{(i)}(K) - \nabla C^{(j)}(K)|| &= ||2E_K^{(i)}\Sigma_K^{(i)} - 2E_K^{(j)}\Sigma_K^{(j)}|| \\ &\leq 2(||E_K^{(i)} - E_K^{(j)}||\underbrace{||\Sigma_K^{(i)}||}_{\beta_1} + \underbrace{||E_K^{(j)}||}_{\beta_2}||\Sigma_K^{(i)} - \Sigma_K^{(j)}||).\end{aligned}$$

From Eq. (6.11) we can upper bound $||\Sigma_K^{(i)}||$ as:

$$||\Sigma_K^{(i)}|| \leq \frac{C^{(i)}(K)}{\sigma_{\min}(Q)}.$$

With the definition of $E_K^{(j)} = RK + B^{(j)\top}P_K^{(j)}B^{(j)}K - B^{(j)\top}P_K^{(j)}A^{(j)}$, we can use triangle inequality to write,

$$\begin{aligned}||E_K^{(j)}|| &\leq ||RK|| + ||B^{(j)}||||P_K^{(j)}||||B^{(j)}K|| + ||B^{(j)}||||P_K^{(j)}||||A^{(j)}|| \\ &\leq ||RK|| + \frac{||B^{(j)}||C^{(j)}(K)}{\mu}(||B^{(j)}K|| + ||A^{(j)}||),\end{aligned}$$

where $||P_K^{(j)}|| \leq \frac{C^{(j)}(K)}{\mu}$ from Eq. (6.11), with $\mu = \sigma_{\min}(\Sigma_0^{(j)})$.

With the notation that we introduced previously, we can write

$$\beta_1 = ||\Sigma_K^{(i)}|| \leq ||\Sigma_K||_{\max} \leq \frac{C_{\max}(K)}{\sigma_{\min}(Q)},$$

and,

$$\beta_2 = ||E_K^{(j)}|| \leq ||E_K||_{\max} \leq ||R||||K|| + \frac{||B||_{\max} C_{\max}(K)}{\mu}(||B||_{\max} + ||A||_{\max}),$$

where $C_{\max}(K) := \max_i C^{(i)}(K)$.

Next we will derive an upper bound for $||E_K^{(i)} - E_K^{(j)}||$.

Upper bound for $||E_K^{(i)} - E_K^{(j)}||$: We can first use the definition of $E_K^{(i)}$ and $E_K^{(j)}$ to write,

$$\begin{aligned}
E_K^{(i)} - E_K^{(j)} &= B^{(j)\top} P_K^{(j)} (A^{(j)} - B^{(j)} K) - B^{(i)\top} P_K^{(i)} (A^{(i)} - B^{(i)} K) \\
&= -B^{(i)\top} P_K^{(i)} (A^{(i)} - B^{(i)} K) + B^{(i)\top} P_K^{(i)} (A^{(j)} - B^{(j)} K) - B^{(i)\top} P_K^{(i)} (A^{(j)} - B^{(j)} K) \\
&+ B^{(i)\top} P_K^{(j)} (A^{(j)} - B^{(j)} K) - B^{(i)\top} P_K^{(j)} (A^{(j)} - B^{(j)} K) + B^{(j)\top} P_K^{(j)} (A^{(j)} - B^{(j)} K).
\end{aligned}$$

Then, by using triangle inequality, we obtain the following expression:

$$\begin{aligned}
||E_K^{(i)} - E_K^{(j)}|| \leq & ||\underbrace{B^{(i)\top} P_K^{(i)} (A^{(i)} - B^{(i)} K) - B^{(i)\top} P_K^{(i)} (A^{(j)} - B^{(j)} K)}_{H_1}|| \\
& + ||\underbrace{B^{(i)\top} P_K^{(i)} (A^{(j)} - B^{(j)} K) - B^{(i)\top} P_K^{(j)} (A^{(j)} - B^{(j)} K)}_{H_2}|| \\
& + ||\underbrace{B^{(i)\top} P_K^{(j)} (A^{(j)} - B^{(j)} K) - B^{(j)\top} P_K^{(j)} (A^{(j)} - B^{(j)} K)}_{H_3}||.
\end{aligned}$$

Incorporating the heterogeneity bounds from assumption 1 gives

$$||H_1|| \leq ||B^{(i)}||||P_K^{(i)}||(\epsilon_1 + \epsilon_2||K||),$$

to which we apply the max-norm definition to arrive at

$$||H_1|| \leq ||B||_{\max}(\epsilon_1 + \epsilon_2||K||)||P_K||_{\max}. \tag{6.17}$$

Similarly, we can also derive upper bounds for $||H_2||$ and $||H_3||$, as follows,

$$||H_2|| \leq ||B^{(i)}||||P_K^{(i)} - P_K^{(j)}||||A^{(j)} - B^{(j)}K|| \leq ||B||_{\max}||P_K^{(i)} - P_K^{(j)}||||A - BK||_{\max} \quad (6.18)$$

and

$$||H_3|| \leq \epsilon_2||A^{(i)} - B^{(i)}K||||P_K^{(j)}|| \leq \epsilon_2||A - BK||_{\max}||P_K||_{\max}. \quad (6.19)$$

To bound H_2, we need to derive an upper bound for $||P_K^{(i)} - P_K^{(j)}||$. For this purpose, we have that for any fixed system $i \in [M]$

$$||P_K^{(i)} - P_K^{(j)}|| = \left\| \mathcal{T}_K^{(i)}\left(Q + K^\top RK\right) - \mathcal{T}_K^{(j)}\left(Q + K^\top RK\right) \right\|.$$

Thus, by using Lemma 33, we can write,

$$||P_K^{(i)} - P_K^{(j)}|| \leq 2\left\|\mathcal{T}_K^{(i)}\right\|^2 \left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\| \left\|Q + K^\top RK\right\|,$$

where $||\mathcal{T}_K^{(i)}|| \leq \frac{C^{(i)}(K)}{\sigma_{\min}(Q)\mu} \leq \frac{C_{\max}(K)}{\sigma_{\min}(Q)\mu}$ (detailed in Lemma 17 of [50]). With the following upper bound for $\left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\|$:

$$\begin{aligned} ||(\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)})(X)|| &= ||(A^{(i)} - B^{(i)}K)X(A^{(i)} - B^{(i)}K)^\top \\ &\quad - (A^{(j)} - B^{(j)}K)X(A^{(j)} - B^{(j)}K)^\top|| \\ &\leq 2(\epsilon_1 + \epsilon_2||K||)||X||||A - BK||_{\max}, \end{aligned}$$

we have

$$||P_K^{(i)} - P_K^{(j)}|| \leq 4\left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)\mu}\right)^2 (\epsilon_1 + \epsilon_2||K||)||A - BK||_{\max}(||Q|| + ||R||||K||^2), \quad (6.20)$$

Plugging in Eq. (6.20) into H_2 and adding the upper bounds of H_1 (Eq. 6.17), H_2 (Eq. 6.18) and H_3 (Eq. 6.19) together, we have

$$||E_K^{(i)} - E_K^{(j)}|| \leq g_1(\epsilon_1, \epsilon_2, K),$$

where g_1 is a linear in ϵ_1, ϵ_2 and polynomial in the remaining problem data. Specifically,

$$g_1(\epsilon_1, \epsilon_2, K) = \epsilon_1 \left(\frac{||B||_{\max} C_{\max}(K)}{\mu} \left[1 + 4 \left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)\mu} \right) (||A - BK||_{\max})^2 \left(||Q|| + ||R|| ||K||^2 \right) \right] \right)$$
$$+ \epsilon_2 \left(\frac{||B||_{\max} ||K|| C_{\max}(K)}{\mu} \left[1 + 4 \left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)\mu} \right) (||A - BK||_{\max})^2 \left(||Q|| + ||R|| ||K||^2 \right) \right] + ||A - BK||_{\max} \right).$$

In what follows, we will derive an upper bound for $||\Sigma_K^{(i)} - \Sigma_K^{(j)}||$:

Upper bound for $||\Sigma_K^{(i)} - \Sigma_K^{(j)}||$: From the previous definitions in Eq.(6.12) and Lemma 33, we have,

$$\begin{aligned}||\Sigma_K^{(i)} - \Sigma_K^{(j)}|| = ||\mathcal{T}_K^{(i)}(\Sigma_0) - \mathcal{T}_K^{(j)}(\Sigma_0)|| &\leq 2\left\|\mathcal{T}_K^{(i)}\right\|^2 \left\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\right\| \| \Sigma_0\| \\ &\leq 4\left(\frac{C_{\max}}{\sigma_{\min}(Q)\mu}\right)^2 (\epsilon_1 + \epsilon_2||K||)||A - BK||_{\max}||\Sigma_0||\end{aligned}$$

where $\Sigma_0 = \mathbb{E}_{x_0^{(i)} \sim \mathcal{D}}\left[x_0^{(i)} x_0^{(i)\top}\right]$.

Thus, we have the following upper bound for $||\Sigma_K^{(i)} - \Sigma_K^{(j)}||$,

$$||\Sigma_K^{(i)} - \Sigma_K^{(j)}|| \leq g_2(\epsilon_1, \epsilon_2, K)$$

with,

$$\boxed{g_2(\epsilon_1, \epsilon_2, K) = \epsilon_1\left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)\mu}\right)^2 (4||A - BK||_{\max}||\Sigma_0||) + \epsilon_2||K||\left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)\mu}\right)^2 (4||A - BK||_{\max}||\Sigma_0||)\,.}$$

Therefore, we can finally write an upper bound for $||\nabla C^{(i)}(K) - \nabla C^{(j)}(K)||$, which is:

$$||\nabla C^{(i)}(K) - \nabla C^{(j)}(K)|| \leq f(\epsilon_1, \epsilon_2, K)$$

where,

$$f(\epsilon_1, \epsilon_2, K) = 2(\beta_1 g_1(\epsilon_1, \epsilon_2, K) + \beta_2 g_2(\epsilon_1, \epsilon_2, K)).$$

After some rearrangement, we have that

$$f(\epsilon_1, \epsilon_2, K) = \epsilon_1 h_{\text{het}}^1(K) + \epsilon_1 h_{\text{het}}^2(K),$$

where $h_{\text{het}}^1 = h_{1f} + h_{2f}$ and $h_{\text{het}}^2 = h_{3f} + h_{4f}$, and

$$h_{1f} = \frac{2||B||_{\max}(C_{\max}(K))^2}{\sigma_{\min}(Q)\mu}\left[1+4\left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)\mu}\right)(||A-BK||_{\max})^2\left(||Q||+||R||||K||^2\right)\right],$$
$$h_{2f} = \frac{2}{\mu}\left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)}\right)^3(4||A-BK||_{\max}||\Sigma_0||),$$

$$h_{3f} = 2\left(||R||||K||+\frac{||B||_{\max}C_{\max}(K)}{\mu}(||B||_{\max}+||A||_{\max})\right)$$
$$\times\left(\frac{||B||_{\max}||K||C_{\max}(K)}{\mu}\left[1+4\left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)\mu}\right)(||A-BK||_{\max})^2\left(||Q||+||R||||K||^2\right)\right]+||A-BK||_{\max}\right),$$
$$h_{4f} = 2\left(||R||||K||+\frac{||B||_{\max}C_{\max}(K)}{\mu}(||B||_{\max}+||A||_{\max}\right)||K||\left(\frac{C_{\max}(K)}{\sigma_{\min}(Q)\mu}\right)^2(4||A-BK||_{\max}||\Sigma_0||).$$

b) Proof of Theorem 9

In this theorem, we consider the setting where $(\epsilon_1\bar{h}^1_{\text{het}}+\epsilon_2\bar{h}^2_{\text{het}})^2 \le \bar{h}^3_{\text{het}}$ with

$$\bar{h}^3_{\text{het}} := \min_{j\in[M]}\left\{\frac{\mu^2\sigma_{\min}(R)\left(C^{(j)}(K_0)-C^{(j)}(K_j^*)\right)}{4||\Sigma_{K_j^*}||\min\{n_x,n_u\}}\right\}. \tag{6.21}$$

Outline: To prove Theorem 9, we first introduce some lemmas: Lemma 34 establishes stability of the local policies; Lemma 35 provides the drift analysis; Lemma 36 quantifies the per-round progress of our `FedLQR` algorithm. As a result, we are able to present the iterative stability guarantees and convergence analysis of `FedLQR` in the model-based setting.

Lemma 34. *(Stability of the local policies) Suppose $K_n \in \mathcal{G}^0$. If the local step-size satisfies $\eta_l \le \min\{\frac{\bar{h}_\Delta}{\bar{h}_1}, \frac{1}{4\bar{h}_{grad}}\}$ and the heterogeneity level satisfies $(\epsilon_1\bar{h}^1_{het}+\epsilon_2\bar{h}^2_{het})^2 \le \bar{h}^3_{het}$, then $K^{(i)}_{n,l} \in \mathcal{G}^0$ holds for all $i \in [M]$ and $l \in [L]$.*

Proof: Since $K_n \in \mathcal{G}^0$, based on the local Lipschitz property in Lemma 25, we have:

$$\begin{aligned} C^{(j)}(K_{n,1}^{(i)}) - C^{(j)}(K_n) &\leq \left\langle \nabla C^{(j)}(K_n), K_{n,1}^{(i)} - K_n \right\rangle + \frac{h_{\text{grad}}(K_n)}{2} \left\| K_{n,1}^{(i)} - K_n \right\|_F^2 \\ &\leq -\left\langle \nabla C^{(j)}(K_n), \eta_l \nabla C^{(i)}(K_n) \right\rangle + \frac{h_{\text{grad}}(K_n)}{2} \left\| \eta_l \nabla C^{(i)}(K_n) \right\|_F^2 \end{aligned} \tag{6.22}$$

holds for any $i, j \in [M]$, if $\left\| \eta_l \nabla C^{(i)}(K_n) \right\|_F \leq \underline{h}_\Delta \leq h_\Delta(K_n)$, which holds when

$$\left\| \eta_l \nabla C^{(i)}(K_n) \right\|_F \overset{(a)}{\leq} \eta_l h_1(K_n) \leq \eta_l \bar{h}_1 \overset{(b)}{\leq} \underline{h}_\Delta,$$

where (a) comes from Lemma 30 and (b) holds because of the requirement on η_l in the statement of the lemma.

Following the analysis in Eq (6.22), we have

$$\begin{aligned} C^{(j)}(K_{n,1}^{(i)}) - C^{(j)}(K_n) \leq & -\eta_l \left\langle \nabla C^{(j)}(K_n), \nabla C^{(j)}(K_n) \right\rangle \\ & - \eta_l \underbrace{\left\langle \nabla C^{(j)}(K_n), \nabla C^{(i)}(K_n) - \nabla C^{(j)}(K_n) \right\rangle}_{T_1} \\ & + \frac{h_{\text{grad}}(K_n)}{2} \left\| \eta_l \nabla C^{(i)}(K_n) \right\|_F^2. \end{aligned} \tag{6.23}$$

Now T_1 can be bounded as

$$\begin{aligned} T_1 &\leq \eta_l \left\| \nabla C^{(j)}(K_n) \right\|_F \left\| \nabla C^{(i)}(K_n) - \nabla C^{(j)}(K_n) \right\|_F \\ &\leq \eta_l \sqrt{\min\{n_x, n_u\}} \left\| \nabla C^{(j)}(K_n) \right\|_F \left\| \nabla C^{(i)}(K_n) - \nabla C^{(j)}(K_n) \right\| \\ &\overset{(c)}{\leq} \eta_l \sqrt{\min\{n_x, n_u\}} \left\| \nabla C^{(j)}(K_n) \right\|_F (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2), \end{aligned} \tag{6.24}$$

where (c) is due to Lemma 27. Plugging in the upper bound of T_1 into (6.22), we have:

$$\begin{aligned} C^{(j)}(K_{n,1}^{(i)}) - C^{(j)}(K_n) \leq & -\eta_l \left\langle \nabla C^{(j)}(K_n), \nabla C^{(j)}(K_n) \right\rangle \\ & + \eta_l \sqrt{\min\{n_x, n_u\}} \left\| \nabla C^{(j)}(K_n) \right\|_F (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2) + \frac{h_{\text{grad}}(K_n)}{2} \left\| \eta_l \nabla C^{(i)}(K_n) \right\|_F^2 \\ \overset{(d)}{\leq} & -\eta_l \left\langle \nabla C^{(j)}(K_n), \nabla C^{(j)}(K_n) \right\rangle + \eta_l \sqrt{\min\{n_x, n_u\}} \left\| \nabla C^{(j)}(K_n) \right\|_F (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2) \end{aligned}$$

$$
\begin{aligned}
&+ h_{\text{grad}}(K_n)\left\|\eta_l \nabla C^{(j)}(K_n)\right\|_F^2 + h_{\text{grad}}(K_n)\left\|\eta_l \nabla C^{(i)}(K_n) - \eta_l \nabla C^{(j)}(K_n)\right\|_F^2 \\
&\overset{(e)}{\leq} -\eta_l \left\langle \nabla C^{(j)}(K_n), \nabla C^{(j)}(K_n)\right\rangle + \eta_l \sqrt{\min\{n_x, n_u\}}\left\|\nabla C^{(j)}(K_n)\right\|_F (\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}}) \\
&+ \eta_l^2 h_{\text{grad}}(K_n)\left\|\nabla C^{(j)}(K_n)\right\|_F^2 + \eta_l^2 h_{\text{grad}}(K_n) \min\{n_x, n_u\}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}})^2 \\
&\overset{(f)}{\leq} -\eta_l \left\langle \nabla C^{(j)}(K_n), \nabla C^{(j)}(K_n)\right\rangle + \frac{\eta_l}{4}\left\|\nabla C^{(j)}(K_n)\right\|_F^2 + \eta_l \min\{n_x, n_u\}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}})^2 \\
&+ \eta_l^2 \bar{h}_{\text{grad}}\left\|\nabla C^{(j)}(K_n)\right\|_F^2 + \eta_l^2 \bar{h}_{\text{grad}} \min\{n_x, n_u\}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}})^2 \\
&= -\eta_l \left\langle \nabla C^{(j)}(K_n), \nabla C^{(j)}(K_n)\right\rangle + (\frac{\eta_l}{4} + \eta_l^2 \bar{h}_{\text{grad}})\left\|\nabla C^{(j)}(K_n)\right\|_F^2 \\
&+ (\eta_l + \eta_l^2 \bar{h}_{\text{grad}}) \min\{n_x, n_u\}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}})^2 \\
&\overset{(g)}{\leq} -\frac{\eta_l}{2}\left\|\nabla C^{(j)}(K_n)\right\|_F^2 + 2\eta_l \min\{n_x, n_u\}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}})^2,
\end{aligned}
$$

which implies

$$
\begin{aligned}
C^{(j)}(K^{(i)}_{n,1}) - C^{(j)}(K^*) &\overset{(h)}{\leq} \left(1 - \frac{2\eta_l \mu^2 \sigma_{\min}(R)}{\left\|\Sigma_{K^*_j}\right\|}\right)(C^{(j)}(K_0) - C^{(j)}(K^*_j)) \\
&+ 2\eta_l \min\{n_x, n_u\}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}})^2, \qquad (6.25)
\end{aligned}
$$

where (d) is due to Eq. (6.8); (e) is due to Lemma 27; (f) is due to Eq.(6.10) with $\zeta = \frac{1}{2}$; (g) is due to the choice of step-size such that $\eta_l^2 \bar{h}_{\text{grad}} \leq \frac{\eta_l}{4}$, which holds when $\eta_l \leq \frac{1}{4\bar{h}_{\text{grad}}}$; and (h) is due to Lemma 26 and the fact that $K_n \in \mathcal{G}^0$. If ϵ_1 and ϵ_2 are small enough that

$$
(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}})^2 \leq \min_{j \in [M]} \left\{ \frac{\mu^2 \sigma_{\min}(R)\left(C^{(j)}(K_0) - C^{(j)}(K^*_j)\right)}{4||\Sigma_{K^*_j}||\min\{n_x, n_u\}} \right\},
$$

we have that

$$
C^{(j)}(K^{(i)}_{n,1}) - C^{(j)}(K_n) \leq C^{(j)}(K_0) - C^{(j)}(K^*_j),
$$

holds for any $j \in [M]$.

The above inequality implies $K^{(i)}_{n,1} \in \mathcal{G}^0$ as long as $K_n \in \mathcal{G}^0$. Then we can use the induction method to obtain that $K^{(i)}_{n,2} \in \mathcal{G}^0$ since $K^{(i)}_{n,1} \in \mathcal{G}^0$. As a result, an identical argument can be used

from $K_{n,1}^{(i)}$ to $K_{n,2}^{(i)}$. Therefore, by repeating this step for L times, we have that all the local polices $K_{n,l}^{(i)} \in \mathcal{G}^0$ holds for all $i \in [M]$ and $l = 1, \cdots, L$, when the global policy $K_n \in \mathcal{G}^0$. □

Lemma 35. *(Drift term analysis) If* $\eta_l \leq \min\left\{\frac{1}{4\bar{h}_{grad}}, \frac{1}{2}, \frac{h_\Delta}{\bar{h}_1}, \frac{\log 2}{L(3\bar{h}_{grad}+1)}\right\}$ *and* $K_n \in \mathcal{G}^0$*, the difference between the local policy and global policy can be bounded as follows* $\forall i \in [M]$ *and* $l \in [L]$*:*

$$\left\|K_{n,l}^{(i)} - K_n\right\|_F^2 \leq 2\eta_l L \left\|\nabla C^{(i)}(K_n)\right\|_F^2 = \frac{2\eta}{\eta_g}\left\|\nabla C^{(i)}(K_n)\right\|_F^2.$$

Proof: We have

$$\begin{aligned}
\left\|K_{n,l}^{(i)} - K_n\right\|_F^2 &= \left\|K_{n,l-1}^{(i)} - K_n - \eta_l \nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&= \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 - 2\eta_l \left[\left\langle \nabla C^{(i)}(K_{n,l-1}^{(i)}), K_{n,l-1}^{(i)} - K_n \right\rangle\right] \\
&+ \left\|\eta_l \nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&= \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 - 2\eta_l \left[\left\langle \nabla C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_n), K_{n,l-1}^{(i)} - K_n \right\rangle\right] \\
&- 2\eta_l \left[\left\langle \nabla C^{(i)}(K_n), K_{n,l-1}^{(i)} - K_n \right\rangle\right] + \left\|\eta_l \nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&\leq \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 + 2\eta_l \left\|\nabla C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_n)\right\|_F \left\|K_{n,l-1}^{(i)} - K_n\right\|_F \\
&+ 2\eta_l \left\|\nabla C^{(i)}(K_n)\right\|_F \left\|K_{n,l-1}^{(i)} - K_n\right\|_F + \left\|\eta_l \nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&\overset{(a)}{\leq} \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 + 2\eta_l h_{\text{grad}}(K_n) \left\|K_{n,l-1}^{(i)} - K_n\right\|_F \left\|K_{n,l-1}^{(i)} - K_n\right\|_F \\
&+ \eta_l \left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \eta_l \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 + \left\|\eta_l \nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&\leq (1 + 2\eta_l h_{\text{grad}}(K_n) + \eta_l) \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 + (\eta_l + 2\eta_l^2) \left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
&+ 2\eta_l^2 \left\|\nabla C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_n)\right\|_F^2 \\
&\overset{(b)}{\leq} (1 + 2\eta_l h_{\text{grad}}(K_n) + \eta_l) \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 + (\eta_l + 2\eta_l^2) \left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
&+ 2\eta_l^2 h_{\text{grad}}^2(K_n) \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2
\end{aligned}$$

$$\overset{(c)}{\leq} \left(1 + 2\eta_l \bar{h}_{\text{grad}} + \eta_l + 2\eta_l^2 \bar{h}_{\text{grad}}^2\right) \left\| K_{n,l-1}^{(i)} - K_n \right\|_F^2 + (\eta_l + 2\eta_l^2) \left\| \nabla C^{(i)}(K_n) \right\|_F^2$$
$$\overset{(d)}{\leq} \left(1 + 3\eta_l \bar{h}_{\text{grad}} + \eta_l\right) \left\| K_{n,l-1}^{(i)} - K_n \right\|_F^2 + 2\eta_l \left\| \nabla C^{(i)}(K_n) \right\|_F^2, \tag{6.26}$$

where (a) and (b) are due to Lemma 25; (c) is due to the fact that $K_n \in \mathcal{G}^0$; (d) is due to the choice of step-size such that $2\eta_l^2 \bar{h}_{\text{grad}}^2 \leq \eta_l \bar{h}_{\text{grad}}$ and $2\eta_l^2 \leq \eta_l$, which hold when $\eta_l \leq \min\{\frac{1}{2\bar{h}_{\text{grad}}}, \frac{1}{2}\}$. Therefore, we have

$$\begin{aligned}
\left\| K_{n,l}^{(i)} - K_n \right\|_F^2 &\leq (1 + 3\eta_l \bar{h}_{\text{grad}} + \eta_l) \left\| K_{n,l-1}^{(i)} - K_n \right\|_F^2 + 2\eta_l \left\| \nabla C^{(i)}(K_n) \right\|_F^2 \\
&\leq (1 + 3\eta_l \bar{h}_{\text{grad}} + \eta_l)^l \underbrace{\left\| K_{n,0}^{(i)} - K_n \right\|_F^2}_{=0} \\
&\quad + 2 \sum_{j=0}^{l-1} \left(1 + 3\eta_l \bar{h}_{\text{grad}} + \eta_l\right)^j \eta_l \left\| \nabla C^{(i)}(K_n) \right\|_F^2 \\
&\leq \frac{\left(1 + 3\eta_l \bar{h}_{\text{grad}} + \eta_l\right)^l - 1}{\left(1 + 3\eta_l \bar{h}_{\text{grad}} + \eta_l\right) - 1} 2\eta_l \left\| \nabla C^{(i)}(K_n) \right\|_F^2 \\
&\overset{(a)}{\leq} 2 \times \frac{1 + l(3\eta_l \bar{h}_{\text{grad}} + \eta_l) - 1}{3\bar{h}_{\text{grad}} + 1} \left\| \nabla C^{(i)}(K_n) \right\|_F^2 \\
&\leq 2\eta_l L \left\| \nabla C^{(i)}(K_n) \right\|_F^2,
\end{aligned}$$

where, for (a), we used the fact that $(1 + x)^{\tau+1} \leq 1 + 2x(\tau + 1)$ holds for $x \leq \frac{\log 2}{\tau}$. In other words, $\left(1 + 3\eta_l \bar{h}_{\text{grad}} + \eta_l\right)^l \leq 1 + l(3\eta_l \bar{h}_{\text{grad}} + \eta_l)$ holds when $3\eta_l \bar{h}_{\text{grad}} + \eta_l \leq \frac{\log 2}{l}$, i.e., when $\eta_l \leq \frac{\log 2}{L(3\bar{h}_{\text{grad}}+1)}$. □

Lemma 36. *(Per round progress) Suppose $K_n \in \mathcal{G}^0$. If we choose the local step-size as*

$$\eta_l = \frac{1}{2} \min \left\{ \frac{1}{4\bar{h}_{grad}}, \frac{1}{2}, \frac{\underline{h}_\Delta}{\bar{h}_1}, \frac{\log 2}{L(3\bar{h}_{grad} + 1)}, \frac{1}{80L\bar{h}_{grad}^2} \right\},$$

choose $\eta = \frac{1}{2} \min\{\frac{\underline{h}_\Delta}{\bar{h}_1}, 1, \frac{2}{3\bar{h}_{grad}}\}$, and the global step-size as $\eta_g = \frac{\eta}{L\eta_l}$, then, for all $i \in [M]$, it

holds that

$$C^{(i)}(K_{n+1}) - C^{(i)}(K_n) \leq -\frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|_{\max}}(C^{(i)}(K_n) - C^{(i)}(K_i^*)) + 3\eta\min\{n_x, n_u\}(\epsilon_1\bar{h}_{het}^1 + \epsilon_2\bar{h}_{het}^2)^2. \tag{6.27}$$

Proof:

$$\begin{aligned}
C^{(i)}(K_{n+1}) - C^{(i)}(K_n) &\overset{(a)}{\leq} \langle \nabla C^{(i)}(K_n), K_{n+1} - K_n\rangle + \frac{h_{\text{grad}}(K_n)}{2}\|K_{n+1} - K_n\|_F^2 \\
&= -\left\langle \nabla C^{(i)}(K_n), \frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\nabla C^{(j)}(K_{n,l}^{(j)})\right\rangle + \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\nabla C^{(j)}(K_{n,l}^{(j)})\right\|_F^2 \\
&= -\left\langle \nabla C^{(i)}(K_n), \frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left[\nabla C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(i)}(K_n)\right]\right\rangle - \eta\left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
&+ \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\nabla C^{(j)}(K_{n,l}^{(j)})\right\|_F^2 \\
&= -\left\langle \nabla C^{(i)}(K_n), \frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left[\nabla C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_n) + \nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n)\right]\right\rangle \\
&- \eta\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\nabla C^{(j)}(K_{n,l}^{(j)})\right\|_F^2 \\
&\leq \eta\left\|\nabla C^{(i)}(K_n)\right\|_F\left\|\frac{1}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left[\nabla C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_n)\right]\right\|_F \\
&+ \frac{\eta}{M}\sum_{j=1}^{M}\left\|\nabla C^{(i)}(K_n)\right\|_F\left\|\nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n)\right\|_F \\
&- \eta\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\nabla C^{(j)}(K_{n,l}^{(j)})\right\|_F^2 \\
&\overset{(b)}{\leq} \eta\left\|\nabla C^{(i)}(K_n)\right\|_F\left[\frac{h_{\text{grad}}(K_n)}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left\|K_{n,l}^{(j)} - K_n\right\|_F\right] \\
&+ \frac{\eta}{4}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \frac{\eta}{M}\sum_{j=1}^{M}\left\|\nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n)\right\|_F^2 - \eta\left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
&+ h_{\text{grad}}(K_n)\frac{3\eta^2}{2ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left\|\nabla C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_n)\right\|_F^2
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{3\eta^2 h_{\text{grad}}(K_n)}{2M} \sum_{j=1}^{M} \left\| \nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n) \right\|_F^2 + \frac{3\eta^2 h_{\text{grad}}(K_n)}{2} \left\| \nabla C^{(i)}(K_n) \right\|_F^2 \\
&\overset{(c)}{\leq} \frac{\eta}{4} \left\| \nabla C^{(i)}(K_n) \right\|_F^2 + \frac{\eta \bar{h}_{\text{grad}}^2}{ML} \sum_{i=1}^{M} \sum_{l=0}^{L-1} \left\| K_{n,l}^{(i)} - K_n \right\|_F^2 \\
&+ \frac{\eta}{4} \left\| \nabla C^{(i)}(K_n) \right\|_F^2 + \left(\eta + \frac{3\eta^2 \bar{h}_{\text{grad}}}{2} \right) \frac{1}{M} \sum_{j=1}^{M} \left\| \nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n) \right\|_F^2 \\
&- \eta \left\| \nabla C^{(i)}(K_n) \right\|_F^2 + \frac{3\eta^2 \bar{h}_{\text{grad}}^2}{2ML} \sum_{j=1}^{M} \sum_{l=0}^{L-1} \left\| K_{n,l}^{(j)} - K_n \right\|_F^2 + \frac{\eta}{8} \left\| \nabla C^{(i)}(K_n) \right\|_F^2 \\
&\overset{(d)}{\leq} -\frac{3\eta}{8} \left\| \nabla C^{(i)}(K_n) \right\|_F^2 + \frac{5\eta^2 \bar{h}_{\text{grad}}^2}{\eta_g M} \sum_{j=1}^{M} \left\| \nabla C^{(j)}(K_n) \right\|_F^2 \\
&+ 2\eta \min\{n_x, n_u\} (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2)^2 \\
&\overset{(e)}{\leq} -\frac{3\eta}{8} \left\| \nabla C^{(i)}(K_n) \right\|_F^2 + \frac{10\eta^2 \bar{h}_{\text{grad}}^2}{\eta_g M} \sum_{j=1}^{M} \left\| \nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n) \right\|_F^2 \\
&+ \frac{10\eta^2 \bar{h}_{\text{grad}}^2}{\eta_g} \left\| \nabla C^{(i)}(K_n) \right\|_F^2 + 2\eta \min\{n_x, n_u\} (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2)^2 \\
&\overset{(f)}{\leq} -\frac{\eta}{4} \left\| \nabla C^{(i)}(K_n) \right\|_F^2 + 3\eta \min\{n_x, n_u\} (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2)^2 \\
&\leq -\frac{\eta \mu^2 \sigma_{\min}(R)}{\left\| \Sigma_{K_i^*} \right\|} (C^{(i)}(K_n) - C^{(i)}(K_i^*)) + 3\eta \min\{n_x, n_u\} (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2)^2.
\end{aligned}
$$

In the above steps, (a) is due to the choice of step-size η such that

$$
\|K_{n+1} - K_n\| = \|\frac{\eta}{ML} \sum_{i=1}^{M} \sum_{l=0}^{L-1} \nabla C^{(i)}(K_{n,l}^{(i)})\| \leq \eta \bar{h}_1 \leq \underline{h}_{\Delta},
$$

holds when $\eta \leq \frac{\underline{h}_{\Delta}}{\bar{h}_1}$. For (b), we use the Lipschitz property of the gradient (Lemma 25) in the first line, and use Eq.(6.10) with $\zeta = \frac{1}{2}$ in the second line, and for the third and forth lines we use Eq. (6.8); (c) is due to Lemma 25 and $\frac{3\eta^2 \bar{h}_{\text{grad}}}{2} \leq \frac{\eta}{8}$; (d) is due to Lemma 27, Lemma 35 and the choice of step-size such that $\frac{3\eta^2 \bar{h}_{\text{grad}}}{2} \leq \frac{\eta}{8} \leq \eta$; (e) is due to Eq.(6.8); and for (f), we use the fact that $\frac{10\eta^2 \bar{h}_{\text{grad}}^2}{\eta_g} \leq \frac{\eta}{8} \leq \eta$, which holds when $\eta_l \leq \frac{1}{80L\bar{h}_{\text{grad}}^2}$. We use the gradient domination property (Lemma 26) in the last equality. □

With this lemma, we are now ready to provide the convergence guarantees for `FedLQR` in the model-based setting.

Proof of the iterative stability guarantees of `FedLQR`: Here we leverage the method of induction to prove `FedLQR`'s iterative stability guarantees. First, we start from an initial policy $K_0 \in \mathcal{G}^0$. At round n, we assume $K_n \in \mathcal{G}^0$. According to Lemma 34, we can show that all the local policies $K_{n,l}^{(i)} \in \mathcal{G}^0$. Furthermore, by choosing the step-sizes properly in Lemma 36, we have that

$$\begin{aligned} C^{(i)}(K_{n+1}) - C^{(i)}(K_n) \leq & -\frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|_{\max}}(C^{(i)}(K_n) - C^{(i)}(K_i^*)) \\ & + 3\eta \min\{n_x, n_u\}(\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2)^2. \end{aligned}$$

for any $i \in [M]$.

Then, for any fixed system $i \in [M]$, with $(\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2)^2 \leq \bar{h}_{\text{het}}^3$, we have that

$$\begin{aligned} C^{(i)}(K_{n+1}) - C^{(i)}(K_i^*) \leq & \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)(C^{(i)}(K_n) - C^{(i)}(K_i^*)) \\ & + 3\eta \min\{n_x, n_u\}(\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2)^2 \\ & \leq \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)(C^{(i)}(K_0) - C^{(i)}(K_i^*)) \\ & + \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}(C^{(i)}(K_0) - C^{(i)}(K_i^*)) \\ & \leq C^{(i)}(K_0) - C^{(i)}(K_i^*). \end{aligned}$$

With this, we can easily see that the global policy K_{n+1} at the next round $n+1$ is also stabilizing, i.e., $K_{n+1} \in \mathcal{G}^0$, by using the definition of $\mathcal{G}^0$ (Definition 2). Therefore, we can complete proving `FedLQR`'s iterative stability property by inductive reasoning.

Proof of `FedLQR`'s convergence: From Eq.(6.27), we have

$$C^{(i)}(K_{n+1}) - C^{(i)}(K_i^*) \leq \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)(C^{(i)}(K_n) - C^{(i)}(K_i^*))$$

$$+ 3\eta \min\{n_x, n_u\}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}})^2.$$

Under the assumptions in Lemma 36, `FedLQR` thus enjoys the following convergence guarantee after N rounds:

$$C^{(i)}(K_N) - C^{(i)}(K_i^*) \leq \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)^N (C^{(i)}(K_0) - C^{(i)}(K_i^*)) + \frac{3\min\{n_x, n_u\}\left\|\Sigma_{K_i^*}\right\|}{\mu^2\sigma_{\min}(R)}(\epsilon_1 h^1_{\text{het}} + \epsilon_1 h^2_{\text{het}})^2.$$

Thus, we finish the proof of Theorem 9 with $c_{\text{uni},1} = 12$ and $\mathcal{B}(\epsilon_1, \epsilon_2) := \frac{v\left\|\Sigma_{K_i^*}\right\|}{4\mu^2\sigma_{\min}(R)}(\epsilon_1 h^1_{\text{het}} + \epsilon_1 h^2_{\text{het}})^2$. □

c) Proof of Theorem 8

Proof: First, we provide the analysis for per-round cost function decrease with one local update, i.e., $L = 1$. For any fixed system $i \in [M]$, the cost decrease $C^{(i)}(K_{n+1}) - C^{(i)}(K_n)$ can be bounded as

$$C^{(i)}(K_{n+1}) - C^{(i)}(K_n) = \underbrace{(C^{(i)}(K_{n+1}) - C^{(i)}(\tilde{K}_{n+1}))}_{T_1} + \underbrace{(C^{(i)}(\tilde{K}_{n+1}) - C^{(i)}(K_n))}_{T_2}$$

where

$$\tilde{K}_{n+1} = K_n - \eta\nabla C^{(i)}(K_n),$$
$$K_{n+1} = K_n - \frac{\eta}{M}\sum_{i=1}^{M} \nabla C^{(i)}(K_n).$$

The term T_2 can be bounded as

$$T_2 \leq -\frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|_{\max}}(C^{(i)}(K_n) - C^{(i)}(K_i^*)),$$

based on the gradient domination property in Lemma 26. It is evident that $\tilde{K}_{n+1} \in \mathcal{G}^0$ holds.

By using a small step-size η such that $\left\|K_{n+1} - \tilde{K}_{n+1}\right\| \leq \underline{h}_{\Delta}$, we can bound T_1 as follows:

$$\begin{aligned} T_1 = C^{(i)}\left(K_{n+1}\right) - C^{(i)}(\tilde{K}_{n+1}) &\overset{(a)}{\leq} \bar{h}_{\text{cost}}\left\|K_{n+1} - \tilde{K}_{n+1}\right\| \\ &\leq \frac{\eta \bar{h}_{\text{cost}}}{M} \sum_{j=1}^{M} \left\|\nabla C^{(i)}(K_n) - \nabla C^{(j)}(K_n)\right\| \\ &\overset{(b)}{\leq} \eta \bar{h}_{\text{cost}}(K_n)(\epsilon_1 \bar{h}^1_{\text{het}}(K_n) + \epsilon_2 \bar{h}^2_{\text{het}}(K_n)) \end{aligned}$$

where (a) is due to the smoothness of the cost function in Lemma 25, and (b) is due to the bound on the policy gradient heterogeneity in Lemma 27.

Plugging in the upper bounds of T_1 and T_2, after some rearrangement, we have

$$\begin{aligned} C^{(i)}(K_{n+1}) - C^{(i)}(K_i^*) \leq &\left(1 - \frac{\eta \mu^2 \sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|_{\max}}\right)\left(C^{(i)}(K_n) - C^{(i)}(K_i^*)\right) \\ &+ \eta \bar{h}_{\text{cost}}(K_n)(\epsilon_1 \bar{h}^1_{\text{het}}(K_n) + \epsilon_2 \bar{h}^2_{\text{het}}(K_n)). \end{aligned}$$

By properly choosing the step-size η, we can ensure that the sequence of control gains (K_n) remains inside the sub-level set $\mathcal{G}^0$. Thus, for any $i \in [M]$, we have the sequence $\left\{C^{(i)}(K_n)\right\}_{n=0}^{\infty}$ is bounded, based on the definition of the stabilizing set $\mathcal{G}^0$. Then, we have:

$$\limsup_{n\to\infty} C^{(i)}(K_n) - C^{(i)}(K_i^*) \leq \frac{\bar{h}_{\text{cost}}\left\|\Sigma_{K_i^*}\right\|_{\max}}{\mu^2 \sigma_{\min}(R)}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}}). \tag{6.28}$$

From the gradient domination Lemma 11 in [50], we know that

$$\begin{aligned} C^{(i)}(K^*) - \limsup_{n\to\infty} C^{(i)}\left(K_n\right) &= \liminf_{n\to\infty}\left[C^{(i)}(K^*) - C^{(i)}\left(K_n\right)\right] \\ &= \liminf_{n\to\infty}\left[-\mathbb{E}\sum_t A^{(i)}_{K^*}\left(x_t^{K_n}, u_t^{K_n}\right)\right] \end{aligned} \tag{6.29}$$

where $\{x_t^{K_n}, u_t^{K_n}\}$ denotes the system's state and input induced by the control action $u_t = -K_n x_t$. Moreover, for any x, the advantage function $A_K\left(x, K'x\right)$ is defined as

$$A_K\left(x, K'x\right) := 2x^\top\left(K' - K\right)^\top E_K x + x^\top\left(K' - K\right)^\top\left(R + B^\top P_K B\right)\left(K' - K\right)x.$$

Following the analysis above in Eq. (6.29), we have that

$$
\begin{aligned}
C^{(i)}(K^*) - \limsup_{n\to\infty} C^{(i)}(K_n) &\leq \liminf_{n\to\infty} \mathbb{E}\sum_t \mathrm{Tr}\left(x_t^{K_n}\left(x_t^{K_n}\right)^\top E_{K^*}^{(i)\top}\left(R + B^{(i)\top}P_{K^*}^{(i)}B^{(i)}\right)^{-1} E_{K^*}^{(i)}\right) \\
&= \liminf_{n\to\infty} \mathrm{Tr}\left(\Sigma_{K_n}^{(i)} E_{K^*}^{(i)\top}\left(R + B^{(i)\top}P_{K^*}^{(i)}B^{(i)}\right)^{-1} E_{K^*}^{(i)}\right) \\
&\leq \liminf_{n\to\infty} \left\|\Sigma_{K_n}^{(i)}\right\| \mathrm{Tr}\left(E_{K^*}^{(i)\top}\left(R + B^{(i)\top}P_{K^*}^{(i)}B^{(i)}\right)^{-1} E_{K^*}^{(i)}\right) \\
&\overset{(a)}{\leq} \liminf_{n\to\infty} \frac{\bar{C}_{\max}}{\sigma_{\min}(Q)} \left\|\left(R + B^{(i)\top}P_{K^*}^{(i)}B^{(i)}\right)^{-1}\right\| \mathrm{Tr}\left(E_{K^*}^{(i)\top}E_{K^*}^{(i)}\right) \\
&\leq \frac{\bar{C}_{\max}}{\sigma_{\min}(R)\sigma_{\min}(Q)} \mathrm{Tr}\left(E_{K^*}^{(i)\top}E_{K^*}^{(i)}\right) \\
&= \frac{\bar{C}_{\max}}{\sigma_{\min}(R)\sigma_{\min}(Q)} \mathrm{Tr}\left((\Sigma_{K^*}^{(i)})^{-1}\nabla C^{(i)\top}(K^*)\nabla C^{(i)}(K^*)(\Sigma_{K^*}^{(i)})^{-1}\right) \\
&\overset{(b)}{\leq} \frac{\bar{C}_{\max}}{\sigma_{\min}(R)\sigma_{\min}(Q)\sigma_{\min}\left(\Sigma_{K^*}^{(i)}\right)^2} \mathrm{Tr}\left(\nabla C^{(i)\top}(K^*)\nabla C^{(i)}(K^*)\right) \\
&\leq \frac{\bar{C}_{\max}}{\mu^2\sigma_{\min}(R)\sigma_{\min}(Q)} \|\nabla C^{(i)}(K^*)\|_F^2 \\
&\overset{(c)}{\leq} \frac{\bar{C}_{\max}}{\mu^2\sigma_{\min}(R)\sigma_{\min}(Q)} \left(\underbrace{\left\|\frac{1}{M}\sum_{j=1}^M \nabla C^{(j)}(K^*)\right\|_F}_{\mathcal{T}_1=0} + \left\|\nabla C^{(i)}(K^*) - \frac{1}{M}\sum_{j=1}^M \nabla C^{(j)}(K^*)\right\|_F\right)^2 \\
&\overset{(d)}{\leq} \frac{\min\{n_x, n_u\}\bar{C}_{\max}}{\mu^2\sigma_{\min}(R)\sigma_{\min}(Q)} \left\|\nabla C^{(i)}(K^*) - \frac{1}{M}\sum_{j=1}^M \nabla C^{(j)}(K^*)\right\|^2 \\
&\leq \frac{\min\{n_x, n_u\}\bar{C}_{\max}}{\mu^2\sigma_{\min}(R)\sigma_{\min}(Q)} (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2)^2. \qquad (6.30)
\end{aligned}
$$

Here we use the uniform upper bound of $\left\|\Sigma_{K_n}^{(i)}\right\|$, i.e., $||\Sigma_{K_n}^{(i)}|| \leq \frac{C^{(i)}(K_n)}{\sigma_{\min}(Q)} \leq \frac{\bar{C}_{\max}}{\sigma_{\min}(Q)}$ in Eq.(6.11) for (a); we use $\left\|\Sigma_{K^*}^{(i)}\right\| \geq \left\|\mathbb{E}[x_0^{(i)}x_0^{(i)\top}]\right\| \geq \mu$ for (b); we bound the L_2 norm with Frobenius norm for (c); we use the policy gradient heterogeneity bound in Lemma 27 for (d). Note that $\mathcal{T}_1 = 0$ since K^* is the optimal solution to the FL problem in Eq. (6.2). Therefore, by adding Eq. (6.30)

and Eq. (6.28) together, we have that

$$C^{(i)}(K^*) - C^{(i)}(K_i^*) \leq \frac{\bar{h}_{\text{cost}} \left\| \Sigma_{K_i^*} \right\|_{\max}}{\mu^2 \sigma_{\min}(R)} (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2) + \frac{\min\{n_x, n_u\} \bar{C}_{\max}}{\mu^2 \sigma_{\min}(R) \sigma_{\min}(Q)} (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2)^2.$$

Thus, we complete the proof of Theorem 9. □

6.9.8 Zeroth-order optimization

To prepare for the model-free setting where the controllers only have access to the system's trajectories, we first quickly recap the basic idea behind zeroth-order optimization. Say our goal is to minimize a loss function $f(x)$, where $x \in \mathbb{R}^d$. When one has access to exact deterministic gradients of this loss function via an oracle, the standard approach for minimization would be to query the gradient oracle at each iteration, and run gradient descent. Concretely, one would run the following iterative scheme: $x_{t+1} = x_t - \eta \nabla f(x_t)$, where η is a suitably chosen learning-rate/step-size. While such first-order optimization schemes have a rich history, there has also been a growing interest in understanding the behavior of derivative-free (zeroth-order) methods that can *only query function values*, as opposed to the gradients. Two immediate reasons (among many) for studying zeroth-order optimization are as follows: (i) in practice, one may only have access to a black-box procedure that cannot evaluate gradients; and (ii) computing gradients might prove to be too computationally-expensive.

Given two or more function evaluations, the basic idea behind zeroth-order algorithms is to construct an estimate of the true gradient for evaluating and updating model parameters. For instance, a typical zeroth-order scheme with single-point function evaluation would take the following form [156]:

$$x_{t+1} = x_t - \eta_t \left(\frac{f(x_t + \mu_t u) - f(x_t)}{\mu_t} \right) u.$$

In the expression above, $\{\eta_t\}$ is the learning-rate sequence, $\{\mu_t\}$ is a sequence typically chosen in a way such that $\mu_t \to 0$, and u is a random vector distributed uniformly over the unit sphere. For details about the convergence of zeroth-order optimization algorithms such as the one above, we refer the interested reader to [8, 41, 144].

We now turn to briefly describing the model-free setup for our LQR problem. [50] propose a zeroth-order-based algorithm (Algorithm 1 in [50]) to compute an estimation $\widehat{\nabla C(K)}$ and $\widehat{\Sigma_K}$ for both $\nabla C(K)$ and Σ_K, for a given K. Algorithm 1 in [50] exploits a multiple-trajectory-based technique that uses a Gaussian perturbed cost function (i.e., producing a Gaussian smoothing function) to estimate $\nabla C(K)$ from cost function perturbed values. That is, given the cost function $C(K)$, we can define its perturbed function as,

$$C_r(K) = \mathbb{E}_{U \sim \mathcal{B}_r}[C(K+U)]$$

where $\mathcal{B}_r$ is the uniform distribution over all matrices with Frobenius norm at most r and U is a random matrix with proper dimension and generated from $\mathcal{B}_r$. For small r, the smooth cost $C_r(K)$ is a good approximation to the original cost $C(K)$. Due to the Gaussian smoothing, the gradient has a particularly simple functional form [65]:

$$\nabla C_r(K) = \frac{n_x n_u}{r^2} \mathbb{E}_{U \sim \mathbb{B}_r}[C(K+U)U].$$

Therefore, this expression implies a straightforward method to obtain an unbiased estimate of $\nabla C_r(K)$, through obtaining the infinite-horizon rollouts. However, in practice, we can only obtain the finite-horizon rollouts to approximate the gradient. Thanks to [50], they showed that the approximation error of the exact gradient can be reduced to arbitrary accuracy if the number of sample trajectories n_s and the length of each rollout τ are sufficiently large, and the smoothing radius r is small enough.

6.9.9 The model-free setting

For notational brevity we rewrite $\widehat{\nabla C^{(i)}(K)}$ as $\tilde{\nabla} C^{(i)}(K)$ where

$$\widehat{\nabla C^{(i)}(K)} = \tilde{\nabla} C^{(i)}(K) := \frac{1}{n_s} \sum_{s=1}^{n_s} \frac{n_x n_u}{r^2} \tilde{C}^{(i),(\tau)} \left(K + U_s^{(i)}\right) U_s^{(i)},$$

and introduce two new gradient-based terms:

$$\nabla' C^{(i)}(K) := \frac{1}{n_s} \sum_{s=1}^{n_s} \frac{n_x n_u}{r^2} C^{(i),(\tau)} \left(K + U_s^{(i)}\right) U_s^{(i)},$$
$$\hat{\nabla} C^{(i)}(K) := \frac{1}{n_s} \sum_{s=1}^{n_s} \frac{n_x n_u}{r^2} C^{(i)} \left(K + U_s^{(i)}\right) U_s^{(i)},$$

where $\tilde{C}^{(i),(\tau)} \left(K + U_s^{(i)}\right) := \sum_{t=0}^{\tau-1} \left(x_t^{(i)\top} Q x_t^{(i)} + u_t^{(i)\top} R u_t^{(i)}\right)$ with $x_t^{(i)} = (K + U_s^{(i)}) u_t^{(i)}$, $C^{(i),(\tau)} \left(K + U_s^{(i)}\right) := \mathcal{E}_{x_0^{(i)} \sim \mathcal{D}} \sum_{t=0}^{\tau-1} \left(x_t^{(i)\top} Q x_t^{(i)} + u_t^{(i)\top} R u_t^{(i)}\right)$ and

$$C^{(i)} \left(K + U_s^{(i)}\right) := \mathcal{E}_{x_0^{(i)} \sim \mathcal{D}} \sum_{t=0}^{\infty} \left(x_t^{(i)\top} Q x_t^{(i)} + u_t^{(i)\top} R u_t^{(i)}\right).$$

a) Auxiliary Lemmas

Lemma 37. *(Approximating $C^{(i)}(K)$ and $\Sigma_K^{(i)}$ with finite horizon) Suppose K is such that $C^{(i)}(K)$ is finite. Define the finite horizon estimates,*

$$\Sigma_K^{(i),(\tau)} := \mathbb{E}\left[\sum_{t=0}^{\tau-1} x_t^{(i)} x_t^{(i)\top}\right] \quad \textit{and} \quad C^{(i),(\tau)}(K) := \mathbb{E}\left[\sum_{t=0}^{\tau-1} x_t^{(i)\top} Q x_t^{(i)} + u_t^{(i)\top} R u_t^{(i)}\right],$$

for all systems $i \in [M]$. Now, let ϵ be an arbitrarily small constant such that

$$\tau \geq h_\tau^1(\epsilon) := \max_{i \in [M]} \left\{ \frac{n_x \cdot (C^{(i)}(K))^2}{\epsilon \mu(\sigma_{min}(Q))^2} \right\} = \frac{n_x \cdot (C_{\max}(K))^2}{\epsilon \mu(\sigma_{min}(Q))^2},$$

such that

$$\left\| \Sigma_K^{(i),(\tau)} - \Sigma_K^{(i)} \right\| \leq \epsilon.$$

If

$$\tau \geq h_\tau^2(\epsilon) := \max_{i\in[M]} \left\{ \frac{n_x \cdot (C^{(i)}(K))^2(\|Q\|+\|R\|\|K\|^2)}{\epsilon\mu(\sigma_{min}(Q))^2} \right\} = \frac{n_x \cdot (C_{\max}(K))^2(\|Q\|+\|R\|\|K\|^2)}{\epsilon\mu(\sigma_{min}(Q))^2},$$

we have

$$\left|C^{(i),(\tau)}(K) - C^{(i)}(K)\right| \leq \epsilon,$$

where $C_{\max}(K) := \max_{i\in[M]} C^{(i)}(K)$.

Proof: The proof for this lemma is detailed in the proof of Lemma 23 in [50]. □

Lemma 38. *(Estimating $\nabla C^{(i)}(K)$ with finitely many infinite-horizon rollouts) Given an arbitrary tolerance ϵ and probability δ, suppose the radius r satisfies*

$$r \leq h_r\left(\frac{\epsilon}{2}\right) := \min\left\{ \underline{h}_\Delta, \frac{\bar{C}_{\max}}{\bar{h}_{cost}}, \frac{\epsilon}{2\bar{h}_{grad}} \right\},$$

and the number of samples n_s satisfies,

$$n_s \geq h_{sample}\left(\frac{\epsilon}{2}, \delta\right) := \frac{8\sigma_{\hat{\nabla}}^2 \min(n_x, n_u)}{\epsilon^2} \log\left[\frac{n_x + n_u}{\delta}\right]$$
$$\sigma_{\hat{\nabla}}^2 := \left(\frac{2n_x n_u \bar{C}_{\max}}{r}\right)^2 + \left(\frac{\epsilon}{2} + \bar{h}_1\right)^2$$

Then with a high probability of at least $1-\delta$, the estimate

$$\hat{\nabla} C^{(i)}(K) = \frac{1}{n_s} \sum_{s=1}^{n_s} \frac{n_x n_u}{r^2} C^{(i)}\left(K + U_s^{(i)}\right) U_s^{(i)}$$

satisfies

$$\|\hat{\nabla} C^{(i)}(K) - \nabla C^{(i)}(K)\|_F \leq \epsilon$$

for any system $i \in [M]$ and $K \in \mathcal{G}^0$.

Proof: The proof for this lemma is detailed in Lemma B.6 of [65]. It is worthwhile to mention that, in [65], the number of samples n_s satisfies

$$n_s \geq \left[\underbrace{\frac{8\sigma_{\hat{\nabla}}^2 \min(n_x, n_u)}{\epsilon^2}}_{T_1} + \underbrace{\frac{8\min(n_x, n_u)}{\epsilon^2} \frac{R_{\hat{\nabla}}\epsilon}{6\sqrt{\min(n_x, n_u)}}}_{T_2} \right] \log\left[\frac{n_x + n_u}{\delta}\right]$$

with $R_{\hat{\nabla}} = \frac{2n_x n_u \bar{C}_{\max}}{r} + \frac{\epsilon}{2} + \bar{h}_1$. In the analysis throughout the paper, we only keep the dominant term T_1 in n_s, since T_1 is in the order $\mathcal{O}(\epsilon^{-2})$ while T_2 is in the order $\mathcal{O}(\epsilon^{-1})$.

By taking the maximum over K inside $\mathcal{G}^0$, we make the local parameters become the global parameters, e.g., $\bar{C}_{\max} := \sup_{K\in\mathcal{G}^0, i\in[M]} C^{(i)}(K)$. □

Lemma 39. *(Estimating $\nabla C^{(i)}(K)$ with finitely many finite-horizon rollouts): Given an arbitrary tolerance ϵ and probability δ, suppose that the smoothing radius r satisfies,*

$$r \leq h_r\left(\frac{\epsilon}{4}\right) = \min\left\{\bar{h}_\Delta, \frac{\bar{C}_{\max}}{\bar{h}_{cost}}, \frac{\epsilon}{4\bar{h}_{grad}}\right\},$$

and the trajectory length τ satisfies

$$\tau \geq h_\tau\left(\frac{r\epsilon}{4n_x n_u}\right) = \frac{4n_u n_x^2 (C_{\max}(K))^2 \left(\|Q\| + \|R\|\|K\|^2\right)}{r\epsilon\mu\sigma_{\min}(Q)^2}.$$

According to Assumption 7, the distribution of the initial states satisfies $x_0^{(i)} \sim \mathcal{D}$ and $\left\|x_0^{(i)}\right\| \leq H$ almost surely. Thus, for any given realization $x_{0,s}^{(i)}$[16] of $x_0^{(i)}$, and for any system $i \in [M]$, we have

$$\left\|x_{0,s}^{(i)}\right\| \leq H, \quad \left(x_{0,s}^{(i)}\right)\left(x_{0,s}^{(i)}\right)^\top \preceq \frac{H^2}{\mu}\mathbb{E}\left[x_0^{(i)} x_0^{(i)\top}\right].$$

As a result, the summation over the finite-time horizon

$$\sum_{t=0}^{\tau-1}\left(x_{t,j}^{(i)\top} Q x_{t,j}^{(i)} + u_{t,j}^{(i)\top} R u_{t,j}^{(i)}\right) \leq \frac{H^2}{\mu} C^{(i)}\left(K + U_j^{(i)}\right).$$

[16]The notation $x_{0,s}^{(i)}$ denotes s-th sample of the initial state from i-th system.

Furthermore, suppose the number of samples n_s satisfies

$$n_s \geq h_{\textit{sample,trunc}}\left(\frac{\epsilon}{4}, \delta, \frac{H^2}{\mu}\right) := \frac{32\sigma_{\tilde{\nabla}}^2 \min(n_x, n_u)}{\epsilon^2} \log\left[\frac{n_x + n_u}{\delta}\right],$$

where

$$\sigma_{\tilde{\nabla}}^2 := \left(\frac{2n_x n_u H^2 \bar{C}_{\max}}{r\mu}\right)^2 + \left(\frac{\epsilon}{2} + \bar{h}_1\right)^2,$$

then, with a high probability of at least $1-\delta$, the estimated gradient

$$\tilde{\nabla} C^{(i)}(K) := \frac{1}{n_s}\sum_{s=1}^{n_s} \frac{n_u n_x}{r^2} \tilde{C}^{(i),(\tau)}\left(K + U_s^{(i)}\right) U_s^{(i)}$$

satisfies

$$\|\tilde{\nabla} C^{(i)}(K) - \nabla C^{(i)}(K)\|_F \leq \epsilon$$

for any system $i \in [M]$ and $K \in \mathcal{G}^0$.

Proof: The proof for this lemma is detailed in Lemma B.7 of [65]. As in Lemma 38, we only keep the dominant term in the requirement of sample size n_s. By taking the maximum over K inside $\mathcal{G}^0$, all the local parameters inside the polynomials such as $h_r(\frac{\epsilon}{4})$ become global parameters. □

b) Proof of Lemma 28

Proof: For our subsequent analysis, we will use $\mathcal{F}_l^n$ to denote the filtration that captures all the randomness up to the l-th local step in round n. We have

$$\begin{aligned}\left\|\frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L-1}\left[\widehat{\nabla C^{(i)}(K_{n,l}^{(i)})} - \nabla C^{(i)}(K_{n,l}^{(i)})\right]\right\|_F &= \left\|\frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L-1}\left[\tilde{\nabla} C^{(i)}(K_{n,l}^{(i)}) - \nabla C^{(i)}(K_{n,l}^{(i)})\right]\right\|_F \\ &\leq \underbrace{\left\| \frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L}\left[\tilde{\nabla} C^{(i)}(K_{n,l}^{(i)}) - \nabla' C^{(i)}(K_{n,l}^{(i)})\right] \right\|_F}_{T_1}\end{aligned}$$

$$+\underbrace{\left\|\frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L-1}\left[\nabla' C^{(i)}(K_{n,l}^{(i)})-\hat{\nabla}C^{(i)}(K_{n,l}^{(i)})\right]\right\|_F}_{T_2}$$

$$+\underbrace{\left\|\frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L-1}\left[\hat{\nabla}C^{(i)}(K_{n,l}^{(i)})-\nabla C^{(i)}(K_{n,l}^{(i)})\right]\right\|_F}_{T_3}.$$

Next, we will bound T_1, T_2, and T_3, respectively.

Bounding T_2: From the proof of Lemma B.7 in [65], we have

$$T_2 \leq \frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L-1}\left\|\nabla' C^{(i)}(K_{n,l}^{(i)})-\hat{\nabla}C^{(i)}(K_{n,l}^{(i)})\right\|_F \leq \frac{\epsilon}{4} \tag{6.31}$$

holds as long as $\tau \geq h_\tau\left(\frac{r\epsilon}{4n_x n_u}\right)$.

Bounding T_3 : To precede, we bound T_3 as

$$\begin{aligned} T_3 &= \left\|\frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L-1}\left[\hat{\nabla}C^{(i)}(K_{n,l}^{(i)})-\nabla C^{(i)}(K_{n,l}^{(i)})\right]\right\|_F \\ &\leq \underbrace{\left\|\frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L-1}\left[\hat{\nabla}C^{(i)}(K_{n,l}^{(i)})-\nabla C_r^{(i)}(K_{n,l}^{(i)})\right]\right\|_F}_{\text{Variance term } \textcircled{2}} + \frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L-1}\underbrace{\left\|\nabla C_r^{(i)}(K_{n,l}^{(i)})-\nabla C^{(i)}(K_{n,l}^{(i)})\right\|_F}_{\text{Bias term } \textcircled{1}} \end{aligned} \tag{6.32}$$

where $\nabla C_r^{(i)}(K_{n,l}^{(i)}) := \mathcal{E}_{U_{n,l}^{(i)}\sim\mathbb{B}_r}\left[\nabla C^{(i)}(K_{n,l}^{(i)}+U_{n,l}^{(i)})\right]$.

For the bias term ①, since the smoothing radius $r \leq h_r\left(\frac{\epsilon}{4}\right)$, we have that

$$\textcircled{1} = \left\|\nabla C_r^{(i)}(K_{n,l}^{(i)})-\nabla C^{(i)}(K_{n,l}^{(i)})\right\|_F \leq h_{\text{grad}}(K_{n,l}^{(i)})r \leq \bar{h}_{\text{grad}}r \leq \frac{\epsilon}{4}. \tag{6.33}$$

For the variance term, ②, we will exploit the matrix Freedman inequality (Lemma 31) to bound it.

For simplicity, we denote

$$e_l^{(i)} := \frac{1}{ML}\left[\hat{\nabla}C^{(i)}(K_{n,l}^{(i)})-\nabla C_r^{(i)}(K_{n,l}^{(i)})\right], \quad e_l := \sum_{i=1}^{M} e_l^{(i)},$$

Then, we have

$$\frac{1}{ML}\sum_{i=1}^{M}\sum_{l=0}^{L-1}\left[\hat{\nabla}C^{(i)}(K_{n,l}^{(i)})-\nabla C_r^{(i)}(K_{n,l}^{(i)})\right]=\sum_{l=0}^{L-1}e_l.$$

Next, we aim to prove the following claims:

Claim I: $Y_t:=\sum_{l=0}^{t}e_l$ is a martingale w.r.t $\mathcal{F}_{t-1}^n$ for $t=1,\cdots,L-1$ and $e_l:=\sum_{i=1}^{M}e_l^{(i)}$ is a martingale difference sequence.

Proof: Note that $\mathcal{E}[\hat{\nabla}C^{(i)}(K_{n,l}^{(i)})]=\nabla C_r^{(i)}(K_{n,l}^{(i)})$. Then we can easily have $\mathcal{E}[e_l]=0$ for $l=0,\cdots,L-1$. As a result, we have $\mathcal{E}[Y_t\mid\mathcal{F}_{t-1}^n]=Y_{t-1}$ since $Y_t=Y_{t-1}+e_t$. In other words, $Y_t:=\sum_{l=0}^{t}e_l$ is a martingale w.r.t $\mathcal{F}_{t-1}^n$ for $t=1,\cdots,L-1$.

Claim II: $\left\|\mathcal{E}\left[e_le_l^\top\;\middle|\;\mathcal{F}_{l-1}^n\right]\right\|\leq\frac{\sigma_{\hat{\nabla}}^2}{n_sML^2}$ where $\sigma_{\hat{\nabla}}^2$ is as defined in Lemma 38.

Proof: From Lemma B.7 in [65], we can write

$$\left\|\mathcal{E}\left[e_l^{(i)}e_l^{(i)\top}\;\middle|\;\mathcal{F}_{l-1}^n\right]\right\|\leq\frac{\sigma_{\hat{\nabla}}^2}{n_sM^2L^2},\left\|\mathcal{E}\left[e_l^{(i)\top}e_l^{(i)}\;\middle|\;\mathcal{F}_{l-1}^n\right]\right\|\leq\frac{\sigma_{\hat{\nabla}}^2}{n_sM^2L^2},$$

and based on this fact, we have

$$\begin{aligned}\left\|\mathcal{E}\left[e_le_l^\top\;\middle|\;\mathcal{F}_{l-1}^n\right]\right\|&=\left\|\mathcal{E}\left[\left(\sum_{i=1}^{M}e_l^{(i)}\right)\left(\sum_{i=1}^{M}e_l^{(i)\top}\right)\;\middle|\;\mathcal{F}_{l-1}^n\right]\right\|\\&\leq\sum_{i=1}^{M}\left\|\mathcal{E}\left[e_l^{(i)}e_l^{(i)\top}\;\middle|\;\mathcal{F}_{l-1}^n\right]\right\|+\underbrace{\sum_{i\neq j}^{M}\left\|\mathcal{E}\left[e_l^{(i)}e_l^{(j)\top}\;\middle|\;\mathcal{F}_{l-1}^n\right]\right\|}_{T_4=0}\leq\frac{\sigma_{\hat{\nabla}}^2}{n_sML^2},\end{aligned}$$

where we use the fact that $T_4=0$ because $e_l^{(i)}$ and $e_l^{(j)}$ are independent, if we conditioned on $\mathcal{F}_l^n$. An identical argument holds for $\left\|\mathcal{E}\left[e_l^\top e_l\;\middle|\;\mathcal{F}_{l-1}^n\right]\right\|$.

Define $W_{\text{col},t}:=\sum_{l=0}^{t}\mathcal{E}\left[e_le_l^\top\;\middle|\;\mathcal{F}_{l-1}^n\right]$ and $W_{\text{row},t}:=\sum_{l=0}^{t}\mathcal{E}\left[e_l^\top e_l\;\middle|\;\mathcal{F}_{l-1}^n\right]$, then we have

$$\|W_{\text{col},t}\|\leq\frac{\sigma_{\hat{\nabla}}^2}{n_sML},\quad\|W_{\text{row},t}\|\leq\frac{\sigma_{\hat{\nabla}}^2}{n_sML}.$$

Claim III: $\|e_l\| \leq \frac{R_{\hat{\nabla}}}{n_s L}$ where $R_{\hat{\nabla}} = \frac{2n_x n_u \bar{C}_{\max}}{r} + \frac{\epsilon}{2} + \bar{h}_1$.

Proof: From Lemma B.7 in [65], we have $\|e_l^{(i)}\| \leq \frac{n_s R_{\hat{\nabla}}}{ML}$. With this fact, we have

$$\|e_l\| \leq \sum_{i=1}^{M} \left\| e_l^{(i)} \right\| \leq \frac{R_{\hat{\nabla}}}{n_s L}.$$

With **Claim I, II** and **Claim III** and the matrix Freedman inequality (31), we have, for all $\epsilon \geq 0$,

$$\mathbb{P}\left\{ \exists t \geq 0 : \lambda_{\max}(Y_t) \geq \epsilon \text{ and } \max\left\{ \|W_{\text{col},t}\|, \|W_{\text{row},t}\| \right\} \leq \frac{\sigma_{\hat{\nabla}}^2}{n_s ML} \right\} \leq (n_x + n_u) \exp\left\{ -\frac{-\epsilon^2/2}{\frac{\sigma_{\hat{\nabla}}^2}{n_s ML} + \frac{R_{\hat{\nabla}}\epsilon}{3n_s L}} \right\}. \tag{6.34}$$

Therefore, rephrasing Eq.(6.34), if

$$n_s \geq \left(\underbrace{\frac{32\sigma_{\hat{\nabla}}^2 \min(n_x, n_u)}{ML\epsilon^2}}_{T_5} + \underbrace{\frac{32LR_{\hat{\nabla}}\sqrt{\min(n_x, n_u)}}{12ML\epsilon}}_{T_6} \right) \log\left[\frac{ML(n_x + n_u)}{\delta} \right], \tag{6.35}$$

we have that

$$\|Y_L\|_F = \| \frac{1}{ML} \sum_{i=1}^{M} \sum_{l=0}^{L-1} \left[\hat{\nabla} C^{(i)}(K_{n,l}^{(i)}) - \nabla C_r^{(i)}(K_{n,l}^{(i)}) \right] \|_F \leq \frac{\epsilon}{4}, \tag{6.36}$$

holds with probability $1 - \delta$. As we discussed in Lemma 38, we only keep the dominant term T_5 in the requirement of the sample size n_s (as in Eq.(6.35)). Because T_5 is in the order $\mathcal{O}(\epsilon^{-2})$ while T_6 is in the order $\mathcal{O}(\epsilon^{-1})$. Then, T_6 when compared to T_5 is negligible.

In summary, if $n_s \geq \frac{32\sigma_{\hat{\nabla}}^2 \min(n_x, n_u)}{ML\epsilon^2} \left[\frac{ML(n_x+n_u)}{\delta} \right] = \frac{h_{\text{sample}}\left(\frac{\epsilon}{4}, \frac{\delta}{ML}\right)}{ML}$,

$$② = \left\| \frac{1}{ML} \sum_{i=1}^{M} \sum_{l=0}^{L-1} \left[\hat{\nabla} C^{(i)}(K_{n,l}^{(i)}) - \nabla C_r^{(i)}(K_{n,l}^{(i)}) \right] \right\|_F \leq \frac{\epsilon}{4} \tag{6.37}$$

holds with probability $1 - \delta$.

As a result, we have $T_3 \leq \frac{\epsilon}{2}$ holds with probability $1 - \delta$, when $r \leq h_r\left(\frac{\epsilon}{4}\right)$ and $n_s \geq \frac{h_{\text{sample}}\left(\frac{\epsilon}{4}, \frac{\delta}{ML}\right)}{ML}$. In what follows, we will provide an upper bound on the term T_1.

Bounding T_1: We can follow the same analysis of bounding ② in T_3 to bound T_1. Different from the filtration we define in analyzing ②, we need to define a new filtration $\tilde{\mathcal{F}}_{l-1}^n$, where $\tilde{\mathcal{F}}_{l-1}^n := \mathcal{F}_{l-1}^n \cup U_l^n$ and $U_l^n := \left\{ U_{n,l,s}^{(i)} \right\}_{s=1,\cdots,n_s}^{i=1,\cdots,N}$. Note that U_l^n is the sigma-field generated by the randomness of all random smoothing matrices $U_{n,l,s}^{(i)}$ [17] from all the systems at the n-th global iteration and l-th local iteration. Replacing

[17] Here we use the index s to denote s-th sample. Note that in each local iteration l, we need to generate the random smoothing matrices n_s times.

$\sigma^2_{\hat{\nabla}}$ Eq.(6.35) with into $\sigma^2_{\tilde{\nabla}}$ and $R_{\hat{\nabla}}$ with $R_{\tilde{\nabla}}$, we have that

$$T_1 = \left\| \frac{1}{ML} \sum_{i=1}^{M} \sum_{l=0}^{L} \left[\tilde{\nabla} C^{(i)}(K_{n,l}^{(i)}) - \nabla' C^{(i)}(K_{n,l}^{(i)}) \right] \right\|_F \leq \frac{\epsilon}{4} \tag{6.38}$$

holds with probability $1 - \delta$ when

$$n_s \geq \frac{32\sigma^2_{\tilde{\nabla}} \min(n_x, n_u)}{ML\epsilon^2} \log \left[\frac{ML(n_x + n_u)}{\delta} \right] = \frac{h_{\text{sample,trunc}} \left(\frac{\epsilon}{4}, \frac{\delta}{ML}, \frac{H^2}{\mu} \right)}{ML}.$$

Combing the upper bound of T_1 (Eq.(6.38)), T_2 (Eq.(6.31)) and T_3 (Eq.(6.33) and (6.37)), we have

$$\left\| \frac{1}{ML} \sum_{i=1}^{M} \sum_{l=0}^{L-1} \left[\widehat{\nabla C^{(i)}(K_{n,l}^{(i)})} - \nabla C^{(i)}(K_{n,l}^{(i)}) \right] \right\|_F \leq T_1 + T_2 + T_3 \leq \epsilon$$

when the trajectory length τ satisfies $\tau \geq h_\tau \left(\frac{r\epsilon}{4n_x n_u} \right)$, the smoothing radius satisfies $r \leq h_r \left(\frac{\epsilon}{4} \right)$ and the size of samples satisfies $n_s \geq \max \left\{ \frac{h_{\text{sample,trunc}} \left(\frac{\epsilon}{4}, \frac{\delta}{ML}, \frac{H^2}{\mu} \right)}{ML}, \frac{h_{\text{sample}} \left(\frac{\epsilon}{4}, \frac{\delta}{ML} \right)}{ML} \right\} = \frac{h_{\text{sample,trunc}} \left(\frac{\epsilon}{4}, \frac{\delta}{ML}, \frac{H^2}{\mu} \right)}{ML}$. Thus, we complete the proof of Lemma 28.

c) Proof of Theorem 10

Outline: To prove Theorem 10, we first introduce some lemmas: Lemma 40 establishes stability of the local policies; Lemma 41 provides the drift analysis; Lemma 42 quantifies the per-round progress of our `FedLQR` algorithm. As a result, we are able to present the iterative stability guarantees and convergence analysis of `FedLQR` in the model-free setting.

Lemma 40. *(Stability of the local policies) Suppose $K_n \in \mathcal{G}^0$ and the heterogeneity level satisfies $(\epsilon_1 \bar{h}^1_{het} + \epsilon_2 \bar{h}^2_{het})^2 \leq \bar{h}^3_{het}$, where $\bar{h}^3_{het}$ is as defined in Eq.*(6.21)*. If the local step-size η_l satisfies*

$$\eta_l \leq \min\left\{\frac{\underline{h}_\Delta \mu}{H^2\left(h_1 + \sqrt{\bar{\epsilon}}\right)}, \frac{1}{9\bar{h}_{grad}}\right\},$$

the smoothing radius satisfies

$$r \leq \min\left\{\frac{\min_{i\in[M]} C^{(i)}(K_0)}{\bar{h}_{cost}}, \underline{h}_\Delta, h_r\left(\frac{\sqrt{\bar{\epsilon}}}{4}\right)\right\},$$

the trajectory length satisfies $\tau \geq h_\tau\left(\frac{r\sqrt{\bar{\epsilon}}}{4n_x n_u}\right)$, and the number of the sample size satisfies

$$n_s \geq \max\left\{h_{sample,trunc}\left(\frac{\sqrt{\bar{\epsilon}}}{4}, \frac{\delta}{L}, \frac{H^2}{\mu}\right), h_{sample}\left(\frac{\sqrt{\bar{\epsilon}}}{2}, \frac{\delta}{L}\right)\right\}$$

where we choose a fixed error tolerance $\bar{\epsilon}$ to be

$$\bar{\epsilon} := \min_{j\in[M]}\left\{\frac{3\mu^2\sigma_{\min}(R)\left(C^{(j)}(K_0) - C^{(j)}(K_j^*)\right)}{5||\Sigma_{K_j^*}||}\right\},$$

then with probability $1-\delta$,, where $\delta \in (0,1)$, $K^{(i)}_{n,l} \in \mathcal{G}^0$ holds for all $i \in [M]$ and $l = 0, 1, \cdots, L-1$.

Proof: For any $i, j \in [M]$, according to the local Lipschitz property in Lemma 25, we have that

$$\begin{aligned} C^{(j)}(K^{(i)}_{n,1}) - C^{(j)}(K_n) &\leq \left\langle \nabla C^{(j)}(K_n), K^{(i)}_{n,1} - K_n\right\rangle + \frac{h_{\text{grad}(K_n)}}{2}\left\|K^{(i)}_{n,1} - K_n\right\|_F^2 \quad \text{(Local lipschitz)} \\ &= -\left\langle \nabla C^{(j)}(K_n), \eta_l \tilde{\nabla} C^{(i)}(K_n)\right\rangle + \frac{h_{\text{grad}(K_n)}}{2}\left\|\eta_l \tilde{\nabla} C^{(i)}(K_n)\right\|_F^2, \end{aligned}$$

holds if $\left\|\eta_l \tilde{\nabla} C^{(i)}(K_n)\right\|_F \leq \underline{h}_\Delta \leq h_\Delta(K_n)$. Note that this inequality holds when η_l satisfies

$$\left\|\eta_l \tilde{\nabla} C^{(i)}(K_n)\right\|_F = \eta_l \left\|\frac{1}{n_s}\sum_{s=1}^{n_s} \frac{n_u n_x}{r^2} \tilde{C}^{(i),(\tau)}\left(K_n + U_s^{(i)}\right) U_s^{(i)}\right\|_F$$

$$\begin{aligned}
&\overset{(a)}{\leq} \eta_l \frac{H^2}{\mu} \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{n_x n_u}{r^2} C^{(i)} \left(K_n + U_s^{(i)} \right) U_s^{(i)} \right\|_F \\
&= \frac{\eta_l H^2}{\mu} \left\| \hat{\nabla} C^{(i)}(K) \right\|_F \\
&\leq \frac{\eta_l H^2}{\mu} \left[\left\| \nabla C^{(i)}(K) \right\|_F + \left\| \hat{\nabla} C^{(i)}(K) - \nabla C^{(i)}(K) \right\|_F \right] \\
&\overset{(b)}{\leq} \frac{\eta_l H^2}{\mu} \left[\left\| \nabla C^{(i)}\left(K_n\right) \right\|_F + \sqrt{\bar{\epsilon}} \right] \\
&\leq \frac{\eta_l H^2}{\mu} \left(h_1 + \sqrt{\bar{\epsilon}} \right)
\end{aligned} \tag{6.39}$$

where[18] (a) is due to Lemma 39; according to Lemma 38, (b) holds with high probability, when the number of the sample size satisfies $n_s \geq h_{\text{sample}} \left(\frac{\sqrt{\bar{\epsilon}}}{2}, \frac{\delta}{L} \right)$. The last inequality follows from the uniform upper gradient bound in Lemma 30. Then we can easily conclude that $\left\| \eta_l \tilde{\nabla} C^{(i)}(K_n) \right\|_F \leq \underline{h}_\Delta$ holds when $\eta_l \leq \frac{\underline{h}_\Delta \mu}{H^2 \left(h_1 + \sqrt{\bar{\epsilon}} \right)}$.

Following the analysis in Eq (6.22), we have

$$\begin{aligned}
C^{(j)}(K_{n,1}^{(i)}) - C^{(j)}(K_n) \leq & -\eta_l \left\langle \nabla C^{(j)}(K_n), \nabla C^{(j)}(K_n) \right\rangle \\
& - \eta_l \underbrace{\left\langle \nabla C^{(j)}(K_n), \nabla C^{(i)}(K_n) - \nabla C^{(j)}(K_n) \right\rangle}_{T_1} \\
& \underbrace{-\eta_l \left\langle \nabla C^{(j)}(K_n), \tilde{\nabla} C^{(i)}(K_n) - \nabla C^{(i)}(K_n) \right\rangle}_{T_2} + \frac{h_{\text{grad}(K_n)}}{2} \left\| \eta_l \tilde{\nabla} C^{(i)}(K_n) \right\|_F^2,
\end{aligned}$$

where T_1 can be upper bounded as

$$\begin{aligned}
T_1 &\leq \eta_l \left\| \nabla C^{(j)}(K_n) \right\|_F \left\| \nabla C^{(i)}(K_n) - \nabla C^{(j)}(K_n) \right\|_F \\
&\leq \eta_l \sqrt{\min\{n_x, n_u\}} \left\| \nabla C^{(j)}(K_n) \right\|_F (\epsilon_1 \bar{h}_{\text{het}}^1 + \epsilon_2 \bar{h}_{\text{het}}^2),
\end{aligned}$$

where we use the policy gradient heterogeneity bound in Lemma 27 and the fact that $K_n \in \mathcal{G}^0$.

We can bound T_2 as follows

$$\begin{aligned}
T_2 &\leq \eta_l \left\| \nabla C^{(j)}(K_n) \right\|_F \left\| \tilde{\nabla} C^{(i)}(K_n) - \nabla C^{(i)}(K_n) \right\|_F \\
&\leq \eta_l \left\| \nabla C^{(j)}(K_n) \right\|_F \sqrt{\bar{\epsilon}},
\end{aligned}$$

[18] For sake of the notation, we ignore the dependence on the local iteration l and global iteration n when we index $U_s^{(i)}$ in this part.

where it holds with probability $1-\delta$. Here we use the Cauchy-Schwarz inequality in the first inequality, and the second inequality is due to Lemma 39 since $n_s \geq h_{\text{sample,trunc}}\left(\frac{\sqrt{\epsilon}}{4}, \delta, \frac{H^2}{\mu}\right)$, the smoothing radius satisfies $r \leq h_r\left(\frac{\sqrt{\epsilon}}{4}\right)$ and the length of trajectories satisfies $\tau \geq h_\tau\left(\frac{r\sqrt{\epsilon}}{4n_x n_u}\right)$.

Plugging the upper bounds of T_1 and T_2 in Eq (6.23), we have:

$$
\begin{aligned}
C^{(j)}(K_{n,1}^{(i)}) - C^{(j)}(K_n) &\overset{(a)}{\leq} -\eta_l \left\|\nabla C^{(j)}(K_n)\right\|_F^2 + \eta_l\sqrt{\min\{n_x,n_u\}}\left\|\nabla C^{(j)}(K_n)\right\|_F(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_2\bar{h}_{\text{het}}^2) \\
&+ \eta_l\left\|\nabla C^{(j)}(K_n)\right\|_F\sqrt{\epsilon} + \frac{3h_{\text{grad}(K_n)}\eta_l^2}{2}\left\|\tilde{\nabla} C^{(i)}(K_n) - \nabla C^{(i)}(K_n)\right\|_F^2 \\
&+ \frac{3h_{\text{grad}(K_n)}\eta_l^2}{2}\left\|\nabla C^{(i)}(K_n) - \nabla C^{(j)}(K_n)\right\|_F^2 \\
&+ \frac{3h_{\text{grad}(K_n)}\eta_l^2}{2}\left\|\nabla C^{(j)}(K_n)\right\|_F^2 \\
&\overset{(b)}{\leq} -\eta_l \left\|\nabla C^{(j)}(K_n)\right\|_F^2 + \eta_l\sqrt{\min\{n_x,n_u\}}\left\|\nabla C^{(j)}(K_n)\right\|_F(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_2\bar{h}_{\text{het}}^2) \\
&+ \eta_l\left\|\nabla C^{(j)}(K_n)\right\|_F\sqrt{\epsilon} + \frac{3\bar{h}_{\text{grad}}\eta_l^2}{2}\bar{\epsilon} \\
&+ \frac{3\bar{h}_{\text{grad}}\eta_l^2\min\{n_x,n_u\}}{2}(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_2\bar{h}_{\text{het}}^2)^2 + \frac{3\bar{h}_{\text{grad}}\eta_l^2}{2}\left\|\nabla C^{(j)}(K_n)\right\|_F^2,
\end{aligned}
$$

where (a) follows from Eq.(6.8); (b) follows from the same reasoning as we bound T_1 and T_2 and the fact that $K_n \in \mathcal{G}^0$. If we choose the local step-size η_l satisfies $\eta_l \leq \frac{1}{9\bar{h}_{\text{grad}}}$, i.e., $\frac{3\bar{h}_{\text{grad}}\eta_l^2}{2} \leq \frac{\eta_l}{6}$, we have

$$
\begin{aligned}
C^{(j)}(K_{n,1}^{(i)}) - C^{(j)}(K_n) &\overset{(a)}{\leq} -\eta_l \left\|\nabla C^{(j)}(K_n)\right\|_F^2 + \frac{\eta_l}{6}\left\|\nabla C^{(j)}(K_n)\right\|_F^2 \\
&+ \frac{3\eta_l\min\{n_x,n_u\}}{2}(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_2\bar{h}_{\text{het}}^2)^2 + \frac{\eta_l\left\|\nabla C^{(j)}(K_n)\right\|_F^2}{6} + \frac{3\eta_l\bar{\epsilon}}{2} + \frac{\eta_l}{6}\bar{\epsilon} \\
&+ \frac{\eta_l\min\{n_x,n_u\}}{6}(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_2\bar{h}_{\text{het}}^2)^2 + \frac{\eta_l}{6}\left\|\nabla C^{(j)}(K_n)\right\|_F^2 \\
&\leq -\frac{\eta_l}{2}\left\|\nabla C^{(j)}(K_n)\right\|_F^2 + \frac{5\eta_l\min\{n_x,n_u\}}{3}(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_2\bar{h}_{\text{het}}^2)^2 + \frac{5\eta_l}{3}\bar{\epsilon} \\
&\overset{(b)}{\leq} -\frac{2\eta_l\sigma_{\min}(R)\mu^2}{||\Sigma_{K_j^*}||}(C^{(j)}(K_n) - C^{(j)}(K_j^*)) + \frac{5\eta_l\min\{n_x,n_u\}}{3}(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_2\bar{h}_{\text{het}}^2)^2 + \frac{5\eta_l}{3}\bar{\epsilon},
\end{aligned}
$$

where (a) follows from the Young's inequality in Eq.(6.9); and (b) follows from the gradient domination in Lemma 26.

Therefore, if the heterogeneity satisfies $(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_2\bar{h}_{\text{het}}^2)^2 \leq \bar{h}_{\text{het}}^3$, then we have

$$
(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_2\bar{h}_{\text{het}}^2)^2 \leq \min_{j\in[M]}\left\{\frac{3\mu^2\sigma_{\min}(R)\left(C^{(j)}(K_0) - C^{(j)}(K_j^*)\right)}{5||\Sigma_{K_j^*}||\min\{n_x,n_u\}}\right\}.
$$

Since the error tolerance

$$\bar{\epsilon} = \min_{j\in[M]} \left\{ \frac{3\mu^2\sigma_{\min}(R)\left(C^{(j)}(K_0) - C^{(j)}(K_j^*)\right)}{5||\Sigma_{K_j^*}||} \right\},$$

we have

$$\begin{aligned}
C^{(j)}(K_{n,1}^{(i)}) - C^{(j)}(K_j^*) &\leq \left(1 - \frac{2\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_j^*}\right\|}\right)(C^{(j)}(K_n) - C^{(j)}(K_j^*)) \\
&+ \frac{\eta_l\mu^2\sigma_{\min}(R)\left(C^{(j)}(K_0) - C^{(j)}(K_j^*)\right)}{||\Sigma_{K_j^*}||} + \frac{\eta_l\mu^2\sigma_{\min}(R)\left(C^{(j)}(K_0) - C^{(j)}(K_j^*)\right)}{||\Sigma_{K_j^*}||} \\
&\overset{(a)}{\leq} \left(1 - \frac{2\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_j^*}\right\|}\right)(C^{(j)}(K_0) - C^{(j)}(K_j^*)) + \frac{2\eta_l\mu^2\sigma_{\min}(R)\left(C^{(j)}(K_0) - C^{(j)}(K_j^*)\right)}{||\Sigma_{K_j^*}||} \\
&= C^{(j)}(K_0) - C^{(j)}(K_j^*), \forall j \in [M],
\end{aligned}$$

where we use the fact that $K_n \in \mathcal{G}^0$ in (a). The above inequality implies $K_{n,1}^{(i)} \in \mathcal{G}^0$ with high probability $1-\delta$ when $K_n \in \mathcal{G}^0$. Then we can use the induction method to obtain that $K_{n,2}^{(i)} \in \mathcal{G}^0$, since $K_{n,1}^{(i)} \in \mathcal{G}^0$. By repeating this step for L times, we have that all the local polices $K_{n,l}^{(i)} \in \mathcal{G}^0$ holds for all $i \in [M]$ and $l = 0, 1, \cdots, L-1$, when the global policy $K_n \in \mathcal{G}^0$. □

Lemma 41. *(Drift term analysis) Suppose $K_n \in \mathcal{G}^0$. If $\eta_l \leq \min\left\{\frac{1}{4\bar{h}_{grad}}, \frac{1}{4}, \frac{\log 2}{L(3\bar{h}_{grad}+2)}\right\}$, the number of the sample size n_s satisfies*

$$n_s \geq \frac{h_{sample,trunc}\left(\frac{\sqrt{\epsilon}}{4}, \frac{\delta}{L}, \frac{H^2}{\mu}\right)}{ML},$$

the smoothing radius satisfies $r \leq h_r\left(\frac{\sqrt{\epsilon}}{4}\right)$ and the length of trajectories satisfies $\tau \geq h_\tau\left(\frac{r\sqrt{\epsilon}}{4n_x n_u}\right)$, given any $\delta \in (0,1)$, the difference between the local policy and global policy can be bounded by

$$\left\|K_{n,l}^{(i)} - K_n\right\|_F^2 \leq 2\eta_l L\left[\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + ML\epsilon\right] = \frac{2\eta}{\eta_g}\left[\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + ML\epsilon\right]$$

holds, with probability $1-\delta$, for all $i \in [M]$ and $l = 0, 1, \cdots, L-1$.

Proof:

$$\left\|K_{n,l}^{(i)} - K_n\right\|_F^2 = \left\|K_{n,l-1}^{(i)} - K_n - \eta_l\tilde{\nabla}C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2$$

$$
\begin{aligned}
&= \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 - 2\eta_l \left[\left\langle \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}), K_{n,l-1}^{(i)} - K_n \right\rangle\right] \\
&+ \left\|\eta_l \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&= \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 - 2\eta_l \left[\left\langle \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)}), K_{n,l-1}^{(i)} - K_n \right\rangle\right] \\
&- 2\eta_l \left[\left\langle \nabla C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_n), K_{n,l-1}^{(i)} - K_n \right\rangle\right] - 2\eta_l \left[\left\langle \nabla C^{(i)}(K_n), K_{n,l-1}^{(i)} - K_n \right\rangle\right] \\
&+ \left\|\eta_l \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&\overset{(a)}{\leq} \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 - 2\eta_l \left[\left\langle \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)}), K_{n,l-1}^{(i)} - K_n \right\rangle\right] \\
&+ 2\eta_l \left\|\nabla C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_n)\right\|_F \left\|K_{n,l-1}^{(i)} - K_n\right\|_F + 2\eta_l \left\|\nabla C^{(i)}(K_n)\right\|_F \left\|K_{n,l-1}^{(i)} - K_n\right\|_F \\
&+ \left\|\eta_l \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&\overset{(b)}{\leq} \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 - 2\eta_l \left[\left\langle \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)}), K_{n,l-1}^{(i)} - K_n \right\rangle\right] \\
&+ 2\eta_l h_{\text{grad}}(K_n) \left\|K_{n,l-1}^{(i)} - K_n\right\|_F \left\|K_{n,l-1}^{(i)} - K_n\right\|_F + \eta_l \left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \eta_l \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 \\
&+ \left\|\eta_l \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2
\end{aligned} \tag{6.40}
$$

where we use Cauchy–schwarz inequality for (a); and for (b), we use Eq. (6.9).

Following the analysis in Eq.(6.40), we have

$$
\begin{aligned}
&\left\|K_{n,l}^{(i)} - K_n\right\|_F^2 \leq \left(1 + 2\eta_l h_{\text{grad}}(K_n) + \eta_l\right) \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 + \eta_l \left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
&- 2\eta_l \left[\left\langle \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)}), K_{n,l-1}^{(i)} - K_n \right\rangle\right] + \left\|\eta_l \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&\overset{(a)}{\leq} \left(1 + 2\eta_l h_{\text{grad}}(K_n) + \eta_l\right) \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 + \eta_l \left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
&+ 2\eta_l \left[\left\|\tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F \left\|K_{n,l-1}^{(i)} - K_n\right\|_F\right] \\
&+ 2\eta_l^2 \left\|\tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 + 2\eta_l^2 \left\|\nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&\overset{(b)}{\leq} \left(1 + 2\eta_l h_{\text{grad}}(K_n) + \eta_l\right) \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 + \eta_l \left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
&+ \eta_l \left\|\tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 + \eta_l \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 \\
&+ 2\eta_l^2 \left\|\tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)})\right\|_F^2 \\
&+ 4\eta_l^2 \left\|\nabla C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_n)\right\|_F^2 + 4\eta_l^2 \left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
&\overset{(c)}{\leq} \left(1 + 2\eta_l h_{\text{grad}}(K_n) + \eta_l\right) \left\|K_{n,l-1}^{(i)} - K_n\right\|_F^2 + (\eta_l + 4\eta_l^2) \left\|\nabla C^{(i)}(K_n)\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \eta_l \left\| \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)}) \right\|_F^2 + \eta_l \left\| K_{n,l-1}^{(i)} - K_n \right\|_F^2 \\
&+ 2\eta_l^2 \left\| \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)}) \right\|_F^2 + 4\eta_l^2 h_{\text{grad}}(K_n)^2 \left\| K_{n,l-1}^{(i)} - K_n \right\|_F^2 \\
&\overset{(d)}{=} \left(1 + 2\eta_l h_{\text{grad}}(K_n) + 2\eta_l + 4\eta_l^2 h_{\text{grad}}(K_n)^2\right) \left\| K_{n,l-1}^{(i)} - K_n \right\|_F^2 + (\eta_l + 4\eta_l^2) \left\| \nabla C^{(i)}(K_n) \right\|_F^2 \\
&+ \left(\eta_l + 2\eta_l^2\right) \left\| \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)}) \right\|_F^2 \\
&\leq \left(1 + 2\eta_l \bar{h}_{\text{grad}} + 2\eta_l + 4\eta_l^2 \bar{h}_{\text{grad}}^2\right) \left\| K_{n,l-1}^{(i)} - K_n \right\|_F^2 + (\eta_l + 4\eta_l^2) \left\| \nabla C^{(i)}(K_n) \right\|_F^2 \\
&+ \left(\eta_l + 2\eta_l^2\right) \underbrace{\left\| \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)}) \right\|_F^2}_{T_1},
\end{aligned}
$$

where we use Cauchy-Schwarz inequality and Eq.(6.6) for (a); for (b), we use Eq.(6.6) and (6.8); for (c), we use the gradient smoothness lemma in Lemma 25; and for (d), we use the fact that $K_n \in \mathcal{G}^0$.

From Lemma 39, we can bound T_1 term as follows

$$
T_1 = \left\| \tilde{\nabla} C^{(i)}(K_{n,l-1}^{(i)}) - \nabla C^{(i)}(K_{n,l-1}^{(i)}) \right\|_F^2 \leq ML\epsilon,
$$

where it holds with probability $1 - \delta$, since $n_s \geq \frac{h_{\text{sample,trunc}}\left(\frac{\sqrt{\epsilon}}{4}, \delta, \frac{H^2}{\mu}\right)}{ML}$, the smoothing radius satisfies $r \leq h_r\left(\frac{\sqrt{\epsilon}}{4}\right)$ and the length of trajectories satisfies $\tau \geq h_\tau\left(\frac{r\sqrt{\epsilon}}{4n_x n_u}\right)$.

Then we have

$$
\begin{aligned}
\left\| K_{n,l}^{(i)} - K_n \right\|_F^2 &\leq \left(1 + 2\eta_l \bar{h}_{\text{grad}} + 2\eta_l + 4\eta_l^2 \bar{h}_{\text{grad}}^2\right) \left\| K_{n,l-1}^{(i)} - K_n \right\|_F^2 + (\eta_l + 4\eta_l^2) \left\| \nabla C^{(i)}(K_n) \right\|_F^2 \\
&\quad + \left(\eta_l + 2\eta_l^2\right) ML\epsilon \\
&\overset{(a)}{\leq} (1 + 3\eta_l \bar{h}_{\text{grad}} + 2\eta_l) \left\| K_{n,l-1}^{(i)} - K_n \right\|_F^2 + 2\eta_l \left\| \nabla C^{(i)}(K_n) \right\|_F^2 + 2\eta_l ML\epsilon \\
&\leq (1 + 3\eta_l \bar{h}_{\text{grad}} + 2\eta_l)^l \underbrace{\left\| K_{n,0}^{(i)} - K_n \right\|_F^2}_{=0} \\
&\quad + 2\eta_l \sum_{j=0}^{l-1} \left(1 + 3\eta_l \bar{h}_{\text{grad}} + 2\eta_l\right)^j \left[\left\| \nabla C^{(i)}(K_n) \right\|_F^2 + ML\epsilon \right] \\
&\leq 2\eta_l \times \frac{\left(1 + 3\eta_l \bar{h}_{\text{grad}} + 2\eta_l\right)^l - 1}{\left(1 + 3\eta_l \bar{h}_{\text{grad}} + 2\eta_l\right) - 1} \left[\left\| \nabla C^{(i)}(K_n) \right\|_F^2 + ML\epsilon \right] \\
&\leq 2 \times \frac{\left(1 + 3\eta_l \bar{h}_{\text{grad}} + 2\eta_l\right)^l - 1}{3\bar{h}_{\text{grad}} + 2} \left[\left\| \nabla C^{(i)}(K_n) \right\|_F^2 + ML\epsilon \right] \\
&\overset{(b)}{\leq} 2 \times \frac{1 + l(3\eta_l \bar{h}_{\text{grad}} + 2\eta_l) - 1}{3\bar{h}_{\text{grad}} + 2} \left[\left\| \nabla C^{(i)}(K_n) \right\|_F^2 + ML\epsilon \right]
\end{aligned}
$$

$$\leq 2\eta_l L \left[\left\| \nabla C^{(i)}(K_n) \right\|_F^2 + ML\epsilon \right],$$

where (a) is due to the choice of local step-size which satisfies $2\eta_l \bar{h}_{\text{grad}} + 2\eta_l + 4\eta_l^2 \bar{h}_{\text{grad}}^2 \leq 3\eta_l \bar{h}_{\text{grad}} + 2\eta_l$ and $\eta_l + 2\eta_l^2 \leq \eta_l + 4\eta_l^2 \leq 2\eta_l$, i.e., $\eta_l \leq \min\left\{ \frac{1}{4\bar{h}_{\text{grad}}}, \frac{1}{4} \right\}$. For (b), we used the fact that $(1+x)^{\tau+1} \leq 1 + 2x(\tau+1)$ holds for $x \leq \frac{\log 2}{\tau}$. In other words, $\left(1 + 3\eta_l \bar{h}_{\text{grad}} + 2\eta_l\right)^l \leq 1 + l(3\eta_l \bar{h}_{\text{grad}} + 2\eta_l)$ when $3\eta_l \bar{h}_{\text{grad}} + 2\eta_l \leq \frac{\log 2}{l}$, i.e., $\eta_l \leq \frac{\log 2}{L(3\bar{h}_{\text{grad}}+2)}$. □

Lemma 42. *(Per round progress) Suppose $K_n \in \mathcal{G}^0$. If we choose the local step-size as*

$$\eta_l = \frac{1}{2} \min\left\{ \frac{\underline{h}_\Delta \mu}{H^2\left(h_1 + \sqrt{\epsilon}\right)}, \frac{1}{9\bar{h}_{grad}}, \frac{1}{4}, \frac{\log 2}{L(3\bar{h}_{grad} + 2)}, \frac{1}{256 L \bar{h}_{grad}^2} \right\},$$

with step-size $\eta := L\eta_l\eta_g = \frac{1}{2}\min\{\frac{\underline{h}_\Delta \mu}{H^2\left(h_1+\sqrt{\epsilon}\right)}, 1, \frac{1}{32\bar{h}_{grad}}\}$, and the smoothing radius[19]

$$r \leq \min\left\{ \frac{\min_{i\in[M]} C^{(i)}(K_0)}{\bar{h}_{cost}}, \underline{h}_\Delta, h_r\left(\frac{\sqrt{\epsilon}}{4}\right) \right\},$$

where the trajectory length satisfies $\tau \geq h_\tau\left(\frac{r\sqrt{\epsilon}}{4n_x n_u}\right)$, and the number of the sample size satisfies

$$n_s \geq \frac{h_{sample,trunc}\left(\frac{\sqrt{\epsilon}}{4}, \frac{\delta}{L}, \frac{H^2}{\mu}\right)}{ML},$$

then with probability $1-\delta$, for any small $\delta \in (0,1)$, the `FedLQR` *algorithm provides the following convergence guarantee:*

$$C^{(i)}(K_{n+1}) - C^{(i)}(K_i^*) \leq \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)\left(C^{(i)}(K_n) - C^{(i)}(K_i^*)\right) + 2\eta\epsilon$$
$$+ 2\eta \min\{n_x, n_u\}(\epsilon_1 \bar{h}_{het}^1 + \epsilon_1 \bar{h}_{het}^2).^2 \tag{6.41}$$

Proof: For any $i \in [M]$, according to the local Lipschitz property in Lemma 25, we have that

$$C^{(i)}(K_{n+1}) - C^{(i)}(K_n) \leq \langle \nabla C^{(i)}(K_n), K_{n+1} - K_n \rangle + \frac{h_{\text{grad}}(K_n)}{2} \|K_{n+1} - K_n\|_F^2$$

[19]The exact requirement of r is $r \leq \min\left\{ \frac{\min_{i\in[M]} C^{(i)}(K_0)}{\bar{h}_{\text{cost}}}, \underline{h}_\Delta, h_r\left(\frac{\sqrt{\epsilon}}{4}\right), h_r\left(\frac{\sqrt{\epsilon}}{4}\right) \right\}$. Here, without loss of generality, we drop the $h_r\left(\frac{\sqrt{\epsilon}}{4}\right)$ term from the $\min$ expression. This can be done because the error tolerance ϵ is usually small, and so $h_r\left(\frac{\sqrt{\epsilon}}{4}\right) \leq h_r\left(\frac{\sqrt{\epsilon}}{4}\right)$ holds. The assumptions on τ and n_s follow similarly.

$$= -\left\langle \nabla C^{(i)}(K_n), \frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla}C^{(j)}(K_{n,l}^{(j)})\right\rangle + \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla}C^{(j)}(K_{n,l}^{(j)})\right\|_F^2, \quad (6.42)$$

holds when $\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla}C^{(j)}(K_{n,l}^{(j)})\right\|_F \le \underline{h}_\Delta \le h_\Delta(K_n)$. Following the same analysis as Eq.(6.39), this inequality holds when

$$\eta \le \frac{\underline{h}_\Delta \mu}{H^2\left(h_1+\sqrt{\epsilon}\right)}, \quad r \le \min\left\{\frac{\min_{i\in[M]} C^{(i)}(K_0)}{\bar{h}_{\text{cost}}}, \underline{h}_\Delta\right\}.$$

Following the analysis in Eq.(6.42), we have

$$\begin{aligned}
C^{(i)}(K_{n+1}) - C^{(i)}(K_n) \le & -\left\langle \nabla C^{(i)}(K_n), \frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla}C^{(j)}(K_{n,l}^{(j)}) - \frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\nabla C^{(j)}(K_{n,l}^{(j)})\right\rangle \\
& - \left\langle \nabla C^{(i)}(K_n), \frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\nabla C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_n)\right\rangle \\
& - \left\langle \nabla C^{(i)}(K_n), \frac{\eta}{M}\sum_{j=1}^{M}\nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n)\right\rangle - \eta\left\|C^{(i)}(K_n)\right\|^2 \\
& + \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla}C^{(j)}(K_{n,l}^{(j)})\right\|_F^2 \\
\overset{(a)}{\le} & \eta\left\|\nabla C^{(i)}(K_n)\right\|_F\left\|\frac{1}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left[\tilde{\nabla}C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_{n,l}^{(j)})\right]\right\|_F \\
& + \eta\left\|\nabla C^{(i)}(K_n)\right\|_F\left\|\frac{1}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left[\nabla C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_n)\right]\right\|_F \\
& + \eta\left\|\nabla C^{(i)}(K_n)\right\|_F\left\|\frac{1}{M}\sum_{j=1}^{M}\left[\nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n)\right]\right\|_F - \eta\left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
& + \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla}C^{(j)}(K_{n,l}^{(j)})\right\|_F^2 \\
\overset{(b)}{\le} & \frac{\eta}{4}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \eta\left\|\frac{1}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left[\tilde{\nabla}C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_{n,l}^{(j)})\right]\right\|_F^2 \\
& + \frac{\eta}{8}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \frac{2\eta h_{\text{grad}}(K_n)^2}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left\|K_{n,l}^{(j)} - K_n\right\|_F^2 \\
& + \frac{\eta}{4}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \frac{\eta}{M}\sum_{j=1}^{M}\left\|\nabla C^{(j)}(K_n) - C^{(i)}(K_n)\right\|_F^2 \\
& - \eta\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla}C^{(j)}(K_{n,l}^{(j)})\right\|_F^2,
\end{aligned}$$

where (a) is due to Cauchy–Schwarz inequality; and (b) is due to Cauchy–Schwarz inequality and Eq.(6.7). Moreover, we have

$$\begin{aligned}
C^{(i)}(K_{n+1}) - C^{(i)}(K_n) &\overset{(b)}{\leq} \frac{\eta}{4}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \eta\left\|\frac{1}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left[\tilde{\nabla} C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_{n,l}^{(j)})\right]\right\|_F^2 \\
&+ \frac{\eta}{8}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \frac{2\eta h_{\text{grad}}(K_n)^2}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\left\|K_{n,l}^{(j)} - K_n\right\|_F^2 \\
&+ \frac{\eta}{4}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \frac{\eta}{M}\sum_{j=1}^{M}\left\|\nabla C^{(j)}(K_n) - C^{(i)}(K_n)\right\|_F^2 \\
&- \eta\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla} C^{(j)}(K_{n,l}^{(j)})\right\|_F^2 \\
&\overset{(c)}{\leq} -\frac{3\eta}{8}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \eta\epsilon + \frac{4\eta^2\bar{h}_{\text{grad}}^2}{\eta_g M}\sum_{j=1}^{M}\left[\left\|\nabla C^{(j)}(K_n)\right\|_F^2 + ML\epsilon\right] \\
&+ \eta\min\{n_x, n_u\}(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_1\bar{h}_{\text{het}}^2)^2 + \frac{h_{\text{grad}}(K_n)}{2}\left\|\frac{\eta}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla} C^{(j)}(K_{n,l}^{(j)})\right\|_F^2,
\end{aligned} \tag{6.43}$$

where (b) follows from the gradient Lipschitz property in Lemma 25; and (c) follows from the policy gradient heterogeneity property in Lemma 27, Lemma 28 and Lemma 41.

Following the analysis in Eq.(6.43), we have

$$\begin{aligned}
C^{(i)}(K_{n+1}) - C^{(i)}(K_n) &\overset{(d)}{\leq} -\frac{3\eta}{8}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + \eta\epsilon + \frac{4\eta^2\bar{h}_{\text{grad}}^2}{\eta_g M}\sum_{j=1}^{M}\left[\left\|\nabla C^{(j)}(K_n)\right\|_F^2 + ML\epsilon\right] \\
&+ \eta\min\{n_x, n_u\}(\epsilon_1 h_{\text{het}}^1 + \epsilon_1 h_{\text{het}}^2)^2 \\
&+ \frac{4\eta^2\bar{h}_{\text{grad}}}{2}\left\|\frac{1}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\tilde{\nabla} C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_{n,l}^{(j)})\right\|_F^2 \\
&+ \frac{4\eta^2\bar{h}_{\text{grad}}}{2}\left\|\frac{1}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\nabla C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_n)\right\|_F^2 \\
&+ \frac{4\eta^2\bar{h}_{\text{grad}}}{2M}\sum_{j=1}^{M}\left\|\nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n)\right\|_F^2 + \frac{4\eta^2\bar{h}_{\text{grad}}}{2}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 \\
&\overset{(e)}{\leq} -\frac{3\eta}{8}\left\|\nabla C^{(i)}(K_n)\right\|_F^2 + (\eta + 2\eta^2\bar{h}_{\text{grad}})\epsilon + \frac{4\eta^2\bar{h}_{\text{grad}}^2}{\eta_g M}\sum_{j=1}^{M}\left[\left\|\nabla C^{(j)}(K_n)\right\|_F^2 + ML\epsilon\right] \\
&+ (\eta + 2\eta^2\bar{h}_{\text{grad}})\min\{n_x, n_u\}(\epsilon_1 h_{\text{het}}^1 + \epsilon_1 h_{\text{het}}^2)^2 + 2\eta^2\bar{h}_{\text{grad}}\left\|\nabla C^{(i)}(K_n)\right\|_F^2
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{4\eta^2\bar{h}_{\text{grad}}}{2}\Big\|\frac{1}{ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\nabla C^{(j)}(K_{n,l}^{(j)}) - \nabla C^{(j)}(K_n)\Big\|_F^2 \\
&\overset{(f)}{\leq} -\left(\frac{3\eta}{8} + 2\eta^2\bar{h}_{\text{grad}}\right)\Big\|\nabla C^{(i)}(K_n)\Big\|_F^2 + (\eta + 2\eta^2\bar{h}_{\text{grad}})\epsilon \\
&+ \frac{4\eta^2\bar{h}_{\text{grad}}^2}{\eta_g M}\sum_{j=1}^{M}\left[\Big\|\nabla C^{(j)}(K_n)\Big\|_F^2 + ML\epsilon\right] + (\eta + 2\eta^2\bar{h}_{\text{grad}})\min\{n_x, n_u\}(\epsilon_1 h_{\text{het}}^1 + \epsilon_1 h_{\text{het}}^2)^2 \\
&+ \frac{4\eta^2\bar{h}_{\text{grad}}^2}{2ML}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\Big\|K_{n,l}^{(j)} - K_n\Big\|_F^2 \\
&\overset{(g)}{\leq} -\left(\frac{3\eta}{8} + 2\eta^2\bar{h}_{\text{grad}}\right)\Big\|\nabla C^{(i)}(K_n)\Big\|_F^2 + (\eta + 2\eta^2\bar{h}_{\text{grad}})\epsilon \\
&+ \frac{4\eta^2\bar{h}_{\text{grad}}^2 + 4\eta^3\bar{h}_{\text{grad}}^2}{\eta_g M}\sum_{j=1}^{M}\left[\Big\|\nabla C^{(j)}(K_n)\Big\|_F^2 + ML\epsilon\right] \\
&+ (\eta + 2\eta^2\bar{h}_{\text{grad}})\min\{n_x, n_u\}(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_1\bar{h}_{\text{het}}^2)^2,
\end{aligned} \tag{6.44}
$$

where (d) is due to Eq.(6.8); (e) is due to variance reduction property in Lemma 28 and policy gradient heterogeneity in Lemma 27; (f) is due to gradient Lipschitz property in Lemma 25; (g) is due to drift term analysis in Lemma 41.

Continuing the analysis in Eq.(6.44), we have that

$$
\begin{aligned}
C^{(i)}(K_{n+1}) - C^{(i)}(K_n) &\leq -\left(\frac{3\eta}{8} + 2\eta^2\bar{h}_{\text{grad}}\right)\Big\|\nabla C^{(i)}(K_n)\Big\|_F^2 + (\eta + 2\eta^2\bar{h}_{\text{grad}})\epsilon \\
&+ \frac{4\eta^2\bar{h}_{\text{grad}}^2 + 4\eta^3\bar{h}_{\text{grad}}^2}{\eta_g M}\sum_{j=1}^{M}\left[\Big\|\nabla C^{(j)}(K_n)\Big\|_F^2 + ML\epsilon\right] \\
&+ (\eta + 2\eta^2\bar{h}_{\text{grad}})\min\{n_x, n_u\}(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_1\bar{h}_{\text{het}}^2)^2 \\
&\overset{(a)}{\leq} -\left(\frac{3\eta}{8} + 2\eta^2\bar{h}_{\text{grad}}\right)\Big\|\nabla C^{(i)}(K_n)\Big\|_F^2 + (\eta + 2\eta^2\bar{h}_{\text{grad}})\epsilon \\
&+ \frac{4\eta^2\bar{h}_{\text{grad}}^2 + 4\eta^3\bar{h}_{\text{grad}}^2}{\eta_g M}\sum_{j=1}^{M}\left[2\Big\|\nabla C^{(j)}(K_n) - \nabla C^{(i)}(K_n)\Big\|_F^2 + 2\Big\|\nabla C^{(i)}(K_n)\Big\|_F^2 + ML\epsilon\right] \\
&+ (\eta + 2\eta^2\bar{h}_{\text{grad}})\min\{n_x, n_u\}(\epsilon_1\bar{h}_{\text{het}}^1 + \epsilon_1\bar{h}_{\text{het}}^2)^2 \\
&\overset{(b)}{\leq} -\left(\frac{3\eta}{8} + 2\eta^2\bar{h}_{\text{grad}} + \frac{8\eta^2\bar{h}_{\text{grad}}^2 + 8\eta^3\bar{h}_{\text{grad}}^2}{\eta_g}\right)\Big\|\nabla C^{(i)}(K_n)\Big\|_F^2 \\
&+ \left(\eta + 2\eta^2\bar{h}_{\text{grad}} + \frac{4\eta^2\bar{h}_{\text{grad}}^2 + 4\eta^3\bar{h}_{\text{grad}}^2}{\eta_g}ML\right)\epsilon
\end{aligned}
$$

$$
\begin{aligned}
&+\left(\eta+2\eta^2\bar{h}_{\text{grad}}+\frac{8\eta^2\bar{h}^2_{\text{grad}}+8\eta^3\bar{h}^2_{\text{grad}}}{\eta_g}\right)\min\{n_x,n_u\}(\epsilon_1\bar{h}^1_{\text{het}}+\epsilon_1\bar{h}^2_{\text{het}})^2\\
&\overset{(c)}{\leq}-\frac{\eta}{4}\left\|\nabla C^{(i)}(K_n)\right\|_F^2+2\eta\epsilon+2\eta\min\{n_x,n_u\}(\epsilon_1\bar{h}^1_{\text{het}}+\epsilon_1\bar{h}^2_{\text{het}})^2\\
&\overset{(d)}{\leq}-\frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}(C^{(i)}(K_n)-C^{(i)}(K_i^*))+2\eta\epsilon+2\eta\min\{n_x,n_u\}(\epsilon_1\bar{h}^1_{\text{het}}+\epsilon_1\bar{h}^2_{\text{het}})^2,
\end{aligned}
\tag{6.45}
$$

where (a) is due to Eq.(6.8); (b) is due to policy gradient heterogeneity in Lemma 27; and (c) is due to the choice of step-size such that $\frac{3\eta}{8}+2\eta^2\bar{h}_{\text{grad}}+\frac{8\eta^2\bar{h}^2_{\text{grad}}+8\eta^3\bar{h}^2_{\text{grad}}}{\eta_g}\leq\frac{\eta}{4}$ and

$$
\eta+2\eta^2\bar{h}_{\text{grad}}+\frac{8\eta^2\bar{h}^2_{\text{grad}}+8\eta^3\bar{h}^2_{\text{grad}}}{\eta_g}\leq 2\eta,
$$

which holds when $\eta\leq\min\{\frac{1}{32\bar{h}_{\text{grad}}},1\}$ and $\eta_l\leq\frac{1}{256L\bar{h}^2_{\text{grad}}}$; for (d) we use the gradient domination lemma in Lemma 26.

In conclusion, we have that

$$
\begin{aligned}
C^{(i)}(K_{n+1})-C^{(i)}(K_i^*)\leq&\left(1-\frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)(C^{(i)}(K_n)-C^{(i)}(K_i^*))+2\eta\epsilon\\
&+2\eta\min\{n_x,n_u\}(\epsilon_1\bar{h}^1_{\text{het}}+\epsilon_1\bar{h}^2_{\text{het}})^2,
\end{aligned}
$$

holds when the step-size, smoothing radius, trajectory length, and sample size satisfy the requirements mentioned above and those in Lemma 40 and Lemma 41. □

With this lemma, we are now ready to provide the convergence guarantees for the `FedLQR` under the model-free setting.

Proof of the iterative stability guarantees of `FedLQR`: Here, we leverage the method of induction to prove `FedLQR`'s iterative stability guarantees. First, we start from an initial policy $K_0\in\mathcal{G}^0$. At round n, we assume $K_n\in\mathcal{G}^0$. According to Lemma 40, we have that all the local policies $K^{(i)}_{n,l}\in\mathcal{G}^0$. Furthermore, frame the hypotheses of in Lemma 42, we have that

$$
\begin{aligned}
C^{(i)}(K_{n+1})-C^{(i)}(K_i^*)\leq&\left(1-\frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)(C^{(i)}(K_n)-C^{(i)}(K_i^*))+2\eta\epsilon\\
&+2\eta\min\{n_x,n_u\}(\epsilon_1\bar{h}^1_{\text{het}}+\epsilon_1\bar{h}^2_{\text{het}})^2.
\end{aligned}
$$

Since $(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_2 \bar{h}^2_{\text{het}})^2 \leq \bar{h}^3_{\text{het}}$, we have

$$\begin{aligned} C^{(i)}(K_{n+1}) - C^{(i)}(K_i^*) &\leq \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)(C^{(i)}(K_0) - C^{(i)}(K_i^*)) + 2\eta\epsilon \\ &+ \frac{\eta\mu^2\sigma_{\min}(R)}{2\left\|\Sigma_{K_i^*}\right\|}(C^{(i)}(K_0) - C^{(i)}(K_i^*)) \\ &\overset{(a)}{\leq} C^{(i)}(K_0) - C^{(i)}(K_i^*), \end{aligned}$$

where (a) follows from the fact that ϵ can be arbitrarily small by choosing a small smoothing radius, sufficient long trajectory length, and enough samples. □

With this, we can easily have that the global policy K_{n+1} at the next round $n+1$ is also stabilizing, i.e., $K_{n+1} \in \mathcal{G}^0$. Therefore, we can finish proving `FedLQR`'s iterative stability property by inductively reasoning.

Proof of `FedLQR`'s convergence: From Eq.(6.41), we have

$$\begin{aligned} C^{(i)}(K_{n+1}) - C^{(i)}(K_i^*) \leq \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)(C^{(i)}(K_n) - C^{(i)}(K_i^*)) + 2\eta\epsilon \\ + 2\eta\min\{n_x, n_u\}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_1 \bar{h}^2_{\text{het}})^2, \end{aligned}$$

Using the above inequality recursively, `FedLQR` enjoys the following convergence guarantee after N rounds:

$$\begin{aligned} C^{(i)}(K_N) - C^{(i)}(K_i^*) \leq \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)^N (C^{(i)}(K_0) - C^{(i)}(K_i^*)) + \frac{2\left\|\Sigma_{K_i^*}\right\|}{\mu^2\sigma_{\min}(R)}\epsilon \\ + \frac{2\min\{n_x, n_u\}\left\|\Sigma_{K_i^*}\right\|}{\mu^2\sigma_{\min}(R)}(\epsilon_1 \bar{h}^1_{\text{het}} + \epsilon_1 \bar{h}^2_{\text{het}})^2. \end{aligned}$$

Suppose the trajectory length satisfies $\tau \geq h_\tau\left(\frac{r\epsilon'}{4n_x n_u}\right)$, the smoothing radius satisfies $r \leq h'_r\left(\frac{\epsilon'}{4}\right)$, where

$$h'_r\left(\frac{\epsilon'}{4}\right) := \min\left\{\frac{\min_{i\in[M]} C^{(i)}(K_0)}{\bar{h}_{\text{cost}}}, \underline{h}_\Delta, h_r\left(\frac{\epsilon'}{4}\right)\right\},$$

and the number of the sample size of each agent n_s satisfies

$$n_s \geq \frac{h_{\text{sample,trunc}}\left(\frac{\epsilon'}{4}, \frac{\delta}{ML}, \frac{H^2}{\mu}\right)}{ML},$$

with $\epsilon' = \sqrt{\frac{\mu^2\sigma_{\min}(R)}{4\left\|\Sigma_{K_i^*}\right\|} \cdot \epsilon}$.

When the number of rounds $N \geq \frac{c_{\text{uni},4}\left\|\Sigma_{K_i^*}\right\|}{\eta\mu^2\sigma_{\min}(R)} \log\left(\frac{2(C^{(i)}(K_0)-C^{(i)}(K_i^*))}{\epsilon'}\right)$, our `FedLQR` algorithm enjoys the following convergence guarantee:

$$\begin{aligned} C^{(i)}(K_N) - C^{(i)}(K_i^*) &\leq \left(1 - \frac{\eta\mu^2\sigma_{\min}(R)}{\left\|\Sigma_{K_i^*}\right\|}\right)^N (C^{(i)}(K_0) - C^{(i)}(K_i^*)) + \frac{\epsilon'}{2} \\ &\quad + \frac{2\min\{n_x, n_u\}\left\|\Sigma_{K_i^*}\right\|}{\mu^2\sigma_{\min}(R)}(\epsilon_1 h_{\text{het}}^1 + \epsilon_1 h_{\text{het}}^2)^2 \\ &\leq \epsilon' + \frac{2\min\{n_x, n_u\}\left\|\Sigma_{K_i^*}\right\|}{\mu^2\sigma_{\min}(R)}(\epsilon_1 h_{\text{het}}^1 + \epsilon_1 h_{\text{het}}^2)^2. \end{aligned}$$

Thus, we complete the proof with $c_{\text{uni},2} = 2$, $c_{\text{uni},3} = 1$ and $c_{\text{uni},4} = 1$. □

Chapter 7

Personalized System Identification

7.1 Introduction

System identification is the data-driven process of estimating a dynamic model of a system based on observations of the system trajectories. It plays a crucial role in aiding our understanding of complex systems and is a fundamental problem in numerous fields, including time-series analysis, control theory, robotics, and reinforcement learning [7, 125]. The effective utilization of available data is pivotal in obtaining an accurate model estimate with a measure of uncertainty quantification. Traditional system identification, methods [125] have focused on asymptotic analysis, which, although insightful, is restrictive when dealing with small to medium sized data sets. Motivated by this, and the fact that data generation is often costly and time consuming, modern approaches focus on developing sample complexity bounds (i.e., non-asymptotic convergence analysis).

Results on the estimation of both fully [35, 168, 179] and partially [146, 177, 186, 206, 253] observed LTI systems have demonstrated that a more precise characterization of error bounds is essential for designing efficient and robust control systems [35, 206, 254]. These studies provide non-asymptotic bounds that are functions of the number of observed trajectories (see Table 1 of [253] for a summary of the bounds).

A recent body of work has begun to formalize methods for improving sample efficiency by considering data (or models generated from data) from multiple systems [27, 193, 212, 229, 230, 245, 246]. Leveraging data from similar systems provides a promising approach although clarifying the effect of the heterogeneity in the systems and their environments is crucial. The aforementioned work have demonstrated that the benefit

of collaboration typically reduces the sample complexity by a factor of the number of collaborators, when compared to the single-agent setting where each system estimate its dynamics from its own observations.

However, the approaches discussed in [212, 229, 230] compute a common estimation for all participants, thereby restricting the ability to obtain personalized estimations. Furthermore, the sample complexity bounds achieved in those studies are subject to an unavoidable heterogeneity bias that cannot be controlled by the number of trajectories or systems, thus leading to an estimation error that scales with the measure of heterogeneity among the considered systems. Specifically in [212, 229, 230] the error of the system identification process is shown to be of order $\mathcal{O}(\frac{1}{\sqrt{N}} + \epsilon_{\text{het}})$ where ϵ_{het} characterizes the worst case heterogeneity and N is the number of trajectories across all systems.

Personalization in collaborative settings aims to provide tailored solutions (e.g. model estimates) to individual agents with distinct objectives, while enabling inter-agent collaboration (e.g. model sharing). This encompasses diverse topics such as representation learning [12, 30, 52, 246] and clustering [233], both widely studied in machine learning and data analysis. The present work address the aforementioned challenges by leveraging clustering techniques to achieve personalized model estimations. The idea is simple: cluster systems into groups that have identical system dynamics, and then apply collaborative learning algorithms to the clusters in order to improve sample complexity (by reducing the heterogeneity induced error ϵ_{het}) and achieve personalization even for heterogeneous settings.

Recent work on clustered federated learning that includes [60], [62], [170] have shown the potential of clustering techniques to collaboratively train models in heterogeneous settings with non-i.i.d. data. Building upon this success, this paper aims to apply clustering to the system identification problem, which poses unique challenges due to the dynamical nature of the system that results in non-i.i.d. data. This is in contrast to the linear regression and model training settings explored in the aforementioned work. Further details on these challenges are discussed later.

Specifically, we investigate the scenario where we have M dynamical systems, with each of them belonging to one of K different system types (which we refer to as a "cluster"). Which cluster a system belongs to is not initially disclosed. Our objective is to simultaneously identify the correct cluster identities for each of the M systems and obtain a system model by collaboratively learning with the systems in the same cluster. Our approach can lead to significant reductions in the amount of data required to accurately estimate the system models, as illustrated in the following theorem.

Theorem 11. *(main result, informal) Suppose the K system types are sufficiently different, and we observe*

the same number of trajectories from each system. Then, for a given cluster, with high probability, the estimation error between the learned and ground truth model is bounded by:

$$\textit{estimation error} \lesssim \frac{1}{\sqrt{\#\textit{ systems} \times \#\textit{ trajectories}}} + \textit{misclass. rate} ,$$

with

$$\textit{misclass. rate} \lesssim \exp(-\#\textit{ trajectories} \times \textit{misclass. const.}).$$

where #systems denotes the number of systems in the cluster, and #trajectories represents the number of trajectories observed by each of them.

The first term captures the error in learning the system dynamics from systems' observations within the same cluster. It shows what one would hope; as the number of systems and observations increase, the error decreases. However, this speedup does not come for free. The second term is the penalty paid for assigning one of the M systems to one of the incorrect K clusters. One of the main results from our work is to show that *both terms* can be controlled by adjusting the number of observed trajectories. Moreover, the misclassification rate is dominated by the first term, thus leading to a an approximate sample complexity that is scale inversely with the number of system within the cluster. This is in stark contrast to [212, 229, 230] which is where the heterogeneity introduces a bias ϵ which is not a function of the number of systems or the volume of data at our disposal. Our work shows that by controlling both sources of error, our approach can accurately estimate the system dynamics with fewer samples, when compared to the single agent case, and provides better estimation in heterogeneous settings when compared to [212, 229, 230].

Contributions: This is the first work to introduce clustering in order to provide sample complexity gains to the collaborative system identification problem. We derive an upper bound on the estimation error (Theorem 12) that decomposes into two terms (as shown above), where each term can be controlled by adjusting the number of observed trajectories. We offer theoretical guarantees on the probability of cluster identity misclassification (Lemma 43) and thus convergence (Corollary 3). We show that under a mild assumption on the number of observed trajectories, our approach correctly estimates the cluster identities, with high probability. Moreover, we show that our method achieves an improved convergence rate when compared to the single-agent system identification process. In contrast to the federated setting [27, 212] and that of [229, 230], we are able to provide personalized models as opposed to a single generic model, thus expanding the use cases for collaborative system identification.

Refer to [199] for all proofs in this Chapter.

7.1.1 Notation

Given a matrix $G \in \mathbb{R}^{m\times n}$, the Frobenius norm of G is denoted by $\|G\|_F = \sqrt{Tr(GG^\top)}$. $\|G\| = \sigma_{\max}(G)$, where $\sigma_{\max}(G)$ is the largest singular value of G. Consider a symmetric matrix Σ, $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote its minimum and maximum eigenvalues, respectively. For systems, we use superscript (i) to denote the system index and subscript t for time. For models, subscript denotes the cluster identity, and superscript (r) is the iteration counter.

7.2 Problem Formulation and Algorithm

Consider M linear time-invariant (LTI) systems

$$x_{t+1}^{(i)} = A^{(i)}x_t^{(i)} + B^{(i)}u_t^{(i)} + w_t^{(i)}, \;\; t = 0, 1, \ldots, T-1 \tag{7.1}$$

where $x_t^{(i)} \in \mathbb{R}^{n_x}$, $u_t^{(i)} \in \mathbb{R}^{n_u}$ and $w_t^{(i)} \in \mathbb{R}^{n_x}$ are the state, input, and process noise at time t, for system $i \in [M]$. We assume that $\{u_t^{(i)}\}_{t=1}^{T-1}, \{w_t^{(i)}\}_{t=1}^{T-1}$ are random vectors distributed according to $u_t^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_{u,i}^2 I_{n_u}\right)$ and $w_t^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_{w,i}^2 I_{n_x}\right)$. Furthermore, it is assumed that $x_0^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_{x,i}^2 I_{n_x}\right)$.

We consider the setting where we have access to M datasets corresponding to observed system trajectories. Each of the datasets is generated by one of K different systems. We consider the case where $K \ll M$. We will from now on refer to the K types of different systems as "clusters", which we label as $\mathcal{C}_1, \ldots, \mathcal{C}_K$. We denote (A_j, B_j) as the ground truth system matrices of cluster $j \in [K]$. That is, $A^{(i)} = A_j$, and $B^{(i)} = B_j$, for any $i \in \mathcal{C}_j$. Note that due to the noise in model (7.1), two datasets generated by cluster $\mathcal{C}_j$ will be different.

The state-input pair of a single trajectory $\{x_t^{(i)}, u_t^{(i)}\}$ of system $i \in \mathcal{C}_j$ is referred to as *rollout*. We consider the setting where multiple rollouts of length T are collected and stored as $\left\{x_{l,t}^{(i)}, u_{l,t}^{(i)}\right\}_{t=0}^{T-1}$, for $l = 1, \ldots N_i$, with l denoting the l-th rollout and t the t-th time-step of the corresponding rollout. Thus, for any system $i \in \mathcal{C}_j$ and cluster $j \in [K]$, the system dynamics is described by:

$$x_{l,t+1}^{(i)} = \Theta_j z_{l,t}^{(i)} + w_{l,t}^{(i)} \quad \forall\, 1 \le l \le N_i \text{ and } 0 \le t \le T-1, \tag{7.2}$$

where $z_{l,t}^{(i),\top} \triangleq \begin{bmatrix} x_{l,t}^{(i),\top} & u_{l,t}^{(i),\top} \end{bmatrix} \in \mathbb{R}^{n_x+n_u}$ corresponds to the augmented state-input pair of system $i \in \mathcal{C}_j$ over rollout l at time t, and $\Theta_j \triangleq [A_j \;\; B_j]$ denotes the concatenation of the ground truth system matrices A_j

and B_j. The state update $x_{l,t+1}^{(i)}$ can be expanded recursively as follows:

$$x_{l,t}^{(i)} = G_t^{(i)} \begin{bmatrix} u_{l,0}^{(i)} \\ \vdots \\ u_{l,t-1}^{(i)} \end{bmatrix} + F_t^{(i)} \begin{bmatrix} w_{l,0}^{(i)} \\ \vdots \\ w_{l,t-1}^{(i)} \end{bmatrix} + A_j^t x_{l,0}^{(i)},$$

where, $G_t^{(i)} \triangleq \begin{bmatrix} A_j^{t-1}B_j & A_j^{t-2}B_j & \cdots & B_j \end{bmatrix}$ and $F_t \triangleq \begin{bmatrix} A_j^{t-1} & A_j^{t-2} & \cdots & I_{n_x} \end{bmatrix}$ for all $t \geq 1$.

The state-input pair $z_{l,t}^{(i)}$ is distributed according to a Gaussian distribution with zero mean and covariance matrix $\Sigma_t^{(i)}$, where,

$$\Sigma_0^{(i)} \triangleq \begin{bmatrix} \sigma_{x,i}^2 I_{n_x} & 0 \\ 0 & \sigma_{u,i}^2 I_{n_u} \end{bmatrix} \succ 0, \quad \text{for } t = 0,$$

and

$$\Sigma_t^{(i)} \triangleq \begin{bmatrix} \sigma_{u,i}^2 G_t^{(i)} G_t^{(i),\top} + \sigma_{w,i}^2 F_t^{(i)} F_t^{(i),\top} + \sigma_{x,i}^2 A_j^t (A_j^t)^\top & 0 \\ 0 & \sigma_{u,i}^2 I_{n_u} \end{bmatrix},$$

for all $t \geq 1$ and $i \in \mathcal{C}_j$, $\forall j \in [K]$, as detailed in [212].

Next, we define the offline batch matrices for each system $i \in \mathcal{C}_j$, $\forall j \in [K]$. For a single rollout l, the data is concatenated according to $X_l^{(i)} = \begin{bmatrix} x_{l,T}^{(i)} & \cdots & x_{l,1}^{(i)} \end{bmatrix} \in \mathbb{R}^{n_x \times T}$, $Z_l^{(i)} = \begin{bmatrix} z_{l,T-1}^{(i)} & \cdots & z_{l,0}^{(i)} \end{bmatrix} \in \mathbb{R}^{(n_x+n_u)\times T}$, and $W_l^{(i)} = \begin{bmatrix} w_{l,T-1}^{(i)} & \cdots & w_{l,0}^{(i)} \end{bmatrix} \in \mathbb{R}^{n_x \times T}$. This is then further stacked to construct the batch matrices

$$X^{(i)} = \begin{bmatrix} X_1^{(i)} & \cdots & X_{N_i}^{(i)} \end{bmatrix} \in \mathbb{R}^{n_x \times N_i T}, \quad Z^{(i)} = \begin{bmatrix} Z_1^{(i)} & \cdots & Z_{N_i}^{(i)} \end{bmatrix} \in \mathbb{R}^{(n_x+n_u)\times N_i T},$$

and $W^{(i)} = \begin{bmatrix} W_1^{(i)} & \cdots & W_{N_i}^{(i)} \end{bmatrix} \in \mathbb{R}^{n_x \times N_i T}$. Therefore, for each system $i \in \mathcal{C}_j$, $\forall j \in [K]$, its state, input, noise, and model parameters are related according to

$$X^{(i)} = \Theta_j Z^{(i)} + W^{(i)}, \tag{7.3}$$

where each column of $Z^{(i)}$ and $W^{(i)}$ are sampled according to Gaussian distributions with zero means and covariance matrices $\Sigma_t^{(i)}$, $\sigma_{w,i}^2 I_{n_x}$, respectively. With that said, we are now able to introduce the clustered system identification problem.

Problem 1. *We consider M dynamical systems as in* (7.1) *that are equipped with batch matrices $X^{(i)}, Z^{(i)}$, and $W^{(i)}$. Each system $i \in [M]$ is associated with its own cost function $C^{(i)}(\Theta) = \|X^{(i)} - \Theta Z^{(i)}\|_F^2$, and is unaware of its cluster identity. We aim to estimate the systems' cluster identities $\widehat{\mathcal{C}}_1, \ldots, \widehat{\mathcal{C}}_K$ and use it to estimate a model $\widehat{\Theta}_j = [\widehat{A}_j \ \ \widehat{B}_j]$ which is close to the ground truth Θ_j, $\forall j \in [K]$.*

To obtain a faster and more accurate estimation, we frame the system identification problem in the setting where systems within the same cluster can leverage data from each other. Further in this paper, we provide theoretical guarantees to support these statements.

The problem described above can be framed into an alternating optimization problem, as the actual cluster identity of each system (i.e., $\mathcal{C}_1, \ldots, \mathcal{C}_K$) is not disclosed to the systems in advance. Therefore, our objective is twofold: firstly, we aim to classify the correct cluster identities of the systems by employing the Mean Square Error (MSE) as the clustering criterion, with the resulting output being the cluster estimation (CE); secondly, we use that estimation to identify the model dynamics of each cluster with a model estimation (ME) step. Next, we introduce our clustered system identification algorithm to solve this problem.

Algorithm 9 `Clustered System Identification`

1: **Initialization:** number of clusters K, step-size η_j, and model initialization $\widehat{\Theta}_j^{(0)}$ $\forall j \in [K]$,
2: **for** each iteration $r = 0, 1, \ldots, R-1$ **do**
3: The systems receive the models $\{\widehat{\Theta}_1^{(r)}, \ldots, \widehat{\Theta}_K^{(r)}\}$, $\forall j \in [K]$,
4: **Cluster estimation (CE):**
5: **for** each system $i \in [M]$
6: $\hat{j} = \text{argmin}_{j \in [K]} \|X^{(i)} - \widehat{\Theta}_j^{(r)} Z^{(i)}\|_F^2$,
7: define $e_i = \{e_{i,j}\}_{j=1}^K$ with $e_{i,j} = \mathbb{1}\{j = \widehat{j}\}$,
8: **end for**
9: **Model estimation (ME):**
10: $\widehat{\Theta}_j^{(r+1)} = \widehat{\Theta}_j^{(r)} + \frac{2\eta_j}{\sum_{i \in [M]} e_{i,j}} \sum_{i \in [M]} e_{i,j} (X^{(i)} - \widehat{\Theta}_j^{(r)} Z^{(i)}) Z^{(i),\top}$ for all $j \in [K]$
11: **end for**
12: **Return** $\widehat{\Theta}_j^{(R)}$ for all $j \in [K]$.

The initial step of Algorithm 9 involves the initialization of the number of clusters and the provision of an initial guess for the dynamics of each cluster. Subsequently, the algorithm iterates from line 2 to 11, during which each system estimates its corresponding cluster identity and stores this information in the form of a one-hot encoding vector denoted by e_i. The one-hot encoding vector comprises K elements, with one in the position of the estimated cluster identity and zero elsewhere. After the estimation of the cluster identity, the cluster model is updated by performing a single gradient descent iteration in line 10, with the gradient being

the average of the gradients of each individual system's cost function that belongs to the cluster.

Remark 5. *Note that Algorithm 9 is an alternating minimization algorithm, where it performs an iterative clustering step followed by a model estimation process. Prior to the start of collaboration, each system $i \in [M]$ collects data and stores it in batch matrices $X^{(i)}, Z^{(i)}$, and $W^{(i)}$. Moreover, it is worth noting that Algorithm 9 uses the same batch matrices for both cluster identity and model estimation.*

The following definitions and assumptions are required in order to analyze Algorithm 1. Subsequently, we provide the intuition behind them.

Definition 3. *The minimum and maximum separation between the clusters are defined as*

$$\Delta_{min} \triangleq \min_{j \neq j'} \|\Theta_j - \Theta_{j'}\| \quad \textit{and} \quad \Delta_{max} \triangleq \max_{j \neq j'} \|\Theta_j - \Theta_{j'}\|,$$

respectively.

We define $\rho^{(i)} \triangleq \frac{\Delta_{\min}^2}{\sigma_{w,i}^2}$ as the signal-to-noise ratio $\forall i \in [M]$.

Assumption 9. *The initial model estimate $\widehat{\Theta}_j^{(0)}$ satisfy $\|\widehat{\Theta}_j^{(0)} - \Theta_j\| \leq \left(\frac{1}{2} - \alpha^{(0)}\right) \Delta_{\min}, \forall j \in [K]$, where $0 < \alpha^{(0)} < \frac{1}{2}$.*

Assumption 10. *For any fixed and small δ, the number of trajectories satisfies $N_i n_x \gtrsim \left(\frac{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}{\alpha^{(0)}\rho^{(i)}\|\Sigma_t^{(i)}\|}\right)^2 \log(\frac{MT}{\delta})$, for all $i \in [M]$. We also assume that*

$$\Delta_{\min} \gtrsim 1 + \Delta_{\max} \sum_{i \in [M]} \sum_{t=0}^{T-1} \exp\left(-cN_i n_x \left(\frac{\alpha^{(0)}\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right)$$

for some constant c.

Assumption 9 implies that the initial guess for the model estimates is superior to a random initialization. This assumption is standard for alternating minimization algorithms, particularly for learning mixture models [9]. The condition on the number of trajectories in Assumption 10 is a common requirement in the concentration bound analysis. This is used to guarantee that the cluster estimation procedure of Algorithm 9 correctly estimate the cluster identities, with high probability. Note that this is a mild assumption since for well-behaved systems where $\Sigma_t^{(i)}$ is well conditioned, $N_i n_x$ is typically in the same or superior to the order of $\log\left(\frac{MT}{\delta}\right)$. The condition on $\Delta_{\min}$ in Assumption 10 is to ensure that any two clusters are well-separated. This is a standard assumption in the literature of clustering [43, 99]. Similar assumptions are exploited in [60] in the context of the linear regression problem.

7.3 Theoretical Guarantees

We begin our analysis by examining a single iteration of Algorithm 9. For simplicity, we omit the superscript r that denotes the iteration counter. Let us assume that we have the current estimated model $\widehat{\Theta}_j$ for all clusters $j \in [K]$ at a given iteration, such that $\|\widehat{\Theta}_j - \Theta_j\| \leq \left(\frac{1}{2} - \alpha\right) \Delta_{\min}$ for all $j \in [K]$, with $0 < \alpha < \frac{1}{2}$.

7.3.1 Probability of Cluster Identity Misclassification

Consider a system $i \in [M]$ within cluster $\mathcal{C}_j$. Let $\mathcal{M}_i^{j,j'}$ be the event in which system i is inaccurately classified as belonging to cluster $\mathcal{C}_{j'}$. The event when system i is *correctly* classified is denoted as $\mathcal{M}_i^{j,j}$. The following lemma provides an upper bound on the probability of misclassification.

Lemma 43. *Suppose that $i \in \mathcal{C}_j$. There exist universal constants c_1 and c_2, such that for any $j' \neq j$,*

$$\mathbb{P}\left\{\mathcal{M}_i^{j,j'}\right\} \leq c_1 \sum_{t=0}^{T-1} \exp\left(-c_2 N_i n_x \left(\frac{\alpha \rho^{(i)} \|\Sigma_t^{(i)}\|}{\rho^{(i)} \|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right).$$

By combining Lemma 43 with the condition on $N_i n_x$ from Assumption 10, our algorithm can ensure that the probability of misclassifying system i to cluster $\mathcal{C}_{j'}$ is at most δ, where δ can be arbitrarily small. Moreover, it is noteworthy that if we assume the data $X^{(i)}$, $Z^{(i)}$, and $W^{(i)}$ to be i.i.d. with $T = 1$ and $n_x = 1$, and the columns of $Z^{(i)}$ to have an identity covariance matrix, we can recover the probability of misclassification in the linear regression problem, as discussed in [60].

7.3.2 Convergence Analysis

We now examine the convergence of Algorithm 9. The theorem below is a single-iteration convergence analysis of our algorithm. Here we assume that, at a given iteration, an estimation $\widehat{\Theta}_j$ is obtained, which closely approximates the true model Θ_j, i.e., $\|\widehat{\Theta}_j - \Theta_j\| \leq \left(\frac{1}{2} - \alpha\right) \Delta_{\min}$, $\forall j \in [K]$ and $0 < \alpha < \frac{1}{2}$. We demonstrate that $\widehat{\Theta}_j$ converges to Θ_j up to a small bias.

Theorem 12. *For any fixed $0 < \delta < 1$, with*

$$N_i \geq \max\left\{8(n_x + n_u) + 16 \log \frac{2MT}{\delta}, (4n_x + 2n_u) \log \frac{MT}{\delta}\right\},$$

$\forall i \in [M]$, and selected step-size $\eta_j = \frac{|\mathcal{C}_j|}{\lambda_{\min}\left(\sum_{i \in \mathcal{C}_j} N_i \sum_{t=0}^{T-1} \Sigma_t^{(i)}\right)}$, with probability at least $1 - 3\delta$, it holds that,

$$\|\widehat{\Theta}_j^+ - \Theta_j\| \leq \frac{1}{2}\|\widehat{\Theta}_j - \Theta_j\| + \bar{c}_0 \times \frac{1}{\sqrt{\sum_{i \in \widehat{\mathcal{C}}_j} N_i}} \tag{7.4}$$

$$+ \bar{c}_1 \Delta_{\max} \sum_{i \in [M]} \sum_{t=0}^{T-1} \exp\left(-\bar{c}_2 N_i n_x \left(\frac{\alpha \rho^{(i)} \|\Sigma_t^{(i)}\|}{\rho^{(i)} \|\Sigma_t^{(i)}\| + \sqrt{n_x}} \right)^2 \right), \tag{7.5}$$

for all $j \in [K]$, where $\bar{c}_0$, $\bar{c}_1$, $\bar{c}_2 > 0$ are problem dependent constants.

This theorem provides an upper bound for the estimation error per iteration of our algorithm. Specifically, this bound consists of three terms. The first term is a contraction term that decreases to zero as the number of iterations increases. The second term is a constant error that decreases as the total number of observed trajectories by the systems within the cluster increases. The final term is the misclassification rate, which decays exponentially with the number of observed trajectories.

Note that although our setting is different from [60], which leads to a different estimation error expression, our per-iteration estimation error is also composed of a contractive term added to a constant error that can be controlled by the amount of data (i.e., the number of observed trajectories). We proceed to show the convergence of our algorithm by demonstrating that $\alpha^{(r)}$ is non-decreasing throughout iterations and using Assumptions 9 and 10 to show that $\|\widehat{\Theta}_j^{(r+1)} - \Theta_j\| \leq \|\widehat{\Theta}_j^{(r)} - \Theta_j\|$ for all $r \in [R]$.

Therefore, equipped with the aforementioned result, the following corollary characterizes the convergence of Algorithm 9 by providing the number of iterations required to attain a certain small and near optimal error ϵ, i.e., $\|\widehat{\Theta}_j^{(R)} - \Theta_j\| \leq \epsilon$, for all clusters $j \in [K]$.

Corollary 3. *Frame the hypotheses of Theorem 12 and Assumptions 9 and 10. Select the step-size as $\eta_j = \frac{|\mathcal{C}_j|}{\lambda_{\min}\left(\sum_{i \in \mathcal{C}_j} N_i \sum_{t=0}^{T-1} \Sigma_t^{(i)}\right)}$ for all $j \in [K]$. Then, after $R \geq 2 + \log(\frac{\Delta_{\min}}{4\epsilon})$ parallel iterations, we have $\|\widehat{\Theta}_j^{(R)} - \Theta_j\| \leq \epsilon$, with*

$$\epsilon = \tilde{c}_0 \times \frac{1}{\sqrt{\sum_{i \in \mathcal{C}_j} N_i}} + \tilde{c}_1 \Delta_{\max} \sum_{i \in [M]} \sum_{t=0}^{T-1} \exp\left(-\tilde{c}_2 N_i n_x \left(\frac{\rho^{(i)} \|\Sigma_t^{(i)}\|}{\rho^{(i)} \|\Sigma_t^{(i)}\| + \sqrt{n_x}} \right)^2 \right), \tag{7.6}$$

for all $j \in [K]$, where $\tilde{c}_0$, $\tilde{c}_1$, $\tilde{c}_2 > 0$ are problem dependent constants.

Our proof builds upon similar arguments as in [60], which considers the linear regression setting. To establish the non-decreasing property of $\alpha^{(r)}$ for all $r \in [R]$ and a decrease in the additive error term over the iterations, we rely on Assumptions 9 and 10. Furthermore, we demonstrate that our algorithm achieves a sufficiently large value of $\alpha^{(r)} \geq \frac{1}{4}$ after only a small number of iterations $R \geq 2$. This indicates that after a suitable number of iterations, our Algorithm 9 produces an estimation error that scales down with the number of systems within the cluster, and is independent of the initial closeness parameter $\alpha^{(0)}$.

This corollary highlights the benefits of collaboration. It demonstrates that the estimation error scales inversely with the number of agents within a cluster, implying that as the number of systems in the cluster increases, this error decreases. This leads to a smaller error when compared to the single agent setting, where each system estimates its dynamics using *only* its own observations.

Importantly, the presented error bound differs from that of [212]. Here the misclassification rate exponentially decays with the number of observed trajectories, whereas the heterogeneity bias ϵ_{het} in [212] cannot be controlled by the number of trajectories. This indicates that under heterogeneous settings where the systems are significantly different, our clustering-based approach outperforms [212] by providing better control over the sources of error. However, it is worth mentioning that when the systems are similar and personalization is not required, the approaches introduced in [212, 229, 230] may be more favorable as their error bounds scale down with the total number of systems and do not necessitate a clustering step.

7.4 Numerical Results

The following simulations[1] illustrate the efficiency of Algorithm 9. Our analysis considers $M = 50$ systems, each described by an LTI model as in (7.1) where $K = 3$ clusters and the number of systems in each cluster is $|\mathcal{C}_1|= 10$, $|\mathcal{C}_2|= 24$, and $|\mathcal{C}_3|= 16$. The systems matrices for each cluster are described as follows:

$$A_1 = \begin{bmatrix} 0.5 & 0.3 & 0.1 \\ 0.0 & 0.2 & 0.0 \\ 0.1 & 0.0 & 0.3 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -0.3 & 0.0 & 0.0 \\ 0.1 & 0.4 & 0.0 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}, \quad A_3 = \begin{bmatrix} -0.1 & 0.1 & 0.1 \\ 0.1 & 0.15 & 0.1 \\ 0.1 & 0.0 & 0.2 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 1 & 0.5 \\ 0.1 & 1 \\ 0.75 & 1.5 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 & 0.5 \\ 0.1 & 1 \\ 0.75 & 1.5 \end{bmatrix}, \quad B_3 = \begin{bmatrix} 0.8 & 0.1 \\ 0.1 & 1.5 \\ 0.4 & 0.8 \end{bmatrix},$$

where the initial state, input, and process noise standard deviations, for each cluster, are set to $\sigma_{x,i} = \sigma_{u,i} = \sigma_{w,i} = 0.11$, $\forall i \in \mathcal{C}_1$, $\sigma_{x,i} = \sigma_{u,i} = \sigma_{w,i} = 0.12$, $\forall i \in \mathcal{C}_2$, and $\sigma_{x,i} = \sigma_{u,i} = \sigma_{w,i} = 0.05$, $\forall i \in \mathcal{C}_3$. We consider the same number of trajectories $N_i = 100$ for all $i \in [M]$. Moreover, the trajectory length is set to $T = 50$. We use a fixed step-size $\eta_j = 10^{-3}$, $\forall j \in [3]$. For each iteration r, the estimation error $e_r^{(j)}$ is defined as the spectral norm distance between the estimated model $\widehat{\Theta}_j^{(r)}$ and the ground truth model Θ_j, i.e.,

[1] Code can be downloaded from `https://github.com/jd-anderson/cluster-sysID`

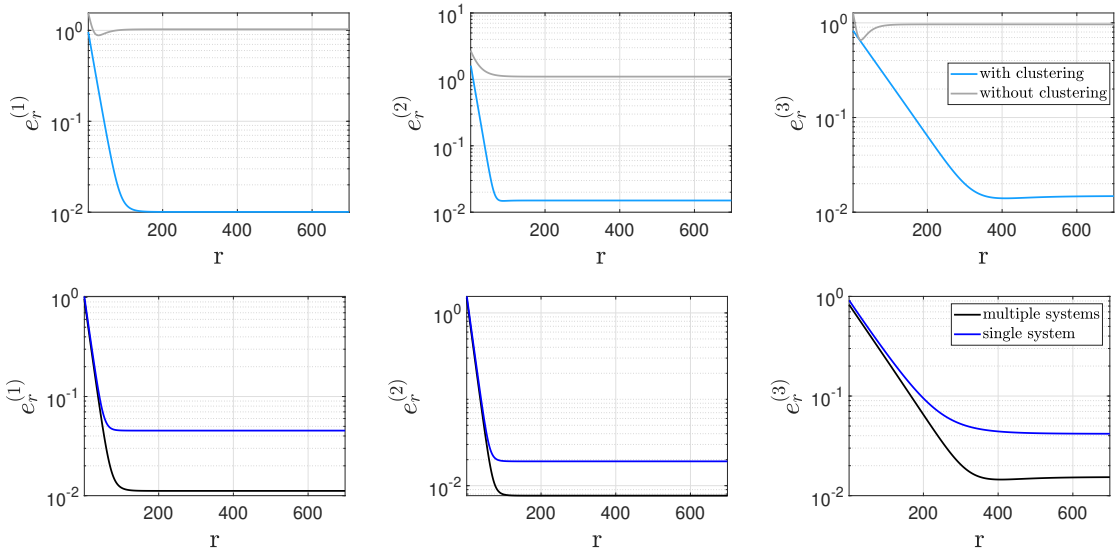


Figure 7.1: Estimation error as a function of iteration count. The plot on the top considers Algorithm 9 with and without clustering, whereas the bottom plot consider the single and multiple agents settings.

$e_r^{(j)} = \|\widehat{\Theta}_j^{(r)} - \Theta_j\|$, for all clusters $j \in [K]$.

Figure 7.1 depicts the estimation error $e_r^{(j)}$ as a function of the number of iterations r for all the three considered clusters. The top plots compare the performance of Algorithm 9 with and without the clustering procedure (i.e., line 5 of Algorithm 9). These plots reveals that the estimation error decreases significantly when systems with the same model are clustered and cooperate to estimate their dynamics. Conversely, in the absence of clustering, the significant heterogeneity level across the systems leads to a poor common estimation, resulting in a large estimation error and unpersonalized solutions. This confirms our theoretical results, showing that the misclassification rate in (7.6) outperforms the heterogeneity constant of [212, 229, 230], when dealing with heterogeneous settings.

The bottom plots of Figure 7.1 demonstrates the benefits of collaboration among systems to learn their dynamics. This shows that the estimation error is considerably reduced when multiple systems within the same cluster (i.e., $|\mathcal{C}_1| = 10$, $|\mathcal{C}_2| = 24$, and $|\mathcal{C}_3| = 16$) leverage the data from each other to identify their dynamics, compared to the case where a single system estimate its dynamics by using its own observations. This also confirms our theoretical results, where the statistical error in (7.6) scales down with the number of systems in the cluster, thus highlighting the benefit of collaboration in improving estimation accuracy in a multi-system setting.

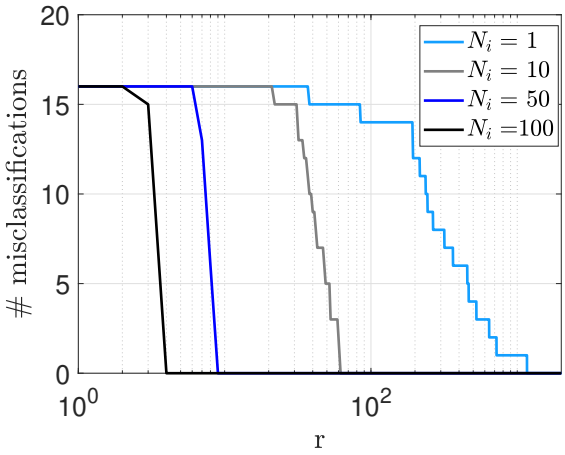


Figure 7.2: Number of misclassification as a function of iteration count.

Figure 7.2 presents the misclassifications of Algorithm 9 as a function of iterations r. It depicts the number of systems whose cluster identity is incorrectly estimated. The figure illustrates the effect of the number of observed trajectories on the misclassification rate. As anticipated and consistent with our theoretical results, an increase in the number of trajectories leads to a considerable reduction in the number of iterations needed to correctly classify all the systems into their respective clusters.

7.5 Chapter Summary and Future Work

We presented an approach to address the system identification problem through the use of clustering. Our method involves partitioning different systems that observe multiple trajectories into disjoint clusters based on the similarity of their dynamics. This approach enjoys an improved convergence rate that scales inversely with the number of systems in the cluster, along with an additive misclassification rate that has been shown to be negligible under mild assumptions. Our approach enables systems within the same cluster to learn their dynamics more efficiently. Future work will involve extending the proposed formulation to online system identification and proposing an adaptive clustering approach that eliminates the necessity for the warm initialization and well-separated clusters assumptions.

7.6 Omitted Proofs

7.6.1 Proof of Lemma 43

Without loss of generality, we can analyze only the first cluster $\mathcal{M}_i^{1,j}$ for some $j \neq 1$. By definition, we have

$$\mathcal{M}_i^{1,j} = \left\{ \|X^{(i)} - \widehat{\Theta}_j Z^{(i)}\|_F^2 \leq \|X^{(i)} - \widehat{\Theta}_1 Z^{(i)}\|_F^2 \right\}$$

where the batch matrices $X^{(i)}, Z^{(i)}$ and $W^{(i)}$ are related according to $X^{(i)} = \Theta_1 Z^{(i)} + W^{(i)}$. Note that $z_{l,t}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \Sigma_t^{(i)}\right)$ and $w_{l,t}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_{w,i}^2 I_{n_x}\right)$ are independent across trajectories (i.e., the columns of $Z^{(i)}$ and $W^{(i)}$ are independent). Thus, we can write

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{M}_i^{1,j}\right\} &= \mathbb{P}\left\{ \left\|(\Theta_1 - \widehat{\Theta}_1)Z^{(i)} + W^{(i)}\right\|_F^2 \geq \left\|(\Theta_1 - \widehat{\Theta}_j)Z^{(i)} + W^{(i)}\right\|_F^2 \right\} \\
&= \mathbb{P}\left\{ \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} m_{l,t}^{(i),\top} m_{l,t}^{(i)} \geq \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} n_{l,t}^{(i),\top} n_{l,t}^{(i)} \right\},
\end{aligned}$$

where $m_{l,t}^{(i)} = (\Theta_1 - \widehat{\Theta}_1) z_{l,t}^{(i)} + w_{l,t}^{(i)} \sim \mathcal{N}\left(0, \bar{\Sigma}_t^{(i)}\right)$, $n_{l,t}^{(i)} = (\Theta_1 - \widehat{\Theta}_j) z_{l,t}^{(i)} + w_{l,t}^{(i)} \sim \mathcal{N}\left(0, \tilde{\Sigma}_t^{(i)}\right)$, with

$$\bar{\Sigma}_t^{(i)} = (\Theta_1 - \widehat{\Theta}_1)\Sigma_t^{(i)}(\Theta_1 - \widehat{\Theta}_1)^\top + \sigma_{w,i}^2 I_{n_x}, \;\; \tilde{\Sigma}_t^{(i)} = (\Theta_1 - \widehat{\Theta}_j)\Sigma_t^{(i)}(\Theta_1 - \widehat{\Theta}_j)^\top + \sigma_{w,i}^2 I_{n_x}.$$

Therefore, we obtain

$$\mathbb{P}\left\{\mathcal{M}_i^{1,j}\right\} = \mathbb{P}\left\{ \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top} \bar{\Sigma}_t^{(i)} v_{l,t}^{(i)} \geq \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} u_{l,t}^{(i),\top} \tilde{\Sigma}_t^{(i)} u_{l,t}^{(i)} \right\},$$

with $m_{l,t}^{(i)} = (\bar{\Sigma}_t^{(i)})^{\frac{1}{2}} v_{l,t}^{(i)}$ and $n_{l,t}^{(i)} = (\bar{\Sigma}_t^{(i)})^{\frac{1}{2}} u_{l,t}^{(i)}$ for some standard normal random vectors $v_{l,t}^{(i)}$, $u_{l,t}^{(i)} \sim \mathcal{N}(0, I_{n_x})$. Then, the above expression can be rewritten as follows

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{M}_i^{1,j}\right\} &= \mathbb{P}\left\{ \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top} \bar{\Sigma}_t^{(i)} v_{l,t}^{(i)} \geq \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} \|\tilde{\Sigma}_t^{(i)}\| u_{l,t}^{(i),\top} u_{l,t}^{(i)} \right\} \\
&= \mathbb{P}\left\{ \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top} \bar{\Sigma}_t^{(i)} v_{l,t}^{(i)} \geq \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} c_t^{(i)} u_{l,t}^{(i),\top} u_{l,t}^{(i)} \right\}
\end{aligned}$$

with $c_t^{(i)} = \|\Theta_1 - \widehat{\Theta}_j\|^2 \|\Sigma_t^{(i)}\| + \sigma_{w,i}^2 \sqrt{n_x}$, which implies

$$\mathbb{P}\left\{\mathcal{M}_i^{1,j}\right\} = \mathbb{P}\left\{ \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top} \bar{\Sigma}_t^{(i)} v_{l,t}^{(i)} \geq \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} c_t^{(i)} u_{l,t}^{(i),\top} u_{l,t}^{(i)} \right\}$$

$$\leq \mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{l=1}^{N_i} c_t^{(i)} u_{l,t}^{(i),\top} u_{l,t}^{(i)} \leq \bar{t}\right\} + \mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top} \bar{\Sigma}_t^{(i)} v_{l,t}^{(i)} > \bar{t}\right\},$$

for any $\bar{t} \geq 0$. Therefore , by using $v_{l,t}^{(i),\top} \bar{\Sigma}_t^{(i)} v_{l,t}^{(i)} \leq d_t^{(i)} v_{l,t}^{(i),\top} v_{l,t}^{(i)}$ with $d_t^{(i)} = \|\Theta_1 - \widehat{\Theta}_1\|^2 \|\Sigma_t^{(i)}\| + \sigma_{w,i}^2 \sqrt{n_x}$ we obtain

$$\mathbb{P}\left\{\mathcal{M}_i^{1,j}\right\} \leq \mathbb{P}\left\{\sum_{t=0}^{T-1} c_t^{(i)} V_t^{(i)} \leq \bar{t}\right\} + \mathbb{P}\left\{\sum_{t=0}^{T-1} d_t^{(i)} V_t^{(i)} > \bar{t}\right\},$$

where $V_t^{(i)}$ are standard Chi-squared distributions with $N_i n_x$ degrees of freedom, for all $t \in \{0, 1, \ldots, T-1\}$. Moreover, by using Definition 3 and Assumption 9,

$$\mathbb{P}\left\{\mathcal{M}_i^{1,j}\right\} \leq \mathbb{P}\left\{\sum_{t=0}^{T-1} f_t^{(i)} V_t^{(i)} \leq \bar{t}\right\} + \mathbb{P}\left\{\sum_{t=0}^{T-1} g_t^{(i)} V_t^{(i)} > \bar{t}\right\},$$

with $f_t^{(i)} = (\frac{1}{2} + \alpha)^2 \Delta_{\min}^2 \|\Sigma_t^{(i)}\| + \sigma_{w,i}^2 \sqrt{n_x}$ and $g_t^{(i)} = (\frac{1}{2} - \alpha)^2 \Delta_{\min}^2 \|\Sigma_t^{(i)}\| + \sigma_{w,i}^2 \sqrt{n_x}$, since $c_t^{(i)} = \|\Theta_1 - \widehat{\Theta}_j\|^2 \|\Sigma_t^{(i)}\| + \sigma_{w,i}^2 \sqrt{n_x} \geq (\frac{1}{2} + \alpha)^2 \Delta_{\min}^2 \|\Sigma_t^{(i)}\| + \sigma_{w,i}^2 \sqrt{n_x}$, with $\|\Theta_j - \widehat{\Theta}_1\| \geq \|\Theta_j - \Theta_1\| - \|\widehat{\Theta}_j - \Theta_j\| = (\frac{1}{2} + \alpha)\Delta_{\min}$ and $d_t^{(i)} = \|\Theta_1 - \widehat{\Theta}_1\|^2 \|\Sigma_t^{(i)}\| + \sigma_{w,i}^2 \sqrt{n_x} \leq (\frac{1}{2} - \alpha)^2 \Delta_{\min}^2 \|\Sigma_t^{(i)}\| + \sigma_{w,i}^2 \sqrt{n_x}$, where $\|\Theta_1 - \widehat{\Theta}_1\| \leq (\frac{1}{2} - \alpha)\Delta_{\min}$ according to Assumption 9. Therefore, to characterize the above tail bounds, we can exploit well-established concentration inequalities as detailed in [17, 207]. To this end, we can use union bound to write

$$\mathbb{P}\left\{\mathcal{M}_i^{1,j}\right\} \leq \sum_{t=0}^{T-1} \mathbb{P}\left\{f_t^{(i)} V_t^{(i)} \leq \bar{t}\right\} + \mathbb{P}\left\{g_t^{(i)} V_t^{(i)} > \bar{t}\right\},$$

where $\mathbb{P}\left\{f_t^{(i)} V_t^{(i)} \leq \bar{t}\right\}$ can be rewritten as follows

$$\mathbb{P}\left\{f_t^{(i)} V_t^{(i)} \leq \bar{t}\right\} = \mathbb{P}\left\{V_t^{(i)} \leq \frac{4\bar{t}}{\sigma_{w,i}^2 \sqrt{n_x}\left((1+2\alpha)^2 \rho^{(i)} \frac{\|\Sigma_t^{(i)}\|}{\sqrt{n_x}} + 4\right)}\right\},$$

thus, by choosing $\bar{t} = N_i n_x \left((\frac{1}{4} + \alpha^2)\Delta_{\min}^2 \|\Sigma_t^{(i)}\| + \sigma_{w,i}^2 \sqrt{n_x}\right)$ we obtain

$$\mathbb{P}\left\{f_t^{(i)} V_t^{(i)} \leq \bar{t}\right\} = \mathbb{P}\left\{\frac{V_t^{(i)}}{N_i n_x} - 1 \leq \frac{-4\alpha \|\Sigma_t^{(i)}\|}{(1+2\alpha)^2 \rho^{(i)} \|\Sigma_t^{(i)}\| + 4\sqrt{n_x}}\right\},$$

as per the concentration of standard Chi-squared distributions in [208], it is established that there exist universal constants c_1 and c_2, such that

$$\mathbb{P}\left\{f_t^{(i)} V_t^{(i)} \leq \bar{t}\right\} \leq c_1 \exp\left(-c_2 N_i n_x \left(\frac{\alpha \rho^{(i)} \|\Sigma_t^{(i)}\|}{\rho^{(i)} \|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right). \tag{7.7}$$

Similarly, $\mathbb{P}\left\{g_t^{(i)}V_t^{(i)} > \bar{t}\right\}$ can be rewritten as follows

$$\mathbb{P}\left\{g_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} = \mathbb{P}\left\{\frac{V_t^{(i)}}{N_i n_x} - 1 \leq \frac{4\alpha\|\Sigma_t^{(i)}\|}{(1-2\alpha)^2\rho^{(i)}\|\Sigma_t^{(i)}\|+4\sqrt{n_x}}\right\},$$

and by the concentration of Chi-squared distribution

$$\mathbb{P}\left\{g_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} \leq c_3 \exp\left(-c_4 N_i n_x \left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\|+\sqrt{n_x}}\right)^2\right), \tag{7.8}$$

where the proof is completed by combining (7.7) and (7.8) to obtain

$$\mathbb{P}\left\{\mathcal{M}_i^{1,j}\right\} \leq c_1 \sum_{t=0}^{T-1} \exp\left(-c_2 N_i n_x \left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\|+\sqrt{n_x}}\right)^2\right).$$

7.6.2 Proof of Theorem 12

Without loss of generality, we analyze only the first cluster. Recall that the model is updated as follows:

$$\widehat{\Theta}_1^+ = \frac{1}{|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1}\tilde{\Theta}_i = \frac{1}{|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1\cap\mathcal{S}_1}\tilde{\Theta}_i + \frac{1}{|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1\cap\overline{\mathcal{S}_1}}\tilde{\Theta}_i \tag{7.9}$$

with $\tilde{\Theta}_i = \widehat{\Theta}_1 + 2\eta_1(X^{(i)} - \widehat{\Theta}_1 Z^{(i)})Z^{(i),\top}$. Here $\widehat{\mathcal{C}}_1 \cap \mathcal{C}_1$ corresponds to the set of systems correctly classified to the first cluster and $\widehat{\mathcal{C}}_1 \cap \overline{\mathcal{C}_1}$ represents the set of systems that are misclassified to the first cluster, with $\overline{\mathcal{C}_1}$ denoting the complement of $\mathcal{C}_1$. The above expression can be rewritten as follows

$$\widehat{\Theta}_1^+ = \widehat{\Theta}_1 + \frac{2\eta_1}{|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1\cap\mathcal{C}_1}(X^{(i)} - \widehat{\Theta}_1 Z^{(i)})Z^{(i),\top} + \frac{2\eta_1}{|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}}(X^{(i)} - \widehat{\Theta}_1 Z^{(i)})Z^{(i),\top},$$

where $X^{(i)} = \Theta_1 Z^{(i)} + W^{(i)}$ for $i \in \widehat{\mathcal{C}}_1 \cap \mathcal{C}_1$, and $X^{(i)} = \Theta_j Z^{(i)} + W^{(i)}$ for $i \in \widehat{\mathcal{C}}_1 \cap \overline{\mathcal{C}_1}$, with $j \neq 1 \in [K]$. Therefore, by manipulating the above expression, we have

$$\begin{aligned}\widehat{\Theta}_1^+ - \Theta_1 = (\widehat{\Theta}_1 - \Theta_1)\left(I - \frac{2\eta_1}{|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1} Z^{(i)}Z^{(i),\top}\right) + \frac{2\eta_1}{|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1} W^{(i)}Z^{(i),\top} \\ + (\Theta_j - \Theta_1)\frac{2\eta_1}{|\widehat{\mathcal{C}}_1|}|\widehat{\mathcal{C}}_1 \cap \overline{\mathcal{C}_1}| \sum_{i\in\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}} Z^{(i)}Z^{(i),\top},\end{aligned}$$

and thus, we obtain

$$\|\widehat{\Theta}_1^+ - \Theta_1\| \leq \|\mathcal{H}_1\| + \|\mathcal{H}_2\|,$$

with,

$$\|\mathcal{H}_1\| = \|\widehat{\Theta}_1 - \Theta_1\| \left\| I - \frac{2\eta_1}{|\widehat{\mathcal{C}}_1|} Z Z^\top \right\| + \frac{2\eta_1}{|\widehat{\mathcal{C}}_1|} \sum_{i \in \widehat{\mathcal{C}}_1} \|W Z^\top\|,$$

$$\|\mathcal{H}_2\| = \|\Theta_j - \Theta_1\| \frac{2\eta_1}{|\widehat{\mathcal{C}}_1|} |\widehat{\mathcal{C}}_1 \cap \overline{\mathcal{C}_1}| \|\bar{Z}\bar{Z}^\top\|.$$

We now concatenate the batch matrices $Z^{(i)}, W^{(i)}$ of the systems classified to the first cluster in $Z \in \mathbb{R}^{(n_x+n_u)\times N_i T|\widehat{\mathcal{C}}_1|}$ and $W \in \mathbb{R}^{n_x \times N_i T|\widehat{\mathcal{C}}_1|}$, and similarly the batch matrices $Z^{(i)}$ of the systems incorrectly classified to the first cluster are concatenated in $\bar{Z} \in \mathbb{R}^{(n_x+n_u)\times N_i T|\widehat{\mathcal{C}}_1 \cap \overline{\mathcal{C}_1}|}$. We proceed with our analysis by controlling both terms separately. To upper bound the first term, we introduce the following propositions.

Proposition 3. *[212, Proposition 8] For any fixed $0 < \delta < 1$, let $N_i \geq (4n_x + 2n_u) \log \frac{T|\widehat{\mathcal{C}}_1|}{\delta}$. It holds, with probability at least $1 - \delta$, that*

$$\left\| W Z^\top \right\| \leq 4\sigma_{w,i} \sqrt{N_i (2n_x + n_u) \log \frac{9|\widehat{\mathcal{C}}_1|T}{\delta}} \sum_{t=0}^{T-1} \left\| (\Sigma_t^{(i)})^{\frac{1}{2}} \right\|. \tag{7.10}$$

Proposition 4. *(Adapted from [212, Proposition 6]) For any fixed $0 < \delta < 1$, let $N_i \geq 8(n_x + n_u) + 16 \log \frac{2|\widehat{\mathcal{C}}_1|T}{\delta}$. It holds, with probability at least $1 - \delta$, that*

$$Z Z^\top \succeq \frac{1}{4} \sum_{i \in \widehat{\mathcal{C}}_1} N_i \sum_{t=0}^{T-1} \Sigma_t^{(i)}, \tag{7.11}$$

$$\|\bar{Z}\bar{Z}^\top\| \leq \frac{9}{4} \sum_{i \in \widehat{\mathcal{C}}_1 \cap \overline{\mathcal{C}_1}} \sum_{t=0}^{T-1} N_i \left\| \Sigma_t^{(i)} \right\|. \tag{7.12}$$

Proof. Expression (7.11) follows direct from Proposition 6 in [212]. For expression (7.12), we can first write

$$\begin{aligned} \|\bar{Z}\bar{Z}^\top\| &= \left\| \sum_{i \in \widehat{\mathcal{C}}_1 \cap \overline{\mathcal{C}_1}} \sum_{l=1}^{N_i} \sum_{t=0}^{T-1} z_{l,t}^{(i)} z_{l,t}^{(i),\top} \right\| \\ &\leq \sum_{i \in \widehat{\mathcal{C}}_1 \cap \overline{\mathcal{C}_1}} \left\| \sum_{l=1}^{N_i} \sum_{t=0}^{T-1} z_{l,t}^{(i)} z_{l,t}^{(i),\top} \right\| \end{aligned}$$

where $\chi_{l,t}^{(i)} = (\Sigma_t^{(i)})^{-\frac{1}{2}} z_{l,t}^{(i)}$ for any fixed l, t, and i, where $\chi_{l,t}^{i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{n_x+n_u})$, for all $l \in \{1, 2, \ldots, N_i\}$, we obtain

$$\|\bar{Z}\bar{Z}^\top\| \leq \sum_{i \in \widehat{\mathcal{C}}_1 \cap \overline{\mathcal{C}_1}} \sum_{t=0}^{T-1} \|\Sigma_t^{(i)}\| \left\| \sum_{l=1}^{N_i} \chi_{l,t}^{(i)} \chi_{l,t}^{(i),\top} \right\|,$$

thus, by using Proposition 6 of [212], with probability $1-\frac{\delta}{T}$, we have

$$\left\|\sum_{l=1}^{N_i}\chi_{l,t}^{(i)}\chi_{l,t}^{(i),\top}\right\| \leq \frac{9N_i}{4},$$

which implies

$$\|\bar{Z}\bar{Z}^\top\| \leq \frac{9}{4}\sum_{i\in\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}}\sum_{t=0}^{T-1}N_i\|\Sigma_t^{(i)}\|.$$

□

Therefore, with probability $1-2\delta$, we have

$$\begin{aligned}\|\mathcal{H}_1\| &\leq \|\widehat{\Theta}_1-\Theta_1\|\left\|I-\frac{\eta_1}{2|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1}N_i\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right\|+\frac{2\eta_1}{|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1}\|WZ^\top\|,\\ &= \|\widehat{\Theta}_1-\Theta_1\|\left(1-\frac{\eta_1}{2|\widehat{\mathcal{C}}_1|}\lambda_{\min}\left(\sum_{i\in\widehat{\mathcal{C}}_1}N_i\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)\right)+\frac{2\eta_1}{|\widehat{\mathcal{C}}_1|}\sum_{i\in\widehat{\mathcal{C}}_1}\|WZ^\top\|.\end{aligned}$$

Hence, by selecting $\eta_1=\frac{|\widehat{\mathcal{C}}_1|}{\lambda_{\min}\left(\sum_{i\in\widehat{\mathcal{C}}_1}N_i\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)}$, we obtain

$$\begin{aligned}\|\mathcal{H}_1\| &\leq \frac{1}{2}\|\widehat{\Theta}_1-\Theta_1\|+\frac{8\sum_{i\in\widehat{\mathcal{C}}_1}\sigma_{w,i}\sqrt{N_i(2n_x+n_u)\log\frac{9|\widehat{\mathcal{C}}_1|T}{\delta}}\sum_{t=0}^{T-1}\left\|(\Sigma_t^{(i)})^{\frac{1}{2}}\right\|}{\lambda_{\min}\left(N_i\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)}\\ &\leq \frac{1}{2}\|\widehat{\Theta}_1-\Theta_1\|+\frac{8\sqrt{(2n_x+n_u)\log\frac{9|\widehat{\mathcal{C}}_1|T}{\delta}}\sqrt{\sum_{i\in\mathcal{S}_1}\sigma_{w,i}^2\left(\sum_{t=0}^{T-1}\left\|(\Sigma_t^{(i)})^{\frac{1}{2}}\right\|\right)^2}}{\sqrt{\sum_{i\in\widehat{\mathcal{C}}_1}N_i}\times\min_{i\in\widehat{\mathcal{C}}_1}\lambda_{\min}\left(\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)}\\ &= \frac{1}{2}\|\widehat{\Theta}_1-\Theta_1\|+\bar{c}_0\times\frac{1}{\sqrt{\sum_{i\in\widehat{\mathcal{C}}_1}N_i}},\end{aligned} \tag{7.13}$$

with $N_i\geq\max\{8(n_x+n_u)+16\log\frac{2|\widehat{\mathcal{C}}_1|T}{\delta},(4n_x+2n_u)\log\frac{|\widehat{\mathcal{C}}_1|T}{\delta}\}$, for all $i\in\widehat{\mathcal{C}}_1$. To control the second term $\|\mathcal{H}_2\|$, we first use the Definition 3 to write

$$\|\mathcal{H}_2\|\leq\Delta_{\max}|\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}|\frac{9\sum_{i\in\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}}N_i\sum_{t=0}^{T-1}\|\Sigma_t^{(i)}\|}{2\lambda_{\min}\left(\sum_{i\in\widehat{\mathcal{C}}_1}N_i\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)},$$

which implies

$$\|\mathcal{H}_2\|\leq c_5\Delta_{\max}|\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}|,$$

by using Jensen and Cauchy-Schwartz inequalities in the denominator and numerator, respectively, where we define $c_5 = \frac{9\sum_{i\in\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}}\sum_{t=0}^{T-1}\|\Sigma_t^{(i)}\|}{2\min_{i\in\widehat{\mathcal{C}}_1}\left(\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)}$. Therefore, we proceed with our analysis to control $|\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}|$. To do so, we use Lemma 43 and obtain

$$\mathbb{E}\left[|\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}|\right] \leq c_6\sum_{i\in[M]}\sum_{t=0}^{T-1}\exp\left(-c_7N_in_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\|+\sqrt{n_x}}\right)^2\right),$$

which yields

$$\begin{aligned}\mathbb{P}\left\{|\widehat{\mathcal{C}}_1\cap\overline{\mathcal{C}_1}|\leq c_6\sum_{i\in[M]}\sum_{t=0}^{T-1}\exp\left(-\frac{c_7}{2}N_in_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\|+\sqrt{n_x}}\right)^2\right)\right\}\\ \geq 1-\sum_{i\in[M]}\sum_{t=0}^{T-1}\exp\left(-\frac{c_7}{2}N_in_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\|+\sqrt{n_x}}\right)^2\right)\geq 1-\delta,\end{aligned}$$

by using Markov's inequality and Assumption 10 with $N_in_x \geq c\left(\frac{\rho^{(i)}\|\Sigma_t^{(i)}\|+\sqrt{n_x}}{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}\right)^2\log(\frac{MT}{\delta})$, for some large enough constant c such that $\frac{1}{c}<c_7$, with $0<\delta<1$ for all $i\in[M]$. Thus, we obtain

$$\|\mathcal{H}_2\|\leq\bar{c}_1\Delta_{\max}\sum_{i\in[M]}\sum_{t=0}^{T-1}\exp\left(-\bar{c}_2N_in_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\|+\sqrt{n_x}}\right)^2\right), \tag{7.14}$$

with probability at least $1-\delta$. The proof is completed by combining (7.13) and (7.14).

7.6.3 Proof of Corollary 3

We first recall that at iteration r we posses an estimation for the model such that $\|\widehat{\Theta}_j^{(r)}-\Theta_j\|\leq(\frac{1}{2}-\alpha^{(r)})\Delta_{\min}$, for all $j\in[K]$ with $\alpha^{(r)}\in\mathbb{R}$. Moreover, according to Theorem 12, we have

$$\begin{aligned}\|\widehat{\Theta}_j^{(r+1)}-\Theta_j\|\leq\frac{1}{2}\|\widehat{\Theta}_j^{(r)}-\Theta_j\|+\bar{c}_0\times\frac{1}{\sqrt{\sum_{i\in\widehat{\mathcal{C}}_j^{(r)}}N_i}}\\ +\bar{c}_1\Delta_{\max}\sum_{i\in[M]}\sum_{t=0}^{T-1}\exp\left(-\bar{c}_2N_in_x\left(\frac{\alpha^{(r)}\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\|+\sqrt{n_x}}\right)^2\right),\end{aligned}$$

where by using Assumption 10 and $0<\alpha^{(0)}<\frac{1}{2}$ we can guarantee that $\|\widehat{\Theta}_j^{(r+1)}-\Theta_j\|\leq\|\widehat{\Theta}_j^{(r)}-\Theta_j\|$ for any $r\in[R]$. This implies that $\alpha^{(r+1)}\geq\alpha^{(r)}$, for any $r\in[R]$. First, we aim to show that after a small

number of iterations, we obtain a sufficiently large value of $\alpha^{(r)} \geq \frac{1}{4}$. To do so, let

$$\epsilon_r := \bar{c}_0 \times \frac{1}{\sqrt{\sum_{i \in \widehat{\mathcal{C}}_j^{(r)}} N_i}} + \bar{c}_1 \Delta_{\max} \sum_{i \in [M]} \sum_{t=0}^{T-1} \exp\left(-\bar{c}_2 N_i n_x \left(\frac{\alpha^{(r)} \rho^{(i)} \|\Sigma_t^{(i)}\|}{\rho^{(i)} \|\Sigma_t^{(i)}\| + \sqrt{n_x}} \right)^2 \right), \tag{7.15}$$

be the error at iteration r, and note that $\epsilon_{r+1} \leq \epsilon_r$ for any $r \in [R]$ since $\alpha^{(r+1)} \geq \alpha^{(r)}$. Then, after R' iterations of Algorithm 9, we obtain

$$\|\widehat{\Theta}_j^{(R')} - \Theta_j\| \leq (1 - \mu_j)^{R'} \left(\frac{1}{2} - \alpha^{(0)} \right) \Delta_{\min} + 2\epsilon_0$$

for $R' \geq 2$. Therefore, we need to guarantee that after $R' \geq 2$ parallel iterations, the right hand side of the above expression is upper bounded by $\frac{1}{4}\Delta_{\min}$. For the first term, since $0 < \alpha^{(0)} < \frac{1}{2}$, it suffices to show that $(\frac{1}{2})^{R'} \leq \frac{1}{4}$, which is satisfied for any $R' \geq 2$. On the other hand, $2\epsilon_0 \leq \frac{1}{8}\Delta_{\min}$ follows directly from the minimum separation condition of Assumption 10. Therefore, we have $\|\widehat{\Theta}_j^{(r)} - \Theta_j\| \leq \frac{1}{4}\Delta_{\min}$, for any $r \geq R'$. Then, after $R'' \geq R'$, we have

$$\|\widehat{\Theta}_j^{(R'')} - \Theta_j\| \leq \left(\frac{1}{2} \right)^{R''} \frac{\Delta_{\min}}{4} + 2\epsilon_0$$

which implies $\|\widehat{\Theta}_j^{(R)} - \Theta_j\| \leq \epsilon$ after $R = R' + R'' \geq 2 + \log(\frac{\Delta_{\min}}{4\epsilon})$, with ϵ as defined in (7.6).

Chapter 8

SUMMARY AND FUTURE DIRECTIONS

In this thesis, we introduced methods to address FL challenges by developing robust, efficient algorithms for supervised learning, reinforcement learning (RL), control, and personalized system identification. By tackling fundamental issues related to data heterogeneity, communication efficiency, and convergence stability, our research pushes the boundaries of FL across diverse applications and environments.

In Chapter 3, we introduced an algorithm for supervised learning problems to address data heterogeneity and ensure stable convergence, even with partial client participation. This algorithm lays the groundwork for more resilient federated models, capable of handling real-world heterogeneity across distributed data sources.

In Chapter 4 and 5, we developed federated reinforcement learning (FRL) algorithms that leverage similarities across heterogeneous environments, demonstrating improvements in sample efficiency and the acceleration of policy learning. Our rigorous theoretical analyses established the efficiency of these algorithms, showcasing the potential of FRL to accelerate policy evaluation and policy optimization across diverse agent environments.

In Chapter 6, we extended FL into control systems through the development of the `FedLQR` algorithm, enabling multiple agents with similar but unknown dynamics to collaboratively learn stabilizing policies. This work highlights the potential of federated approaches to ensure stability and optimize control systems, even in the presence of heterogeneity among agents.

To further advance FL's application, we explored techniques for personalized system identification in Chapter 7. Our approach leverages clustering methods to enable clients to obtain customized models, improving convergence and adaptability to individual system dynamics. This personalized framework

enhances FL's flexibility, making it suitable for complex applications where each client requires a customized solution.

Now, let us briefly outline certain problems that are a subject of future research.

• **Meta and Fine-tuning Reinforcement Learning (Meta-RL):** When a RL agent encounters a change in task or environment, conventional methods often discard the pretrained policy and start training a new policy from scratch, which is usually computationally expensive. Meta-RL, as empirically investigated in [11], offers a solution to this problem by allowing the agent to adapt quickly to new tasks/environments. However, the empirical advances lack solid theoretical guarantees and systematic guidance. Our future work aims to *provide systematic fine-tuning methodologies and develop more effective meta-RL algorithms* by leveraging principles from Information Theory, Network Control, high-dimensional probability, and Optimization. We believe that exploring this direction will have a profound impact on fields like autonomous self-driving cars and recommendation systems.

• **Robust Reinforcement Learning:** Drawing on our previous works on FRL, we aim to provide more general and theoretically grounded solutions to improve the robustness of RL algorithms. The main focus of FRL is environmental heterogeneity among agents, while robust RL emphasizes the environmental uncertainty. This raises a crucial question: Is there an intrinsic link between this environmental heterogeneity and the robustness of RL agents? Revealing this connection could be key to develop more robust RL algorithms that remain effective despite the presence of noise, misinformation, or even adversarial attacks. The significance of this research lies in its potential to create more reliable and resilient AI systems for complex, real-world decision-making tasks.

• **Future Research: More Complex Models and Modalities** In our prior work [193], we explored a collaborative FL framework geared towards rapidly acquiring knowledge for linear-time-invariant (LTI) controllers in the context of basic LTI systems. However, the real-world systems we encounter are often *nonlinear* and subject to inherent *safety constraints*. Extending our research to encompass nonlinear systems with safety constraints is an enticing prospect. To tackle this challenge, we envision integrating Model Predictive Control (MPC) with techniques like sequential linearization or lifting and enforce safety through barrier certificates. Furthermore, dealing with multiple systems that provide *observations in different dimensions and modalities* presents a significant challenge. For instance, one system may generate image data while another captures speech or audio data. This disparity in data modalities poses a complex problem within FL for control, particularly in data aggregation on the server side. Addressing this challenge is a main

focus of my future research, promising to enhance the utility of FL for control across diverse real-world scenarios.

• **Personalization in Robust Multi-Agent control in Networked Systems:** In a network of interconnected and diverse agents, the challenge is to facilitate collaborative learning of personalized and resilient controllers through limited information exchange. This problem is intricate due to factors like agent heterogeneity, dynamic network structures, model uncertainties, and potential adversarial disruptions. Our future research aims to assess the impact of agent diversity, refine the definition of heterogeneity, and find ways to mitigate its effects. Leveraging personalization techniques such as classification, transfer learning, and representation learning, my work seeks to enhance our understanding of personalization in robust multi-agent control problems. Ultimately, this will lead to more efficient and resilient control strategies in complex networked systems.

Bibliography

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

[2] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

[3] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

[4] Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Carmen Amo Alonso, Jing Shuang, James Anderson, and Nikolai Matni. Distributed and localized model predictive control. part i: Synthesis and implementation. *arXiv preprint arXiv:2110.07010*, 2021.

[6] Brian DO Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[7] Karl Johan Åström and Peter Eykhoff. System identification—a survey. *Automatica*, 7(2):123–162, 1971.

[8] Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory*, pages 257–283. PMLR, 2016.

[9] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.*, 45:77–120, 2017.

[10] Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *journal of artificial intelligence research*, 15:319–350, 2001.

[11] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*, 2023.

[12] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[13] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.

[14] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1:374–388, 2019.

[15] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[16] Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

[17] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[18] Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

[19] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[20] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

[21] Zachary Charles and Jakub Konečnỳ. On the outsized importance of learning rates in local update methods. *arXiv preprint arXiv:2007.00878*, 2020.

[22] Zachary Charles and Jakub Konečnỳ. On the outsized importance of learning rates in local update methods. *arXiv preprint arXiv:2007.00878*, 2020.

[23] Zachary Charles and Jakub Konečnỳ. Convergence and Accuracy Trade-Offs in Federated Learning and Meta-Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2575–2583. PMLR, 2021.

[24] Zachary Charles and Jakub Konečnỳ. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2575–2583. PMLR, 2021.

[25] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pages 2722–2730, 2015.

[26] Jingdi Chen, Tian Lan, and Nakjung Choi. Distributional-utility actor-critic for network slice performance guarantee. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 161–170, 2023.

[27] Yiting Chen, Ana M Ospina, Fabio Pasqualetti, and Emiliano Dall'Anese. Multi-Task System Identification of Similar Linear Time-Invariant Dynamical Systems. *arXiv preprint arXiv:2301.01430*, 2023.

[28] Ziheng Cheng, Xinmeng Huang, and Kun Yuan. Momentum benefits non-iid federated learning simply and provably. *International Conference on Learning Representations*, 2024.

[29] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.

[30] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.

[31] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Fedavg with fine tuning: Local updates lead to representation learning. *arXiv preprint arXiv:2205.13692*, 2022.

[32] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.

[33] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

[34] Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for TD (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[35] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

[36] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

[37] Yuhao Ding, Junzi Zhang, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv preprint arXiv:2110.10117*, 2021.

[38] Thinh Doan, Siva Maguluri, and Justin Romberg. Finite-time analysis of distributed TD (0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 1626–1635. PMLR, 2019.

[39] Jim Douglas and Henry H Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439, 1956.

[40] J.C. Doyle, K. Glover, P.P. Khargonekar, and B.A. Francis. State-space solutions to standard h/sub 2/ and h/sub infinity / control problems. *IEEE Transactions on Automatic Control*, 34(8):831–847, 1989.

[41] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[42] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

[43] Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

[44] Jonathan Eckstein. *Splitting methods for monotone operators with applications to parallel optimization*. PhD thesis, Massachusetts Institute of Technology, 1989.

[45] Lawrence C. Evans. *An Introduction to Mathematical Optimal Control Theory*. University of California, Department of Mathematics, 2005.

[46] Nicolò Dal Fabbro, Aritra Mitra, and George J Pappas. Federated td learning over finite-rate erasure channels: Linear speedup under markovian sampling. *arXiv preprint arXiv:2305.08104*, 2023.

[47] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

[48] Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. Fault-tolerant federated reinforcement learning with theoretical guarantee. *Advances in Neural Information Processing Systems*, 34:1007–1021, 2021.

[49] Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. *arXiv preprint arXiv:2302.01734*, 2023.

[50] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

[51] Claude-Nicolas Fiechter. Pac adaptive control of linear systems. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 72–80, 1997.

[52] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[53] Georg Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, and Germany Mathematician. Über matrizen aus nicht negativen elementen. 1912.

[54] Masao Fukushima. Application of the alternating direction method of multipliers to separable convex programming problems. *Computational Optimization and Applications*, 1(1):93–111, 1992.

[55] Thomas Furmston, Guy Lever, and David Barber. Approximate Newton methods for policy search in markov decision processes. *Journal of Machine Learning Research*, 17, 2016.

[56] Daniel Gabay. Chapter ix applications of the method of multipliers to variational inequalities. In *Studies in mathematics and its applications*, volume 15, pages 299–331. Elsevier, 1983.

[57] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976.

[58] Matilde Gargiani, Andrea Zanelli, Andrea Martinelli, Tyler Summers, and John Lygeros. PAGE-PG: A simple and loopless variance-reduced policy gradient method with probabilistic gradient estimation. In *International Conference on Machine Learning*, pages 7223–7240. PMLR, 2022.

[59] Konstantinos Gatsis. Federated reinforcement learning at the edge: Exploring the learning-communication tradeoff. In *2022 European Control Conference (ECC)*, pages 1890–1895. IEEE, 2022.

[60] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

[61] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

[62] Avishek Ghosh, Arya Mazumdar, et al. An Improved Algorithm for Clustered Federated Learning. *arXiv preprint arXiv:2210.11538*, 2022.

[63] Pontus Giselsson and Stephen Boyd. Linear convergence and metric selection for Douglas-Rachford splitting and ADMM. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2016.

[64] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local SGD: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR, 2021.

[65] Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11):5283–5298, 2020.

[66] F. Haddadpour and M. Mahdavi. On the convergence of local sgd methods. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2301–2311, 2019.

[67] Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems*, pages 11082–11094, 2019.

[68] Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32, 2019.

[69] Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

[70] Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

[71] Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021.

[72] Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33:2304–2315, 2020.

[73] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[74] Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Towards a theoretical foundation of policy optimization for learning control policies. *arXiv preprint arXiv:2210.04810*, 2022.

[75] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *International conference on machine learning*, pages 4422–4433. PMLR, 2020.

[76] Xinmeng Huang, Donghwan Lee, Edgar Dobriban, and Hamed Hassani. Collaborative learning of discrete distributions under heterogeneity and communication constraints. In *Advances in Neural Information Processing Systems*, 2022.

[77] Xinmeng Huang, Ping Li, and Xiaoyun Li. Stochastic controlled averaging for federated learning with communication compression. *International Conference on Learning Representations*, 2024.

[78] Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated Reinforcement Learning with Environment Heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 18–37. PMLR, 2022.

[79] Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 18–37. PMLR, 2022.

[80] Zeyu Jin, Johann Michael Schmitt, and Zaiwen Wen. On the analysis of model-free methods for the linear quadratic regulator. *arXiv preprint arXiv:2007.03861*, 2020.

[81] Gangshan Jing, He Bai, Jemin George, Aranya Chakrabortty, and Piyush K Sharma. Learning distributed stabilizing controllers for multi-agent systems. *IEEE Control Systems Letters*, 6:301–306, 2021.

[82] Caleb Ju, Georgios Kotsalis, and Guanghui Lan. A model-free first-order method for linear quadratic regulator with $\tilde{O}(1/\varepsilon)$ sampling complexity. *arXiv preprint arXiv:2212.00084*, 2022.

[83] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[84] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[85] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

[86] A. Khaled, K. Mishchenko, and P. Richtárik. Tighter theoretical guarantees for local sgd in federated learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1234–1244, 2020.

[87] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local gd on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

[88] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local GD on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

[89] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

[90] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

[91] Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated Reinforcement Learning: Linear Speedup Under Markovian Sampling. In *International Conference on Machine Learning*, pages 10997–11057. PMLR, 2022.

[92] Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under Markovian sampling. In *International Conference on Machine Learning*, pages 10997–11057. PMLR, 2022.

[93] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U Stich. A unified theory of decentralized SGD with changing topology and local updates. *arXiv preprint arXiv:2003.10422*, 2020.

[94] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[95] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

[96] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[97] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[98] Nathaniel Korda and Prashanth La. On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *International conference on machine learning*, pages 626–634. PMLR, 2015.

[99] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE, 2010.

[100] Yassine Laguel, Krishna Pillutla, Jerôme Malick, and Zaid Harchaoui. A superquantile approach

to federated learning with heterogeneous devices. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2021.

[101] Chandrashekar Lakshminarayanan and Csaba Szepesvári. Linear stochastic approximation: Constant step-size and iterate averaging. *arXiv preprint arXiv:1709.04073*, 2017.

[102] Andrew Lamperski. Computing stabilizing linear controllers via policy iteration. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1902–1907. IEEE, 2020.

[103] Guangchen Lan, Dong-Jun Han, Abolfazl Hashemi, Vaneet Aggarwal, and Christopher G Brinton. Asynchronous federated reinforcement learning with policy gradient updates: Algorithm design and convergence analysis. *arXiv preprint arXiv:2404.08003*, 2024.

[104] Guangchen Lan, Xiao-Yang Liu, Yijing Zhang, and Xiaodong Wang. Communication-efficient federated learning for resource-constrained edge devices. *IEEE Transactions on Machine Learning in Communications and Networking*, 2023.

[105] Guangchen Lan, **Han Wang**, James Anderson, Christopher Brinton, and Vaneet Aggarwal. Improved Communication Efficiency in Federated Natural Policy Gradient via ADMM-based Gradient Updates. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[106] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[107] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[108] Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015.

[109] Guoyin Li and Ting Kei Pong. Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical programming*, 159(1):371–401, 2016.

[110] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[111] X. Li, J. Liu, and C. Wang. Advances in distributed optimization algorithms. *Journal of Applied Mathematics*, 56(3):401–414, 2019.

[112] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

[113] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

[114] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2019.

[115] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and Statistics*, pages 983–992. PMLR, 2019.

[116] Xinle Liang, Yang Liu, Tianjian Chen, Ming Liu, and Qiang Yang. Federated transfer reinforcement learning for autonomous driving. In *Federated and Transfer Learning*, pages 357–371. Springer, 2022.

[117] Hyun-Kyo Lim, Ju-Bong Kim, Joo-Seong Heo, and Youn-Hee Han. Federated reinforcement learning for training control policies on multiple iot devices. *Sensors*, 20(5):1359, 2020.

[118] Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. Multi-agent reinforcement learning in stochastic networked systems. *Advances in Neural Information Processing Systems*, 34:7825–7837, 2021.

[119] Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[120] Boyi Liu, Lujia Wang, and Ming Liu. Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters*, 4(4):4555–4562, 2019.

[121] Boyi Liu, Lujia Wang, and Ming Liu. Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters*, 4(4):4555–4562, 2019.

[122] Rui Liu and Alex Olshevsky. Distributed TD (0) with almost no communication. *arXiv preprint arXiv:2104.07855*, 2021.

[123] Rui Liu and Alex Olshevsky. Temporal difference learning as gradient splitting. In *International Conference on Machine Learning*, pages 6905–6913. PMLR, 2021.

[124] Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

[125] Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[126] W-M Lu, Kemin Zhou, and John C Doyle. Stabilization of uncertain linear systems: An lft approach. *IEEE Transactions on Automatic Control*, 41(1):50–65, 1996.

[127] Weimin Lyu, Songzhu Zheng, Tengfei Ma, and Chao Chen. A study of the attention abnormality in trojaned berts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4727–4741, 2022.

[128] Weimin Lyu, Songzhu Zheng, Lu Pang, Haibin Ling, and Chao Chen. Attention-enhancing backdoor attacks against bert-based models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10672–10690, 2023.

[129] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd international conference on artificial intelligence and statistics*, pages 2916–2925. PMLR, 2019.

[130] Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtarik. From local SGD to local fixed-point methods for federated learning. In *International Conference on Machine Learning*, pages 6692–6701. PMLR, 2020.

[131] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

[132] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[133] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[134] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019.

[135] Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. ProxSkip: Yes! Local Gradient Steps Provably Lead to Communication Acceleration! Finally! *arXiv preprint arXiv:2202.09357*, 2022.

[136] Aritra Mitra, Rayana Jaafar, George Pappas, and Hamed Hassani. Linear Convergence in Federated Learning: Tackling Client Heterogeneity and Sparse Gradients. *Advances in Neural Information Processing Systems*, 34, 2021.

[137] Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

[138] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[139] Zhaobin Mo, Wangzhi Li, Yongjie Fu, Kangrui Ruan, and Xuan Di. Cvlight: Decentralized learning for adaptive traffic signal control with connected vehicles. *Transportation research part C: emerging technologies*, 141:103728, 2022.

[140] Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem. *IEEE Transactions on Automatic Control*, 67(5):2435–2450, 2021.

[141] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

[142] C Narayanan and Csaba Szepesvári. Finite time bounds for temporal difference learning with function approximation: Problems with some "state-of-the-art" results. Technical report, Technical report, 2017.

[143] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[144] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

[145] Colm Art O'cinneide. Entrywise perturbation theory and error analysis for markov chains. *Numerische Mathematik*, 65(1):109–120, 1993.

[146] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of LTI systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

[147] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pages 4026–4035. PMLR, 2018.

[148] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[149] Reese Pathak and Martin J Wainwright. FedSplit: An algorithmic framework for fast federated optimization. *Advances in Neural Information Processing Systems*, 33:7057–7066, 2020.

[150] Reese Pathak and Martin J Wainwright. FedSplit: An algorithmic framework for fast federated optimization. *arXiv preprint arXiv:2005.05238*, 2020.

[151] Donald W Peaceman and Henry H Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for industrial and Applied Mathematics*, 3(1):28–41, 1955.

[152] Juan Perdomo, Jack Umenberger, and Max Simchowitz. Stabilizing dynamical systems via policy gradient methods. *Advances in Neural Information Processing Systems*, 34:29274–29286, 2021.

[153] Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. *Advances in Neural Information Processing Systems*, 26, 2013.

[154] Hossein Pishro-Nik. Introduction to probability, statistics, and random processes. 2016.

[155] Author(s) Placeholder. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. *Journal/Conference Placeholder*, page Pages Placeholder, Year Placeholder.

[156] Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1:32, 1987.

[157] Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. Federated reinforcement learning: techniques, applications, and open challenges. *arXiv preprint arXiv:2108.11887*, 2021.

[158] Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. Federated reinforcement learning: Techniques, applications, and open challenges. *arXiv preprint arXiv:2108.11887*, 2021.

[159] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

[160] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[161] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A communication-efficient federated learning method with periodic averaging and

quantization. In *International Conference on Artificial Intelligence and Statistics*, pages 2021–2031. PMLR, 2020.

[162] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pages 2021–2031. PMLR, 2020.

[163] Zhaolin Ren, Aoxiao Zhong, Zhengyuan Zhou, and Na Li. Federated lqr: Learning through sharing. *arXiv preprint arXiv:2011.01815v1*, 2020.

[164] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1):433–484, 2016.

[165] Kangrui Ruan, Junzhe Zhang, Xuan Di, and Elias Bareinboim. Causal imitation learning via inverse reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

[166] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3:3, 2018.

[167] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3, 2018.

[168] Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

[169] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

[170] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic

distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

[171] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[172] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[173] Jacob H Seidman, Mahyar Fazlyab, Victor M Preciado, and George J Pappas. A control-theoretic approach to analysis and parameter selection of Douglas–Rachford splitting. *IEEE Control Systems Letters*, 4(1):199–204, 2019.

[174] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

[175] Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *International conference on machine learning*, pages 5729–5738. PMLR, 2019.

[176] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[177] Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

[178] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

[179] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

[180] Artin Spiridonoff, Alex Olshevsky, and Ioannis Ch Paschalidis. Local SGD With a Communication Overhead Depending Only on the Number of Workers. *arXiv preprint arXiv:2006.02582*, 2020.

[181] Artin Spiridonoff, Alex Olshevsky, and Ioannis Ch Paschalidis. Local sgd with a communication overhead depending only on the number of workers. *arXiv preprint arXiv:2006.02582*, 2020.

[182] Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and td learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

[183] Sebastian U Stich. Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[184] Sebastian U Stich. Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[185] Lili Su, Jiaming Xu, and Pengkun Yang. Global convergence of federated learning for mixed regression. *arXiv preprint arXiv:2206.07279*, 2022.

[186] Jun Sun, Gang Wang, Georgios B Giannakis, Qinmin Yang, and Zaiyue Yang. Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 4485–4495. PMLR, 2020.

[187] Yue Sun and Maryam Fazel. Learning optimal controllers by policy gradient: Global optimality via convex parameterization. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 4576–4581. IEEE, 2021.

[188] Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

[189] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[190] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

[191] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[192] **Han Wang**, Aritra Mitra, Hamed Hassani, George J Pappas, and James Anderson. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *Transactions on Machine Learning Research*, 2023.

[193] **Han Wang**, Leonardo F Toso, Aritra Mitra, and James Anderson. Model-free learning with heterogeneous dynamical systems: A federated LQR approach. *arXiv preprint arXiv:2308.11743*, 2023.

[194] **Han Wang**, Leonardo Felipe Toso, and James Anderson. Fedsysid: A federated approach to sample-efficient system identification. In *Learning for Dynamics and Control Conference*, pages 1308–1320. PMLR, 2023.

[195] Andreas Themelis and Panagiotis Patrinos. Douglas–Rachford splitting and ADMM for nonconvex optimization: Tight convergence results. *SIAM Journal on Optimization*, 30(1):149–181, 2020.

[196] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[197] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[198] Leonardo F Toso, **Han Wang**, and James Anderson. Learning personalized models with clustered system identification. *arXiv e-prints*, pages arXiv–2304, 2023.

[199] Leonardo F. Toso, Han Wang, and James Anderson. Learning Personalized Models with Clustered System Identification. 2023.

[200] Leonardo F Toso, Han Wang, and James Anderson. Asynchronous heterogeneous linear quadratic regulator design. *arXiv preprint arXiv:2404.09061*, 2024.

[201] Leonardo F Toso, Han Wang, and James Anderson. Oracle complexity reduction for model-free lqr: A

stochastic variance-reduced policy gradient approach. In *2024 American Control Conference (ACC)*, pages 4032–4037. IEEE, 2024.

[202] Leonardo F Toso, Donglin Zhan, James Anderson, and Han Wang. Meta-learning linear quadratic regulators: A policy gradient maml approach for the model-free lqr. *arXiv preprint arXiv:2401.14534*, 2024.

[203] Quoc Tran Dinh, Nhan Pham, Dzung Phan, and Lam Nguyen. FedDR–randomized Douglas-Rachford splitting algorithms for nonconvex federated composite optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[204] Joel Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab.*, 2011.

[205] John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. In *IEEE Transactions on Automatic Control*, 1997.

[206] Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

[207] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[208] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[209] Han Wang and James Anderson. Large-scale system identification using a randomized svd. In *2022 American Control Conference (ACC)*, pages 2178–2185. IEEE, 2022.

[210] Han Wang, Sihong He, Zhili Zhang, Fei Miao, and James Anderson. Momentum for the win: Collaborative federated reinforcement learning across heterogeneous environments. *International Conference on Machine Learning, 2024*, 2024.

[211] Han Wang, Siddartha Marella, and James Anderson. Fedadmm: A federated primal-dual algorithm allowing partial participation. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 287–294. IEEE, 2022.

[212] Han Wang, Leonardo F Toso, and James Anderson. Fedsysid: A federated approach to sample-efficient system identification. *arXiv preprint arXiv:2211.14393*, 2022.

[213] Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

[214] Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

[215] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms. *The Journal of Machine Learning Research*, 22(1):9709–9758, 2021.

[216] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[217] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[218] Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. SlowMo: Improving communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*, 2019.

[219] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

[220] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

[221] Xiaofei Wang, Yiwen Han, Chenyang Wang, Qiyang Zhao, Xu Chen, and Min Chen. In-Edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning. *IEEE Network*, 33(5):156–165, 2019.

[222] Xinrui Wang and Yan Jin. Exploring causalworld: Enhancing robotic manipulation via knowledge transfer and curriculum learning. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 88360, page V03AT03A013. American Society of Mechanical Engineers, 2024.

[223] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[224] Jiin Woo, Gauri Joshi, and Yuejie Chi. The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. *arXiv preprint arXiv:2305.10697*, 2023.

[225] Blake Woodworth, Kumar Kshitij Patel, Sebastian U Stich, Zhen Dai, Brian Bullins, H Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is Local SGD Better than Minibatch SGD? *arXiv preprint arXiv:2002.07839*, 2020.

[226] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local SGD for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020.

[227] Zhijie Xie and SH Song. Client selection for federated policy optimization with environment heterogeneity. *arXiv preprint arXiv:2305.10978*, 2023.

[228] Zhijie Xie and Shenghui Song. FedKL: Tackling data heterogeneity in federated reinforcement learning by penalizing KL divergence. *IEEE Journal on Selected Areas in Communications*, 41(4):1227–1242, 2023.

[229] Lei Xin, Lintao Ye, George Chiu, and Shreyas Sundaram. Identifying the Dynamics of a System by Leveraging Data from Similar Systems. *arXiv preprint arXiv:2204.05446*, 2022.

[230] Lei Xin, Lintao Ye, George Chiu, and Shreyas Sundaram. Learning Dynamical Systems by Leveraging Data from Similar Systems. *arXiv preprint arXiv:2302.04344*, 2023.

[231] Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.

[232] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551. PMLR, 2020.

[233] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[234] Ming Yan and Wotao Yin. Self equivalence of the alternating direction method of multipliers. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 165–194. Springer, 2016.

[235] Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

[236] A. Yu, B. Chen, and D. Sun. Federated learning with privacy protection. *IEEE Transactions on Neural Networks*, 30(4):845–859, 2019.

[237] Shuai Yu, Xu Chen, Zhi Zhou, Xiaowen Gong, and Di Wu. When deep reinforcement learning meets federated learning: Intelligent multitimescale resource management for multiaccess edge computing in 5g ultradense network. *IEEE Internet of Things Journal*, 8(4):2238–2251, 2020.

[238] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[239] Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. In *International Conference on Machine Learning*, pages 12253–12266. PMLR, 2021.

[240] Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic recursive momentum for policy gradient methods. *arXiv preprint arXiv:2003.04302*, 2020.

[241] Sihan Zeng, Malik Aqeel Anwar, Thinh T Doan, Arijit Raychowdhury, and Justin Romberg. A decentralized policy gradient approach to multi-task reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 1002–1012. PMLR, 2021.

[242] Chenyu Zhang, **Han Wang**, James Anderson, and Aritra Mitra. Finite-Time Analysis of On-Policy Heterogeneous Federated Reinforcement Learning. In *Twelfth International Conference on Learning Representations*, 2024.

[243] Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. *International Conference on Learning Representations*, 2024.

[244] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

[245] Thomas T Zhang, Katie Kang, Bruce D Lee, Claire Tomlin, Sergey Levine, Stephen Tu, and Nikolai Matni. Multi-Task Imitation Learning for Linear Dynamical Systems. *arXiv preprint arXiv:2212.00186*, 2022.

[246] Thomas TCK Zhang, Leonardo F Toso, James Anderson, and Nikolai Matni. Meta-Learning Operators to Optimality from Multi-Task Non-IID Data. *arXiv preprint arXiv:2308.04428*, 2023.

[247] Xinwei Zhang and Mingyi Hong. On the Connection Between FedDyn and FedPD, 2021.

[248] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. FedPD: A federated learning framework with adaptivity to non-iid data. *IEEE Transactions on Signal Processing*, 69:6055–6070, 2021.

[249] Feiran Zhao, Xingyun Fu, and Keyou You. On the sample complexity of stabilizing linear systems via policy gradient methods. *arXiv preprint arXiv:2205.14335*, 2022.

[250] Shipu Zhao, Laurent Lessard, and Madeleine Udell. An automatic system to detect equivalence between iterative algorithms. *arXiv preprint arXiv:2105.04684*, 2021.

[251] Shipu Zhao, Laurent Lessard, and Madeleine Udell. An automatic system to detect equivalence between iterative algorithms. *arXiv preprint arXiv:2105.04684*, 2021.

[252] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[253] Yang Zheng and Na Li. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, 2020.

[254] K Zhou, JC Doyle, and K Glover. *Robust and optimal control*. Prentice hall, 1996.